UTBoost: A Tree-boosting based System for Uplift Modeling

Anonymous authors
Paper under double-blind review

Abstract

Uplift modeling is a suite of machine learning techniques that managers can use to predict the incremental impact of an action on a particular customer outcome. These models enable decision-makers to pinpoint which customer segments will most likely respond positively to a specific intervention, thereby enhancing the efficiency of resource distribution and boosting total gains. However, accurately estimating this incremental impact presents tangible difficulties since it requires assessing the difference between two mutually exclusive outcomes for each individual. In our research, we introduce two novel modifications to the established Gradient Boosting Decision Trees (GBDT) technique, which learn the causal effect in a sequential way and overcome the counter-factual nature. Both approaches innovate existing techniques in terms of ensemble learning method and learning objective, respectively. Experiments on large-scale datasets demonstrate the usefulness of the proposed methods, which often yielding remarkable improvements over baseline models.

1 Introduction

Uplift modeling, a machine learning technique used to estimate the net effect of a particular action, has recently drawn considerable attention. In contrast to traditional supervised learning, uplift models concentrate on modeling the effect of a treatment on individual outcomes and generate uplift scores that show the "probability of persuasion" for each instance. This technique has proven particularly useful in personalized medicine, performance marketing, and social science. These scenarios are typically presented using tabular data, with each entity having an indicator variable representing whether the subject was treated or not. Additionally, each entity includes a corresponding label.

However, one significant challenge in the uplift problem is the absence of observations for both treated and control responses for an individual in exactly the same context. This hinders the direct application of standard classification techniques to model individual treatment effects. Various tree-based studies have tackled this challenge by creating measures of outcome differences between treated and untreated observations and maximizing heterogeneity between groups (Rzepakowski & Jaroszewicz, 2012; Athey & Imbens, 2016). Several researches provide further generalization to bagging ensemble methods on the idea of random forests (Guelman et al., 2012; 2015), which aim to address the challenges of tree model performance decay with an increasing number of covariates. Despite the success of these nonparametric methods in prediction, we have experimentally discovered that random-forest based methods still suffer from significant degradation of power in predicting causal effects as the number of variables increases.

An alternative to a random forest is boosted trees. In this paper, we propose an extension to the nonparametric method drawn from a boosting perspective. The subsequent trees are fitted based on the inherited causal effect learned by preceding models. We found that the two ensemble learning methods perform similarly for low-dimensional problems. However, when it comes to modeling extremely high-dimensional data, boosting shows a significant superiority.

However, the split criterion which aim at maximizing heterogeneity neglects the learning of outcomes and focuses solely on uplift signals. This results in limited prediction accuracy and restricts the application of these models in domains where outcome prediction plays a crucial role. In scenarios such as performance

marketing, to balance the fairness and profitability of campaigns, the manager selects the objects with respect to both the natural probability of conversion and the predicted incremental score. This prevents customers with a high propensity to purchase from not receiving discounts.

Various studies provide solutions from a potential outcomes estimation standpoint. Among them, meta-learning has gained popularity owing to its ability to utilize any machine learning estimator as base model. Single-Model (Athey & Imbens, 2015) and Two-Model (Radcliffe, 2007) approaches are two commonly adopted meta learning paradigms. The first model is trained over an entire space, with the treatment indicator serving as an additive input feature. However, this method's drawback is that a model may not select the treatment indicator if it only uses a subset of features for predictions, such as a tree model. Consequently, the causal effect is estimated as zero for all subjects. An improvement to the Single-Model method is to model the two potential outcomes using two separate models. However, this approach do not utilise the information shared by the control and treated subjects suffer from cumulative errors.

Recently, neural network-based methods becomes the mainstream in uplift modeling. TARNet (Shalit et al., 2017) estimates the causal effect by learning two functions parameterized by two neural networks, utilizing a shared feature representation across treated and control subjects. This method overcomes the limitations of the Two-Models approach through novel network structure. However, since TARNet still does not explicitly consider the heterogeneity of causal effects on the loss objective during training, and derives individual causal effects as a by-product of estimating the outcomes. In scenarios with weak causal effects and very similar label distributions across groups, we observe a performance decline in predicting individual incremental effect. This decline is evident not only with this method but also with other approaches that only use the potential outcome as the learning objective.

In this paper, we propose a new nonparametric method called CausalGBM (Causal Gradient Boosting Machine). The proposed method uses the tree model as the base learner and learns both the causal effects and potential outcomes through loss optimization. The method explicitly computes the contribution of causal effects to the loss function at each split selection, avoiding the defect that the treatment indicator may not be picked up in some scenarios where the causal effects are weak. We additionally introduce the second-order gradient information to improve the convergence speed of the model. Experimental comparisons on four large-scale datasets demonstrate that our model outperforms baseline methods and shows better robustness.

With the widespread use of automated A/B testing platforms in internet scenarios, the dramatically expanding scale of data puts great pressure on model training. In the context of randomized trial, we enhance the computational efficiency of CausalGBM through a multi-objective approximation method. This technique improves computational efficiency and scalability by approximating the binary quadratic optimization problem into a sequence of solving two separate univariate quadratic problems. We have implemented the newly proposed techniques, as well as other baseline approaches, in the UTBoost (Uplift Tree Boosting system) software, which is freely available under the MIT license.

We highlight our contributions as follows:

1. We innovates the uplift tree method that focuses on maximizing heterogeneity by extending the ensemble of trees from bagging to boosting.

2. We fuse potential outcomes and causal effects into the classical GBDT theoretical framework for the first time, and a second-order method is adopted to fit the multi-objective.

3. In the context of randomized trial, we propose a approximate algorithm that reduces the computational complexity of the CausalGBM algorithm.

4. We conduct extensive experiments on large-scale realworld and public datasets. Our models outperform baseline methods and shows better robustness.

The rest of the paper is organized as follows. In Section 2, we review the existing literature on uplift modeling. Section 3 introduces the causal inference framework and notations. Sections 4 and 5 explain the uplift boosting algorithms. In Section 6, we describe the experiments and report the results. Finally, Section 7 presents our conclusions.

2 Related Work

Existing works on uplift modeling can be categorized into different frameworks. One common framework is meta learning, which utilizes various machine learning estimators as base learners to estimate treatment effect (Athey & Imbens, 2015; Radcliffe, 2007; Jaskowski & Jaroszewicz, 2012). Notable examples of meta learning paradigms include Single-Model and Two-Model approaches. However, when the sizes of the two groups are uneven, the models may perform inconsistently, which could undermine the accuracy of causal effect estimates. To address this issue, the X-learner was proposed, which leverages information learned from the control group to improve the estimation of the treated group and vice versa (Künzel et al., 2019).

In the realm of neural networks, representation learning has become a key strategy for addressing biases in treatment distribution, a technique that is also applied in uplift modeling. These methods utilize networks with layers that are either specific to each group or shared between them. An example is the Balancing Neural Network (BNN) (Johansson et al., 2016), which aims to reduce the differences between the treatment and control group distributions by applying the Integral Probability Metric (IPM). Building on the BNN, the Treatment-Agnostic Representation Network (TARNet) and Counterfactual Regression (CFR) (Shalit et al., 2017) use a dual-headed structure to facilitate multi-task learning, developing two functions from a common, balanced representation of features.

Tree-based methods have also attracted considerable attention. Rzepakowski & Jaroszewicz (2012) put forward decision tree-based methods for uplift modeling that use one of the following splitting criteria: Kullback-Leibler (KL) divergence, squared Euclidean distance (ED), and chi-squared divergence. The authors compare the KL and ED-based models to a few baseline approaches, including the Two Model method, and found that their proposed methods outperform the baselines in the binary-treatment case. Athey & Imbens (2016) introduced the causal tree algorithm, which relies on a criterion centered on treatment effects for splitting. Moving forward, Rafla et al. (2023) proposed a Bayesian criterion for evaluating uplift decision trees, with the goal of improving their generalizability and reducing the risk of overfitting.

Furthermore, ensemble methods have seen advancements in improving model performance. Many of these methods use tree-based models as their core learners. Inspired by traditional random forests, Guelman et al. (2015) introduced an uplift random forest framework. Similarly, Wager & Athey (2018) presented a causal forest algorithm that estimates treatment effects using causal trees as the base learner. An alternative to a random forest is boosted trees. This type of methods builds up a function approximation by successively fitting weak learners to the residuals of the model at each step. Powers et al. (2018) developed a causal boosting algorithm that employed causal trees as base learners, estimating treatment-specific means rather than treatment effects. This algorithm updates the residuals for each sample in every iteration, which are then used to train new models. However, recalculating residuals for the entire sample set can carry over biases from previous iterations, potentially leading to compounded errors in later models. We have refined this process by limiting residual calculations to the treated group only, reducing dependencies between learners and ensuring that subsequent models have the necessary information to adjust for bias. Furthermore, we have optimized the loss function, extending the reach of the standard gradient boosting algorithm into causal effect estimation and thus narrowing the divide between these two methodologies.

3 Notation and assumptions

Uplift modeling targets on predicting the incremental change in certain outcomes caused by an intervention for each individual (Gubela et al., 2019; Devriendt et al., 2018). Suppose that we have access to n independent and identically distributed training examples $\{(\boldsymbol{X}_i, y_i, w_i)\}_{i=1}^n$, each of which consists of p features $\boldsymbol{X}_i \in \mathcal{X}$, an observed outcome $y_i \in \mathbb{R}$, and a binary treatment indicator w_i indicating that unit i is under treatment ($w_i = 1$) or control ($w_i = 0$).

In uplift modeling literature (Zhang et al., 2021), it is typical to assume that the treatment w_i is randomly assigned (*i.e.* $w \perp \boldsymbol{X}$) and we commonly take the following quantity as the learning objective:

$$\tau(\boldsymbol{x}) = \mathbb{E}[y_i | w_i = 1, \boldsymbol{X_i} = \boldsymbol{x}] - \mathbb{E}[y_i | w_i = 0, \boldsymbol{X_i} = \boldsymbol{x}], \tag{1}$$

which signifies the uplift on y caused by treatment w for the subjective with feature $\boldsymbol{x}$.

Notice that although we expect to interpret $\tau(x)$ from a causal perspective, equation 1 is not defined under a causal framework. In the following, we briefly show that under certain assumptions, estimating the uplift $\tau(x)$ is equal to inferring the individual treatment effect, a causal quantity defined in the potential outcomes framework (Neyman, 1923; Rubin, 1974). Let $y_i(1)$ and $y_i(0)$ be the potential outcomes that would be observed for unit i when $w_i = 1$ and 0, respectively. The individual treatment effect (ITE) is defined as

$$\tau_i \coloneqq y_i(1) - y_i(0), \tag{2}$$

and $\mathbb{E}[y_i(1) - y_i(0)|\boldsymbol{X}_i]$ is the best estimator in terms of the mean squared error (Künzel et al., 2019). In the literature of causal inference (Imbens & Rubin, 2015; Imbens & Wooldridge, 2009), it is common to assume the following Assumptions 1-2.

Assumption 1 (consistency) $y_i = y_i(1)w_i + y_i(0)(1 - w_i)$

Assumption 2 (Ignorability) $\{y_i(1), y_i(0)\} \perp w_i|X_i$

Under Assumptions 1-2, we have

$$\mathbb{E}[y(w)|\boldsymbol{X}] \overset{\text{Asp } 1}{=} \mathbb{E}[y(w)|\boldsymbol{X}, w] = \mathbb{E}[y(1)w + y(0)(1 - w)|\boldsymbol{X}, w] \overset{\text{Asp } 2}{=} \mathbb{E}[y|\boldsymbol{X}, w],$$

where the 1st and 3rd equation relies on Assumption 1 and 2, respectively, and the 2nd equation can be verified by taking $w = 1, 0$. Then we have $\boldsymbol{E}[y(1) - y(0)|\boldsymbol{X}] = \boldsymbol{E}[y|\boldsymbol{X}, w = 1] - \boldsymbol{E}[y|\boldsymbol{X}, w = 0]$

To summarize, uplift modeling for $\tau(\boldsymbol{x})$ in equation 1 is equivalent to estimating the ITE defined in equation 2 under Assumptions 1-2. Further, if we replace the condition $\{\boldsymbol{X} = \boldsymbol{x}\}$ in equation 1 to a more rough granularity such as $\{\boldsymbol{X} \in \mathcal{X}_t\}$ (*e.g.*, restricting the sample in a leaf node t), then the equivalence between Uplift and ITE requires w being randomly assigned, a stronger one than the ignorability Assumption 1. In the following, we focus on the setting with randomly assigned treatment.

4 Tree Boosting for Treatment Effect Estimation

Our first proposed method adopts a sequential learning approach to fit causal effects without focusing on potential outcomes. As the splitting criterion in training process is similar to the standard delta-delta-p (DDP) algorithm (Hansotia & Rukstales, 2002), which aims to maximize the difference of causal effect between the left and right child nodes, once the boosting method is overlooked. We refer to it as TDDP (Transformed DDP).

4.1 Ensemble Learning via Transformed Label

Taking the decision tree as the base learner, we use a sequence of decision trees to predict the uplift $\tau(\boldsymbol{x})$. As the uplift can not be observed for each sample unit, ensemble method for $\tau(\boldsymbol{x})$ differs from the common supervised-learning scenarios. We explicitly derive the optimizing target for uplift estimation in each iteration of the boosting.

Let $T(\boldsymbol{x}; \theta_j)$ denote a tree model with θ_m denoting its parameters including the way of partitions and estimation within leaf nodes. We aim to predict the uplift via

$$\widehat{\tau}(\boldsymbol{x}) = \sum_{j=1}^{M} T(\boldsymbol{x}; \theta_j). \tag{3}$$

We learn $T(\boldsymbol{x}, \theta_j)$ sequentially to minimize the loss related to $\widehat{\tau}(\boldsymbol{x})$

Let $u_m = \sum_{j=1}^{m} T(\boldsymbol{x}, \theta_j)$ denote the sum of previous m trees. Given u_m at the m step, the current loss is $\mathcal{L}(\tau(x), u_m(\boldsymbol{x}))$, the gradient is

$$\boldsymbol{g}_m \coloneqq \frac{\partial \mathcal{L}(\tau(\boldsymbol{x}), u_m(\boldsymbol{x}))}{\partial u_m(\boldsymbol{x})}$$

Then $-\boldsymbol{g}_m$ suggest a local direction where the loss can be further decreased at the current predictor u_m. Therefore, in a greedy way, we learn $T(\boldsymbol{x};\theta_{m+1})$ to fit $\boldsymbol{g}_m$.

Particularly, in the setting of uplift modeling, the quadratic loss at the $m-$step is

$$\mathcal{L}(\tau(\boldsymbol{x}),u_m(\boldsymbol{x}))=\frac{1}{2}\Big\{\mathbb{E}[y|\boldsymbol{X}=\boldsymbol{x},w=1]-\mathbb{E}[y|\boldsymbol{X}=\boldsymbol{x},w=0]-u_m(x)\Big\}^2$$

where the coefficient $\frac{1}{2}$ is only for the ease of gradient formulation. Then the opposite value of the gradient is

$$-\frac{\partial\mathcal{L}(\tau(\boldsymbol{x}),u_m(\boldsymbol{x}))}{\partial u_m(\boldsymbol{x})}=\mathbb{E}[y|\boldsymbol{X}=\boldsymbol{x},w=1]-\mathbb{E}[y|\boldsymbol{X}=\boldsymbol{x},w=0]-u_m(\boldsymbol{x})=\mathbb{E}[y-u_m(\boldsymbol{x})|\boldsymbol{X}=\boldsymbol{x},w=1]-\mathbb{E}[y|\boldsymbol{X}=\boldsymbol{x},w=0].$$

Therefore, in constructing the $(m+1)$-th tree $T(\boldsymbol{x};\theta_{m+1})$, we may transform the outcome from the original y_i as $y_i-u_m(\boldsymbol{x})$ for the units with $w_i=1$ in the treated group while keep the outcome unchanged for the control group. Then we build the tree based on the transformed outcomes.

To summarize, Algorithm 1 outlines the overall training procedures for TDDP, where the *split criterion* temporarily serve as placeholders and the details will be introduced in the following subsection 4.2.

Algorithm 1 Gradient Tree Boosting for Uplift Modeling

Input: Data: $\mathcal{D}=\{(\boldsymbol{X}_i,y_i,w_i)\}_{i=1}^N$, Shrinkage rate: α
Output: $u_M=\sum_{m=1}^M T(\boldsymbol{x};\theta_m)$

Set $u_0(\boldsymbol{x})=0$.
for $m=1,\cdots,M$ **do**
  Set $y_i=y_i-T(\boldsymbol{x},\theta_{m-1})$ for $\{i\mid w_i=1\}$.
  Build Tree Structure: Recursively Partitions $\mathcal{D}^m$:
  while the *stopping rule* is not satisfied **do**
    Select the optimal split s^* in the candidate splits by the *split criterion*.
    Split the current node into child nodes by s^*.
  end while
  Output the Tree Structure T_m
  Obtain Estimator $T(\boldsymbol{x};\theta_m)$**:**
  for each leaf node t_i of T_m **do**
    Get $D^m(t_i)$: the sample units in D^m that falls into t_i.
    Estimate Weight: $\widehat{\tau}_m(t_i)=\bar{Y}_1(D^m(t_i))-\bar{Y}_0(D^m(t_i))$.
    Output the m-th predictor: $T(x;\theta_m)=\alpha\widehat{\tau}_m(t_{T_m}(x))$, where $t_{T_m}(x)$ denotes the leaf node that x belongs to in T_m.
  end for
end for

4.2 Tree Construction Method

In this subsection, we introduce the split criterion, the core aspect of tree construction.

Split Criterion Recall that in the conventional CART algorithm (Breiman et al., 2017), the optimal split is selected in a greedy way to minimize the mean squared errors (MSE) at each internal node of the regression tree. However, this is not directly applicable because the unit-level uplift τ_i can not be observed (*each unit can only be under treatment or control, rather than both*). Only in a group of units, treated and control are both available, and therefore we may compute the aggregation statistics of uplift such as average or variance. Therefore, we need to convert the original unit-level split criterion into a feasible version that relies on the aggregation-level quantity of uplift.

In the next, we will show that minimizing the MSE is equivalent to maximizing the gaps between the average uplift within the split nodes. Consider the split selection at an internal root node t with data

$D_t \coloneqq \{\boldsymbol{X}_i, y_i, w_i\}_{i=1}^{n_t}$. Let s denote a split, s_L and s_R denote the indices set in the left and right child nodes with sub-sample size n_L and n_R, respectively, under the split s. For example, suppose $s = \{x_j = a\}$ for a numeric variable x_j, then $s_L = \{i|X_{ij} \leq a\}$ and $s_R = \{i|X_{ij} > a\}$. Let $\bar{\tau}_L \coloneqq \sum_{i \epsilon s_L} \frac{\tau_i}{n_L}$ and $\bar{\tau}_R \coloneqq \sum_{i \epsilon s_R} \frac{\tau_i}{n_R}$ denote the average uplift in the left and right child nodes, respectively. Then we have the following proposition

Proposition 1 *Minimizing the mean squared errors of τ_i in the split nodes is equivalent to maximizing the difference between the average uplift within the left and right child nodes, i.e.,*

$$\underset{s}{argmin} \Big\{ \sum_{i|i \epsilon s_L} \big(\tau_i - \bar{\tau}_L\big)^2 + \sum_{i|i \epsilon s_R} \big(\tau_i - \bar{\tau}_R\big)^2 \Big\} = \underset{s}{argmax} \Big\{ \frac{n_L n_R}{n} (\bar{\tau}_L - \bar{\tau}_R)^2 \Big\}.$$

Proposition 1 guides us to a practical split criterion for the uplift modeling as both $\bar{\tau}_L$ and $\bar{\tau}_R$ are aggregated values that can be estimated from data. Taking $\bar{\tau}_L$ as an example, the definition of $\bar{\tau}_L$ involves $\{y_i(1), y_i(0)\}$ as shown in equation equation 4,

$$\bar{\tau}_L = \frac{\sum_{i \epsilon s_L} y_i(1) - y_i(0)}{n_L} = \bar{Y}_L(1) - \bar{Y}_L(0). \tag{4}$$

Under randomly assigned treatment, $\bar{Y}(1)$ and $\bar{Y}(0)$ can be estimated by the sample average of y in the treatment and control groups, respectively. Therefore, the optimal split s^* is selected by the following rule:

$$s^* = \arg\max_s \Big\{ \frac{n_L n_R}{n} \big[\big(\bar{Y}_L^1 - \bar{Y}_L^0\big) - \big(\bar{Y}_R^1 - \bar{Y}_R^0\big) \big]^2 \Big\}, \qquad (\textit{split criterion})$$

where $Y_L^1 \coloneqq \frac{\sum_{i|i \epsilon s_L, w_i = 1} y_i}{n_L^1}$ with n_L^1 denoting the number of treated units in the left child node, and Y_L^0, Y_R^1, Y_R^0 are defined similarly.

Note that TDDP is not strictly a gradient boosting algorithm, because there is no loss function for which we are evaluating the gradient at each step. Rather, TDDP is an adaptation of gradient boosting on the observed outcomes and encourages weak learners to find treatment effect heterogeneities.

5 Causal Gradient Boosting Machine

The TDDP algorithm used to identify the heterogeneity of causal effects skips the fitting of potential outcomes and has limitations in scenarios where potential outcomes are still needed for decision making. Meanwhile, in order to emphasize the importance of causal individual causal effect estimation in the model, we abandon the previous framework of obtaining causal effect estimates derived from outcome prediction. We propose to use CausalGBM (Causal Gradient Boosting Machine) to fit causal effects and outcomes in a single learner. This approach extends the standard gradient boosting algorithm to the field of causal effect estimation, thus bridging the gap between the two classes of methods.

5.1 Learning Objective

In order to realize the simultaneous estimation of the two objectives, we split the original single learning task. Under Assumption equation 1 and Equation equation 2, we can conduct that:

$$y_i = y_i(1) w_i + y_i(0)(1 - w_i) = w_i \tau_i + y_i(0) \tag{5}$$

This equation reveals that for treated samples, the observed labels are equal to the sum of the potential outcomes and the individual causal effects, whereas for control samples, the observed label are equal to the potential outcomes, and enables us to learn both potential outcomes and individual causal effects indirectly from fitting the observation y_i.

For a given data set with n samples and p features, we use a tree ensemble model which contains $2M$ additive functions to predict the output:

$$\hat{y_i} = \sum_{m=1}^{M} f_m(X_i) + w_i \tau_m(X_i), \;\; f_m, \tau_m \in \mathcal{F}$$

where $\mathcal{F} = \{f(X) = v_{q(X)}, \tau(X) = u_{q(X)}\}(q : \mathbb{R}^p \to T, v \in \mathbb{R}^T, u \in \mathbb{R}^T)$ represents the regression trees space. Here q maps each example to the corresponding leaf index in each tree, and T refers to the number of leaves in the tree. Note that leaf weights comprise both u and v in this framework, which is significantly different from the classical regression tree. We will use the decision rules in the trees (given by q) to classify instance to leaves and compute the final predictions by summing up the scores by Equation equation 5 in the corresponding leaves (given by u, v). To learn the set of functions that are employed in the ensemble model, we minimize the following objective function:

$$\mathcal{L}(\theta) = \sum_i l(y_i, \hat{y}_i)$$

Here l is a differentiable convex loss function that measures the difference between the prediction and the observed label. Using the binary decision tree as a meta-learner, we train the ensemble model sequentially to minimize loss. In other words, let ${\hat{y}_i}^t$ be the prediction of the i-th instance at the t-th iteration, we add $f_t + w_i\tau_t$ to minimize the following objective:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, {\hat{y}_i}^{(t-1)} + f_t(x_i) + w_i\tau_t(x_i)) = \sum_{i=1}^{n} l(y_i, \hat{y}_i(0)^{(t-1)} + f_t(x_i) + w_i({\hat{\tau}_i}^{(t-1)} + \tau_t(x_i)))$$

We can conduct the second-order approximation to quickly optimize the objective.

$$\begin{aligned}\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n}[l(y_i, {\hat{y}_i}^{(t-1)}) + \frac{\partial l(y_i, \hat{y}^{(t-1)})}{\partial \hat{\tau}_{t-1}}\tau_t(x_i) + \frac{\partial l(y_i, \hat{y}^{(t-1)})}{\partial \hat{f}_{t-1}}f_t(x_i) + \frac{1}{2}\frac{\partial^2 l(y_i, \hat{y}^{(t-1)})}{\partial \hat{f}_{t-1}^2}f_t^2(x_i) \\ + \frac{1}{2}\frac{\partial^2 l(y_i, \hat{y}^{(t-1)})}{\partial \hat{\tau}_{t-1}^2}\tau_t^2(x_i) + \frac{\partial^2 l(y_i, \hat{y}^{(t-1)})}{\partial \hat{f}_{t-1}\partial \hat{\tau}_{t-1}}f_t(x_i)\tau_t(x_i)]\end{aligned}$$

Under the setting that $w_i \in [0, 1]$, we can remove the constant terms to obtain the following simplified objective at step t:

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n}[w_i g_i \tau_t(x_i) + \frac{1}{2}w_i h_i \tau_t^2(x_i) + g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i) + w_i h_i \tau_t(x_i) f_t(x_i)]$$

where $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ and $h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ are first and second order gradient statistics on the loss function. Note that they are defined in the same way as standard gradient trees.

Define $I_j = \{i | q(X_i) = j\}$ as the instance set of leaf j, we can rewrite the above equation as:

$$\tilde{\mathcal{L}}^{(t)} = \sum_{j=1}^{T}[(\sum_{i \in I_j} w_i g_i + w_i h_i f_j)\tau_j + \frac{1}{2}(\sum_{i \in I_j} w_i h_i)\tau_j^2 + (\sum_{i \in I_j} g_i)f_j + \frac{1}{2}(\sum_{i \in I_j} h_i)f_j^2]$$

We can further derive the optimal values for f_j and τ_j of this dual quadratic function and the corresponding optimal weights v^* and u^*. However, it is important to note that solving for both weights simultaneously will result in a significant decrease in the computing efficiency of the algorithm, compared to the standard regression tree, which has a simpler analytic solution for the quadratic function, during training process. This constrains the practical application of the scheme. We will introduce an approximation method to solve this difficulty in the next section.

5.2 Multi-objective Approximation

We point that if all w_i are equal to 0, i.e., the data set contains only control samples, the above equation is identical to the optimization objective of the regression tree, and we can compute the optimal v_0^* in that specific context. Under the setting that treatments are assigned randomly, we futher assume that $v^* = v_0^*$ on each leaf, which enables us to derive the optimal weights v^* with control instances. After that, the objective function degenerates to a simple quadratic function whith one variable and we can solve optimal u^*. We can compute the optimal weights by

$$v_j^* = -\frac{\sum_{i \in I_j^0} g_i}{\sum_{i \in I_j^0} h_i}, \quad u_j^* = -\frac{\sum_{i \in I_j} w_i g_i + w_i h_i v_j^*}{\sum_{i \in I_j} w_i h_i} = -\frac{\sum_{i \in I_j^1} g_i + h_i v_j^*}{\sum_{i \in I_j^1} h_i}$$

where $I_j^0 = \{i|q(X_i) = j, w_i = 0\}$ is the control instance set and $I_j^1 = \{i|q(X_i) = j, w_i = 1\}$ is the treated instance set of leaf j. It is obvious that, after obtaining v^* from the control group, u^* is only related to the treated samples. This innovation reduces the original solution process to the sequential solving of two quadratic equations in one variable, which is much simpler, and allows the CausalGBM algorithm is scalable to multiple treatment scenarios with very low additional computational resources, since u^* is computed based on the samples of the corresponding group independently of the other groups. We then calculate the corresponding optimal loss by

$$\tilde{\mathcal{L}}_{global}^{(t)}(q) = \sum_{j=1}^{T}[(\sum_{i\in I_j} g_i)v_j^* + \frac{1}{2}(\sum_{i\in I_j} h_i)(v_j^*)^2 - \frac{(\sum_{i\in I_j^1} g_i + h_i v_j^*)^2}{2\sum_{i\in I_j^1} h_i}]$$

Note that the value is obtained by computing all instances on leaf I_j to ensure the global loss is optimized under this approximation method. We provide two other different forms of the optimum value, which are defined as

$$\tilde{\mathcal{L}}_{local}^{(t)} = \sum_{j=1}^{T}[(\sum_{i\in I_j^1} g_i)v_j^* + \frac{1}{2}(\sum_{i\in I_j^1} h_i)(v_j^*)^2 - \frac{(\sum_{i\in I_j^1} g_i + h_i v_j^*)^2}{2\sum_{i\in I_j^1} h_i}]$$

and

$$\tilde{\mathcal{L}}_{\tau}^{(t)} = \sum_{j=1}^{T} - \frac{(\sum_{i\in I_j^1} g_i + h_i v_j^*)^2}{2\sum_{i\in I_j^1} h_i}$$

where $I_j^1 = \{i|q(x_i) = j, w_i = 1\}$ is the treated instance set of leaf j. The first form is only influenced by the treatment group sample, and the other only by the causal effect. We present comparisons of the impact of the different forms for the model in the experimental section.

5.3 Greedy Algorithm for Tree Construction

It is impossible to enumerate all tree structures to find the one with the minimum loss value. Instead, we use a greedy algorithm that recursively divides the child nodes starting from a single parent node. Define:

$$\tilde{\mathcal{L}}_{leaf}^{(t)} = (\sum_{i\in I_{leaf}} g_i)f_j^* + \frac{1}{2}(\sum_{i\in I_{leaf}} h_i)(f_j^*)^2 - \frac{(\sum_{i\in I_{leaf}^1} g_i + h_i f_j^*)^2}{2\sum_{i\in I_{leaf}^1} h_i}$$

Assume that I_L and I_R are the instance sets of left and right nodes after the split. Letting $I = I_L \cup I_R$, then the loss reduction after the split is given by:

$$\tilde{\mathcal{L}}_{split} = \tilde{\mathcal{L}}_I^{(t)} - (\tilde{\mathcal{L}}_L^{(t)} + \tilde{\mathcal{L}}_R^{(t)})$$

The above function will be used to evaluate the candidate split points. Compared with the xgboost algorithm, CausalGBM redefines the computation formulas for weights and evaluation functions in uplift modeling problems. As for the construction of the tree, we follow the computational framework of xgboost but have adjusted the calculation methods pertaining to weights and evaluations.

6 EXPERIMENTS

In order to thoroughly evaluate our proposed methods, we conduct extensive experiments on three large-scale real-world datasets and a synthesized dataset to answer the following research questions:

- **Q1**: How is the overall performance of TDDP and CausalGBM compared with baseline methods?
- **Q2**: How does changing the ensemble mode from boosting to bagging influence model performance?
- **Q3**: Does the performance of the CasualGBM vary when using different loss computation methods?

Table 1: The basic statistics of datasets used in the paper.

Metrics	CRITEO-UPLIFT$_{1M}$	HILLSTROM	VOUCHER-UPLIFT	SYNTHETIC$_m$
Size	1,000,000	42,693	371,730	200,000
Features	12	8	2076	m
Avg. Label	0.047	0.129	0.356	0.600
Treatment Ratio	0.85	0.50	0.85	0.50
Relative Avg. Uplift (%)	26.7	42.6	2.0	50.0

6.1 Experiment Setup

6.1.1 Datasets

We evaluate our proposed models on three large-scale real-world datasets and a synthesized dataset. A summary of these datasets is given in Table 1. Here, we briefly introduce them as follows:

- HILLSTROM(Hillstrom, 2008): The dataset contains information about 64,000 customers who last purchased within at most twelve months. The customers were evenly distributed in two treatment groups and a control group, where the first treatment group was sent a "Men's Advertisement Email" and the second treatment was sent a "Women's Advertisement Email". The control group received no email. The outcome variables were the visit and conversion status of the customers. We report results on the Women's merchandise e-mail versus no e-mail split as in previous research.

- CRITEO-UPLIFT(Diemert et al., 2018): A public dataset consists of 25M rows, with each row representing a user and 11 anonymized features. It also includes a treatment indicator representing whether the customer received a promotional email, as well as two binary labels indicating the visiting and conversion status of the customer. We randomly selected a subset with 1M instances and used the visit status as the target.

- VOUCHER-UPLIFT: We collect a dataset from a e-commerce platform. About 85% of the users were provided with a reminder message with in-account vouchers, with the outcome variable being whether the vouchers was used in the next seven days. The resulting dataset consists of 371,730 samples, each containing 2076-dimensional sparse features, a binary treatment indicator, and a binary label.

- SYNTHETIC$_m$: In order to compare the performances of various methods under ideal data conditions, we adopted the synthesis method proposed by (Chen et al., 2020). The dataset contains m variables, a binary indicator and a binary label. About 70% of the variables are associated with both potential outcomes and uplift, and the remaining variables are redundant. The data set consisted of 200K samples and 5% of the samples are randomly labeled.

6.1.2 Evaluation Protocols

We perform 10-fold cross-validation and use two widely used metrics in prior work, including area under the ROC curve (AUC) and Qini coefficient (Diemert et al., 2018; Gutierrez & Gérardy, 2017; Zhang et al., 2021) (normalized by prefect Qini score) for evaluation, and we perform a grid search on hyperparameters to search for an optimal parameter set that achieved the best performance on the validation dataset, which consisted of 25% of the training dataset in each fold. The parameters we grid-searched included combinations of maximum depth (3, 4, 5) and number of trees (25,50,100,150). Other hyperparameters in algorithms keep their default values. All neural networks of various deep models consist of 128 hidden units with 3 fully connected layers. L2 regularization is applied to mitigate overfitting with a coefficient of 0.01, and the learning rate is set to 0.001 without decay.

Table 2: Model performance evaluated by Qini coefficient on four datasets with corresponding mean and standard error. "S-", "T-", "TO-" and "URF-" stands for instantiations of single-model, two-model, transformed outcome and uplift random forest approaches, respectively. Note that for Qini larger value is better. Best results of all methods are highlighted in boldface.

Model	HILLSTROM	CRITEO-UPLIFT$_{1M}$	VOUCHER-UPLIFT	SYNTHETIC$_{100}$
	Qini Coefficient ($mean \pm s.e.$)			
S-LGB	0.0616 ± 0.0179	0.0933 ± 0.0160	0.0032 ± 0.0053	0.1812 ± 0.0029
T-LGB	0.0567 ± 0.0168	0.0900 ± 0.0178	0.0014 ± 0.0054	0.1831 ± 0.0024
TO-LGB	0.0377 ± 0.0200	0.0941 ± 0.0201	0.0048 ± 0.0061	0.1832 ± 0.0044
X-Learner	0.0619 ± 0.0150	0.0929 ± 0.0246	0.0029 ± 0.0072	0.1824 ± 0.0030
TARNet	0.0636 ± 0.0203	0.0935 ± 0.0109	0.0045 ± 0.0076	0.1803 ± 0.0049
CFRNet$_{wass}$	0.0635 ± 0.0218	0.0909 ± 0.0171	0.0042 ± 0.0082	0.1829 ± 0.0035
CFRNet$_{mmd}$	0.0629 ± 0.0257	0.0923 ± 0.0146	0.0047 ± 0.0073	0.1808 ± 0.0047
Causal Forest	0.0617 ± 0.0139	0.0933 ± 0.0113	0.0055 ± 0.0041	0.1395 ± 0.0039
URF-Chi	0.0623 ± 0.0167	0.0925 ± 0.0125	0.0062 ± 0.0073	0.1003 ± 0.0054
URF-ED	0.0613 ± 0.0183	0.0942 ± 0.0141	0.0070 ± 0.0064	0.1657 ± 0.0057
URF-KL	0.0605 ± 0.0160	0.0926 ± 0.0123	0.0060 ± 0.0055	0.1457 ± 0.0036
URF-DDP	0.0599 ± 0.0161	0.0938 ± 0.0144	0.0072 ± 0.0060	0.1661 ± 0.0054
TDDP	0.0576 ± 0.0123	0.0884 ± 0.0179	0.0088 ± 0.0059	0.1836 ± 0.0045
CausalGBM	**0.0643 ± 0.0246**	**0.0971 ± 0.0142**	**0.0108 ± 0.0042**	**0.1863 ± 0.0039**

6.2 Overall Performance Comparison (Q1)

To verify that our proposed methods can make the uplift prediction model more accurate, we compare TDDP and CausalGBM with different types of baselines and show their prediction performance on four large-scale datasets in Table 2. Here, we summarize key observations and insights as follows:

- **CausalGBM's superior performance**: Our proposed model CausalGBM outperforms all different baseline methods across all datasets. Specifically, it achieves relative performance gains of 1.1%, 3.1%, 22.7%, and 1.5% on four datasets, respectively, comparing to the best baseline. Further, Qini reflects the model's ability to give high predicted probabilities of persuasion to those who are actually more likely to be persuaded, and the improvement implies that our proposed model more accurately finds the target population for which the treatment is effective.

- **CausalGBM's robustness across different scenarios**: Comparing the results between the four datasets with different scales, we find that many baseline models are not robust in different scenarios. For example, URF-based methods perform significantly better on the HILLSTROM and CRITEO-UPLIFT$_{1M}$ than on the SYNTHETIC$_{100}$. A plausible reason is that the first two datasets are much smaller than the last two datasets in terms of feature dimensions. As a result the model's fitting ability is not the key on the first two datasets. In contrast, CausalGBM achieves the best performance in all datasets, which demonstrated our model's robustness across different scenarios.

  On the VOUCHER-UPLIFT dataset, we observe that deep learning and partial metamodeling-type methods driven by fitting potential outcomes under different treatments are weaker than tree models that focus on the heterogeneity of causal effects. Recall that, as seen in Table 1, this dataset has the weakest significance of causal effects of all the datasets. This poses a challenge to methods that focus only on potential outcomes, i.e., the label distributions of the different groups are very similar, which would result in the individual causal effects predicted by this class of methods close to zero. URF-based methods at minimizing heterogeneous differences remain well performing. Since the model of CausalGBM is trained by simultaneously computing both potential outcomes and causal effects, it still performs robustly on the weak causal effect dataset compared to the approaches driven only by potential outcomes.

- **An analysis of the volatility for TDDP on different datasets**: Comparing TDDP and URF-DDP, the only difference between the two methods lies in the ensemble learning method, the former employs boosting, while the latter takes bagging. Comparing the performances of the two methods on four datasets, TDDP performs better on the dataset with high-dimensional features, while URF-DDP performs better on the dataset consisting of low-dimensional features. A plausible reason is that TDDP overfits the data in datasets with low-dimensional features, meanwhile, URF-DDP does not fit the data in high-dimensional datasets adequately.

6.3 Ablation Study

We compare how the ensemble learning method and the loss form contribute to the predictive performance of the model in this section. In order to better visualize the experimental findings and to prevent the derivation of conclusions from misleading information, we select the synthetic dataset and divide 50% as the training set and the remaining part as the test.

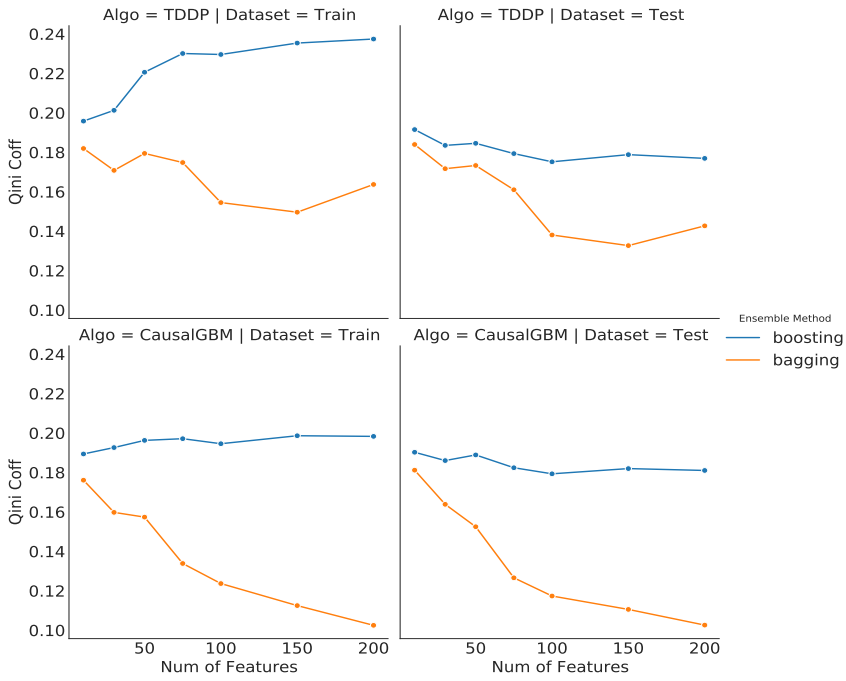

Figure 1: The ablation study on different ensemble methods.

6.3.1 How does ensemble method facilitate the prediction performance? (Q2)

To answer this question, we compare the prediction results of TDDP and CausalGBM with its ablation version, which uses bagging as the ensemble mode. This blocks the updating of the gradient for the CausalGBM algorithm, where the gradient is only computed before the training based on a uniform initial value. The gradient of each sample is constant throughout the training process. Overall, we observe that boosting significantly improves the model's capability to fit the training data compared to bagging across the two algorithms, and the gap widens as the feature dimensionality grows. Similar conclusions hold for the test dataset. This suggests that boosting provides a significant improvement in model performance, and that this approach is particularly suitable for high-dimensional datasets. On the low-dimensional dataset, the difference between the two methods is relatively small. In addition, we observe that the boosting version of the TDDP method is more likely to overfit the training data with increasing feature size than CausalGBM, leading to a weak generalization performance. In light of this finding, we suggest that tree models that

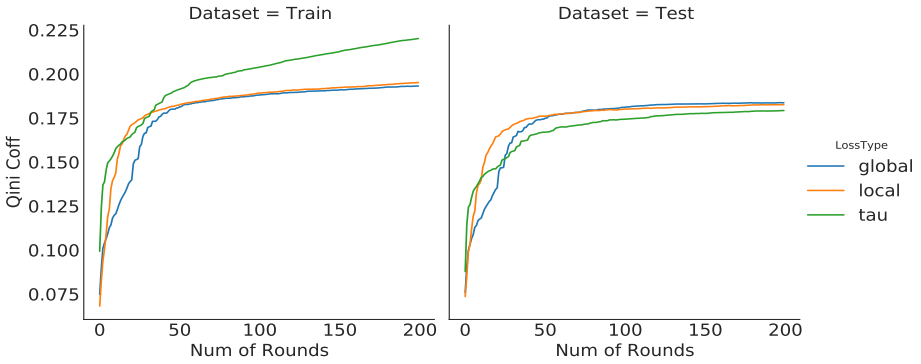

Figure 2: The ablation study on different loss types of CausalGBM.

focus on the heterogeneity of local causal effects require regular methods to avoid overfitting compared to the GBM approach that optimizes the global loss function.

6.3.2 Which type of loss is appropriate for CausalGBM? (Q3)

We compare the performance of the models under different forms of loss. On the training dataset, $\tilde{\mathcal{L}}_{\tau}$, which focuses only on causal effects, shows superiority in terms of optimization speed and optimal score, significantly outperforming the other two formulations. $\tilde{\mathcal{L}}_{local}$ significantly outperforms $\tilde{\mathcal{L}}_{global}$ in terms of initial improvement optimization speed, and both forms perform similarly in the middle and late stages of training. However, on the test dataset, a contrasting pattern emerges. $\tilde{\mathcal{L}}_{\tau}$ significantly underperforms the other two formulations in the middle stage. After 200 iterations, $\tilde{\mathcal{L}}_{global}$ achieves similarity to $\tilde{\mathcal{L}}_{local}$ in terms of the Qini Coefficient, with both surpassing the performance of $\tilde{\mathcal{L}}_{\tau}$.

We emphasize that if we focus only on the contribution of the causal effect prediction to the loss, the model will not be able to determine whether the prediction of the potential outcome is accurate. Since the causal effects are computed after the potential outcomes are obtained, any bias in the prediction of the outcomes will be transferred to the estimation of the causal effects. This invariably impacts model's generalization performance. Regarding $\tilde{\mathcal{L}}_{local}$, we highlight that excluding the control group is equivalent to increasing the contribution of the causal effect to the loss, which will force the model to pay more attention to it. There is no significant difference between the two in terms of final predictive performance. However, $\tilde{\mathcal{L}}_{local}$ may outperform $\tilde{\mathcal{L}}_{global}$ in situations where the sample size of the experimental group is much smaller than that of the control group. In such case, the loss value will come predominantly from the control samples. By employing $\tilde{\mathcal{L}}_{local}$, the model effectively mitigates the risk of disregarding causal effects.

7 Conclusion

In this paper we formulate two novel boosting methods for the uplift modeling problem. The first algorithm we proposed follows the idea of maximizing the heterogeneity of causal effects. This algorithm differs from the standard regression tree, because there is no loss function for which we are evaluating the gradient at each step, in order to minimize this loss. In contrast, the second algorithm we proposed, CausalGBM, fits both potential outcomes and causal effects by optimizing the loss function. This is similar to standard supervised learning methods. We demonstrate that our proposed techniques outperform the baseline model on large-scale real datasets, where the CausalGBM algorithm shows excellent robustness, while the TDDP algorithm needs to blend in some regularization methods to prevent the model from overfitting the training data. The UTBoost package, which is both open source and ready-to-use, integrates our newly proposed algorithms.

Furthermore, both algorithms we propose can be directly extended to multi-treatment problems, and we will continue to explore the applicability of the algorithms with appropriate datasets in the future.

References

S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects: Table 1. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

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

Leo Breiman. Random forests. *Machine Learning*, 45, 2001.

Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. *Classification and regression trees*. Routledge, 2017.

Huigang Chen, Totte Harinen, Jeong-Yoon Lee, Mike Yung, and Zhenyu Zhao. Causalml: Python package for causal machine learning, 2020.

T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. *ACM*, 2016.

Floris Devriendt, Darie Moldovan, and Wouter Verbeke. A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics. *Big Data*, 6(1):13, 2018.

Eustache Diemert, Artem Betlei, Christophe Renaudin, and Massih-Reza Amini. A large scale benchmark for uplift modeling. In *KDD*, 2018.

R. Gubela, A. Beque, F. Gebert, and S. Lessmann. Conversion uplift in e-commerce: A systematic benchmark of modeling strategies. *International Journal of Information Technology & Decision Making (IJITDM)*, 2019.

Leo Guelman, Montserrat Guillén, and Ana M Pérez-Marín. Random forests for uplift modeling: an insurance customer retention case. In *International conference on modeling and simulation in engineering, economics and management*, pp. 123–133. Springer, 2012.

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

Pierre Gutierrez and Jean-Yves Gérardy. Causal inference and uplift modelling: A review of the literature. In *International conference on predictive applications and APIs*, pp. 1–13. PMLR, 2017.

Behram Hansotia and Brad Rukstales. Incremental value modeling. *Journal of Interactive Marketing*, 16 (3):35–46, 2002.

Kevin Hillstrom. The minethatdata e-mail analytics and data mining challenge. *MineThatData blog*, 2008.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Guido W Imbens and Jeffrey M Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86, 2009.

Maciej Jaskowski and Szymon Jaroszewicz. Uplift modeling for clinical trial data. In *ICML Workshop on Clinical Data Analysis*, volume 46, pp. 79–95, 2012.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pp. 3020–3029. PMLR, 2016.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

Jerzy S Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9.(tlanslated and edited by dm dabrowska and tp speed, statistical science (1990), 5, 465-480). *Annals of Agricultural Sciences*, 10:1–51, 1923.

Scott Powers, Junyang Qian, Kenneth Jung, Alejandro Schuler, Nigam H Shah, Trevor Hastie, and Robert Tibshirani. Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in medicine*, 37(11):1767–1787, 2018.

Nicholas Radcliffe. Using control groups to target on predicted lift: Building and assessing uplift model. *Direct Marketing Analytics Journal*, pp. 14–21, 2007.

Mina Rafla, Nicolas Voisine, and Bruno Crémilleux. Parameter-free bayesian decision trees for uplift modeling. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 309–321. Springer, 2023.

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.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pp. 3076–3085. PMLR, 2017.

Michał Sołtys and Szymon Jaroszewicz. Boosting algorithms for uplift modeling. *arXiv preprint arXiv:1807.07909*, 2018.

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

W. Zhang, J. Li, and L. Liu. A unified survey of treatment effect heterogeneity modelling and uplift modelling. *ACM Computing Surveys (CSUR)*, 2021.