Decision Support Systems 72 (2015) 24–32

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

A decision support framework to implement optimal personalized marketing interventions

Leo Guelman [a,*], Montserrat Guillén [b], Ana M. Pérez-Marín [b]

[a] Royal Bank of Canada, RBC Insurance, 6880 Financial Drive, Mississauga, Ontario L5N 7Y5, Canada
[b] Department of Econometrics, Riskcenter, University of Barcelona, Diagonal 690, Barcelona E-08034, Spain

ARTICLE INFO

Article history:
Received 27 November 2013
Received in revised form 31 October 2014
Accepted 27 January 2015
Available online 3 February 2015

Keywords:
Personalized treatment learning
Causal inference
Marketing interventions

ABSTRACT

In many important settings, subjects can show significant heterogeneity in response to a stimulus or "treatment." For instance, a treatment that works for the overall population might be highly ineffective, or even harmful, for a subgroup of subjects with specific characteristics. Similarly, a new treatment may not be better than an existing treatment in the overall population, but there is likely a subgroup of subjects who would benefit from it. The notion that "one size may not fit all" is becoming increasingly recognized in a wide variety of fields, ranging from economics to medicine. This has drawn significant attention to personalize the choice of treatment, so it is optimal for each individual. An optimal personalized treatment is the one that maximizes the probability of a desirable outcome. We call the task of learning the optimal personalized treatment *personalized treatment learning*. From the statistical learning perspective, this problem imposes important challenges, primarily because the optimal treatment is unknown on a given training set. A number of statistical methods have been proposed recently to tackle this problem. However, considering the critical importance of these methods to decision support systems, personalized treatment learning models have received relatively little attention in the literature. The purpose of this paper is to propose a novel method labeled *causal conditional inference trees* and its natural extension to *causal conditional inference forests*. The performance of the new method is analyzed and compared to alternative methods for personalized treatment learning. The results show that our new proposed method often outperforms the alternatives on the numerical settings described in this article. We also illustrate an application of the proposed method using data from a large Canadian insurer for the purpose of selecting the best targets for cross-selling an insurance product.

1. Introduction

In the past two decades, rapid advances in data collection and storage technology have created vast quantities of data. The field of statistics has been revolutionized by the development of algorithmic and data models [5] in response to challenging new problems coming from science and industry, mostly resulting from an increasing size and complexity in the data structures. In this context, the concept of *learning from data* [1] has emerged as the task of extracting "implicit, previously unknown, and potentially useful information from data" [12]. A distinction is usually made between *supervised* and *unsupervised* learning. In the former, the objective is to predict the value of a response variable based on a collection of *observable* covariates. In the latter, there is no response variable to "supervise" the learning process, and the objective is to find structures and patterns among the covariates.

In many important settings, the values of some covariates are not only observable, but they can be chosen at the discretion of a decision maker [53]. For instance, a doctor can choose the medical treatment for a patient among a set of alternatives, a company can decide the type of marketing intervention activity (direct mail, phone call, e-mail, etc.) to make an offer to a client, and a bank can decide the credit limit to offer a client on a credit card. In all these examples, the objective is not necessarily to predict a response variable with high accuracy but to select the optimal action or "treatment" for each subject based on his or her individual characteristics.[1] Optimal is understood here as the treatment that maximizes the probability of a desirable outcome. We call the task of learning the optimal personalized treatment *personalized treatment learning*.

A key challenge in building decision support systems based on personalized treatment learning models is that the quantity we are trying to predict (i.e., the optimal personalized treatment) is unknown on a given training data set. As each subject can only be exposed to a single

* Corresponding author. Tel.: +1 905 606 1175; fax: +1 905 286 4756.
E-mail addresses: leo.guelman@rbc.com (L. Guelman), mguillen@ub.edu (M. Guillén), amperez@ub.edu (A.M. Pérez-Marín).

[1] Domain knowledge can also play an important role in selecting the optimal treatment [42].

http://dx.doi.org/10.1016/j.dss.2015.01.010
0167-9236/

treatment, the value of the subject's response under alternative treatments is unobserved, a problem also known as *the fundamental problem of causal inference* [18]. This aspect makes this problem unique within the discipline of learning from data.

The underlying motivation for personalized treatment learning is that subjects can show significant heterogeneity in response to treatments, so making an accurate treatment decision for each subject becomes essential. For instance, a new treatment may not be better than an existing treatment in the overall population, but it might be beneficial/harmful for a subgroup of subjects. The idea that "one size may not fit all" has been increasingly recognized in a variety of disciplines, ranging from economics to medicine. Alemi et al. [2] argue that improved statistical methods are needed for personalized treatments and propose an adapted version of the *K-nearest-neighbor* (*KNN*) classifier [8]. Imai and Ratkovic [21] propose a method that adapts the *support vector machine* classifier [49] and then apply it to a widely known data set pertaining to the National Supported Work program [26,10] to identify the characteristics of workers who greatly benefit from (or are negatively affected by) a job training program. Tian et al. [48] propose a method designed to deal with high-dimensional covariates and use it to identify breast cancer patients who may or may not benefit from a specific treatment based on the individual patient's gene expression profile. Liang et al. [28] describe a web-based intervention support system to provide tailored interventions to individual patients with chronic diseases. Xu et al. [51] propose a *Bayesian network* model that integrates with other components to better support personalized mobile advertising applications. In the context of insurance, Guelman et al. [14,15] propose a method based on an adapted version of *random forests* to identify policyholders who are positively/negatively impacted by a client retention program. Also, Guelman and Guillén [16] describe a framework to determine the optimal rate change (playing the role of the treatment) for each individual policyholder for the purpose of maximizing the overall expected profitability of an insurance portfolio.

In addition to the methods discussed above, other methods have been proposed in the literature, mostly in the context of clinical trials and direct marketing [44,31,52,22,27,33,39,46]. However, considering the critical importance of these methods to decision support systems, personalized treatment learning models have received relatively little attention in the literature. The purpose of this paper is to propose a novel method labeled *causal conditional inference trees* and its natural extension to *causal conditional inference forests*. The performance of the new method is compared to the existing methods in an extensive numerical study and analyzed on real-world data. We implement all these methods in a package named *uplift* [13], which is now freely available from the CRAN (Comprehensive R Archive Network) repository under the R statistical computing environment.

This paper is organized as follows. Section 2 defines the scope of the personalized treatment learning problem. In Section 3, we discuss our new proposed method. In Section 4, we report the finite sample performance of all methods under an extensive numerical simulation. The results show that our new proposed method often outperforms the alternatives on the numerical settings described in this article. Finally, in Section 5, we describe an empirical application of the proposed method, using data from a major Canadian insurer, to determine which auto insurance policyholders are more likely to be positively stimulated to buy a home policy as a result of a marketing cross-sell intervention activity.

2. Problem formulation

We frame the *personalized treatment learning* problem in the context of Rubin's model of causality [35–38]. Under this model, we conceptualize the learning problem in terms of the potential outcomes under treatment alternatives, only one of which is observed for each subject. The causal effect of a treatment on a subject is defined in terms of the difference between an observed outcome and its counterfactual. The notation introduced below will be used throughout the paper.

In the following, we use uppercase letters to denote random variables and lowercase letters to denote values of the random variables. Assume that a sample of subjects is randomly assigned to two treatment arms, denoted by A, $A \in \{0, 1\}$, also referred as control and treatment states, respectively. Let $Y(a) \in \{0, 1\}$ denote a binary potential outcome of a subject if assigned to treatment $A = a$, $a = \{0, 1\}$. The observed outcome is $Y = AY(1) + (1 - A)Y(0)$. Throughout this paper, we assume a value of $Y = 1$ is more desirable than $Y = 0$. Each subject is characterized by a p − dimensional vector of baseline covariates $\mathbf{X} = (X_1, \ldots, X_p)^T$. We assume the data consists of L independent and identically distributed realizations of $(Y, A, \mathbf{X})$, $\{(Y_\ell, A_\ell, \mathbf{X}_\ell), \ell = 1, \ldots, L\}$.

Under the assumption of randomization, treatment assignment A ignores its possible impact on the outcomes $Y(0)$ and $Y(1)$, and hence they are independent—using the notation of Dawid [9], $\{Y_\ell(0), Y_\ell(1) \perp A_\ell\}$. In this context, the *average treatment effect* (ATE) can be estimated by

$$\begin{aligned} \tau &= E[Y_\ell(1) - Y_\ell(0)] \\ &= E[Y_\ell | A_\ell = 1] - E[Y_\ell | A_\ell = 0]. \end{aligned} \tag{1}$$

In observational studies, subjects assigned to different treatment conditions are not exchangeable, and thus direct comparisons can be misleading [34].

In many circumstances, subjects can show significant heterogeneity in response to treatments, in which case the ATE is of limited value. The problem addressed in this paper is the identification of subgroups of subjects for which the treatment is most beneficial (or most harmful) within the context of experimental data. As discussed by Holland and Rubin [19], the most granular level of causal inference is the *individual treatment effect* (ITE), defined by $Y_\ell(1) - Y_\ell(0)$ for each subject $\ell = \{1, \ldots, L\}$. However, this is an unobserved quantity, as a subject is never observed simultaneously in both treatment states. The best approximation to the ITE that is possible to obtain in practice is the *subpopulation treatment effect* (STE), which is defined for a subject with individual covariate profile $\mathbf{X}_\ell = \mathbf{x}$ by

$$\begin{aligned} \tau(\mathbf{x}) &= E[Y_\ell(1) - Y_\ell(0) | \mathbf{X}_\ell = \mathbf{x}] \\ &= E[Y_\ell | \mathbf{X}_\ell = \mathbf{x}, A_\ell = 1] - E[Y_\ell | \mathbf{X}_\ell = \mathbf{x}, A_\ell = 0]. \end{aligned} \tag{2}$$

Understanding the precise nature of the STE variability can be extremely valuable in personalizing the choice of treatment, so that it is most appropriate for each individual. Henceforth, in this paper, we use the term *personalized treatment effect* (PTE) to refer to the subpopulation treatment effect (2).

A *personalized treatment rule* $\mathcal{H}$ is a map from the space of baseline covariates $\mathbf{X}$ to the space of treatments A, $\mathcal{H}(\mathbf{X}) : \mathbb{R}^p \to \{0, 1\}$. An *optimal treatment rule* is one that maximizes the expected outcome, $E[Y(\mathcal{H}(\mathbf{X}))]$, if the personalized treatment rule is implemented for the whole population. Notice that since Y is binary, this expectation has a probabilistic interpretation. That is, $E[Y(\mathcal{H}(\mathbf{X}))] = P(Y(\mathcal{H}(\mathbf{X})) = 1)$ and thus $\tau(\mathbf{x}) \in [-1, 1]$.

A straightforward calculation gives the optimal personalized treatment rule $\mathcal{H}^* = \operatorname{argmax}_{\mathcal{H}} E[Y(\mathcal{H}(\mathbf{X}))]$ for a subject with covariates $\mathbf{X}_\ell = \mathbf{x}$ as $\mathcal{H}^* = 1$ if $\tau(\mathbf{x}) > 0$, and $\mathcal{H}^* = 0$ otherwise. In many situations, the alternative treatments have unequal costs, in which case the decision rule can simply be replaced by $\mathcal{H}^* = 1$ if $\tau(\mathbf{x}) > c$, and $\mathcal{H}^* = 0$ otherwise, for some constant threshold $c \in [-1, 1]$.

3. Causal conditional inference trees

The most relevant methods discussed in the literature to estimate personalized treatment effects include the so-called *indirect estimation methods*, which are based on a systematic 2-stage procedure to estimate the PTE. In the first stage, they attempt to achieve high accuracy in predicting the outcome Y conditional on the covariates $\mathbf{X}$ and treatment A. In the second stage, they subtract the predicted value of Y under each treatment to obtain a PTE estimate. Indirect estimation methods include the *difference score* method discussed by Larsen [27], the *interaction* approach proposed by Lo [29], and the L2 − SVM method proposed by Imai and Ratkovic [21].

Additionally, other methods have been proposed in the literature to estimate personalized treatment effects, such as the *modified covariate* method proposed by Tian et al. [48], the *modified outcome* method proposed by Jaśkowski and Jaroszewicz [22], and the *causal K-nearest-neighbor* (*CKNN*) discussed by Alemi et al. [2]. More recently, Guelman et al. [15] proposed a tree-based method called *uplift random forests* to estimate personalized treatment effects. Uplift random forests directly predict the expected change in the outcome as a result of the treatment, as opposed to predicting the outcome itself. Further details about uplift random forests can be found in the work of Guelman et al. [15].

We propose here an improved tree-based method to estimate personalized treatment effects. The key idea of this method is to recursively partition the covariate space into meaningful subgroups with heterogeneous treatment effects. The standard decision tree methodology [4,32] is inherited, but the individual trees are grown using a more appropriate split criterion to the problem at hand. These concepts were already implemented in uplift random forests, but there are two fundamental aspects in which this method could be improved: overfitting and the selection bias towards covariates with many possible splits. The development of the framework introduced here to tackle these issues was motivated by the *unbiased recursive partitioning* method proposed by Hothorn et al. [20].

With regards to overfitting, we point out that the individual trees in uplift random forests are grown to maximal depth. While this helps to reduce bias, there is the usual tradeoff with variance. Maximal-depth trees could be highly unstable and this may overemphasize learning patterns and noise in the data which may not recur in future samples. This problem, known as overfitting, can be exacerbated in the context of personalized treatment learning models. In these models, the variability in the response from the treatment heterogeneity effects tends to be small relative to the variability in the response from the main effects. If the fitted model is not able to distinguish well between the relative strength of these two effects and the levels of noise in the data are relatively high, this may easily translate into overfitting problems. In conventional decision trees, such as CART [4] and C4.5 [32], overfitting is solved by a pruning procedure. This consists in traversing the tree bottom up and testing for each (non-terminal) node, whether collapsing the subtree rooted at that node into a single leaf would improve the model's generalization performance. Tree-based methods proposed in the literature to estimate personalized treatment effects [40,45,33] use some sort of pruning. However, the pruning procedures used by these methods are all *ad hoc* and lack a theoretical foundation.

Besides the overfitting problem, the second concern is the bias of variable selection towards covariates with many possible splits or missing values. This problem is also present in conventional decision trees and results from the maximization of the split criterion over all possible splits simultaneously [4,25].

Following the framework proposed by Hothorn et al. [20], we have considerably improved the generalization performance of uplift random forests by solving both the overfitting and biased variable selection problems. The key to the solution is separating the variable selection and the splitting procedure, coupled with a statistically motivated and computationally efficient stopping criteria based on the theory of permutation tests developed by Strasser and Weber [43].

The pseudocode of the proposed *causal conditional inference forest* algorithm is shown in Algorithm 1. The most relevant aspects to discuss are steps 7–12. Specifically, for each terminal node in the tree, we test the global null hypothesis of no interaction effect between the treatment A and *any* of the n covariates selected at random from the set of p covariates. The global hypothesis of no interaction is formulated in terms of n partial hypotheses $H_0^j : E[W|X_j] = E[W], j = \{1, \ldots, n\}$, with the global null hypothesis $H_0 = \cap_{j=1}^{n} H_0^j$, where W is defined as in the modified outcome method, namely,

$$W_\ell = \begin{cases} 1 & \text{if } A_\ell = 1 \text{ and } Y_\ell = 1 \\ 1 & \text{if } A_\ell = 0 \text{ and } Y_\ell = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Thus, a conditional independence test of W and X_j has a causal interpretation for the treatment effect for subjects with baseline covariate X_j. Multiplicity in testing can be handled via Bonferroni-adjusted P values or alternative adjustment procedures [50,41,3]. When we are not able to reject H_0 at a prespecified significance level α, we stop the splitting process at that node. Otherwise, we select the j^*th covariate X_{j^*} with the smallest adjusted P value. The algorithm then induces a partition Ω^* of the covariate X_{j^*} into two disjoint sets $\mathcal{M} \subset X_{j^*}$ and $X_{j^*} \backslash \mathcal{M}$ based on the split criterion discussed below. This statistical approach prevents overfitting, without requiring any form of pruning or cross-validation.

One approach to measuring the independence between W and X_j would be to use a classical statistical test, such as a Pearson's chi-squared. However, the assumed distribution in these tests is only a valid approximation to the actual distribution in the large-sample case, and this does not likely hold near the leaves of the decision tree. Instead, we measure independence based on the theoretical framework of permutation tests, which is admissible for arbitrary sample sizes. Strasser and Weber [43] developed a comprehensive theory based on a general functional form of multivariate linear statistics appropriate for arbitrary independence problems. Specifically, to test the null hypothesis of independence between W and $X_j, j = \{1, \ldots, n\}$, we use linear statistics of the form

$$\mathcal{T}_j = \operatorname{vec}\left(\sum_{\ell=1}^{L} g\left(X_{j\ell}\right) h(W_\ell, (W_1, \ldots, W_L))^T\right) \in \mathbb{R}^{u_j v \times 1} \tag{3}$$

where $g : X_j \rightarrow \mathbb{R}^{u_j \times 1}$ is a transformation of the covariate X_j and $h : W \rightarrow \mathbb{R}^{v \times 1}$ is called the *influence function*. The "vec" operator transforms the $u_j \times v$ matrix into a $u_j v \times 1$ column vector. The distribution of $\mathcal{T}_j$ under the null hypothesis can be obtained by fixing $X_{j1}, \ldots, X_{jL}$ and conditioning on all possible permutations S of the responses $W_1, \ldots, W_L$. A univariate test statistic c is then obtained by standardizing $\mathcal{T}_j \in \mathbb{R}^{u_j v \times 1}$ based on its conditional expectations $\mu_j \in \mathbb{R}^{u_j v \times 1}$ and covariance $\Sigma_j \in \mathbb{R}^{u_j v \times u_j v}$, as derived by Strasser and Weber [43]. A common choice is the maximum of the absolute values of the standardized linear statistic

$$c_{\max}(\mathcal{T}, \mu, \Sigma) = \max\left|\frac{\mathcal{T} - \mu}{\operatorname{diag}(\Sigma)^{1/2}}\right|, \tag{4}$$

or a quadratic form

$$c_{\text{quad}}(\mathcal{T}, \mu, \Sigma) = (\mathcal{T} - \mu)\Sigma^{+}(\mathcal{T} - \mu)^T, \tag{5}$$

where Σ^{+} is the Moore–Penrose inverse of Σ. Many well-known classical tests (e.g., Pearson's chi-squared, Cochran–Mantel–Haenszel, Wilcoxon–Mann–Whitney) can be formulated from Eq. (3) by choosing the appropriate transformation g, influence function h, and test statistic c to map the linear statistic $\mathcal{T}$ into the real line. This sheds light on the extension of the proposed method to response variables measured in arbitrary scales and multi-category or continuous treatment settings.

In step 11 of Algorithm 1, we select the covariate X_{j^*} with smallest adjusted P value. The P value P_j is given by the number of permutations

$s \in S$ of the data with corresponding test statistic exceeding the observed test statistic $t \in \mathbb{R}^{u_j v \times 1}$. That is,

$$P_j = \mathbb{P}\left(c\left(\mathcal{T}_j, \mu_j, \Sigma_j\right) \geq c\left(t_j, \mu_j, \Sigma_j\right) \middle| S\right).$$

For moderate to large samples sizes, it might not be possible to obtain the exact distribution (calculated exhaustively) of the test statistic. However, we can approximate the exact distribution by computing the test statistic from a random sample of the set of all permutations S. In addition, Strasser and Weber [43] showed that the asymptotic distribution of the test statistic given by Eq. (4) tends to multivariate normal with parameters μ and Σ as $L \to \infty$. The test statistic Eq. (5) follows an asymptotic chi-squared distribution with degrees of freedom given by the rank of Σ. Therefore, asymptotic P values can be computed for these test statistics.

Once we select the covariate X_{j^*} to split, we next use a split criterion which explicitly attempts to find subgroups with heterogeneous treatment effects. Specifically, we use the following measure proposed by Su et al. [44], also implemented later by Radcliffe and Surry [33] for assessing the personalized treatment effect from a split Ω:

$$G^2(\Omega) = \frac{(L-4)\left\{\left(\overline{Y}_{n_L}(1)-\overline{Y}_{n_L}(0)\right)-\left(\overline{Y}_{n_R}(1)-\overline{Y}_{n_R}(0)\right)\right\}^2}{\hat{\sigma}^2\left\{1/L_{n_L}(1)+1/L_{n_L}(0)+1/L_{n_R}(1)+1/L_{n_R}(0)\right\}} \tag{6}$$

where n_L and n_R denotes the left and right child nodes, respectively, $L_{i \in \{n_L, n_R\}}(A)$ denotes the number of observations in child node i exposed to treatment $A \in \{0, 1\}$, and

$$\overline{Y}_{i \in \{n_L, n_R\}}(1) = \frac{\sum_{\forall \ell \in i} Y_\ell A_\ell}{\sum_{\forall \ell \in i} A_\ell}, \tag{7}$$

$$\overline{Y}_{i \in \{n_L, n_R\}}(0) = \frac{\sum_{\forall \ell \in i} Y_\ell (1-A_\ell)}{\sum_{\forall \ell \in i} (1-A_\ell)}, \tag{8}$$

$$\hat{\sigma}^2 = \sum_{A \in \{0,1\}} \sum_{i \in \{n_L, n_R\}} L_i(A)\overline{Y}_i(A)\left(1-\overline{Y}_i(A)\right). \tag{9}$$

The best split is given by $G^2(\Omega^*) = \max_\Omega G^2(\Omega)$; that is, the split that maximizes the criterion $G^2(\Omega)$ among all permissible splits. It can be seen [44] that the split criterion given in Eq. (6) is equivalent to a chi-squared test of the interaction effect between the treatment and the covariate X_{j^*} dichotomized at the value given by the split Ω.

Algorithm 1. Causal conditional inference forests

for $b = 1$ to B **do**
  Draw a sample with replacement from the training observations L such that $P(A = 1) = P(A = 0) = 1/2$
  Grow a conditional causal inference tree $CCIT_b$ to the sampled data:
  for each terminal node **do**
    repeat
      Select n covariates at random from the p covariates
      Test the global null hypothesis of no interaction effect between the treatment A and any of the n covariates (i.e., $H_0 = \cap_{j=1}^{n} H_0^j$, where $H_0^j : E[W|X_j] = E[W]$) at a level of significance α based on a permutation test
      if the null hypothesis H_0 cannot be rejected **then**
        Stop
      else
        Select the j^*th covariate X_{j^*} with the strongest interaction effect (i.e., the one with the smallest adjusted P value)
        Choose a partition Ω^* of the covariate X_{j^*} into two disjoint sets $\mathcal{M} \subset X_{j^*}$ and $X_{j^*} \setminus \mathcal{M}$ based on the $G^2(\Omega)$ split criterion
      end if
    until a minimum node size $l_{\min}$ is reached
  end for
end for
Output the ensemble of causal conditional inference trees $CCIT_b;\ b = \{1, \ldots, B\}$
The predicted personalized treatment effect for a new data point $\mathbf{x}$, is obtained by averaging the predictions of the individual trees in the ensemble: $\hat{\tau}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} CCIT_b(\mathbf{x})$

4. Simulation studies

In this section, we conduct a numerical study for the purpose of assessing the finite sample performance of the new method introduced in Section 3 and other alternative methods also mentioned in the same section. Most of these methods require specialized software for implementation. We have developed a software package in R named **uplift** [13] that implements a variety of algorithms for building and testing personalized treatment learning models. Currently, the following methods are implemented: uplift random forests (upliftRF), causal conditional inference forests (ccif), causal K-nearest-neighbor (cknn), modified covariate method (mcm), and modified outcome method (mom). **uplift** is available from the Comprehensive R Archive Network at: http://www.cran.r-project.org/package=uplift. We also used the package **FindIt**, which implements the L2-SVM method (l2svm) and was developed by the authors of the method [21]. Finally, the difference score (dsm) and interaction (int) methods can be implemented straightforwardly using readily available software.

Our simulation framework is based on the one described by Tian et al. [48], but with a few modifications. We evaluate the performance of the aforementioned methods in eight simulation settings, by varying i) the relative strength of the main effects relative to the treatment heterogeneity effects, ii) the degree of correlation among the covariates, and iii) the noise levels in the response.

We generated L independent binary samples from the regression model

$$Y = I\left(\left[\sum_{j=1}^{p}\eta_j X_j + \sum_{j=1}^{p}\delta_j X_j A_j^* + \epsilon\right] \geq 0\right), \tag{10}$$

where the covariates $(X_1, \ldots, X_p)$ follow a mean-zero multivariate normal distribution with covariance matrix $(1 - \rho)\mathbf{I}_p + \rho\mathbf{1}^T\mathbf{1}$, $A_\ell^* = 2A_\ell - 1 \in \{-1, 1\}$ was generated with equal probability at random, and $\epsilon \sim N(0, \sigma_0^2)$. We let $L = 200$, $p = 20$, and $(\delta_1, \delta_2, \delta_3, \delta_4, \delta_5, \ldots, \delta_p) = (1/2, -1/2, 1/2, -1/2, 0, \ldots, 0)$.

Table 1 shows the simulation scenarios. The first four scenarios model a situation in which the variability in the response from the main effects is twice as big as that from the treatment heterogeneity effects, whereas in the last four scenarios, the variability in the response from the main effects is four times as big as that from the treatment heterogeneity effects. Each of these scenarios were tested under zero and moderate correlation among the covariates ($\rho = 0$ and $\rho = 0.5$) and two levels of noise ($\sigma_0 = \sqrt{2}$ and $\sigma_0 = 2\sqrt{2}$).

The key benefit of simulations in the context of personalized treatment effects is that the "true" treatment effect is known for each subject, a value which is not observed in empirical data. The performance of the analytical methods was measured using the *Spearman's rank correlation*

Table 1
Simulation scenarios.

Scenario	η_j	ρ	σ_0
1	$(-1)^{(j+1)}I(3 \leq j \leq 10)/2$	0	$\sqrt{2}$
2	$(-1)^{(j+1)}I(3 \leq j \leq 10)/2$	0	$2\sqrt{2}$
3	$(-1)^{(j+1)}I(3 \leq j \leq 10)/2$	0.5	$\sqrt{2}$
4	$(-1)^{(j+1)}I(3 \leq j \leq 10)/2$	0.5	$2\sqrt{2}$
5	$(-1)^{(j+1)}I(3 \leq j \leq 10)$	0	$\sqrt{2}$
6	$(-1)^{(j+1)}I(3 \leq j \leq 10)$	0	$2\sqrt{2}$
7	$(-1)^{(j+1)}I(3 \leq j \leq 10)$	0.5	$\sqrt{2}$
8	$(-1)^{(j+1)}I(3 \leq j \leq 10)$	0.5	$2\sqrt{2}$

Note. This table displays the numerical settings considered in the simulations. Each scenario is parameterized by the strength of the main effects, η_j, the correlation among the covariates, ρ, and the magnitude of the noise, σ_0.

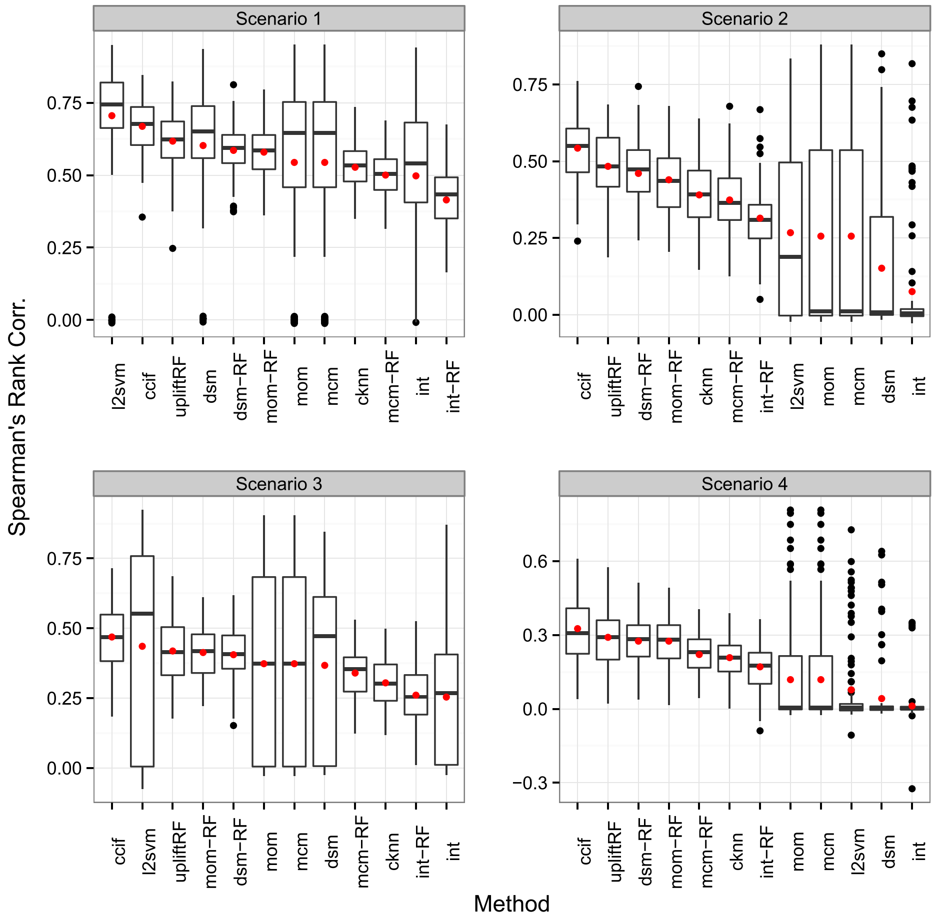

Fig. 1. Box plots of the Spearman's rank correlation coefficient between the estimated treatment effect $\hat{\tau}(X)$ and the "true" treatment effect $\tau(X)$ for all methods. The plots illustrate the results for simulation scenarios 1–4, which model a situation with "stronger" treatment heterogeneity effects, under none and moderate correlation among the covariates ($\rho = 0$ and $\rho = 0.5$) and two levels of noise ($\sigma_0 = \sqrt{2}$ and $\sigma_0 = 2\sqrt{2}$). The box plots within each simulation scenario are shown in decreasing order of performance based on the average correlation. The dots outside the box plots represent outliers. We used the "1.5 rule" for determining if a data point is an outlier: less than $Q1 - 1.5 \times (Q3 - Q1)$ or greater than $Q3 + 1.5 \times (Q3 - Q1)$, where $Q1$ and $Q3$ represent the first and third quartiles, respectively.

coefficient between the estimated treatment effect $\hat{\tau}(X)$ derived from each model, and the "true" treatment effect

$$\begin{aligned}\tau(\mathbf{X}) &= E[Y(1) - Y(0)|\mathbf{X}] \\ &= P\left(\sum_{j=1}^{p}\left(\eta_j + \delta_j\right)X_j \le \epsilon\right) - P\left(\sum_{j=1}^{p}\left(\eta_j - \delta_j\right)X_j \le \epsilon\right) \\ &= F\left(\sum_{j=1}^{p}\left(\eta_j + \delta_j\right)X_j\right) - F\left(\sum_{j=1}^{p}\left(\eta_j - \delta_j\right)X_j\right),\end{aligned} \tag{11}$$

in an independently generated test set with a sample size of 10,000. In Eq. (11), F denotes the cumulative distribution function of a normal random variable with mean zero and variance σ_0^2.

Variable selection for the mcm, mom, dsm, and int methods was performed using the LASSO logistic regression via a 10-fold cross-validation procedure. Based on this selection method, we found cases where the LASSO procedure could not select any non-zero covariate based on cross-validation. Similar to Tian et al. [48], in these cases, we simply forced the correlation coefficient to be zero in the test set since the method did not find anything informative. For this reason, we alternatively fitted these methods based on random forests [6] using their default settings.[2] We refer to these methods based on random forest fits as mcm-RF, mom-RF, dsm-RF, and int-RF. The optimal values for the LASSO penalties for the l2svm method, and the number K of subjects within the neighborhood of the target subject for the cknn method, were also selected via 10-fold cross-validation. Lastly, the methods upliftRF and ccif were fitted using their default settings.[3]

The results over 100 repetitions of the simulation for the first and last four simulation scenarios are shown in Figs. 1 and 2, respectively. These figures illustrate the box plots of the Spearman's rank correlation coefficient between $\hat{\tau}(X)$ and $\tau(X)$. The box plots within each simulation scenario are shown in decreasing order of performance based on the average correlation. The ccif method performed either the best or next to the best in all eight scenarios.

5. An insurance cross-sell application

In this section, we apply the new proposed causal conditional inference forest method to an insurance marketing application. The data used for this analysis are based on a direct mail campaign implemented by a large Canadian insurer between June 2012 and May 2013. The objective of the campaign was to drive more business from the existing portfolio of auto insurance clients by cross-selling them a home insurance policy with the company. The standard savings via multiproduct discount was prominently featured and positioned as the key element

[2] Specifically, we fitted the models using $B = 500$ trees and $n = \sqrt{p}$ as the number of variables randomly sampled as candidates at each split.

[3] In both cases, we used $B = 500$ trees and $n = p/3$ as the number of variables randomly sampled as candidates at each split. For ccif, we set the P value $= 0.05$.

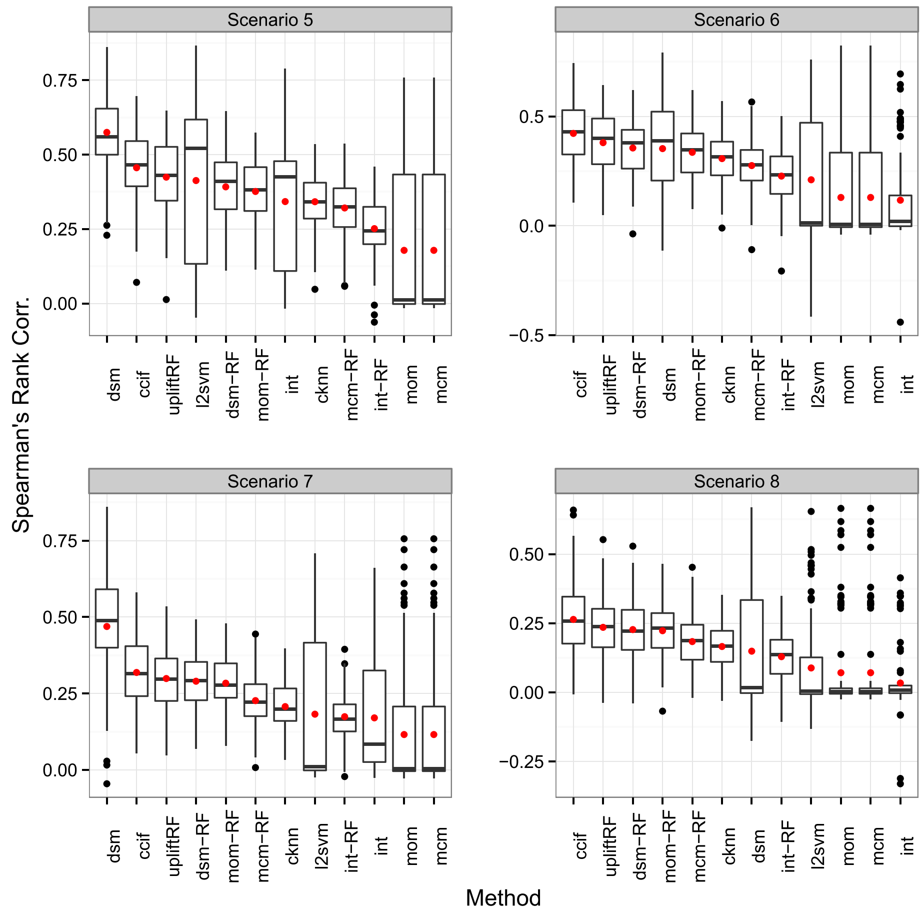

Fig. 2. Box plots of the Spearman's rank correlation coefficient between the estimated treatment effect $\hat{\tau}(X)$ and the "true" treatment effect $\tau(X)$ for all methods. The plots illustrate the results for simulation scenarios 5–8, which model a situation with "weaker" treatment heterogeneity effects, under none and moderate correlation among the covariates ($\rho = 0$ and $\rho = 0.5$) and two levels of noise ($\sigma_0 = \sqrt{2}$ and $\sigma_0 = 2\sqrt{2}$). The box plots within each simulation scenario are shown in decreasing order of performance based on the average correlation.

in the offer to the clients. In addition to the direct mail, the same clients were also contacted over the phone to further motivate them to initiate a home policy quote. A randomly selected control group was included as part of the campaign design, consisting of clients who were not mailed or called. The response variable is determined by whether the client purchased the home policy between the mail date and 3 months thereafter. In addition to the response, the data set contains approximately 50 covariates related to the auto policy, including driver and vehicle characteristics and general policy information.

Table 2 shows the cross-sell rates by group. The average treatment effect (ATE) of 0.34% (2.55% − 2.21%) is not statistically significant, with a *P* value of 0.23 based on a chi-squared test. However, as discussed above, the average treatment effect would be of limited value if policyholders show significant heterogeneity in response to the marketing intervention activity. Our objective is to estimate the personalized treatment effect and use it to construct an optimal treatment rule for the auto insurance portfolio, namely, the policyholder-treatment assignment that maximizes the expected profits from the campaign.

Table 2
Cross-sell rates by group.

	Treatment	Control
Purchased home policy = N	30,184	3,322
Purchased home policy = Y	789	75
Cross-sell rate	2.55%	2.21%

Note. This table displays the cross-sell rate for the treatment and control groups. The average treatment effect (ATE) is 0.34% (2.55% − 2.21%), which is not statistically significant (*P* value = 0.23).

To objectively examine the performance of the proposed method, we randomly split the data into training and validation sets in a 70/30 ratio. A preliminary analysis showed that model performance is not highly sensitive to the values of its tuning parameters (i.e., number of trees B and number of variables n randomly sampled as candidates at each split), as long as they are specified within a reasonable range. Thus, we fitted a causal conditional inference forest (ccif) to the training data using its default parameter values. Specifically, in Algorithm 1, we used $B = 500$, $n = 16$, and a *P* value of 0.05 as the level of significance α. We next ranked policyholders in the validation data set based on their estimated personalized treatment effect (from high to low), and grouped them into deciles. We then computed the actual average treatment effect within each decile (defined as the difference in cross-sell rates between the treatment and control groups).

Fig. 3 shows the box plots of the actual average treatment effect for each decile based on 100 random training/validation data partitions. The results show that clients with higher estimated personalized treatment effect were, on average, positively influenced to buy as a result of the marketing intervention activity, with ATEs ranging from 1% to 2.5% for the first three deciles as compared with the ATE of 0.34% over all deciles. Also, notice there is a subgroup of clients (deciles 8–10) whose purchase behavior was negatively impacted by the campaign. Negative reactions to sale attempts have been recognized in the literature [17,24,7] and may happen for a variety of reasons. For instance, the marketing activity may trigger a decision to shop for better multi-product rates among other insurers. Moreover, if the client currently owns a home policy with another insurer, she may decide to switch her auto policy to that insurer instead. We found evidence of higher

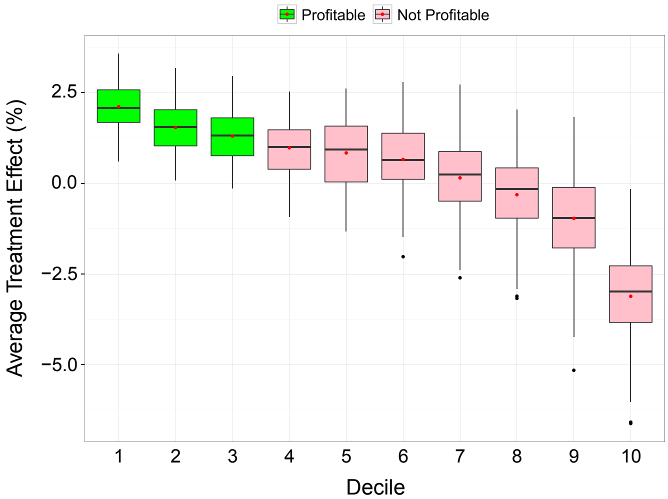

Fig. 3. Box plots of the actual average treatment effect (ATE) for each decile based on 100 random training/validation data splits. The first (tenth) decile represents the 10% of clients with highest (lowest) predicted personalized treatment effect. Clients with higher estimated personalized treatment effect were, on average, positively influenced to buy as a result of the marketing intervention activity.

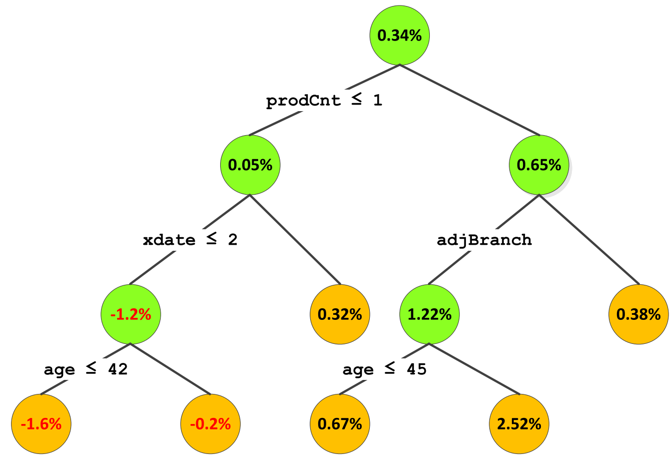

Fig. 4. Prototype causal conditional inference tree from applying Algorithm 1 to the insurance cross-sell data set. Internal nodes are denoted by green circles, and terminal nodes by orange circles. The splitting rule is given under each internal node. Observations satisfying the rule go to the left child node and observations not satisfying it go to the right child node. Within each node, we display the incremental cross-sell rates (i.e., the difference in cross-sell rates between treatment and control groups). The tree provides marketers with further insights in terms of the characteristics of clients with positive/negative personalized treatment effects.

auto policy cancellation rates in the higher deciles. In addition, some clients may perceive the call as intrusive and likely be annoyed by it, generating a negative reaction.

In the context of insurance, it is important to consider not only the personalized treatment effect from the cross-sell activity but also the risk profile of the targeted clients [47,23,11]. To determine the expected profitability from targeting each decile, we first calculated the product between the ATE and the expected lifetime-value of a home policy,[4] and then we subtracted the fixed and variable campaign expenses. Based on these considerations, Fig. 3 shows that only clients in deciles 1–3 have positive expected profits from the marketing activity and should be targeted. The incremental profits from clients in deciles 4–7 is outweighed by the incremental costs, and so the company should avoid targeting these clients. Clients in deciles 8–10 have negative reactions to the campaign and clearly should not be targeted either.

As discussed in Section 3, one of the key challenges in building personalized treatment learning models is that the magnitude of the variability in the response due to the treatment heterogeneity effects is usually much smaller than the variability in the response due to the main effects. For instance, the 2.5% average treatment effect in the top decile (Fig. 3) is the result of a difference in cross-sell response rates of 13% and 10.5% between treatment and control groups, respectively. For most companies with a sizable portfolio of clients,[5] relatively small incremental response rates (as the ones evidenced in this application, which range from 1% to 2.5%) translate into significant profits.

Conventional marketing models developed within a decision support framework are good in predicting which clients have higher propensity to buy a product/service (the so-called *propensity to buy models*), but not in predicting which clients are more likely to buy *as a result* of the marketing intervention activity (given by the difference between the expected responses under the alternative treatments). Propensity to buy models do not attempt to directly maximize the expected profitability of the intervention, as some clients will buy even if they are not targeted, while others might be negatively impacted by the intervention. Our proposed method for targeting clients offers a decision support framework to determine the optimal policyholder-treatment assignment that maximizes the expected profitability from the campaign. Only profitable clients who are positively influenced to buy the additional insurance product as a result of the marketing cross-sell intervention activity are targeted.

To allow marketers gain further insights in terms of the characteristics of clients with positive/negative personalized treatment effects, we illustrate in Fig. 4 a prototype causal conditional inference tree drawn from Algorithm 1 when applied to the insurance cross-sell data set. Internal nodes are denoted by green circles, and terminal nodes by orange circles. The splitting rule is given under each internal node. Observations satisfying the rule go to the left child node and observations not satisfying it go to the right child node. Within each node, we display the incremental cross-sell rates (i.e., the difference in cross-sell rates between treatment and control groups). Clients with highest positive impact from the cross-sell activity are those currently holding more than one product with the company (prodCnt), live near one of the company's branch locations (adjBranch), and older than 45 years (age). This is not surprising as clients already holding more than one product are more engaged with the products offered by the company and are likely to buy more. In addition, clients living near a branch are more likely to personally walk into it and obtain a quote for the additional product. Clients with negative impact are those who only hold a single product, namely, the auto policy, have this policy expiring within the next two months (xdate), and who are relatively younger. For clients with their auto policy expiring shortly, the campaign may be acting as a trigger to shop for better insurance rates in the market. As previously discussed, if the client already owns a home policy with another insurer, she may decide to switch the auto policy to that insurer instead, provided that the competitor offers lower multiproduct rates.

6. Conclusions

The estimation of personalized treatment effects is becoming increasingly important in many scientific disciplines and decision support systems. As subjects can show significant heterogeneity in response to treatments, making an optimal treatment choice at the individual subject level is essential. An optimal personalized treatment is the one that maximizes the probability of a desirable outcome. We call

[4] The expected lifetime-value (LTV) of a home policy in decile $i = \{1, \ldots, 10\}$ is given by $\text{LTV}_i = \left[\text{Prem}_i - \hat{L}C_i - \text{Exp}_i\right] \sum_{t=1}^{5} P(S_{it}) r^t$, where Prem is the average policy premium, $\hat{L}C$ is the predicted insurance loss per policy-year, Exp captures the fixed and variable expenses for servicing the policy (excluding campaign expenses), $P(S_{it})$ is the probability that a policyholder in decile $i = \{1, \ldots, 10\}$ will continue with the home product beyond year $t = \{1, \ldots, 5\}$, and r^t is the interest discount factor.

[5] In the order of hundred thousands or more clients.

the task of learning the optimal personalized treatment *personalized treatment learning*.

From the statistical learning perspective, estimating personalized treatment effects imposes some key challenges, primarily because the optimal treatment is unknown on a given training set. In this paper, we proposed a new approach called *causal conditional inference trees* for personalized treatment learning and compared its performance to seven alternative methods proposed in the literature to tackle this problem. Our method recursively partitions the input space into subgroups with heterogeneous treatment effects. Motivated by the *unbiased recursive partitioning* method proposed by Hothorn et al. [20], the key ingredient of our tree-based method is the separation between the variable selection and the splitting procedure, coupled with a statistically motivated and computationally efficient stopping criteria based on the theory of permutation tests developed by Strasser and Weber [43]. This statistical approach prevents overfitting, without requiring any form of pruning or cross-validation. It also avoids selection bias towards covariates with many possible splits. Performance results measured on synthetic data show that our proposed method often outperforms the alternatives on the numerical settings described in this article.

We have also discussed an application of the proposed method in the context of insurance marketing for the purpose of selecting the best targets for cross-selling an insurance product. Our method was able to identify the policyholders who were positively/negatively motivated to buy as a result of the marketing intervention activity. Based on marketing costs and expected client lifetime-value considerations, we next derived the policyholder-treatment assignment that maximizes the expected profitability from the campaign.

We would also like to acknowledge the limitations of this work. First, we have only considered the case of binary treatments. It would be worthwhile to examine the extent to which the methods discussed in this article can be extended to multi-category or continuous treatment settings. Second, in many situations, the interest may be to estimate the personalized treatment effect when the intervention is not applied on a randomized basis, but we think there are major background variables that influence which treatment is received. Thus, it would be relevant to consider personalized treatment learning models in the context of observational data. Finally, we have only consider the case of personalized treatments in a single-decision setup. In dynamic treatment regimes, the treatment type is repeatedly adjusted according to an ongoing individual response [30]. In this context, the goal is to optimize a set of time-varying personalized treatments for the purpose of maximizing the probability of a long-term desirable outcome.

Acknowledgments

LG thanks Royal Bank of Canada, RBC Insurance. MG and AMP-M thanks ICREA Academia and the Ministry of Science/FEDER grant ECO2013-48326-C02-1. The authors are grateful to the editor and the reviewers for their thoughtful comments that helped us improve a prior version of this article. All errors remain our own responsibility.

References

[1] Y. Abu-Mostafa, M. Magdon-Ismail, L. Hsuan-Tien, Learning From Data, AMLBook. 2012.
[2] F. Alemi, H. Erdman, I. Griva, C. Evans, Improved statistical methods are needed to advance personalized medicine, Open Translational Medicine Journal 1 (2009) 16–20.
[3] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, Journal of the Royal Statistical Society Series B 57 (1) (1995) 289–300.
[4] L. Brieman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Chapman & Hall, New York, 1984.
[5] L. Brieman, Statistical modeling: the two cultures, Statistical Science 16 (3) (2001) 199–231.
[6] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32.
[7] R. Byers, K. So, Note—a mathematical model for evaluating cross-sales policies in telephone service centers, Manufacturing and Service Operations Management 9 (1) (2007) 1–8.
[8] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Transactions on Information Theory 13 (1) (1967) 21–27.
[9] A. Dawid, Conditional independence in statistical theory, Journal of the Royal Statistical Society Series B 41 (1) (1979) 1–31.
[10] R. Dehejia, S. Wahba, Causal effects in non experimental studies: reevaluating the evaluation of training programs, Journal of the American Statistical Association 94 (1999) 1053–1062.
[11] M. Englund, J. Gustafsson, J. Nielsen, F. Thuring, Multidimensional credibility with time effects: an application to commercial business lines, Journal of Risk and Insurance 76 (2) (2009) 443–453.
[12] W. Frawley, G. Piatetsky-Shapiro, C. Matheus, Knowledge discovery in databases—an overview, Knowledge Discovery in Databases 1–30 (1991).
[13] L. Guelman, uplift: Uplift Modeling, R package version 0.3.52014.
[14] L. Guelman, M. Guillén, A.M. Pérez-Marin, Random forests for uplift modeling: an insurance customer retention case, Lecture Notes in Business Information Processing 115 (2012) 123–133.
[15] L. Guelman, M. Guillén, A.M. Pérez-Marin, Uplift random forests, Cybernetics and Systems (2015) (forthcoming).
[16] L. Guelman, M. Guillén, A causal inference approach to measure price elasticity in automobile insurance, Expert Systems with Applications 41 (2014) 387–396.
[17] E. Günes, O. Aksin-Karaesmen, L. Örmeci, H. Özden, Modeling customer reactions to sales attempts: if cross-selling backfires, Journal of Service Research 13 (2) (2010) 168–183.
[18] P. Holland, Statistics and causal inference, Journal of the American Statistical Association 81 (396) (1986) 945–960.
[19] P. Holland, D. Rubin, Causal inference in retrospective studies, Evaluation Review 12 (1988) 203–231.
[20] T. Hothorn, K. Hornik, A. Zeileis, Unbiased recursive partitioning: a conditional inference framework, Journal of Computational and Graphical Statistics 15 (3) (2006) 651–674.
[21] K. Imai, M. Ratkovic, Estimating treatment effect heterogeneity in randomized program evaluation, Forthcoming in Annals of Applied Statistics, 7, The Institute of Mathematical Statistics, 2013, pp. 443–470. http://dx.doi.org/10.1214/12-AOAS593.
[22] M. Jaśkowski, S. Jaroszewicz, Uplift modeling for clinical trial data, ICML 2012 Workshop on Clinical Data Analysis, Edinburgh, Scotland, UK2012.
[23] V. Kaishev, J. Nielsen, F. Thuring, Optimal customer selection for cross-selling of financial services products, Expert Systems with Applications 40 (5) (2013) 1748–1757.
[24] W. Kamakura, Cross-selling: offering the right product to the right customer at the right time, Journal of Relationship Marketing 6 (3–4) (2008) 41–58.
[25] G. Kass, An exploratory technique for investigating large quantities of categorical data, Applied Statistics 29 (2) (1980) 119–127.
[26] R. LaLonde, Evaluating the econometric evaluations of training programs with experimental data, The American Economic Review 76 (4) (1986) 606–620.
[27] K. Larsen, Net lift models, M2009 - 12th Annual SAS Data Mining Conference2009.
[28] H. Liang, Y. Xue, B. Berger, Web-based intervention support system for health promotion, Decision Support Systems 42 (1) (2006) 435–449.
[29] V. Lo, The true lift model, ACM SIGKDD Explorations Newsletter 4 (2) (2002) 78–86.
[30] S. Murphy, An experimental design for the development of adaptive treatment strategies, Statistic Medical 24 (2005) 1455–1481.
[31] M. Qian, S. Murphy, Performance guarantees for individualized treatment rules, The Annals of Statistics 39 (2) (2011) 1180–1210.
[32] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993.
[33] N. Radcliffe, P. Surry, Real-World Uplift Modelling with Significance-Based Uplift Trees, Portrait Technical Report TR-2011–1 (2011).
[34] P. Rosenbaum, D. Rubin, The central role of the propensity score in observational studies for causal effects, Biometrika 70 (1) (1983) 41–55.
[35] D. Rubin, Estimating causal effects of treatments in randomized and nonrandomized studies, Journal of Education and Psychology 66 (5) (1974) 688–701.
[36] D. Rubin, Assignment to treatment group on the basis of a covariate, Journal of Educational Statistics 2 (1977) 1–26.
[37] D. Rubin, Bayesian inference for causal effects: the role of randomization, The Annals of Statistics 6 (1978) 34–58.
[38] D. Rubin, Causal inference using potential outcomes, Journal of the American Statistical Association 100 (469) (2005) 322–330.
[39] D. Rubin, R. Waterman, Estimating the causal effects of marketing interventions using propensity score methodology, Statistical Science 21 (2006) 206–222.
[40] P. Rzepakowski, S. Jaroszewicz, Decision trees for uplift modeling with single and multiple treatments, Knowledge and Information Systems 32 (2) (2012) 303–327.
[41] J. Shaffer, Multiple hypothesis testing, Annual Review of Psychology 46 (1995) 561–584.
[42] A. Sinha, H. Zhao, Incorporating domain knowledge into data mining classifiers: an application in indirect lending, Decision Support Systems 46 (1) (2008) 287–299.
[43] H. Strasser, C. Weber, On the asymptotic theory of permutation statistics, Mathematical Methods of Statistics 8 (1999) 220–250.
[44] X. Su, C. Tsai, H. Wang, D. Nickerson, B. Li, Subgroup analysis via recursive partitioning, Journal of Machine Learning Research 10 (2) (2009) 141–158.
[45] X. Su, J. Kang, J. Fan, R. Levine, X. Yan, Facilitating score and causal inference trees for large observational studies, Journal of Machine Learning Research 13 (10) (2012) 2955–2994.

[46] H. Tang, S. Liao, S. Sun, A prediction framework based on contextual data to support mobile personalized marketing, Decision Support Systems 56 (2013) 234–246. . http://dx.doi.org/10.1016/j.dss.2013.06.004.
[47] F. Thuring, J. Nielsen, M. Guillén, C. Bolancé, Selecting prospects for cross-selling financial products using multivariate credibility, Expert Systems with Applications 39 (10) (2012) 8809–8816.
[48] L. Tian, A. Alizadeh, A. Gentles, R. Tibshirani, A simple method for detecting interactions between a treatment and a large number of covariates, arXiv:1212.2995v1 [stat.ME]. 2012. (Submitted on Dec 2012).
[49] V. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.
[50] P. Wright, Adjusted *p*-values for simultaneous inference, Biometrics 48 (1992) 1005–1013.
[51] D. Xu, S. Liao, Q. Li, Combining empirical experimentation and modeling techniques: a design research approach for personalized mobile advertising applications, Decision Support Systems 44 (3) (2008) 710–724.
[52] Y. Zhao, D. Zeng, J. Rush, M. Kosorok, Estimating individualized treatment rules using outcome weighted learning, Journal of the American Statistical Association 107 (499) (2012) 1106–1118.
[53] I. Žliobaitė, M. Pechenizkiy, Learning with actionable attributes: attention—boundary cases! ICDMW'10 Proceedings of the 2010 IEEE International Conference on Data Mining Workshops: 1021–10282010.

Leo Guelman holds a master's degree in Economics from the University of British Columbia. He is currently a PhD candidate in Economics at the University of Barcelona and Chief Statistician at Royal Bank of Canada (RBC Insurance Division). Leo has several publications in international peer-reviewed Journals.

Montserrat Guillén is full professor of Econometrics at the University of Barcelona and director of the research group Riskcenter. She holds an MSc in Mathematics and a PhD in Economics. She was visiting researcher at the University of Texas at Austin and the University of Paris Panthéon-Assas. In 2011, she was elected president of the European Group of Risk and Insurance Economists.

Ana Maria Pérez-Marín is associate professor of actuarial statistics at the University of Barcelona and member of the research group Riskcenter. She holds an MSc in Actuarial Science and a PhD in Business Studies. Her research interests are related to actuarial statistics, survival analysis and marketing in insurance.