Parameter-Free Bayesian Decision Trees for Uplift Modeling

Mina Rafla, Nicolas Voisine, Bruno Crémilleux

▶ **To cite this version:**

Mina Rafla, Nicolas Voisine, Bruno Crémilleux. Parameter-Free Bayesian Decision Trees for Uplift Modeling. 27th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2023), May 2023, Osaka, Japan. pp.309-321, 10.1007/978-3-031-33377-4_24 . hal-04173558

HAL Id: hal-04173558
https://hal.science/hal-04173558v1

Submitted on 29 Jul 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Parameter-free Bayesian Decision Trees for Uplift Modeling

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

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

Abstract. Uplift modeling aims to estimate the incremental impact of a treatment, such as a marketing campaign or a drug, on an individual's behavior. These approaches are very useful in several applications such as personalized medicine and advertising, as it allows targeting the specific proportion of a population on which the treatment will have the greatest impact. Uplift modeling is a challenging task because data are partially known (for an individual, responses to alternative treatments cannot be observed). In this paper, we present a new tree algorithm named UB-DT designed for uplift modeling. We propose a Bayesian evaluation criterion for uplift decision trees T by defining the posterior probability of T given uplift data. We transform the learning problem into an optimization one to search for the uplift tree model leading to the best evaluation of the criterion. A search algorithm is then presented as well as an extension for random forests. Large scale experiments on real and synthetic datasets show the efficiency of our methods over other state-of-art uplift modeling approaches.

Keywords: Uplift Modeling · Decision trees · Random Forests · Bayesian methods · Machine Learning · Treatment Effect Estimation

1 Introduction

Uplift modeling aims to estimate the incremental impact of a treatment, such as a marketing campaign or a drug, on an individual's behavior. These approaches are very useful in several applications such as personalized medicine and advertising, as it allows targeting the specific proportion of a population on which the treatment will have the greatest impact. Uplift estimation is based on groups of people who have received different treatments. A major difficulty is that data are only partially known: it is impossible to know for an individual whether the chosen treatment is optimal because their responses to alternative treatments cannot be observed. Several works address challenges related to the uplift modeling, among which uplift decision tree algorithms became widely used [15,17].

Despite their usefulness, current uplift decision tree methods have limitations such as local splitting criteria. A split criterion decides whether to divide a

terminal node. However these splits are independent to each other and a pruning step is then used to ensure generalization and avoid overfitting. Moreover, these methods require parameters to set. In this paper, we present UB-DT (Uplift Bayesian Decision Tree) a parameter-free method for uplift decision tree based on the Bayesian paradigm. Contrary to state-of-art uplift decision tree methods, we define a global criterion designed for an uplift decision tree. A major advantage of a global tree criterion is it allows to get rid of the pruning step, since it acts as a regularization to avoid overfitting. We transform the uplift tree learning problem to an optimization problem according to the criterion. Then a search algorithm is used to find the decision tree that optimizes the global criterion. Moreover our approach is easily extended to random forests and we propose UB-RF (Uplift Bayesian Random Forest). We evaluate both UB-DT and UB-RF to state-of-art uplift modeling approaches through a benchmarking study.

This paper is organized as follows. Section 2 introduces an overview of uplift modeling and related work. Section 3 presents UB-DT. We conduct experiments in Section 4 and conclude in Section 5.

2 Context and literature overview

2.1 Uplift problem formulation

Uplift is a notion introduced by Radcliffe and Surry [11] and defined in Rubin's causal inference models [14] as the *Individual Treatment effect (ITE)*.

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

In practice, we will never observe both $Y_i(Z = 1)$ and $Y_i(Z = 0)$ for a same individual and thus τ_i cannot be directly calculated. However, uplift can be empirically estimated by considering two groups: a treatment group (individual with a treatment) and a control group (without treatment). The estimated uplift of an individual i denoted by $\hat{\tau}_i$ is then computed by using the CATE (Conditional Average Treatment Effect)[14]:

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

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

2.2 Related work

Uplift modeling approaches Uplift modeling approaches are divided into two categories. The first one (called *metalearners*) is made up of methods that take advantage of usual machine learning algorithms to estimate the CATE. One of the most intuitive approaches is the *two-model approach*. It consists of fitting two independent classification models, one for the treated group and another for the control group. The estimated uplift is then the difference between the estimations of the two classification models. While this approach is simple, intuitive and allows the usage of any machine learning algorithm, it has also known weaknesses with particular patterns [12]. The causal inference community has also proposed other metalearners such as X-learner [8], R-Learner and DR-learner [7].

The second category is closer to our work. This category gathers tailored methods for uplift modeling such as tree-based algorithms. Trees are built using recursive partitioning to split the root node to child nodes according to a splitting criterion. [15] defines a splitting criterion that compares the probability distributions of the outcome variable in each of the treatment groups using weighted divergence measures like the Kullback-Leibler (KL), the squared euclidean distance (ED) and the chi-squared divergence. [17] proposes the Contextual Treatment Selection algorithm (CTS) where a splitting criterion directly maximizes a performance measure called the expected performance. Causal machine learning algorithms were also developed such as the Causal Trees algorithm [1] and the Causal Forests [2].

Uplift tree splitting criterion and Bayesian approaches. Building an uplift tree requires to discretize variables to detect areas with homogeneous treatment effects. The global criterion of UB-DT to select a variable on a node takes advantage of on a univariate parameter-free Bayesian approach for density estimation through discretization called UMODL [13]. More precisely, UMODL applies a Bayesian approach to select the most probable uplift discretization model M given the data. This implies finding the model M that maximizes the posterior probability $P(M|Data)$, hence maximizing $P(M) \times P(Data|M)$. Finally, a global criterion within the Bayesian framework for decision trees is given in [16] but it does not deal with uplift.

3 UB-DT: uplift decision tree approach

UB-DT is made up of two ingredients: a global criterion $C(T)$ for a binary uplift decision tree T and a tree search algorithm to find the most probable optimal tree. We start by presenting the structure of an uplift tree model. Then we describe the new global criterion for an uplift decision tree and the algorithm to give the best tree. Finally we show how the approach is straightforwardly extended to random forests.

3.1 Parameters of an uplift tree model T

We define a binary uplift decision tree model T by its structure and the distribution of instances and class values in this structure. The structure of T consists

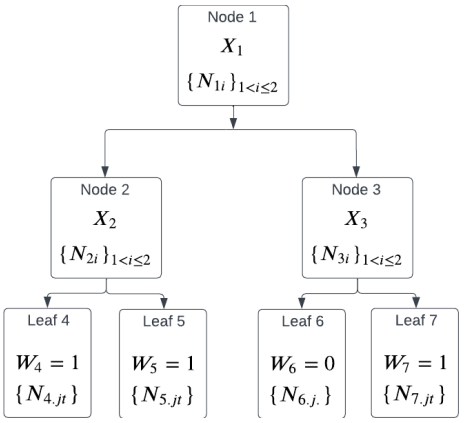


Fig. 1: Example of an uplift tree model. Internal nodes are described by the segmentation variable X_s and the distribution of instances in each of the two children $\{N_{si}\}$. Leaf nodes containing a treatment effect (i.e $W_l = 1$) are described by the class distribution for each treatment. This applies to leaves 4, 5 and 7. Leaf nodes containing no treatment effect (i.e $W_l = 0$) are only described by the class distribution (this is the case of leaf 6).

of the set of internal nodes $\mathbb{S}_T$ and the set of leaf nodes $\mathbb{L}_T$. The distribution of the instances in this structure is described by the partition of the segmentation variable X_s for each internal node s, the class frequency in each leaf node where there is no treatment effect, and the class frequency on each treatment in the leaf nodes with a treatment effect. More precisely, T is defined by:

- the subset of variables $\mathbb{K}_T$ used by model T. This includes the number of the selected variables K_T and their choice among a set of $\mathbb{K}$ variables provided in a dataset, we note $K = |\mathbb{K}|$.
- a binary variable I_n indicating the choice of whether each node n is an internal node ($I_n = 1$) or a leaf node ($I_n = 0$).
- the distribution of instances in each internal node s, which is described by the segmentation variable X_s of the node s and how the instances of s are distributed on its two child nodes.
- a binary variable W_l indicating for each leaf node l if there is a treatment effect ($W_l = 1$) or not ($W_l = 0$). If $W_l = 0$, l is described by the distribution of the output values $\{N_{l.j.}\}_{1\leq j\leq J}$, where $N_{l.j.}$ is the number of instances of output value j in leaf l. If $W_l = 1$, l is described by the distribution of the class values per treatment $\{N_{l.jt}\}_{1\leq j\leq J, 1\leq t\leq 2}$, where $N_{l.jt}$ is the number of instances of output value j and treatment t in leaf l.

These parameters are automatically optimized by the search algorithm (presented in Section 3.4) and not fixed by the user. In the rest of the paper, the following notations $N_{s.}$, $N_{si.}$, $N_{l.}$ and $N_{l..t}$ will additionally be used to respectively designate the number of instances in node s, in the i^{th} child of node s, in the leaf l and treatment t in leaf l.

3.2 Uplift tree evaluation criterion

We now presents the new global criterion $C(T)$ which is an uplift tree model evaluation criterion. UB-DT applies a Bayesian approach to select the most probable uplift tree model T that maximizes the posterior probability $P(T|Data)$.

This is equivalent to maximizing the product of the prior and the likelihood i.e. $P(T) \times P(Data|T)$. Taking the negative log turns the maximization problem into a minimization one: $C(T) = -\log\left(P(T) \times P(Data|T)\right)$, $C(T)$ is the cost of the uplift tree model T. T is optimal if $C(T)$ is minimal. By exploiting the hierarchy of the presented uplift tree parameters and assuming a uniform prior, we express $C(T)$ as follows (cf. Eq. 2):

$$
\begin{aligned}
C(T) = & \underbrace{\log(K+1) + \log\binom{K+K_T-1}{K_T}}_{\text{Variable selection}} \\
& + \underbrace{\sum_{s\in\mathbb{S}_{T_n}} \log 2 + \log K_T + \log(N_{s.}+1) +}_{\text{Prior of internal nodes}} \underbrace{\sum_{l\in\mathbb{L}_T} \log 2}_{\text{Treatment effect W}} \\
& + \underbrace{\sum_{l\in\mathbb{L}_T} \log 2 + \sum_{l\in\mathbb{L}_T} (1-W_l) \log\binom{N_{l.}+J-1}{J-1} + \sum_{l\in\mathbb{L}_T} W_l \sum_t \log\binom{N_{l..t}+J-1}{J-1}}_{\text{Prior of leaf nodes}} \\
& + \underbrace{\sum_{l\in\mathbb{L}_T} (1-W_l) \log \frac{N_{l.}!}{N_{l.1.}!N_{l.2.}!\ldots N_{l.J.}!} + \sum_{l\in\mathbb{L}_T} W_l \sum_t \log \frac{N_{l..t}!}{N_{l.1t}!..N_{l.Jt}!}}_{\text{Tree Likelihood}}
\end{aligned}
\tag{2}
$$

The next section demonstrates Eq. 2.

3.3 $C(T)$: proof of Equation 2

We express the prior and the likelihood of a tree model, resp. $P(T)$ and $P(Data|T)$ according to the hierarchy of the uplift tree parameters. Assuming the independence between all the nodes, the prior probability of an uplift decision tree is thus defined as:

$$
\begin{aligned}
& P(T) = P\left(\mathbb{K}_T\right) \times \\
& \prod_{s\in\mathbb{S}_T} P\left(I_s\right) P\left(X_s \mid \mathbb{K}_T\right) P\left(N_{si.} \mid \mathbb{K}_T, X_s, N_{s.}, I_s\right) \times \\
& P\left(\{W_l\}\right) \times \prod_{l\in\mathbb{L}_T} P\left(I_l\right) \left[(1-W_l) \times p\left(\{N_{l.j}\} \mid \mathbb{K}_T, N_{l.}\right) + W_l \times \prod_t P\left(\{N_{l.jt}\} \mid \mathbb{K}_T, N_{l..t}\right)\right]
\end{aligned}
\tag{3}
$$

The first line is the prior probability of the variable selection, the second line the prior of internal nodes and the third line the prior of the leaf nodes.

Variable selection probability. A hierarichal prior is chosen: first the choice of the number of selected variables K_T, then the choice of the subset $\mathbb{K}_T$ among

$\mathbb{K}$ variables. By using a uniform prior the number K_T can have any value between 0 and K in an equiprobable manner. For the choice of the subset $\mathbb{K}_T$, we assume that every subset has the same probability. Then the prior of the variable selection can be defined as:

$$P(\mathbb{K}_T) = \frac{1}{K+1} \frac{1}{\binom{K + K_T - 1}{K_T}}$$

Prior of internal nodes. Each node can either be an internal node or a leaf node with equal probability. This implies that: $P(I_s) = \frac{1}{2}$

The choice of the segmentation variable is equiprobable between 1 and K_T. We obtain:

$$P(X_s|\mathbb{K}_T) = \frac{1}{K_T}$$

All splits of an internal node s to two intervals are equiprobable. We then obtain:

$$P\left(N_{si.} \mid \mathbb{K}_T, X_s, N_{s.}, I_s\right) = \frac{1}{N_s + 1}$$

Prior of leaf nodes. Similar to the prior of internal nodes, each node can either be internal or a leaf node with equal probability leading to $P(I_l) = \frac{1}{2}$. For each leaf node, we assume that a treatment can have an effect or not, with equal probability, we get:

$$P(\{W_l\}) = \prod_l \frac{1}{2}$$

In the case of a leaf node l where there is not effect of the treatment ($W_l = 0$), UB-DT describes one unique distribution of the class variable. Assuming that each of the class distributions is equiprobable, we end up also with a combinatorial problem:

$$P\left(\{N_{l.j}\} \mid \mathbb{K}_T, N_{l.}\right) = \frac{1}{\binom{N_{l.} + J - 1}{J - 1}}$$

In a leaf node with an effect of the treatment ($W_i = 1$), UB-DT describes two distributions of the outcome variable, with and without the treatment. Given a leaf l and a treatment t, we know the number of instances $N_{l..t}$ Assuming that each of the distributions of class values is equiprobable, we get:

$$P\left(\{N_{l.jt}\} \mid \mathbb{K}_T, N_{l..t}\right) = \frac{1}{\binom{N_{l..t} + J - 1}{J - 1}}$$

Tree likelihood. After defining the tree's prior probability, we establish the likelihood probability of the data given the tree model. The class distributions depend only of the leaf nodes. For each multinomial distribution of the outcome variable (a single or two distinct distributions per leaf depending on whether the treatement has an effect or not), we assume that all possible observed data D_l consistent with the multinomial model are equiprobable. Using multinomial terms, we end up with:

$$P(Data \mid T) = \prod_{l \in L} P(D_l|M)$$
$$\prod_{l \in L} \left[(1 - W_l) \times \frac{1}{N_{l.}!/N_{l.1.}!N_{l.2.}! \ldots N_{l.J.}!} + W_l \times \prod_t \frac{1}{(N_{l..t}!/N_{i.1t}!..N_{i.Jt}!)} \right] \tag{4}$$

By combining the prior and the likelihood (resp. Eq. 3 and 4) and by taking their negative log, we obtain $C(T)$ and thus Eq. 2 is proved.

3.4 Search algorithm

The induction of an optimal uplift decision tree from a data set is NP-hard [10]. Thus, learning the optimal decision tree requires exhaustive search and is limited to very small data sets. As a result, heuristic methods are required to build uplift decision trees. Algorithm 1 (see below) selects the best tree according to the global criterion. Algorithm 1 chooses a split among all possible splits in all terminal nodes only if it minimizes the global criterion of the tree. The algorithm continues as long as the global criterion is improved. Since a decision tree is a partitioning of the feature space, a prediction for a future instance is then the average uplift in its corresponding leaf. This algorithm is deterministic and thus it always leads to the same local optimum. Experiments show the quality of the building trees.

3.5 UB-RF

UB-DT is easily extended to random forests. For that purpose, a split is randomly chosen among all possible splits that improve the global criterion. The number of trees is set by the analyst and the prediction of a forest is the average predictions of all the trees.

4 Experiments

We experimentally evaluate the quality of UB-DT as an uplift estimator and compare UB-DT and UB-RF versus state-of-art uplift modeling approaches [3].

[3] Code, datasets and complementary results are at `https://github.com/MinaWagdi/UB-DT`

We use the following state-of-art methods: (1) metalearners: two-model approach (2M), X-Learner and R-Learner, each with Xgboost; (2) uplift trees: CTS-DT,KL-DT, Chi-DT, ED-DT; (3) uplift random forests: CTS-RF,KL-RF, Chi-RF, ED-RF [15]; (4) and causal forests (all forest methods were used with 10 trees).

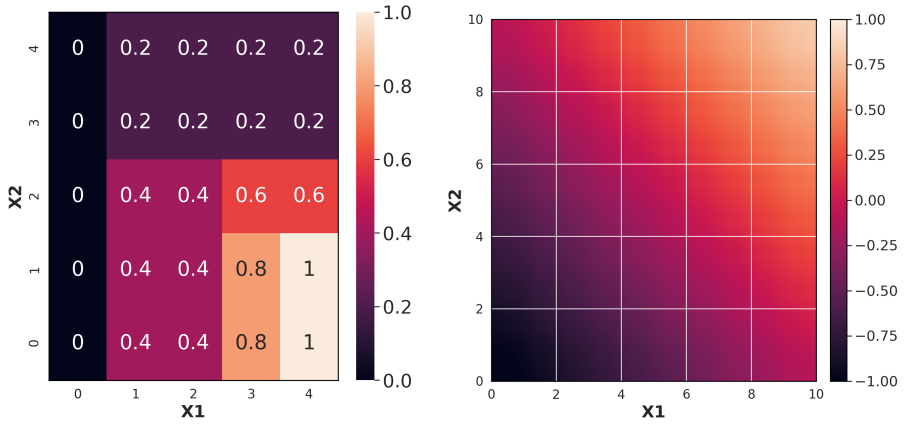
4.1 Is UB-DT a good uplift estimator?

To be able to measure the estimated uplift we need to know the real uplift and therefore we use synthetic data. Fig. 2 depicts two synthetic uplift patterns where $P(Y = 1|X, T = 1)$ and $P(Y = 1|X, T = 0)$ are identified for each instance. The grid pattern can be considered as a tree-friendly pattern whereas the continuous pattern is much more difficult. We generated several datasets according to these patterns with several different numbers of instances (also called data size) ranging from 100 to 100,000 instances. Uplift models were built using 10-fold stratified cross validation and the RMSE (Root Mean Squared Error) was used to evaluate the performance of the models.

Results: Fig. 3 gives the RMSE for the two synthetic patterns according to the data size for different uplift methods. We see that UB-DT is a good estimator for uplift. With UB-DT, RMSE decreases and converges to zero when data sizes increase both for the grid and continuous patterns. This is the expected behavior of a good uplift estimator. This also means that UB-DT, thanks to its global criterion, avoids overfitting of uplift trees. The two-model approach with decision trees also shows competitive performance. UB-DT clearly outperforms the other tree-based methods, these latter having similar performances. With the continuous pattern, KL-DT, Chi-DT, ED-DT and CTS-DT approaches have

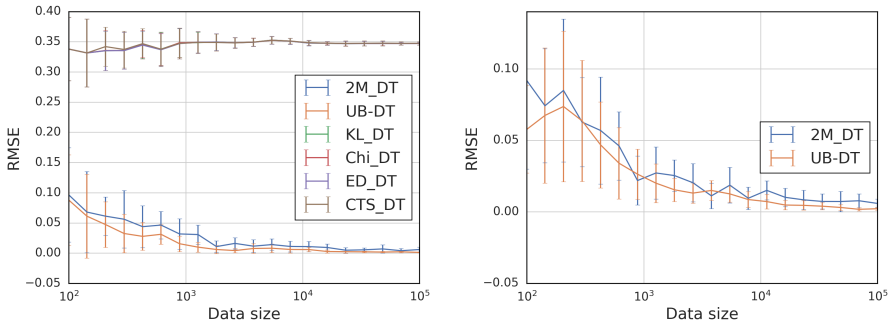
Algorithm 1: UB-DT algorithm

input : T the root tree
output: the tree T^* which minimizes the proposed criterion
$T^* \leftarrow T$
while $C(T^*)$ *decreases*:
  $T' \leftarrow T^*$
  for *leaf* l **in** $\mathbb{L}_T$:
    for X **in** $\mathbb{K}$:
      Get the best Split $S_X(l)$ according to UMODL
      $T_X \leftarrow T^* + S_X(l)$
      if $C(T_X) < C(T')$:
        $T' \leftarrow T_X$
  if $C(T') < C(T^*)$:
    $T^* \leftarrow T'$
Prediction: The output of a tree is a partition of the feature space. The predicted uplift for each instance is the average uplift of its leaf node.



(a) Grid pattern (b) Continuous pattern

Fig. 2: Uplift for 2 synthetic patterns. Fig. 2a (grid pattern): uplift values for each cell. Fig. 2b (continuous pattern): uplift values are $P(Y|T=0, x1, x2) = 1 - (x1 + x2)/20$ while $P(Y|T=1, x1, x2) = (x1 + x2)/20$.



(a) Grid Pattern (b) Continuous Pattern

Fig. 3: The RMSE of tree-based approaches according to data size

lower performances (their RMSE are around 0.5). To avoid a cluttered visualisation, their performances are not included in Fig. 3b.

4.2 UB-DT and UB-RF versus state of the art methods

Datasets. We conducted experiments on 8 real and synthetic datasets widely used in the uplift modeling community: (1) *Hillstrom*[4] (a classical dataset for

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

Dataset	No. Rows	No. Columns	Treatment ratio	Outcome Ratio	Average Uplift	Treatment variable	Outcome variable
Hillstrom-m	42,613	10	0.5	0.145	0.076	*'mens'*	*'visit'*
Hillstrom-w	42,693	10	0.5	0.128	0.045	*'womens'*	*'visit'*
Hillstrom-mw	64,000	10	0.67	0.146	0.06	*'mens' & 'womens'*	*'visit'*
Gerber-N	229,444	16	0.166	0.31	0.081	*'neighbour'*	*'voted'*
Geber-S	229,461	16	0.166	0.304	0.04	*'self'*	*'voted'*
Starbucks	84,534	9	0.5	0.012	0.009	*'promotion'*	*'purchase'*
Information	20,000	69	0.5	0.2	0.0018	*'treatment'*	*'purchase'*
Bank-tel	15,926	17	0.18	0.05	0.09	*'telephone'*	*'Y'*
Bank-cell	42,305	17	0.6	0.115	0.11	*'cellular'*	*'Y'*
Bank-tel-cel	45,211	17	0.71	0.116	0.107	*'telephone'& 'cellular'*	*'Y'*
Megafon	600,000	52	0.5	0.2	-0.18	*'treatment'*	*'conversion'*
Criteo-v	13,979,592	12	0.85	0.047	0.68	*'treatment'*	*'visit'*
Criteo-c	13,979,592	12	0.85	0.0029	0.37	*'treatment'*	*'conversion'*
RHC	5735	62	0.38	0.35	-0.05	*'RHC'*	*'swang1'*

Table 1: Summary of datasets specifications

uplift modeling with data of customers who either received emails featuring men's/ women's products, or received no emails); (2) *Criteo* [5] (a marketing dataset for uplift modeling), (3) *Bank* [9] (a marketing campaign conducted by a bank), (4) *Information*[5] (a marketing dataset in the insurance domain, a part of the Information R package); (5) *Megafon*[6] (a synthetic dataset generated by a telecom company); (6) *Starbucks*[7] (an advertising promotion tested to improve customers purchases); (7) *Gerber* [6] (a policy-relevant dataset used to study the effect of social pressure on voter turnout); (8) *Right Heart Catheterization (RHC)* [3] (a real dataset from the medical domain, the treatment indicates whether a patient received a RHC and the outcome is whether the patient died at any time up to 180 days after admission to the study).

Each dataset was used with different settings of treatment and outcome variables. For all datasets, each treatment and outcome variables are binary. Table 1 provides the most relevant specifications about the data sets.

Results. We evaluate the uplift models by using the qini metric [4]. Qini is a variant of the Gini coefficient. Its values are in $[-1, 1]$, the higher the value, the larger the impact of the predicted optimal treatment. Fig. 4a (resp. Fig. 4b) shows the overall average ranking of tree based methods (resp. meta-learners and forest-based methods) according to its qini performance against each dataset. Compared to other tree-based methods and to the two-model approach with decision trees, Figure 4a shows that UB-DT achieves the best performance. Table 2 reports the results of the experiment for the qini metric. This table shows that UB-DT is also a good estimator of the uplift on real data. Figure 4b shows that both UB-RF and 2M have the best rank. Table 3 indicates that the random forest strategy improves the performance of the uplift models (qini values are higher with UB-RF than UB-DT). UB-RF has the best performance on 4 out the 14 experiments.

[5] `https://cran.r-project.org/web/packages/Information/index.html`

[6] `https://ods.ai/tracks/df21-megafon/competitions/megafon-df21-comp/data`

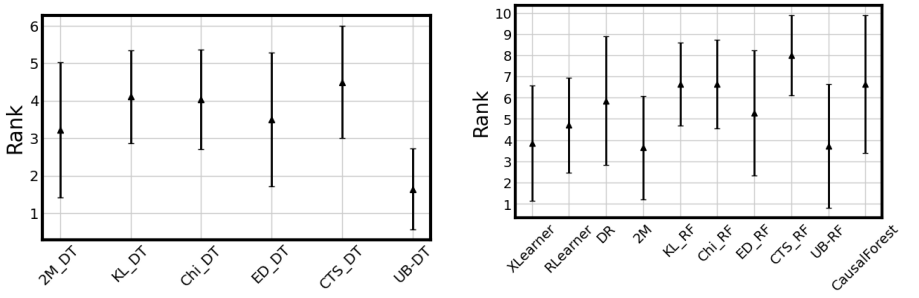
[7] `https://github.com/joshxinjie/Data_Scientist_Nanodegree/tree/master/starbucks_portfolio_exercise`joshxinjie

Dataset	2M_DT	KL_DT	Chi_DT	ED_DT	CTS_DT	UB-DT
Hillstrom-m	0.3(1.0)	1.1(1.9)	1.0(1.9)	0.0(1.4)	0.2(1.0)	**1.6(1.6)**
Hillstrom-w	0.8(1.6)	5.2(2.5)	5.2(2.6)	**6.4(1.2)**	-0.4(2.0)	4.8(2.3)
Hillstrom-mw	-0.6(0.8)	-0.1(1.2)	-0.8(1.1)	**4.4(2.7)**	-0.0(1.0)	-0.4(1.4)
Gerber-n	**5.6(0.8)**	1.3(0.8)	1.2(0.8)	1.1(0.6)	1.3(0.8)	1.9(0.6)
Gerber-s	**5.5(1.1)**	0.4(0.5)	0.4(0.6)	0.5(0.3)	0.4(0.4)	0.8(0.6)
Criteo-c	8.0(1.5)	4.1(1.4)	4.8(1.5)	**15.2(0.3)**	1.7(0.3)	13.7(3.2)
Criteo-v	0.4(0.3)	-1.2(0.2)	-1.1(0.3)	-1.3(0.3)	0.4(1.1)	**3.6(1.2)**
Megafon	5.1(0.6)	4.5(0.9)	4.7(0.9)	4.7(0.9)	4.9(0.8)	**7.8(0.8)**
Bank-tel	5.4(7.6)	-12.5(2.8)	-10.8(7.0)	-10.2(7.8)	-12.8(2.9)	**12.8(8.0)**
Bank-cell	11.1(3.0)	-2.0(1.5)	-1.4(2.5)	-2.2(1.5)	-3.7(1.5)	**38.4(3.4)**
Bank-tel-cell	10.3(1.6)	-1.9(1.2)	-1.2(2.1)	-1.8(1.2)	-3.4(1.4)	**37.1(2.6)**
Information	4.6(3.4)	-6.3(2.8)	-6.3(2.8)	-2.8(1.5)	-5.4(1.5)	**11.8(2.4)**
Starbucks	1.4(1.4)	20.1(3.0)	18.3(3.4)	19.9(3.2)	13.9(3.9)	**20.2(3.5)**
RHC	12.8(1.9)	18.4(3.8)	19.9(4.2)	18.4(3.8)	16.7(2.5)	**20.7(5.0)**

Table 2: Average qini values and standard deviation (multiplied by 100). The best qini value for each dataset is marked in bold.

Dataset	XLearner	RLearner	DR	2M	KL_RF	Chi_RF	ED_RF	CTS_RF	UB-RF	CausalForest
Hillstrom-m	0.3(2.3)	0.3(1.8)	1.2(1.6)	0.7(2.3)	-0.0(2.1)	-0.9(1.5)	0.7(1.5)	1.1(1.9)	**1.8(1.6)**	-0.2(1.6)
Hillstrom-w	6.2(1.7)	6.2(1.4)	6.0(1.4)	4.9(1.1)	6.2(1.1)	**7.0(1.0)**	6.2(1.1)	5.7(1.3)	6.7(1.1)	2.1(1.9)
Hillstrom-mw	3.7(2.3)	**3.9(2.7)**	3.8(2.8)	3.0(2.0)	3.0(1.3)	2.8(1.5)	3.6(2.5)	2.3(2.4)	3.1(1.7)	0.1(1.7)
Gerber-n	**3.7(0.6)**	1.9(0.7)	0.5(0.9)	3.1(0.6)	1.8(1.0)	2.1(1.1)	1.9(0.5)	1.4(1.0)	2.7(0.7)	2.9(1.0)
Gerber-s	2.4(0.9)	1.7(0.7)	0.6(0.9)	2.2(0.8)	1.3(1.0)	1.4(0.6)	1.6(0.8)	1.4(0.7)	1.8(0.8)	**3.1(0.5)**
Criteo-c	**22.3(1.8)**	19.4(1.0)	20.0(0.6)	19.5(1.6)	14.6(3.5)	12.4(4.3)	21.1(2.3)	7.3(3.9)	18.7(1.5)	10.9(2.4)
Criteo-v	0.3(0.8)	5.3(0.5)	4.8(1.5)	3.9(0.5)	5.4(1.2)	4.8(1.7)	**6.1(1.0)**	2.4(0.8)	5.7(0.7)	0.4(0.4)
Megafon	**18.2(0.6)**	2.6(0.5)	2.2(0.9)	16.6(0.9)	11.2(0.7)	11.0(1.2)	10.8(0.8)	9.2(1.1)	12.8(1.0)	9.7(0.7)
Bank-tel	14.5(7.6)	2.8(8.8)	16.0(9.0)	21.1(11.6)	-15.5(6.3)	-6.1(12.6)	-15.8(5.6)	-18.7(2.9)	**26.7(7.2)**	25.4(5.3)
Bank-cell	18.8(4.7)	23.3(3.6)	17.4(6.5)	31.0(3.9)	0.4(2.3)	1.5(2.5)	-2.5(2.6)	-1.0(1.9)	**45.5(2.7)**	20.8(2.6)
Bank-tel-cell	16.2(5.6)	23.8(2.5)	17.0(3.4)	30.5(2.7)	1.4(3.4)	-0.4(5.7)	-1.7(3.1)	-0.5(2.3)	**46.1(2.1)**	23.5(2.9)
Information	**14.9(3.3)**	10.0(3.1)	4.1(2.3)	13.7(4.1)	9.6(2.0)	9.7(3.1)	11.2(2.9)	10.6(2.9)	12.0(3.1)	10.5(3.2)
Starbucks	22.3(4.5)	22.4(3.9)	22.4(3.7)	22.7(4.1)	22.4(2.1)	21.4(3.4)	**23.4(3.2)**	20.8(3.1)	20.2(3.3)	8.1(3.7)
RHC	32.4(3.5)	31.3(4.3)	30.3(5.0)	**34.6(4.3)**	29.6(4.2)	29.7(5.0)	30.0(4.1)	29.1(3.7)	27.2(5.0)	27.6(4.5)

Table 3: Average qini values and standard deviation (multiplied by 100) across datasets and uplift approaches. In bold, the best value for each dataset



(a) tree-based methods (b) meta-learners and forest-based methods

Fig. 4: Overall average ranking of the uplift approaches

5 Conclusion and perspectives

In this paper, we presented a new parameter-free method called UB-DT for uplift decision trees. We have designed a Bayesian approach to select the most probable uplift tree model T that maximizes the posterior probability $P(T|Data)$. Contrary to state-of-art uplift decision tree approaches, UB-DT is characterized by a global criterion to build a tree, so the splits in one node depend on the splits in the other nodes. This approach avoids overfitting and the need for a pruning step. A search algorithm finds the tree that optimizes this criterion. We showed that our approach is easily extended to random forests and we defined UB-RF. Evaluations on real and synthetic data sets show that UB-DT is a good uplift estimator and our tree and forests methods perform competitively with state-of-art uplift modeling approaches including non tree methods.

This work opens several perspectives. Studies on *general* trees (with more than two child nodes) is promising. In addition, studies with multiple treatments are still open work in uplift modeling. Moreover, the search algorithm leads to a local optimum and may create under-fitted uplift trees. To go above this horizon effect, it would be interesting to use a post-pruning algorithm [16].

References

1. Athey, S., Imbens, G.: Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences **113**(27), 7353–7360 (2016)
2. Athey, S., Tibshirani, J., Wager, S.: Generalized random forests. The Annals of Statistics **47**(2), 1148–1178 (2019)
3. Connors, A.F., et al.: The effectiveness of right heart catheterization in the initial care of critically ill patients. support investigators. JAMA **276 11**, 889–97 (1996)
4. Devriendt, F., Van Belle, J., Guns, T., Verbeke, W.: Learning to rank for uplift modeling. IEEE Transactions on Knowledge and Data Engineering pp. 1–1 (2020)
5. Diemert, E., Betlei, A., Renaudin, C., Amini, M.R.: A Large Scale Benchmark for Uplift Modeling. In: KDD. London, United Kingdom (2018)
6. Gerber, A.S., Green, D.P., Larimer, C.W.: Social pressure and voter turnout: Evidence from a large-scale field experiment. American Political Science Review (2008)
7. Kennedy, E.H.: Towards optimal doubly robust estimation of heterogeneous causal effects (2020), `https://arxiv.org/abs/2004.14497`
8. Künzel, S.R., Sekhon, J.S., Bickel, P.J., Yu, B.: Metalearners for estimating heterogeneous treatment effects using machine learning. Proceedings of the National Academy of Sciences **116**(10), 4156–4165 (2019)
9. Moro, S., Cortez, P., Rita, P.: A data-driven approach to predict the success of bank telemarketing. Decis. Support Syst. **62**, 22–31 (2014)
10. Naumov, G.: Np-completeness of problems of construction of optimal decision trees. In: Soviet Physics Doklady. vol. 36, p. 270 (1991)
11. Radcliffe, N., Surry, P.: Differential response analysis: Modeling true responses by isolating the effect of a single action. Credit Scoring and Credit Control IV (1999)
12. Radcliffe, N.J., Surry, P.D.: Real-world uplift modelling with significance-based uplift trees. Stochastic Solutions (2011)
13. Rafla, M., Voisine, N., Crémilleux, B., Boullé, M.: A non-parametric bayesian approach for uplift discretization and feature selection. In: ECML PKDD (2022)

14. Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology **66**, 688–701 (1974)
15. Rzepakowski, P., Jaroszewicz, S.: Decision trees for uplift modeling with single and multiple treatments. Knowl. Inf. Syst. **32**(2), 303–327 (2012)
16. Voisine, N., Boullé, M., Hue, C.: A bayes evaluation criterion for decision trees. In: Advances in knowledge discovery and management. pp. 21–38. Springer (2009)
17. Zhao, Y., Fang, X., Simchi-Levi, D.: Uplift modeling with multiple treatments and general response types. In: SIAM Int. Conf. on Data Mining. SIAM (2017)