A Non-Parametric Bayesian Approach for Uplift Discretization and Feature Selection

Mina Rafla, Nicolas Voisine, Bruno Crémilleux, Marc Boullé

A Non-Parametric Bayesian Approach for Uplift Discretization and Feature Selection

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

[1] Orange Labs, 22300 Lannion, France
{mina.rafla, nicolas.voisine, marc.boulle}@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 outcome. Bank or Telecom uplift data often have hundreds to thousands of features. In such situations, detection of irrelevant features is an essential step to reduce computational time and increase model performance. We present a parameter-free feature selection method for uplift modeling founded on a Bayesian approach. We design an automatic feature discretization method for uplift based on a space of discretization models and a prior distribution. From this model space, we define a Bayes optimal evaluation criterion of a discretization model for uplift. We then propose an optimization algorithm that finds near-optimal discretization for estimating uplift in $O(n \log n)$ time. Experiments demonstrate the high performances obtained by this new discretization method. Then we describe a parameter-free feature selection method for uplift. Experiments show that the new method both removes irrelevant features and achieves better performances than state of the art methods.

Keywords: Uplift Modeling · Feature Selection · Discretization · 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. Uplift models help identify groups of people likely to respond positively to treatment *only because* they received one. This research domain has multiple applications like customer relationship management, personalized medicine, advertising. 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 [14, 26] or the evaluation of uplift models [21].

Many databases are large and contain hundreds of features [12]. Keeping all the features is costly and inefficient to build uplift models. A feature selection

process is then an essential step to remove irrelevant features, improves the estimation accuracy and accelerates the model building. While there are a lot of feature selection methods in machine learning, there are very few propositions for uplift modeling [27]. This observation might be explained since uplift creates new challenges such as the impossibility to observe two treatment outcomes for a same individual. Designing methods for uplift requires overcoming this difficulty. This paper aims to answer the need for feature selection methods for uplift.

We present a parameter-free feature selection method for uplift modeling founded on a Bayesian approach. Following a part of literature on feature selection that performs a discretization of numerical features [16, 25], we first describe an automatic feature discretization method for uplift modeling that we call UMODL (for Uplift MODL). UMODL is based on the Bayesian MODL (Minimum Optimized Description Length) criterion [1] that we have extended to the uplift problem. UMODL defines a space of discretization models and a prior distribution on this model space. We construct a Bayes optimal evaluation criterion of a discretization model for uplift modeling. In practice, the best model according to the criterion cannot be computed due to the complexity of the problem and we present a greedy search algorithm in $\mathcal{O}(n \log n)$ to find near-optimal discretizations. Experiments show that the discretization model found by UMODL gives a good estimator of uplift. Then, based on UMODL, we present UMODL feature selection (UMODL-FS in short) a feature selection method for uplift. UMODL-FS computes a score of the features and automatically selects appropriate features for uplift. Experiments demonstrate that UMODL-FS properly removes irrelevant features and clearly outperforms state of the art methods by providing uplift models with the highest and most stable performance. Being a parameter-free method (neither the number of bins in the discretization nor the number of features to keep or remove are given), UMODL-FS can be used without effort.

The remainder of the paper is organized as follows. In the next section, we introduce uplift modeling, feature selection for uplift, MODL and the literature related to our problem setting. Section 3 presents UMODL which is experimentally evaluated in Section 4. UMODL-FS and the associated experiments are described in Section 5. We conclude in Section 6.

2 Background and literature review

2.1 Uplift modeling

Uplift definition. Uplift is a notion introduced by Radcliffe and Surry [20] and defined in Rubin's causal inference models [23] as the *Individual Treatment Effect*. The uplift modeling literature and a branch of the causal inference literature have recently approached each other [8]. We present the notion of uplift.

Let D be a group of N individuals indexed by $n : 1 \dots N$ where each individual is described by a set of variables $\mathbb{X}$. X_n denotes the set of values of $\mathbb{X}$ for the individual n. Let T 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 $T = 1$) and those without treatment (denoted $T = 0$). Let Y be the outcome variable (for instance, the purchase or not of a product). We note $Y_n(T = 1)$ the outcome of an individual n when he received a treatment and $Y_n(T = 0)$ his outcome without treatment. The uplift of an individual n, denoted by τ_n, is defined as: $\tau_n = Y_n(T = 1) - Y_n(T = 0)$. The main difficulty is that uplift value is not directly measurable, i.e for each individual we can either observe $Y_n(T = 1)$ or $Y_n(T = 0)$ but cannot observe simultaneously both outcomes. However, uplift τ_n can be empirically estimated by considering two groups: a treatment group (individuals who received a treatment) and a control group (individuals who did not). The estimated uplift of an individual n denoted by $\hat{\tau}_n$ is then the difference between response rates in both groups and computed by using the CATE [3] (Conditional Average Treatment Effect) [23]: CATE : $\hat{\tau}_n = \mathbb{E}[Y_n(T = 1)|X_n] - \mathbb{E}[Y_n(T = 0)|X_n]$

As the real value of τ_n cannot be observed, it is impossible to directly use machine learning algorithms such as regression to infer a model to predict τ_n. We sketch below how uplift is modeled in the literature. Simple methods such as considering only individuals having received the treatment fail because they do not detect individuals whose response is always positive even without treatment.

Uplift modeling approaches. In recent years, several studies on uplift models design have been conducted. One of the most classical and intuitive approach is the *two model approach* [11] (also called *T-learner* in the causal community), which consists of two independent predictive models, one on the treatment group to estimate $P(Y|X, T = 1)$ and another on the control group to estimate $P(Y|X, T = 0)$. The estimated uplift of an individual n is the difference between those values for the given individual, i.e. $\hat{\tau}_n = P(Y = 1|X_n, T = 1) - P(Y = 1|X_n, T = 0)$. *Class transformation approach* [14] is another family of methods that maps the uplift modeling problem to a usual supervised learning problem. With the *Direct-approach*, different machine learning algorithms are modified to suit uplift modeling such as methods based on *decision trees* [24, 26], *k nearest neighbors* [7], *logistic regression* [17], etc. The causal inference community defines other methods such as *S-Learner* which includes the outcome variable T in the features with a standard regression, *X-Learner* [13] which performs a two-step regression before the estimation of the CATE, *DR-Learner* [15] which combines a two-model approach and the use of the Inverse Propensity Weighting [18].

Evaluation of uplift models Since real values of uplift cannot be observed, standard performance measures of supervised learning algorithms cannot be used. That is why uplift is evaluated through the ranking of the individuals according to their estimated uplift value. The intuition is that a good uplift model estimates higher uplift values to individuals in the treatment group with positive outcomes than those with negative outcomes and vice versa for the control group. A common approach to evaluate uplift models is the qini measure,

[3] The terms *treatment effect* and *uplift* address the same notion. CATE is an estimation of uplift and we use "CATE" for speaking of the estimated uplift values.

also known as the Area Under Uplift Curve ($AUUC$) [3, 19]. It is a variant of the Gini coefficient. Qini values are in $[-1, 1]$, the higher the value, the larger the impact of the predicted optimal treatment.

2.2 Feature Selection for uplift models

The accessibility of high dimensional datasets with hundreds of features makes the use of feature selection techniques crucial for machine learning tasks and uplift. The goal of feature selection techniques is to select subset of features that could efficiently describe data while expelling irrelevant features [9]. This can significantly improve models performances and computation time [2]. Regarding uplift modeling, studies addressing feature selection are very limited. To the best of our knowledge, only two research papers deal with this challenge.

Zhao et al. [27] propose filter and embedded feature selection methods for uplift. The principle is to remove features that are not correlated either with the outcome variable or uplift. Filter methods are used in a pre-processing step independently of an uplift model while embedded methods perform feature selection during the training of a model and are specific to an uplift algorithm. In [27], the presented filter methods are *bins methods* (inspired from [24]), *F-filter* and *LR-filter*. Experiments in [27] show that bin-based filter methods have the best performances while *F-filter*, *LR-filter* and embedded methods have poor performances.

We give a few words on the above methods providing the best results as well as *F-filter* and *LR-filter*. The principle of a bins method is to discretize a feature into L bins based on the percentiles of the feature (L is given by the user). The importance of a feature regarding uplift is evaluated by a divergence measure of the treatment effect over the bins. Three divergence measures are used: Kullback-Leibler (KL), squared Euclidean distance (ED), and chi-squared (Chi). The *F-filter* and the *LR-filter* are based respectively on the F-statisic [10] and the likelihood ratio statistic [5] for the coefficients of regression models.

On the other hand, a very recent paper [12] uses some of the filtering methods given in [27] as well as a correlation coefficient to remove redundant features. The paper describes an uplift application on a private database in the bank domain.

2.3 MODL approach

The MODL (Minimum Optimized Description Length) approach is a non-parametric Bayesian approach for discretization and conditional probability estimation [1]. It is based on the Minimum Description Length (MDL) principle [6, 22]. In the MODL approach, a space of discretization models is defined. A discretization model is described by a set of parameters: the number of intervals, the boundaries of the intervals and the frequencies of the classes in each interval. Using these parameters, the MODL approach consists of defining a criterion for a discretization model and with the help of a search algorithm, the MODL approach can score all possible discretization models and selects the one with the best score.

3 UMODL

This section presents UMODL a new criterion for uplift discretization modeling and the search algorithm to find the optimal uplift discretization model.

3.1 UMODL Criterion

While MODL properly exploits discretization for density estimation, it is not suitable for uplift modeling since uplift deals with two treatment groups and the estimation of the conditional probabilities of the outcome variable Y given an attribute X also depends on the treatment variable T.

We now introduce the new criterion that we propose to define the best discretization model for uplift. Let M be an uplift discretization model and D denotes data. From a Bayesian point of view, the best uplift discretization model is found by maximizing the posterior probability of the model given the data $P(M|D)$. Let us consider the Bayes rule:

$$P(M \mid D) = \frac{P(M)P(D \mid M)}{P(D)} \tag{1}$$

Given that $P(D)$ is constant, maximizing $P(M|D)$ is equivalent to maximizing $P(M)P(D|M)$, i.e the prior probability and the likelihood of the data given the chosen model. Let us first introduce some notations:

- X : explanatory variable to discretize
- Y : binary outcome variable
- N : number of instances in the dataset
- J : number of classes of Y
- I : number of intervals
- N_i : number of instances in the interval i
- $N_{it.}$: number of instances in the interval i of treatment t
- $N_{i.j}$: number of instances in the interval i of class j
- N_{itj} : number of instances in the interval i of class j and the treatment t
- W_i : boolean term indicating if the treatment has an effect in interval i (W_i=1) or not (W_i=0)

We define an uplift discretization model M by the number of intervals, the bounds of the intervals, the presence or absence of a treatment effect, class frequencies per interval or for each treatment per interval. In other words, a model M is defined by the hierarchy of parameters (cf. Fig. 1):

$$\{I, \{N_i\}, \{W_i\}, \{N_{i.j}\}_{W_i=0}, \{N_{itj}\}_{W_i=1}\}$$

The evaluation criterion $C(M)$ which is the cost of an uplift discretization model M is defined then by: $C(M) = -\log\big(P(M) \times P(D|M)\big)$. Taking the negative log turns the maximization problem to a minimization one. M is optimal if $C(M)$ is minimal.

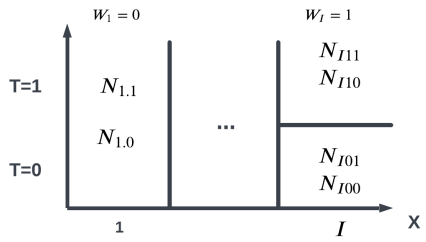


Fig. 1: Parameters of an uplift discretization model. The presence of a treatment effect ($W_i = 1$) in interval i requires describing the distribution of the outcome variable Y separately for each treatment (part right). In contrast, the absence of a treatment effect ($W_i = 0$) indicates to consider the distribution of the outcome variable Y for the interval i independently of the treatment variable (part left).

For the prior distribution of the model parameters, we exploit the hierarchy of the parameters and assume a uniform prior at each stage of the hierarchy with independence across the intervals. Using these assumptions, we express $C(M)$ according to the parameters of an uplift discretization model and we obtain Eq. 2 that we demonstrate below.

$$\begin{aligned} C(M) = \log N + \log \binom{N+I-1}{I-1} + I \times \log 2 \\ + \sum_{i=1}^{I} (1 - W_i) \log \binom{N_i + J - 1}{J-1} + \underbrace{\sum_{i=1}^{I} (1 - W_i) \log \frac{N_i!}{N_{i.1}!..N_{i.J}!}}_{Likelihood} \\ + \sum_{i=1}^{I} W_i \sum_t \log \binom{N_{it.} + J - 1}{J-1} + \underbrace{\sum_{i=1}^{I} W_i \sum_t \log \frac{N_{it.}!}{N_{it1}!..N_{itJ}!}}_{Likelihood} \end{aligned} \tag{2}$$

Proof of Eq 2. We express $P(M)$ and $P(D|M)$ according to the parameters of an uplift discretization model. We introduce a prior distribution by exploiting the hierarchy of the models' parameters. Assuming the independence of the local distributions across the intervals, we obtain:

$$\begin{aligned} P(M) = P(I) \times P(\{N_i\}|I) \times \\ \prod_i P(W_i|I) \left[(1 - W_i) \times P(\{N_{i.j}\}|I, \{N_i\}) + W_i \times \prod_t P(\{N_{itj}\}|I, \{N_{it.}\}) \right] \end{aligned} \tag{3}$$

We express each of the terms of Eq. 3 according to the parameters of M assuming a uniform distribution for each parameter. Assuming that the number

of intervals I is uniformly distributed between 1 and N, the first term in Eq. 3 becomes:

$$P(I) = \frac{1}{N} \tag{4}$$

Given a number of intervals I, all the discretizations into I intervals (i.e. the choices of the bounds) are equiprobable. Computing the probability of an interval set leads to a combinatorial calculation of the number of all possible interval sets or equivalently the number of ways of distributing the N instances in the I intervals, with counts N_i per interval. The second term of Eq. 3 is then:

$$P(\{N_i\}|I) = \frac{1}{\binom{N+I-1}{I-1}} \tag{5}$$

For a given interval i, we assume that a treatment can have an effect or not, with equal probability, i.e. $P(W_i|I) = \frac{1}{2}$. We obtain:

$$\prod_i P(W_i|I) = \left(\frac{1}{2}\right)^I \tag{6}$$

In the case of an interval i where there is not effect of the treatment ($W_i = 0$), UMODL describes one unique distribution of the outcome variable. Given an interval i, its number of examples N_i is known. Assuming that each of the class distributions is equiprobable, we end up also with a combinatorial problem:

$$P(\{N_{i.j}\}|I, N_i) = \frac{1}{\binom{N_i+J-1}{J-1}} \tag{7}$$

In the case of an interval i with an effect of the treatment ($W_i = 1$), UMODL describes two distributions of the outcome variable, with and without the treatment. Given an interval i and a treatment t, we know the number of examples $N_{it.}$. Assuming that each of the distributions of class values is equiprobable, we get:

$$P(\{N_{itj}\}|I, N_{it.}) = \frac{1}{\binom{N_{it.}+J-1}{J-1}} \tag{8}$$

After defining the models' prior, we define the likelihood $P(D|M)$ of the data given the uplift discretization model. For each multinomial distribution of the outcome variable (a single or two distinct distributions per interval depending on whether the treatement has an effect or not), we assume that all possible observed data D_i consistent with the multinomial model are equiprobable. Using multinomial terms, we obtain the following likelihood term:

$$P(D|M) = \prod_i P(D_i|M) \tag{9}$$

$$= \prod_i \left[(1 - W_i) \times \frac{1}{(N_i!/N_{i.1}!..N_{i.J}!)} + W_i \times \prod_t \frac{1}{(N_{it.}!/N_{it1}!..N_{itJ}!)} \right] \tag{10}$$

Combining the prior $P(M)$ (Eq 4 to 8) with the likelihood $P(D|M)$ (Eq. 10), we obtain $P(M)P(D|M)$. Taking the negative log yields to the UMODL criterion presented in Eq. 2. Coming back to Eq. 2, the prior terms of the first line come from Eq. 4 to 6. In the second line of Eq. 2 (modeling a situation w/o a treatment effect) and the third line (situation with a treatment effect), the first terms are prior terms (Eqs 7- 8) and the second terms are likelihood terms (Eq. 10).

Uplift estimation The presented discretization approach is a density estimation approach for uplift modeling. We model the probability of Y conditionally on the explanatory variable X and a binary treatment variable T. The search algorithm we present is looking for the parameters I, $\{W_i\}$, $\{N_i\}$, $\{N_{i.j}\}$, $\{N_{ijt}\}$, and $\{W_i\}$ that minimize the cost of the model. In other words, the search algorithm tries to find the optimal discretization in the Bayes sense that best estimates the real densities of the outcome variable Y conditionally on X and T. Once a discretization and its parameters are defined, the estimation of the CATE for each interval is simple. As shown in Fig.1, assuming a binary outcome variable Y and given $W_i = 1$, we have $P_i(Y = 1|T = 1) = N_{i11}/(N_{i11} + N_{i01})$ and $P_i(Y = 1|T = 0) = N_{i10}/(N_{i10} + N_{i00})$, therefore $CATE_i = P_i(Y = 1|T = 1) - P_i(Y = 1|T = 0)$. For intervals with $W_i = 0$, $CATE_i$ is considered insignificant.

3.2 Search algorithm and post-optimization

We sketch below our search algorithm to find the best model w.r.t. the UMODL criterion. This algorithm finds the optimal values of the parameters that minimize $C(M)$. The principle of this algorithm is inspired by the search algorithm [1] which we adapted to our criterion. As an optimal search algorithm is not practical due to the complexity of the problem, we build a greedy algorithm[4].

Greedy Search algorithm The search algorithm is a greedy bottom-up algorithm with the following steps:

- The algorithm starts by making an elementary discretization such that all examples with the same value have their own interval,
- Compute the costs of all possible merges i.e. try to merge adjacent intervals,
- Merge the two adjacent intervals that decrease $C(M)$ the most,
- Recalculate the cost of all possible adjacent merges and select the merge that reduces $C(M)$ the most,
- Repeat until no merge decreases $C(M)$.

While this algorithm is complex, it can be implemented in $O(n \log n)$ time [1].

Post-optimization This greedy search algorithm can fall into a local minimum, so post-optimization steps are needed to perturb the interval bounds. We used post-optimization steps that consist of recurrent splits, merges, merge splits, and merge merge splits of adjacent intervals, as described in [1] but designed in this work for uplift.

[4] Our implementation is provided at https://github.com/MinaWagdi/UMODL

4 UMODL quality evaluation experiments

This section experimentally evaluates whether UMODL is a good estimator of uplift. The principle of the experiments is to generate data with different synthetic uplift patterns in order that results of UMODL can be compared to true uplift. A *synthetic uplift pattern* is a data pattern where $P(Y = 1|X, T = 1)$ and $P(Y = 1|X, T = 0)$ are identified for each example. Therefore several indicators can be observed: (1) the number of intervals founded by UMODL w.r.t. the characteristics of the uplift pattern, (2) the RMSE (root mean squared error) between the real uplift and the estimated uplift by UMODL computed for each instance and (3) the number of instances needed by UMODL to find the uplift pattern. We generate synthetic uplift patterns of different characteristics for simulating various situations.

4.1 Description

The experimental protocol is made of the following steps:

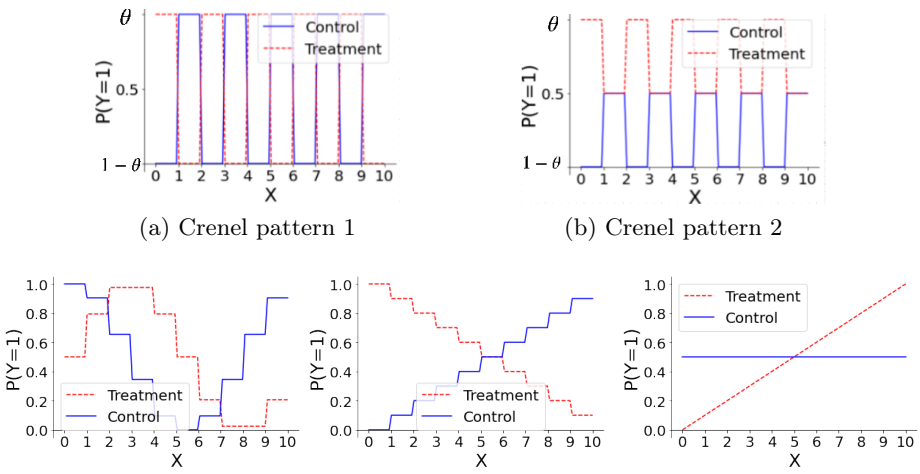
1. Define a particular synthetic uplift pattern of one dimension.
2. Generate several train samples according to the defined pattern with 40 different number of instances (also called *data size*) ranging from 10 to 100,000 instances. For each data size, generate ten datasets. All generated data are uniformly distributed on the $[0, 10]$ numerical domain for each of the treatment ($T = 1$) and control groups ($T = 0$).
3. Generate a test set of 10,000 instances based on the defined uplift pattern.
4. For each training sample, apply the UMODL approach to search for the best discretization model.
5. For each experiment, the obtained discretization model is then applied to the test set, and RMSE is computed by comparing for each data point: (a) the CATE estimation in the found interval and (b) the real CATE value.
6. By observing both the number of found intervals for each dataset and the RMSE values, we can determine whether the UMODL approach manages to find the synthetic pattern or not.
7. Repeat these steps with different synthetic uplift patterns.

4.2 Synthetic uplift patterns

We generate four bin-based patterns and one continuous pattern. We use patterns of different characteristics[5] to evaluate how UMODL performs both in various situations and different rates of uplift. The patterns are illustrated in Fig. 2 and depicted below.

- *Crenel pattern 1* (cf. Fig. 2a): this crenel pattern is made of 10 intervals containing a repeated sequence of a positive treatment effect followed by a negative one. We generated five versions of this pattern with different treatment effects (uplift).

[5] Other patterns can be found using the github link provided previously.



(a) Crenel pattern 1 (b) Crenel pattern 2

(c) Trigonometric pattern (d) Scissors pattern (e) Continuous pattern

Fig. 2: Synthetic uplift patterns. The X-axis represents variable X and the Y-axis represents $P(Y = 1)$. For *Crenel Pattern 1* and *Crenel Pattern 2*, five versions are generated with different values of $\theta \in \{0.6, 0.7, 0.8, 0.9, 1\}$. The difference between $P(Y = 1)$ in the treatment and control groups represents the uplift.

- *Crenel pattern 2* (cf. Fig. 2b): is a slightly different crenel pattern similarly made of 10 intervals containing a repeated sequence of a positive treatment effect followed by no treatment effect. We generated five versions of this pattern with different treatment effects (uplift).
- *Trigonometric pattern* (cf. Fig. 2c) is a particular bin-based pattern with trigonometric shape where: $P(Y = 1|T = 1) = 0.5 + (0.5 \times sin(i \times \frac{2\pi}{10}))$ and $P(Y = 1|T = 0) = 0.5 + (0.5 \times cos(i \times \frac{2\pi}{10}))$
- *Scissors pattern* (cf. Fig. 2d) is a bin-based pattern where $P(Y = 1|T = 1) = \frac{i}{10}$ and $P(Y = 1|T = 0) = 1 - \frac{i}{10}$, where i is the interval number.
- *Continuous pattern* (cf. Fig. 2e) differs from bin-based patterns. Here $P(Y = 1|T = 1) = X/10$ $P(Y = 1|T = 0) = 0.5$.

4.3 Results

Results are given in Figures 3, 4 and 5. We start by the central question – "Is UMODL a good estimator of uplift?" – and provide complementary observations.

Is UMODL a good estimator of uplift? From Figures 3 (left) and 4 (left), we clearly see that even when the treatment effect is very small per interval (grey curves), UMODL is able to find the proper number of intervals of the uplift patterns. This is also illustrated by the RMSE curves (Figures 3 (right) and 4 (right)) showing that RMSE always converges towards 0 for sufficiently large

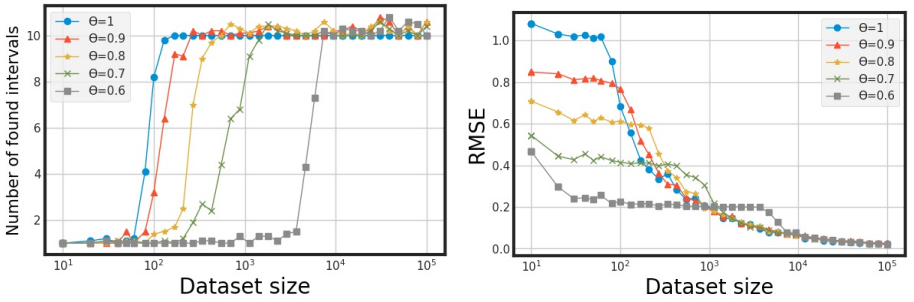


Fig. 3: Results obtained for *Crenel pattern 1*. The left (resp. right) figure shows the mean number of found intervals (resp. the mean value of RMSE) on the test set by UMODL according to the dataset size. Different curve colors correspond to different treatment effects. For example, the blue curve corresponds to the crenel pattern of repeated positive uplift ($= 1$) followed by negative uplift ($= -1$).

datasets. Similar performances are reported with the *trigonometric pattern* (cf. Fig. 5a), the *scissors pattern* (cf. Fig. 5b) and the *continuous pattern* (cf. Fig. 5c) except that the number of estimated intervals is not a relevant indicator for the *continuous pattern* because this pattern is continuous.

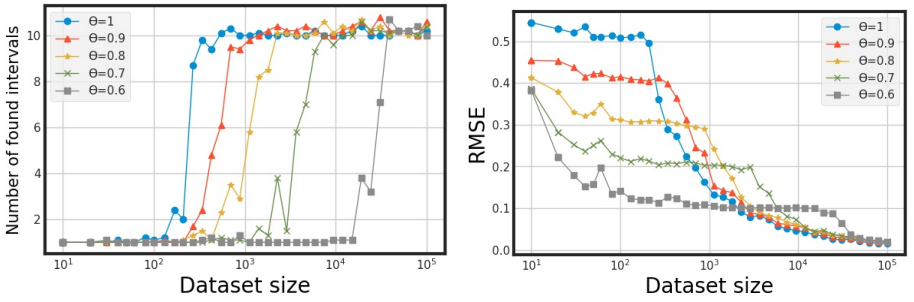
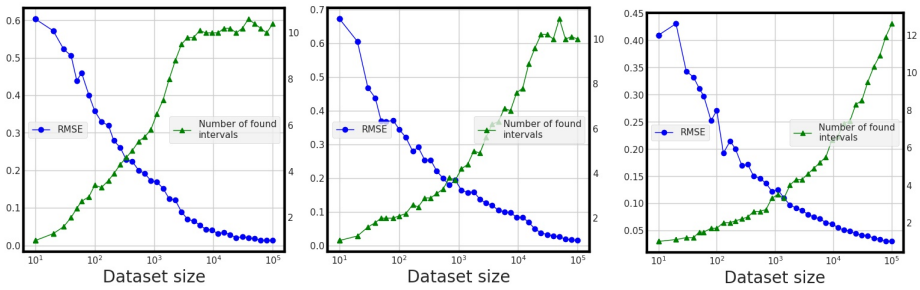


Fig. 4: Results obtained for *Crenel pattern 2*. The left (resp. right) figure shows the mean number of found intervals (resp. the mean value of RMSE) on the test set by UMODL according to the dataset size. Different curve colors correspond to separate treatment effects. For example, the blue curve corresponds to the crenel pattern of repeated positive uplift ($=1$) followed by zero uplift.

How many instances are needed to find the uplift pattern according to its characteristics? When the differences of densities between adjacent intervals get smaller, UMODL needs more instances to give prominence to a model with more intervals. This is typically the case with the *scissors pattern* (cf. Fig. 5b). Analogous behaviors are observed in Figures 3 and 4. For example, in Fig. 3,



(a) Trigonometric pattern (b) Scissors pattern (c) continuous pattern

Fig. 5: Figures 5a, 5b, 5c present the performances obtained with the *trigonometric pattern, scissors pattern and continuous pattern*. Blue curves depict the mean value of the RMSE per dataset size while the green curves indicate the number of found intervals.

the blue curve finds the uplift pattern with less instances than the red curve. Interestingly, UMODL succeeds in finding the proper intervals even when there is no treatment effect (cf. the results with the *crenel pattern 2* in Fig. 4).

Does UMODL overfit? Another important aspect of the UMODL discretization is that the UMODL method does not overfit, i.e. UMODL always finds the ten intervals of the underlying patterns and does not consider extra intervals even when the data size increases significantly (cf. Figures 3 and 4). With the *continuous pattern*, UMODL goes on to consider more intervals as long as the size of the data increases (cf. Fig. 5c) which is appropriate since the pattern is continuous and there is no defined intervals.

5 UMODL Feature Selection

Description of UMODL feature selection. We describe now the method:

1. Given a feature X, we apply the UMODL discretization method to find the optimal uplift discretization model as presented in Section 3.1.
2. Compute for X an importance score (described below), denoted by $imp.s(X)$, which is the divergence measure of the treatment effect over the found intervals.
3. We repeat these steps for each feature of the dataset.
4. All features with $imp.s(X) > 0$ are considered relevant for the uplift estimation, while any feature with $imp.s(X) = 0$ is eliminated.

We define $imp.s(X)$ as follows. Assuming $p_i = P_i(Y = 1|T = 1)$ and $q_i = P_i(Y = 1|T = 0)$. We define:

$$imp.s(X) = \begin{cases} \sum_{i=1}^{I} \frac{N_i}{N} D(p_i : q_i), & \text{if } I > 1 \\ 0, & \text{otherwise .} \end{cases} \tag{11}$$

where the distribution divergence measure D is the squared euclidean distance. We choose the squared euclidean distance for the divergence since it is symmetric and stable [24]. UMODL-FS considers irrelevant for the uplift estimation any feature with $imp.s(X) = 0$ and keeps for the uplift modeling any feature with $imp.s(X) > 0$. When UMODL finds a single interval for a feature, it means there is only one distribution for all instances and thus a non-informative feature (i.e . $imp.s(X) = 0$). Unlike feature selection methods of the literature, our approach does not require parameters to set, and there is no need to give the number of features to keep or delete.

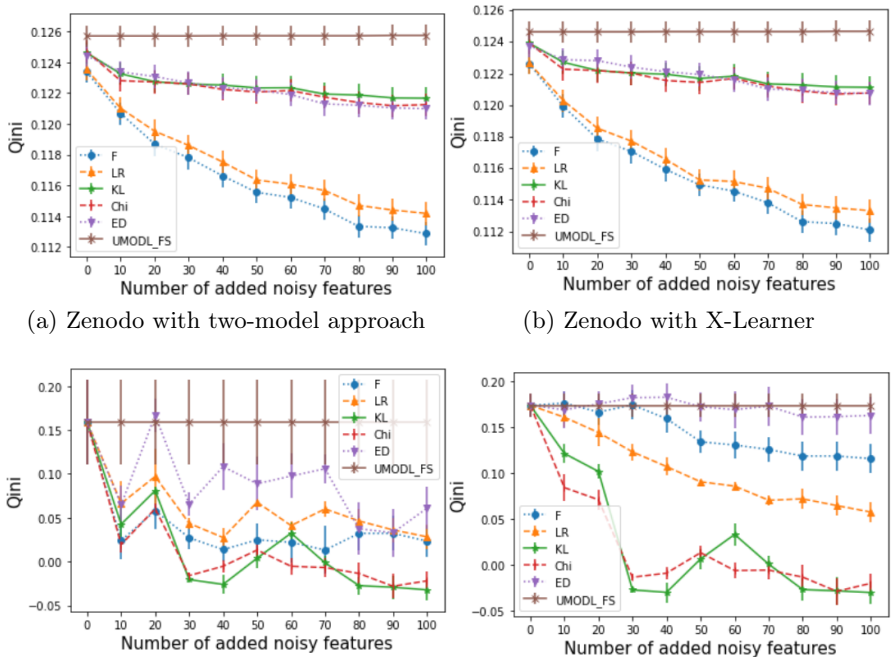
Experimental Protocol. For comparing UMODL-FS to the state-of-art uplift feature selection methods (cf. Section 2.2), we design the following experimental protocol:

1. For each dataset, we generate eleven variants of the dataset, each with an incremental total number (from 0 to 100) of noise features. Noise features are sampled from $\mathcal{N}(0, 1)$ for each of the treatment and control groups.
2. For each variant, we apply the following feature selection methods: (a) KL–filter (b) Chi-filter (c) ED-filter (d) LR-filter (e) F-filter (f) UMODL-FS. For KL-filter, Chi-filter and ED-filter, we set the number of bins to 10.
3. To have the same number of features for each feature selection method and perform a fair comparison, we pick the M most important features, where M is the number of all features deemed informative by UMODL-FS.
4. With these sets of features, we build uplift models: a two-model approach with logistic regression [11] and X-Learner with linear regression [13].
5. The learning process is done with stratified ten-fold cross-validation. Test samples are used to evaluate the performance of uplift models based on the selected features.
6. The qini coefficient metric [3] is used to evaluate the performance of the uplift model.

Datasets. Experiments are conducted on two publicly available continuous datasets which are usual on the uplift community:

1. Criteo dataset [4]: a real large scale dataset constructed by assembling data resulting from several incrementality tests in advertising. In the experiments, we use a random sample of 10,000 instances with the 'visit' variable as outcome variable.
2. Zenodo synthetic dataset [6]: this dataset was created for evaluating feature selection methods for uplift modeling. It has three types of features: (a) uplift features influencing the treatment effect on the conversion probability (outcome variable is 'conversion'); (b) classification features influencing the conversion probability independent of the treatment effect; (c) irrelevant features. This dataset consists of 100 trials of different patterns. Each trial has 10,000 instances and 36 features.

[6] https://doi.org/10.5281/zenodo.3653141



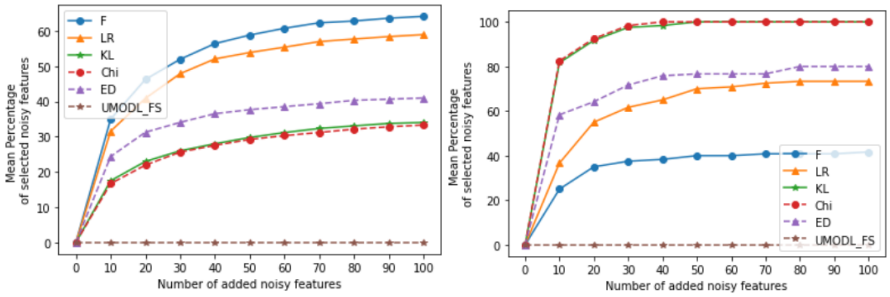
(a) Zenodo with two-model approach (b) Zenodo with X-Learner

(c) Criteo with two-model approach (d) Criteo with X-Learner

Fig. 6: Average qini and its variance according to the number of added noisy features. The X-axis indicates the total number of added noisy features. Y-axis represents the qini values achieved by uplift models.

Results. Fig. 6 presents the results on the use of UMODL-FS for uplift modeling. In all experiments, UMODL-FS selects the set of features leading to the uplift model with the best qini (therefore the best uplift model) whatever the used uplift approach. Remarkably, the more noisy features are added, the more the qini difference between UMODL-FS and other feature selection methods increases.

Fig. 7 indicates the percentage of added noisy features which are selected by the different feature selection methods according to the number of added noisy features. UMDOL-FS never selects a noisy feature. It illustrates the clear ability of UMODL-FS to remove noisy features. On the contrary, all other methods select noisy features and the percentage of the selected noisy ones increases as the number of added noisy features increases. To sum up, the more the number of added noisy features, the more the feature selection methods of the literature select irrelevant features as informative. In contrast, UMODL-FS always neglects irrelevant features and has the most stable qini. Moreover, UMODL-FS does not require to set a parameter giving the number of features to keep.



(a) Zenodo (b) Criteo

Fig. 7: Percentage of selected noisy features according to the number of added noisy features.

6 Conclusion and future work

In this paper, we have proposed a new non-parametric Bayesian approach for uplift discretization and feature selection. We have defined UMODL, a Bayes optimal evaluation criterion of a discretization model for uplift modeling and a search algorithm to find the best model. We have experimentally shown that UMODL is an efficient and accurate uplift estimator through discretization. Then we have presented UMODL-FS, a feature selection method for uplift. Experiments demonstrate that UMODL-FS properly removes irrelevant features and clearly outperforms state of the art methods by providing uplift models with the highest and most stable qini. The method is parameter free, making it easy to use.

This work opens several perspectives. It is promising to study this approach in the case of multiple treatments and multiple outcomes. On the other hand, as decision trees are based on discretized variables, this approach can be investigated to develop tree-based uplift modeling algorithms.

References

1. Boullé, M.: MODL: A bayes optimal discretization method for continuous attributes. Mach. Learn. **65**(1), 131–165 (2006)
2. Chandrashekar, G., Sahin, F.: A survey on feature selection methods. Computers & Electrical Engineering **40**(1), 16–28 (2014)
3. 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)
4. Diemert, E., Betlei, A., Renaudin, C., Amini, M.R.: A Large Scale Benchmark for Uplift Modeling. In: KDD. London, United Kingdom (2018)
5. Glover, S., Dixon, P.: Likelihood ratios: A simple and flexible statistic for empirical psychologists. Psychonomic bulletin and review **11**, 791–806 (11 2004)
6. Grünwald, P.: The minimum description length principle. Adaptive computation and machine learning, MIT Press (2007)

7. Guelman, L.: Optimal personalized treatment learning models with insurance applications. Ph.D. thesis, Universitat de Barcelona (2015)
8. Gutierrez, P., Gérardy, J.Y.: Causal inference and uplift modelling: A review of the literature. In: PAPIs (2016)
9. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2003)
10. Habbema, J., Hermans, J.: Selection of variables in discriminant analysis by f-statistic and error rate. Technometrics **19**(4), 487–493 (1977)
11. Hitsch, G.J., Misra, S.: Heterogeneous treatment effects and optimal targeting policy evaluation. Randomized Social Experiments eJournal (2018)
12. Hu, J.: Customer feature selection from high-dimensional bank direct marketing data for uplift modeling. Journal of Marketing Analytics pp. 1–12 (2022)
13. Jacob, D.: Cate meets ml. Digital Finance **3**(2), 99–148 (2021)
14. Jaskowski, M., Jaroszewicz, S.: Uplift modeling for clinical trial data. In: ICML Workshop On Clinical Data Analysis (2012)
15. Kennedy, E.H.: Towards optimal doubly robust estimation of heterogeneous causal effects (2020), https://arxiv.org/abs/2004.14497
16. Liu, H., Setiono, R.: Feature selection via discretization. IEEE Trans. Knowl. Data Eng. **9**(4), 642–645 (1997)
17. Lo, V., Pachamanova: From predictive uplift modeling to prescriptive uplift analytics: A practical approach to treatment optimization while accounting for estimation risk. Journal of Marketing Analytics (2015)
18. Lunceford, J.K., Davidian, M.: Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in medicine **23 19**, 2937–60 (2004)
19. Radcliffe, N.: Using control groups to target on predicted lift: Building and assessing uplift model. Direct Marketing Analytics Journal pp. 14–21 (2007)
20. 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)
21. Radcliffe, N.J., Surry, P.D.: Real-world uplift modelling with significance-based uplift trees. Stochastic Solutions (2011)
22. Rissanen, J.: Modeling by shortest data description. Autom. **14**(5), 465–471 (1978)
23. Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology **66**, 688–701 (1974)
24. Rzepakowski, P., Jaroszewicz, S.: Decision trees for uplift modeling with single and multiple treatments. Knowl. Inf. Syst. **32**(2), 303–327 (2012)
25. Sharmin, S., Shoyaib, M., Ali, A.A., Khan, M.A.H., Chae, O.: Simultaneous feature selection and discretization based on mutual information. Pattern Recognit. **91**, 162–174 (2019)
26. Zhao, Y., Fang, X., Simchi-Levi, D.: Uplift modeling with multiple treatments and general response types. In: Chawla, N.V., Wang, W. (eds.) SIAM Int. Conf. on Data Mining, Houston, Texas, USA, April 27-29, 2017. pp. 588–596. SIAM (2017)
27. Zhao, Z., Zhang, Y., Harinen, T., Yung, M.: Feature selection methods for uplift modeling. CoRR **abs/2005.03447** (2020), https://arxiv.org/abs/2005.03447