Paper

Uplift Modeling in Direct Marketing

Piotr Rzepakowski and Szymon Jaroszewicz

National Institute of Telecommunications, Warsaw, Poland

Abstract—Marketing campaigns directed to randomly selected customers often generate huge costs and a weak response. Moreover, such campaigns tend to unnecessarily annoy customers and make them less likely to answer to future communications. Precise targeting of marketing actions can potentially results in a greater return on investment. Usually, response models are used to select good targets. They aim at achieving high prediction accuracy for the probability of purchase based on a sample of customers, to whom a pilot campaign has been sent. However, to separate the impact of the action from other stimuli and spontaneous purchases we should model not the response probabilities themselves, but instead, the change in those probabilities caused by the action. The problem of predicting this change is known as uplift modeling, differential response analysis, or true lift modeling. In this work, tree-based classifiers designed for uplift modeling are applied to real marketing data and compared with traditional response models, and other uplift modeling techniques described in literature. The experiments show that the proposed approaches outperform existing uplift modeling algorithms and demonstrate significant advantages of uplift modeling over traditional, response based targeting.

Keywords— decision trees, information theory, marketing tools, uplift modeling.

1. Introduction

When a customer is not completely anonymous, a company can send marketing offers directly to him/her. For example an Internet retailer's product offer can be sent by e-mail or by traditional post; telecommunication operators may advertise their services by SMS, voice calls or other communication channels.

However, to make campaigns effective they should be directed selectively to those who, with high probability, will respond positively (will, e.g., buy a product, or visit a web site). Properly targeted campaign will give a greater return on investment than a randomly targeted one, and, what is even more important, it will not annoy those who are not interested in the offer. It is well known in the direct marketing community that campaigns do put off some customers. There are however few methods available to identify them. See [1]–[4] for more detailed information.

In this paper we experimentally verify the above claims on real direct marketing data. The data is publicly available [5] and comes from an online retailer offering women's and men's merchandise; the next section gives a more detailed description. We test both standard, response based models, as well as uplift approaches described in literature and compare them with decision trees designed especially for uplift modeling, which we introduced in [6], [7]. The experiments verify that the uplift approach gives much better marketing results. Moreover, we demonstrate that our decision trees, designed especially for uplift modeling, outperform other uplift approaches described in literature.

2. Problem Statement

In this section, we describe the marketing data on which we have tested our models. The dataset [5], provided on Kevin Hillstrom's MineThatData blog, contains results of an e-mail campaign for an Internet based retailer. The dataset [5] contains information about 64,000 customers who last purchased within at most twelve months. The customers were subjected to a test e-mail campaign:

- 1/3 were randomly chosen to receive an e-mail campaign featuring men's merchandise,
- 1/3 were randomly chosen to receive an e-mail campaign featuring women's merchandise,
- 1/3 were randomly chosen to not receive an e-mail.

The data describes customer behavior for two weeks after the campaign. The details of the dataset are summarized in Tables 1 and 2.

Table 1
Hillstrom's marketing data: customers' attributes

Attribute	Definition
Recency	Months since last purchase
History_Segm	Categorization of dollars spent in the past year
History	Actual dollar value spent in the past year
Mens	1/0 indicator, 1 = customer purchased mens merchandise in the past year
Womens	1/0 indicator, 1 = customer purchased womens merchandise in the past year
Zip_Code	Classifies zip code as urban, suburban, or rural
Newbie	1/0 indicator, 1 = new customer in the past twelve months
Channel	Describes the channels the customer purchased from in the past year

Table 2
Hillstrom's marketing data: type of e-mail campaign sent and activity in the two following weeks

Attribute	Definition
Segment	E-mail campaign the customer received
Visit	1/0 indicator, 1 = customer visited website in the following two weeks
Conversion	1/0 indicator, 1 = customer purchased merchandise in the following two weeks
Spend	Actual dollars spent in the following two weeks

The author asked several questions to be answered based on the data. Here we address the problem of predicting the people who visited the site within the two-week period (attribute *Visit* in Table 2) because they received the campaign. The estimate was based by comparing customer behavior on the treatment and control groups, i.e., comparing customers who did and did not receive an e-mail.
During an initial analysis we have found that about 10.62% of the people visited the site spontaneously, but after the campaign (combined men's and women's) the visits increased to 16.7%. Men's merchandise campaign outperformed women's, as the increase in visits was about 7.66% (from 10.62% to 18.28%), while the women's merchandise campaign resulted in an increase of only 4.52% (from 10.62% to 15.14%).
Afterward, we used traditional response based targeting, as well as uplift modeling based targeting to select the customers for the campaign. Because there is a large difference in response between treatment groups who received advertisements for men's and women's merchandise, the two campaign types were analyzed, both jointly and separately. In the first case, the treatment group consists of all those who received an e-mail and the control group of those who did not. In the second case, there are two treatment groups, one for man's and one for women's merchandise campaign; both treatment groups are analyzed separately with respect to the same control group. Since the men's merchandise group showed little sensitivity to attribute values, our experiments focused primarily on the women's merchandise group.
The following two sections give the literature overview, describe the uplift modeling methodology used and compare it to the traditional predictive modeling. Section 5 presents experimental results.

3. Uplift Modeling

In this section we give a more detailed overview of uplift modeling and review available literature.
Traditionally used response models are built on a sample of data about the customers. Each record in the dataset represents a customer and the attributes describe his/her characteristics. In the *propensity* models, historical information about purchases (or other success measures like visits) is used, while in the *response* models, all customers have been subject to a pilot campaign. A distinguished class attribute informs on whether a customer responded to the offer or not. Afterward, the data is used to build a model that predicts conditional probability of response *after* the campaign. This model is then applied to the whole customer database to select people with high probability of purchasing the product. The process is illustrated in Fig. 1.

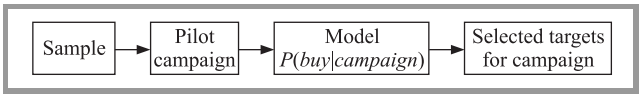

Fig. 1. Response model creation process.

However, in reality, we can divide the customers into four groups, i.e., those who:

- responded *because* of the action,
- responded *regardless* of the action (unnecessary costs),
- did not respond and the action had *no impact* (unnecessary costs),
- did not respond *because* the action had a *negative impact* (*e.g.* a customer got annoyed by the campaign, might even have churned).

Propensity models, as well as traditional response models are not capable of distinguishing those four groups, while uplift models can do that. This is because traditional models predict the conditional class probability

$$P(response|treatment),$$

while uplift models predict the change in behavior resulting from the action

$$P(response|treatment) - P(response|no\,treatment).$$

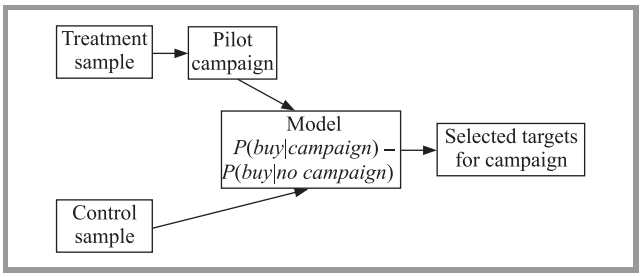

Fig. 2. Uplift model creation process.

To build an uplift model, a random sample (the *treatment* dataset) of customers is selected and subjected to the marketing action. Disjoint sample is also selected (the *control* dataset), to which the action is not applied, and which serves as the background against which the results of the action will be measured. The model is now built for predicting the *difference* between class probabilities on the two sets of data. The process is illustrated in Fig. 2.

3.1. Literature Overview

The problem of uplift modeling has received little attention in literature – a surprising fact, if one considers its practical importance.

There exist two overall approaches to uplift modeling. The first, obvious one is to build two separate classifiers. One on the treatment and another on the control dataset (as shown in Fig. 3). For each classified object class probabilities predicted by the control model are subtracted from those predicted by the treatment model, giving a direct estimate of the difference in behavior caused by the action.

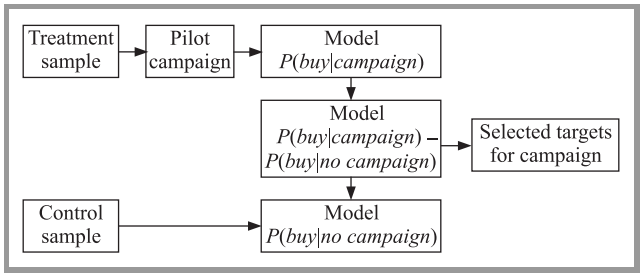

Fig. 3. Uplift model based on two separate classifiers

This approach has a major disadvantage: the behavior of the differences between class probabilities can be very different than the behavior of the class probabilities themselves. Thus, it is possible that the models will focus too much on modeling the class in both datasets, instead of focusing on the differences between them. The problem is exacerbated by the fact that the variation in the difference between class probabilities is usually much smaller than variability in class probabilities themselves. For example, in case of decision trees, the double model approach does not necessarily favor splits, which lead to *different* responses in treatment and control groups, just splits, which lead to predictable outcomes in each of the groups separately, wasting valuable data. See [1], [4], [8], [9] for details.

Let us now look at the second type of approaches, which attempt to model the difference between treatment and control probabilities directly.

One of the first 'native' uplift modeling approaches builds a single decision tree, by trying to maximize the uplift criterion at each step [1]. The splitting criterion used by the algorithm, called $\Delta\Delta P$, selects tests, which maximize the difference between the differences between treatment and control probabilities in the left and right subtrees. This corresponds to maximizing the desired difference, directly in the fashion of greedy algorithms. More formally, suppose we have a test A with outcomes a_0 and a_1. The $\Delta\Delta P$ splitting criterion is defined as

$$\Delta\Delta P(A) = \Big|\big(P^T(y_0|a_0) - P^C(y_0|a_0)\big) - \big(P^T(y_0|a_1) - P^C(y_0|a_1)\big)\Big|,$$

where y_0 is a selected (positive) class. The calculation of the criterion for subtree is illustrated in Fig. 4.

While the original $\Delta\Delta P$ criterion works only for binary trees and two-class problems, we have generalized it in [6], [7] to multiway splits and multiclass problems to make comparisons with other methods easier.

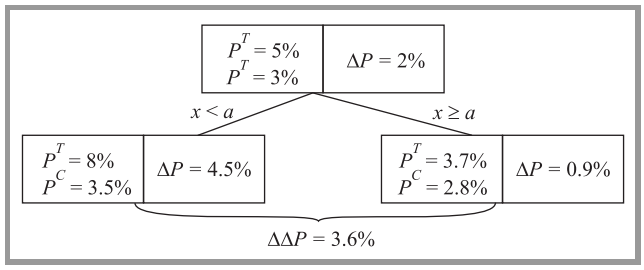

Fig. 4. An example calculation of the $\Delta\Delta P$ criterion

The first paper explicitly discussing uplift modeling was [3]. It presents an extensive motivation including several used cases. Recently, a detailed description of their decision tree learning algorithm has been published in [4]. The decision trees have been adapted to the uplift case by using a splitting criterion, based on statistical tests of the differences between treatment and control probabilities introduced by the split. There is also a variance based pruning technique. See [4] for more details.

Other approaches to uplift modeling include modifications of the naive Bayesian classifier and logistic regression [10], or different approaches to uplift decision tree learning, see e.g., [9].

In [6], [7] we have presented another algorithm for learning uplift decision trees. Our approach follows the more modern tree learning algorithms which use information theory for test selection. We describe it in the next section.

4. Information Theory Based Uplift Decision Trees

In [6], [7] we presented an approach to uplift decision tree learning more in the spirit of modern learning algorithms (such as Quinlan's C4.5 [11]) with tests selected based on information theoretical measures, and overfitting controlled by tree pruning. The first paper presented the case where all customers receive and identical offer, the second extended the approach to the case when multiple treatments are possible. In the remaining part of the paper we only deal with the single treatment case. This section provides a description of those algorithms, which, while being quite thorough, leaves out several details. The reader is referred to [6], [7] for a full description.

4.1. Notation

Let us now introduce the notation used in this section. Recall that nonleaf nodes in a decision trees are labeled with *tests* [11]. We create a single test for each categorical attribute, the outcomes of this test are all attribute's values. For each numerical attribute X we create tests of

the form $X < v$, where v is a real number. Tests will be denoted with uppercase letter A and the class attribute with the letter Y. Values from the domains of attributes and test outcomes will be denoted by corresponding lowercase letters, e.g., a will denote an outcome of a test A, and y a specific class; $\sum_a$ denotes the sum over all outcomes of a test A, and $\sum_y$ the sum over all classes.

We need to introduce special notation reflecting the fact, that, contrary to the standard Machine Learning setting, we now have *two* training datasets: treatment and control. The probabilities estimated from the treatment dataset will be denoted by P^T and those estimated from the control dataset by P^C. We assume that Laplace correction is used while estimating the probabilities P^T and P^C.

Additionally, let N^T and N^C denote the number of records in the treatment and control samples respectively, and $N^T(a)$ and $N^C(a)$, the number of records in which the outcome of a test A is a. Finally let $N = N^T + N^C$ and $N(a) = N^T(a) + N^C(a)$.

4.2. Splitting Criteria

One of the most important aspects of a decision tree learning algorithm is the criterion used to select tests in the nodes of the tree. In this section we present two uplift specific splitting criteria. Instead of using the target quantity directly, we attempt to model the amount of *information* that a test gives about the difference between treatment and control class probabilities. In [6], [7] we stated several postulates which an uplift splitting criterion should satisfy, and proved that our criteria do indeed satisfy them.

The splitting criteria we propose are based on distribution divergences [12]–[15] – information theoretical measures of differences between distributions. We use two distribution divergence measures, the Kullback-Leibler divergence [12], [14] and the squared Euclidean distance [13]. Those divergences, from a distribution $Q = (q_1, \ldots, q_n)$ to a distribution $P = (p_1, \ldots, p_n)$, are defined respectively as

$$\begin{aligned} KL(P:Q) &= \sum_i p_i \log \frac{p_i}{q_i}, \\ E(P:Q) &= \sum_i (p_i - q_i)^2. \end{aligned}$$

Given a divergence measure D, our splitting criterion is

$$D_{gain}(A) = D\left(P^T(Y) : P^C(Y)|A\right) - D\left(P^T(Y) : P^C(Y)\right),$$

where A is a test and $D\left(P^T(Y) : P^C(Y)|A\right)$, the conditional divergence defined below. Substituting for D the KL-divergence and squared Euclidean distance divergence we obtain our two proposed splitting criteria, the KL_{gain} and E_{gain}.

To justify the definition, note that we want to build the tree, in which the distributions in the treatment and control groups differ as much as possible. The first part of the expression picks a test, which leads to most divergent class distributions in each branch; from this value we subtract the divergence between class distributions on the whole dataset in order to measure the increase or *gain* of the divergence resulting from splitting with test A. This is analogous to entropy gain [11] and Gini gain [16] used in standard decision trees. In fact, one of our postulates was that, when the control dataset is missing the splitting criteria should reduce to entropy and Gini gains respectively [6].

Conditional KL-divergences have been used in literature [14] but the definition is not directly applicable to our case, since the probability distributions of the test A differ in the treatment and control groups. We have thus defined conditional divergence as:

$$D(P^T(Y) : P^C(Y)|A) = \sum_a \frac{N(a)}{N} D\left(P^T(Y|a) : P^C(Y|a)\right). \tag{1}$$

The relative influence of each test value is proportional to the total number of training examples falling into its branch in both treatment and control groups.

4.3. Correcting for Tests with Large Number of Splits and Imbalanced Treatment and Control Splits

In order to prevent a bias towards tests with high number of outcomes decision, tree learning algorithms normalize the information gain dividing it by the information value of the test itself [11]. In our case the normalization factor is more complicated, as the information value can be different in the control and treatment groups. Moreover, it is desirable to punish tests, which split the control and treatment groups in different proportions, since such splits indicate that the test is not independent from the assignment of cases to the treatment and control groups.

The proposed normalization value for a test A is given by

$$\begin{aligned} I(A) = H\left(\frac{N^T}{N}, \frac{N^C}{N}\right) KL(P^T(A) : P^C(A)) \\ + \frac{N^T}{N} H(P^T(A)) + \frac{N^C}{N} H(P^C(A)) + \frac{1}{2}, \end{aligned} \tag{2}$$

for the KL_{gain} criterion, and

$$\begin{aligned} J(A) = Gini\left(\frac{N^T}{N}, \frac{N^C}{N}\right) E(P^T(A) : P^C(A)) \\ + \frac{N^T}{N} Gini(P^T(A)) + \frac{N^C}{N} Gini(P^C(A)) + \frac{1}{2}, \end{aligned}$$

for the E_{gain} criterion.

The first term is responsible for penalizing uneven splits. The unevenness of splitting proportions is measured using the divergence between the distributions of the test outcomes in the treatment and control datasets. However, penalizing uneven splits only makes sense if there is enough data in *both* treatment and control groups. The $KL(P^T(A) : P^C(A))$ term is thus multiplied by $H\left(\frac{N^T}{N}, \frac{N^C}{N}\right)$, which is close to zero when there is a large imbalance between the number of data in treatment and control groups (analogous, Gini based measures are used for E_{gain}). The

following two terms penalize tests with large numbers of outcomes, just as in classification decision trees [11]. The final $\frac{1}{2}$ term prevents the division by very small normalization factors from inflating the value of the splitting criterion for tests with highly imbalanced outcome probabilities. Notice that when $N^C = 0$ the criterion reduces to $H(P^T(A)) + \frac{1}{2}$ which is identical to normalization used in standard decision tree learning (except for the extra $\frac{1}{2}$). After taking the normalization into account, the final splitting criteria become

$$\frac{KL_{ratio}(A)}{I(A)}, \qquad \text{and} \qquad \frac{E_{ratio}(A)}{J(A)}.$$

4.4. Application of the Tree

Once the tree has been built, its leaves correspond to subgroups of objects, for which the treatment and control class distributions differ. The question now is how to apply the tree to make decisions on whether the marketing action should be applied to customers falling into a given leaf. To this end, we annotate each leaf with an expected profit, which can also be used for scoring new customers.

The assignment of profits uses an approach similar to [1], [9]. Each class y is assigned to profit v_y, that is, the expected income if a given object (whether treated or not) falls into this class. If each object in a leaf l is targeted, the expected profit (per object) is equal to $-c + \sum_y P^T(y|l)v_y$, where c is the cost of performing the action. If no object in l is targeted, the expected profit is $\sum_y P^C(y|l)v_y$. Combining the two, we get the following expected gain from treating each object falling into l:

$$-c + \sum_y v_y \left(P^T(y|l) - P^C(y|l)\right). \tag{3}$$

4.5. Pruning

Decision tree pruning has decisive influence on the performance of the model. There are several pruning methods, based on statistical tests, Minimum Description Length principle, and others [11], [17]–[19].

We chose the simplest, but nevertheless effective pruning method based on using a separate validation set [17], [18]. For the classification problem, after the full tree has been built on the training set, the method traverses the tree bottom up and tests, for each node, whether replacing the subtree rooted at that node with a single leaf would improve accuracy on the validation set. If this is the case, the subtree is replaced, and the process continues.

Applying this method to uplift modeling required an analogue of classification accuracy. To this end we have devised a measure of improvement called the *maximum class probability difference*, which can be viewed as a generalization of classification accuracy to the uplift case. The idea is to look at the differences between treatment and control probabilities in the root of the subtree and in its leaves, and prune if, overall, the differences in leaves are not greater than the difference in the root. In each node we only look at the class, for which the difference was largest on the training set, and in addition remember the sign of that difference such that only differences, which have the same sign in the training and validation sets contribute to the increase of our criterion.

More formally, while building the tree on the *training* set, for each node t, we store the class $y^*(t)$, for which the difference $\left|P^T(y^*|t) - P^C(y^*|t)\right|$ is maximal, and also remember the sign of this difference $s^*(t) = \text{sgn}(P^T(y^*|t) - P^C(y^*|t))$. During the pruning step, suppose we are examining a subtree with root r and leaves $l_1, \ldots, l_k$. We calculate the following quantities with the stored values of y^* and s^*, and all probabilities computed on the *validation* set:

$$\begin{aligned} d_1(r) &= \sum_{i=1}^{k} \frac{N(l_i)}{N(r)} s^*(l_i) \left(P^T(y^*(l_i)|l_i) - P^C(y^*(l_i)|l_i)\right), \\ d_2(r) &= s^*(r) \left(P^T(y^*(r)|r) - P^C(y^*(r)|r)\right), \end{aligned}$$

where $N(l_i)$ is the number of validation examples (both treatment and control) falling into the leaf l_i. The first quantity is the maximum class probability difference of the unpruned subtree and the second is the maximum class probability difference we would obtain on the validation set, if the subtree was pruned and replaced with a single leaf. The subtree is pruned if $d_1(r) \leq d_2(r)$.

The class y^* is an analogue of the predicted class in standard classification trees. In [7] we describe the relation of maximum class probability difference to classification accuracy.

5. Experimental Evaluation on Direct Marketing Data

We now present an application of uplift models, as well as traditional response models to the problem of selection of customers for an e-mail campaign based on the data described in Section 2. The target is to maximize the num-

Table 3
Models used in the experiments

Response models	
SingleTree.E	Decision tree model based on the E_{ratio} criterion
SingleTree.KL	Decision tree model based on the KL_{ratio} criterion
SingleTree.J48	Decision tree model based on J48 Weka implementation
Uplift models	
UpliftTree.E	Uplift decision tree based on the E_{ratio} criterion
UpliftTree.KL	Uplift decision tree based on the KL_{ratio} criterion
DoubleTree.J48	Separate decision trees for the treatment and control groups (J48 Weka implementation)

ber of visits to the web site that were *driven* by the campaign.

We compared six different models, three response models and three uplift models (Table 3).

The models were evaluated using 10×10 crossvalidation, all figures present results obtained on the test folds.

We begin by building models with both types of campaign e-mails treated jointly. The results for traditional response models are presented in Fig. 5. The figure shows cumulative percent of total page visits for customers sorted from the highest to the lowest score. The area under the curve for each model is included in the legend. The given value is the actual area under the curve, from which the area under the diagonal line corresponding to random selection is subtracted. The greater the area, the better. We can see that all traditional response models perform much better at predicting who will visit the site than random selection.

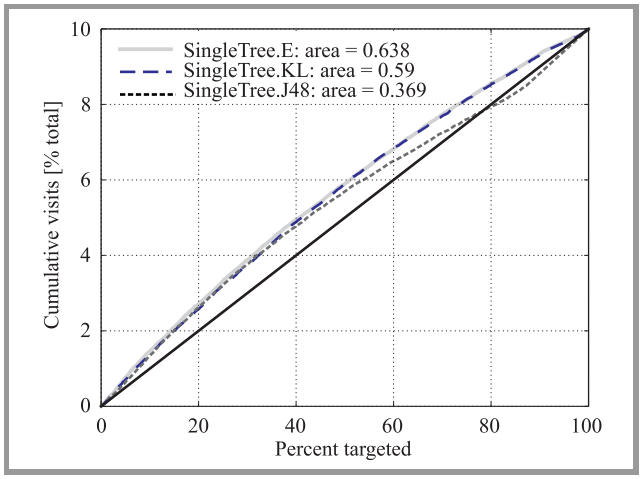

Fig. 5. Cumulative visits (lift) predicted by classification models built just on the treatment dataset.

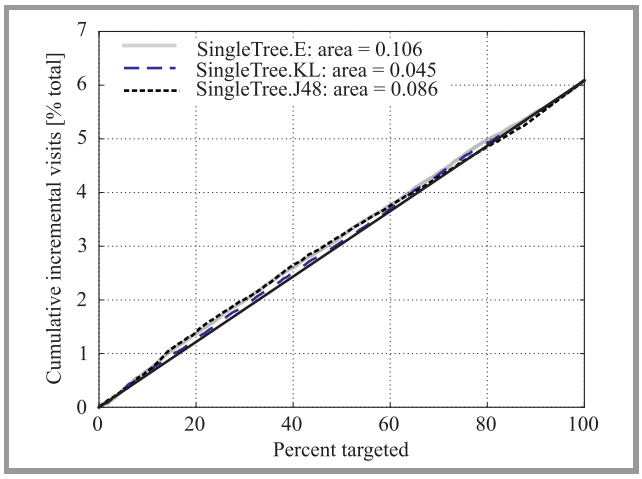

Fig. 6. Cumulative incremental visits (uplift) predicted by classification models built just on the treatment dataset.

Traditional models predict all possible visits, so they indicate as positive customers visit the site spontaneously, as well as those who visit as a result of the campaign. However, those models are not successful in predicting new visits. To indicate this, Fig. 6 shows the cumulative percentage (of the total population) of the *new visits*. The curve is obtained by subtracting two gain curves (such as those used in Fig. 5): the one obtained on the control dataset from the one obtained on the treatment dataset. Areas under those curves are also indicated. Fig. 7 includes the same results for dedicated uplift models.

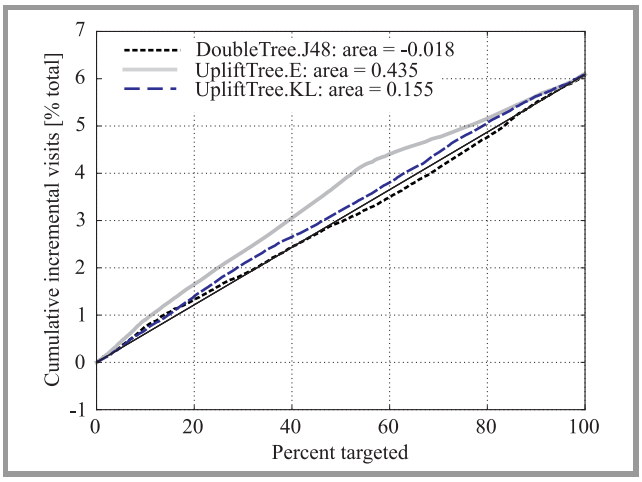

Fig. 7. Cumulative incremental visits (uplift) predicted by uplift models built on treatment and control datasets.

Results presented in Fig. 6 and Fig. 7 show that traditional response models are very poor in predicting uplift, i.e., which customers are likely to visit the site *because* of the campaign (areas under their uplift curves are practically equal to random selection), even though they are highly effective in predicting who will visit the site, i.e., combined spontaneous and campaign induced visits. This is not what a marketer is looking for, because targeting customers, which have high response scores does not generate a tangible increase in the number of visits.

In contrast, uplift models perform much better at predicting new visits. This is especially true for the model based on the E_{ratio} criterion, which very significantly outperformed all response based models. The KL_{ratio} based model performed much worse than the E_{ratio} based, but still outperforms traditional response models. The approach based on two separate models also performed poorly, confirming the superiority of dedicated uplift approaches.

Below, we show two top levels of an uplift decision tree for combined men's and women's merchandise campaigns (UpliftTree.E built on one of the crossvalidation folds). The *womens* attribute gives the most information about the increase in visits, and is placed in the root of the tree. It splits the data more or less in half. In a subgroup of 55.3% of the customers ($womens = 1$) we reached an uplift of 7.9% and in 45% of this subgroup ($zip\_code = Suburban$) an uplift of 8.4%. This is much more than the average uplift of 6.1%. In a small group ($womens = 0$, $history \geq 1621.49$) the uplift is negative (-17.3%); the campaign had a nega-

tive effect on this group (note that these are highly valuable customers who made large purchases before).

UpliftTree.E (Combined campaigns):
Total uplift = 6.1%

- [44.7%] $womens = 0$: $uplift = 3.8\%$
  - [0.1%] $history \geq 1621.49$: $uplift = -17.3\%$
  - [99.9%] $history < 1621.49$: $uplift = 3.9\%$
- [55.3%] $womens = 1$: $uplift = 7.9\%$
  - [14.8%] $zip\_code = Rural$: $uplift = 5.9\%$
  - [45.0%] $zip\_code = Suburban$: $uplift = 8.4\%$
  - [40.2%] $zip\_code = Urban$: $uplift = 8.1\%$

Next, new models were built on women's and men's merchandise campaign data separately. As the results for the men's merchandise campaign showed little dependence on customers' attributes, we show only the results for the women's merchandise campaign. The results are presented in Figs. 8, 9 and 10). The advantage of uplift models is much more pronounced than in the case of both campaigns treated jointly. The KL_{ratio} based model worked very well in this case, its performance was practically identical to that of the E_{ratio} based model, and much better than the performance of the model based on two separate decision trees. It is enough to target just about half of the customers to achieve results almost identical to targeting the whole database.

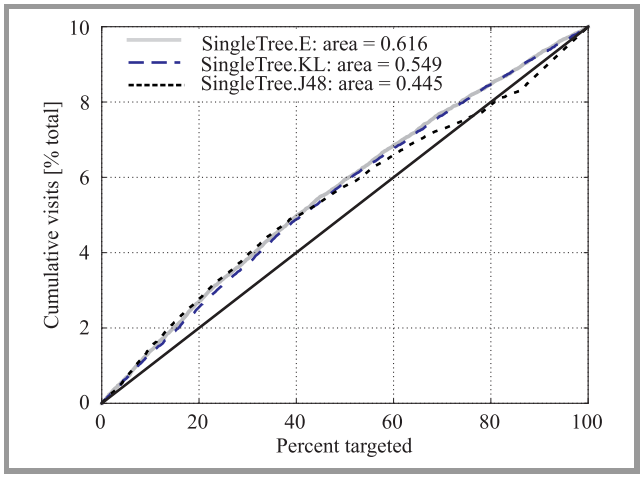

Fig. 8. Cumulative visits (lift) after the women's merchandise campaign predicted by classification models built just on the treatment dataset.

We now look at the top two levels of an uplift tree model build on the data from women's merchandise campaign. We can see that also for this group the *women's* attribute is very important. In a group of 55.3% of the customers ($womens = 1$) the uplift is 7.3%. It means that by directing the campaign to this group we can encourage $55.3\% \times 7.3\% = 4.04\%$ of the total population to visit our site.

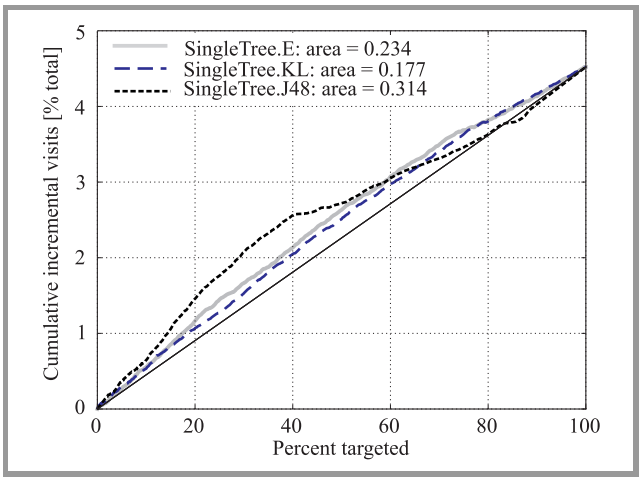

Fig. 9. Cumulative incremental visits (uplift) after women's campaign predicted by classification models built just on the treatment dataset.

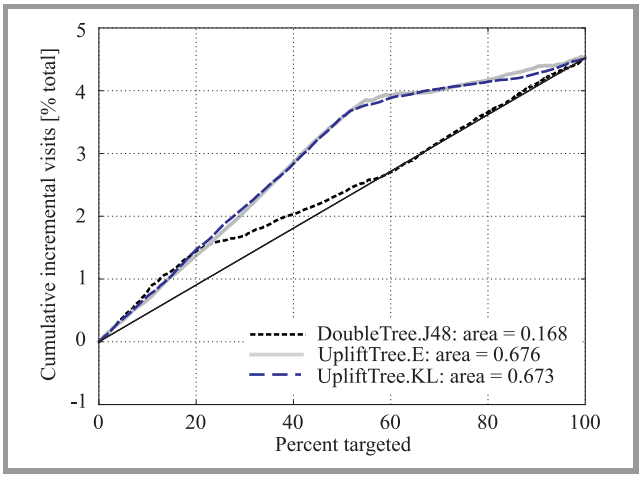

Fig. 10. Cumulative incremental visits (uplift) after women's campaign predicted by uplift models built on the treatment and control datasets.

UpliftTree.E (Women's merchandise campaign):
Total uplift = 4.5%

- [44.9%] $womens = 0$: $uplift = 1.1\%$
  - [0.2%] $history \geq 1618.85$: $uplift = -26.3\%$
  - [99.8%] $history < 1618.85$: $uplift = 1.1\%$
- [55.3%] $womens = 1$: $uplift = 7.3\%$
  - [0.9%] $history \geq 1317.02$: $uplift = -9.4\%$
  - [99.1%] $history < 1317.02$: $uplift = 7.5\%$

6. Conclusions

Our experiments confirm the usefulness of uplift modeling in campaign optimization. Using uplift models, we can predict new buyers much more precisely than using traditional response or propensity approaches. The effectiveness in predicting new visits by response models is low, even if

accuracy of predicting all visits is high. The reason for this is that the response models do not distinguish between spontaneous and new buyers. Quite often, the spontaneous hits are more frequent, and the models tend concentrate on them. Only if the uplift is correlated with the class itself, the response models are able to indicate new buyers.
Additionally, our experiments confirm that dedicated uplift modeling algorithms are more effective than the naive approach based on two separate models.

7. Acknowledgments

This work was supported by Research Grant no. N N516 414938 of the Polish Ministry of Science and Higher Education (Ministerstwo Nauki i Szkolnictwa Wyższego) from research funds for the period 2010–2012.

References

[1] B. Hansotia and B. Rukstales, "Incremental value modeling", *J. Interactive Marketing*, vol. 16, no. 3, pp. 35–46, 2002.

[2] N. J. Radcliffe and R. Simpson, "Identifying who can be saved and who will be driven away by retention activity", *White paper*, Stochastic Solutions Limited, 2007.

[3] N. J. Radcliffe and P. D. Surry, "Differential response analysis: modeling true response by isolating the effect of a single action", in *Proc. Credit Scoring Credit Control VI*, Edinburgh, Scotland, 1999.

[4] N. J. Radcliffe and P. D. Surry, "Real-world uplift modelling with significance-based uplift trees", Portrait Tech. Rep. TR-2011-1, Stochastic Solutions, 2011.

[5] K. Hillstrom, "The MineThatData e-mail analytics and data mining challenge", MineThatData blog, 2008 [Online]. Available: http://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html, retrieved on 02.04.2012.

[6] P. Rzepakowski and S. Jaroszewicz, "Decision trees for uplift modeling", in *Proc. 10th IEEE Int. Conf. Data Mining ICDM-2010*, Sydney, Australia, Dec. 2010, pp. 441–450.

[7] P. Rzepakowski and S. Jaroszewicz, "Decision trees for uplift modeling with single and multiple treatments", *Knowledge and Information Systems*, pp. 1–25, 2011 [Online]. Available: http://www.springerlink.com/content/f45pw0171234524j

[8] C. Manahan, "A proportional hazards approach to campaign list selection", in *Proc. Thirtieth Ann. SAS Users Group Int. Conf. SUGI*, Philadelphia, PA, 2005.

[9] D. M. Chickering and D. Heckerman, "A decision theoretic approach to targeted advertising", in *Proc. 16th Conf. Uncertainty in Artif. Intell. UAI-2000*, Stanford, CA, 2000, pp. 82–88.

[10] V. S. Y. Lo, "The true lift model – a novel data mining approach to response modeling in database marketing", *SIGKDD Explor.*, vol. 4, no. 2, pp. 78–86, 2002.

[11] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kauffman, 1992.

[12] I. Csiszár and P. Shields, "Information theory and statistics: A tutorial", *Found. Trends in Commun. Inform. Theory*, vol. 1, no. 4, pp. 417–528, 2004.

[13] L. Lee, "Measures of distributional similarity", in *Proc. 37th Ann. Meet. Associ. Computat. Linguistics ACL-1999*, Maryland, USA, 1999, pp. 25–32.

[14] T. S. Han and K. Kobayashi, *Mathematics of information and coding*. Boston, USA: American Mathematical Society, 2001.

[15] S. Jaroszewicz and D. A. Simovici, "A general measure of rule interestingness", in *Proc. 5th Eur. Conf. Princ. Data Mining Knowl. Discov. PKDD-2001*, Freiburg, Germany, 2001, pp. 253–265.

[16] L. Brieman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Monterey, USA: Wadsworth Inc., 1984.

[17] T. Mitchell, *Machine Learning*. McGraw Hill, 1997.

[18] J. R. Quinlan, "Simplifying decision trees", *Int. J. Man-Machine Studies*, vol. 27, no. 3, pp. 221–234, 1987.

[19] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005.

Piotr Rzepakowski received his M.Sc. degree in Computer Science from Warsaw University of Technology, Poland, in 2003. Currently, he is a Ph.D. student at the Faculty of Electronics and Information Technology at Warsaw University of Technology and a research assistant at the National Institute of Telecommunications in Warsaw, Poland. His research interests include data mining, data analysis and decision support. He has taken part in several industrial projects related to data warehousing and data analysis.
E-mail: P.Rzepakowski@itl.waw.pl
National Institute of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

Szymon Jaroszewicz is currently an Associate Professor at the National Institute of Telecommunications, Warsaw, Poland and at the Institute of Computer Science of the Polish Academy of Sciences. He received the Master's degree in Computer Science at the Department of Computer Science at the Szczecin University of Technology in 1998 and his Ph.D. at the University of Massachusetts Boston in 2003, where in 1998 and 1999 he was a Fulbright scholar. In 2010, he received his D.Sc. degree at the Institute of Computer Science, Polish Academy of Sciences. His research interests include data analysis, data mining and probabilistic modeling; he is the author of several publications in those fields. He has served as a program committee member for major data mining conferences and he is a member of the editorial board of Data Mining and Knowledge Discovery Journal.
E-mail: S.Jaroszewicz@itl.waw.pl
National Institute of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland