Contents lists available at ScienceDirect

Information Sciences

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

Targeting customers for profit: An ensemble learning framework to support marketing decision-making

Stefan Lessmann[a,*], Johannes Haupt[a], Kristof Coussement[b], Koen W. De Bock[c]

[a] *School of Business and Economics, Humboldt-University of Berlin, Unter den Linden 6, D-10099 Berlin, Germany*
[b] *Department of Marketing, IÉSEG Center for Marketing Analytics (ICMA), IÉSEG School of Management – Université Catholique de Lille (LEM, UMR CNRS 9221), 3 Rue de la Digue, F-59000 Lille, France*
[c] *Audencia Business School, 8 Route de la Jonelière, F-44312 Nantes, France*

ARTICLE INFO

Article history:
Received 30 March 2018
Revised 25 March 2019
Accepted 10 May 2019
Available online xxx

Keywords:
Profit analytics
Marketing decision support
Machine learning
Business value

ABSTRACT

Marketing messages are most effective if they reach the right customers. Deciding which customers to contact is an important task in campaign planning. The paper focuses on empirical targeting models. We argue that common practices to develop such models do not account sufficiently for business goals. To remedy this, we propose profit-conscious ensemble selection, a modeling framework that integrates statistical learning principles and business objectives in the form of campaign profit maximization. Studying the interplay between data-driven learning methods and their business value in real-world application contexts, the paper contributes to the emerging field of profit analytics and provides original insights how to implement profit analytics in marketing. The paper also estimates the degree to which profit-concious modeling adds to the bottom line. The results of a comprehensive empirical study confirm the business value of the proposed ensemble learning framework in that it recommends substantially more profitable target groups than several benchmarks.

1. Introduction

Business analytics revolutionizes the face of decision support. Skepticism toward formal decision aids used to be widespread among executives. Today, we witness an unprecedented interest in quantitative decision aids and analytic models. Vast amounts of data, powerful pattern extraction algorithms, and easy to use software systems fuel this development and promise to improve management support. The paper concentrates on decision support in marketing campaign planning. Campaign planners need to answer three questions [9]: when to make an offer (timing), how often to make an offer (frequency), and whom to contact (target group selection). We focus on the target group selection problem, which has been studied in the direct marketing and churn management literature [e.g., [43]. To target marketing offers, companies use response models, which estimate acceptance probabilities for individual customers. Corresponding predictions facilitate targeting the most likely responders.

Modeling response behavior on the level of an individual customer is a popular use case of business analytics in marketing. Developments in the scope of big data have a sizeable impact on customer response modeling, which we discuss along the well-known four V's volume, variety, velocity, and value that characterize big data. First, the volume dimension

[*] Corresponding author.
E-mail address: stefan.lessmann@hu-berlin.de (S. Lessmann).

https://doi.org/10.1016/j.ins.2019.05.027
0020-0255/© 2019 Elsevier Inc. All rights reserved.

implies that companies have more detailed records of past customer behavior and information related to customer preferences [27]. Such behavioral information enters response models in the form of novel attributes from which acceptance probability predictions are eventually derived. Second, the variety dimension refers to new and often unstructured sources of data, which companies can unlock for gaining business insight. The use of text analytics to extract information from product reviews, postings in social media, etc. illustrates this approach and contributes attitudinal information, which further expands to scope of customer characteristics that enter response models. Third, the velocity dimension postulates that novel data arrives with higher speed and implies a necessity to reduce the latency of decision-making. For example, response model-based targeting decisions in digital advertisement must be made in real-time and the number of application settings that also require real-time decision-making tends to increase in the big data era. Finally, there is much evidence of big data creating considerable value for marketing, which emerges from enhanced decision-making [35].

Response models use a variety of prediction methods including, artificial neural networks, support vector machines, or tree-based approaches. However, prediction methods are designed for generality and support decision-making in many fields such as credit scoring [25] and fraud detection [40]. Developing a prediction model involves minimizing a statistical loss function on a labeled training sample [e.g., [17]. We argue that using an off-the-shelf method for customer targeting suffers a limitation. Contextual information related to the actual decision task does not enter model development. Budget constraints, customer lifetime value, parallel campaigns – relevant information in campaign planning – have no effect on the estimation of the targeting model. Therefore, the objective of the paper is to develop and test a contextualized modeling framework that accounts for business objectives during model development.

Current trends in marketing support this objective. Big data facilitates an increasing degree of personalization in marketing communication [e.g., [15]. Likewise, an increasing amount of information is distributed through digital channels [e.g., [8]. These developments amplify the scale of targeting decisions and require decision-making in real-time. Therefore, marketers need to automate targeting decisions. A high recognition of business goals during model development seems especially important when targeting models operate in a self-governed manner. More generally, our focus on the business value of empirical decision support models echoes the recent call for a higher recognition of managerial objectives in modeling, which gave rise to the emerging field of profit analytics [e.g., [24].

The contribution of the paper to the literature is threefold. First, we propose a new modeling methodology for profit-conscious ensemble selection (PCES). We design PCES in such a way that it integrates established principles of statistical inference with marketing objectives in customer targeting. A related design goal is to mimic the way in managers contextualize recommendation from model-based decision aids [10]. PCES-based targeting models are contextualized in the sense that they account for marketing objectives and constraints at earlier stages of the model development process than existing approaches. We hypothesize that a contextualization of the model development process improves the quality of targeting decisions.

The second contribution stems from a comprehensive empirical analysis, which includes twenty-five real-world marketing data sets from different industries, of the effectiveness of alternative paradigms toward customer targeting. Beyond comparing an arsenal of alternative targeting models, we contrast three fundamentally different modeling philosophies. The first approach, which we refer to as *profit-agnostic,* relies on statistical learning and develops targeting models through minimizing a statistical loss-function [17]. We consider this approach to represent standard practice in predictive analytics. The second approach derives targeting models from maximizing business performance while disregarding statistical learning principles. We consider this approach an extreme form of profit analytics and call corresponding models *profit-centered.* The third approach represents a hybrid solution in the form of PCES, which balances between statistical and economic considerations. This three-facetted empirical design provides novel insight concerning the relative merits of fundamentally different approaches toward predictive modeling.

The empirical design also facilitates the third and last contribution of the paper. In particular, the paper provides an estimate of the degree to which incorporating business goals into prediction model development raises the business performance (e.g., return on marketing) of model-based (targeting) decisions. We achieve this through estimating the campaign profit that emerges from model-based targeting and the marginal profit of PCES-based targeting, respectively. Corresponding results provides a clear and managerially meaningful measure of the business value of a targeting model and the extent to which PCES improves decision quality.

2. Background and related work

Related work splits into three streams. First, prior work on decision support systems (DSS) provide theoretical foundations (*Stream 1*). Second, related studies in forecasting and machine learning consider the interplay between predictive models and their value implications in economic contexts but differ in the methodology they employ and applications they consider (*Stream 2*). We sketch the connections and differences to these streams in the following. Subsequently, we discuss previous research on marketing decision support and customer targeting (*Stream 3*), which is particularly related to this study.

Papers from *Stream 1* examine the antecedents of (model-based) DSS effectiveness and highlight the importance of a DSS exhibiting high fit for the decision task. However, managers can mitigate a lack of fit if given an opportunity to post-process DSS recommendations [10]. Specifically, managers' decision-making is guided by a mental model that enables them to appraise DSS outputs in awareness of a specific problem context, connect DSS outputs to decision quality, and, in this

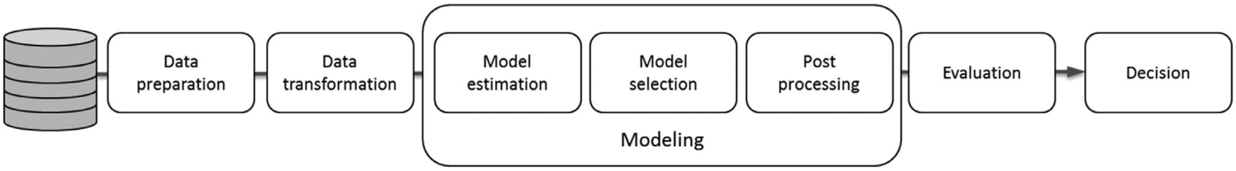


Fig. 1. Simplified process of prediction model development without feedback loops between stages.

way, correct for misleading information from an inadequate model [10]. This theory shows the merit of human supervision in model-based decision support and provides a design goal for the PCES approach developed here. We strive to combine the efficiency of automated, model-based decision-making with the ability of managers to improve decision quality through using contextual, task-specific information.

Prior work in *Stream 2* examines whether and when the development of data-driven prediction methods should account for economic objectives. Granger [16] was the first to criticize the use of quadratic loss functions in forecasting and to propose loss functions that penalize positive and negative residuals differently. Subsequent studies contribute further theoretical insights and empirical evidence concerning asymmetric loss functions in forecasting [e.g., [4]. The cost-sensitive learning literature also studies asymmetric cost of error functions but focuses on classification models [e.g., [42].

Research from *Stream 2* inspires the proposed PCES approach. PCES also employs non-standard, asymmetric loss functions for the development and assessment of predictive models. The main differences lie in the methodology and application. We focus on multivariate machine learning models as opposed to univariate time series models in forecasting. Our focus on decision problems in marketing campaign planning also implies that we study a different business objective (i.e., campaign profit). Specifically, the different errors in campaign planning are soliciting customers who do not respond and failing to contact customers who would respond (e.g., purchase an item) otherwise. This perspective on model errors is similar to cost-sensitive learning. Cost-sensitive learning, however, aims at generality. While generality is a goal worth pursuing, a DSS approach that focuses on a specific application context better reflects the unique characteristics and requirements of this context [23]. PCES is such an approach for decisions in the scope of targeted marketing where campaigns typically solicit only a small fraction of responsive customers. This implies a different notion of model performance compared to cost-sensitive learners, the objective of which is to minimize overall error costs [e.g., [42].

Finally, there is a large body of literature on predictive models for customer targeting (*Stream 3*). Previous work has studied all steps of the predictive modeling process, which we depict in Fig. 1. In interpreting Fig. 1, it is important to note that we deliberately refrain from incorporating feedback loops. Research on data preparation includes endeavors to build an analytic database from past campaigns and test mailings [e.g., [32]. Marketing papers in the field data preparation examine how alternative definitions of the modeling target [e.g., [14] or covariates [e.g., [28] affect model quality. The data transformation step has been studied through the lens of feature selection [e.g., [24] and independent variable projection [e.g., [6]. The estimation of the actual marketing decision model, its tuning, and possible combination with other models (i.e., ensembling) is the process step that has attracted the largest attention in prior literature [e.g., [27] and is also the focus of this paper. Other papers study a post-processing of model prediction to enhance calibration [e.g., [5] or design new indicators to measure the performance of a decision model [e.g., [39].

The majority of previous studies estimate the targeting model using standard prediction methods (neural networks, support vector machines, etc.). We call this approach profit-agnostic because it does not take account of the actual decision context (i.e., customer targeting) and business objective (i.e., profit maximization) during model development. Only a few studies emphasize the inability of statistical accuracy indicators to reflect marketing objectives and propose application specific alternatives such as the (expected) maximum profit criterion for churn modeling [37,39]. We add to this research through using a more general profit function, which enables us to study a broad range of targeting applications beyond churn. Focusing on profit-oriented model development, we also introduce the business goal earlier in the modeling process where corresponding information can exert more influence on the eventual model. To confirm this, we empirically compare PCES to the approach proposed in [37].

To our knowledge, three studies consider a profit-oriented model development in marketing. Using a genetic algorithm (GA), Bhattacharyya [2] estimates the parameters of a linear model so as to maximize profit. Stripling et al. [34] further extends this approach to maximize the expected maximum profit criterion for churn modeling, while Cui et al. [7] select customers with heterogeneous expected returns via partial ordering. PCES differs from these approaches in that it i) uses a more advanced ensemble learning paradigm and ii) adopts a multi-stage approach to balance statistical loss and business goals. To verify the appropriateness of this design, we empirically compare PCES to the GA-based approach of Bhattacharyya [2] and Stripling et al. [34].

Finally, research in information retrieval is concerned with ranking algorithms, for example to identify the top N most relevant search results for a query. Advanced solutions use deep learning in the form of convolutional neural networks to optimize ranking functions directly [12]. Allocating marketing budgets in campaign planning could be framed as a ranking problem, so that corresponding advancements could have much potential to perform profit analytics in fundamentally new ways.[1]

3. Methodology

In the following, we elaborate on our methodology. First, we review the statistical fundamentals of predictive models and explain how standard loss functions disregard application characteristics. Next, we discuss business goals in campaign planning and corresponding objective functions. Last, we elaborate on the PCES framework, which we propose to combine statistical and business objectives.

3.1. Profit-agnostic targeting models

Targeting models belong to the field of supervised learning [e.g., [17]. Assume a marketer wishes to predict the behavior of customer i, characterized by vector $\boldsymbol{x}_i = (x_{1i},\ x_{2i}, \ldots, x_{Mi}) \in \mathbb{R}^M$, where the elements of $\boldsymbol{x}_i$ capture transactional and demographic information, amongst others. Let y_{i} denote the response of customer i to a past marketing action. The response may be continuous (e.g., purchase amount) or discrete (e.g., whether an offer was accepted). We focus on binary classification where $y_i \in \{0,\ 1\}$, with a value of $y_{\mathrm{i}} = 1$ $(y_{\mathrm{i}} = 0)$ indicating that customer i accepted (rejected) a marketing offer. A targeting model, $f(\boldsymbol{x})$, represents a functional mapping from customer records to responses: $f_{\boldsymbol{\Lambda}}(\boldsymbol{x}) : \mathbb{R}^M \leftrightarrow \{0, 1\}$, where $\boldsymbol{\Lambda}$ denotes a vector of model parameters. Model estimation involves fitting model parameters to data. Afterwards, the model allows the marketer to predict customer response (and more generally behavior) from observable customer data.

Targeting model development follows an inductive approach: Given a data set of customer records and corresponding responses, $D = (y_i, \boldsymbol{x}_i)_{i=1}^{N}$, a learning algorithm fits the model parameters, $\boldsymbol{\Lambda}$, so as to minimize the deviation between model estimates and actual responses: $\boldsymbol{\Lambda}' \leftarrow \min_{\boldsymbol{\Lambda}} Q(y_i, f_{\boldsymbol{\Lambda}}(\boldsymbol{x}_i))\ \forall\ i = 1,\ \ldots,\ N$, where $\boldsymbol{\Lambda}'$ denotes the optimal set of parameters and the loss function Q measures the disagreement between model outputs and data. Therefore, model estimation is equivalent to minimizing a loss function over D. A loss function represents a model-internal notion of fit. Considering the logit model as an example, Q equals the negative log-likelihood (NLL). Common statistical loss functions (NLL, cross-entropy, Hinge loss, etc.) implement the principles of statistical learning to ensure that a model is able to generalize to novel data. Prediction models estimated using such loss functions are generic and can be employed in many domains. However, they disregard specific application characteristics unless these are accurately reflected in the loss function. We argue that a close correspondence between a model-internal internal notion of fit and business value should not be taken for granted. Maximizing fit using some statistical loss function may lead to a different model compared to maximizing campaign profit. On the other hand, statistical loss functions have strong theoretical underpinnings and exhibit desirable properties related to generalization [e.g., [17]. It is imperative to build on this theory when developing a prediction model. This motivates our PCES approach to integrate statistical considerations (in the form of established loss functions and estimation principles) and business value (in the form of campaign profit) during target model development.

3.2. Target group selection and model assessment in marketing campaign planning

Campaign planning aims at maximizing the efficiency of resource utilization. Contacting customers with a marketing message entails a cost so that it is typically inefficient to target the whole customer base. Instead, marketers use targeting models to estimate response probabilities on a customer level. This facilitates restricting solicitations to likely responders. Applications of targeting models include the mail-order industry, churn management, and cross-selling. Recently, targeting models are increasingly used in real-time settings such as digital marketing [e.g., [30] and social media [e.g., [21].

From a managerial point of view, the business value of a targeting model depends on the degree to which it increases the profitability of targeted marketing actions. We model the profit of a marketing campaign, Ω, as follows [26]:

$$\Omega(l(\tau), \tau) = N \cdot \tau \cdot (\pi_+ \cdot l(\tau) \cdot r - c), \tag{1}$$

where N denotes the size of the customer base, τ the fraction of targeted customers (i.e., campaign size), and π_+ the base rate of customers willing to accept the marketing offer in the customer base. The parameters r and c represent the return and cost associated with an accepted offer and making the offer, respectively. The quantity $l(\tau)$, called the lift, is a marketing specific measure of predictive accuracy, which depends on the size of the campaign, τ. With π_τ denoting the fraction of responses in the target group, the lift is given as:

$$l(\tau) = \frac{\pi_\tau}{\pi_+} \tag{2}$$

A campaign that targets customers at random reaches a fraction of π_+ actual responders. Thus, the lift assesses the degree to which a model-based targeting improves over a random benchmark.

Revised versions of (1) have been proposed to capture the characteristics of specific marketing applications. Neslin et al. [29] devise a profit function for models that target retention actions to customers with high churn probability. The expected

[1] The authors would like to thank an anonymous reviewer for suggesting this approach toward profit analytics.

maximum profit criterion further refines this approach [39]. The advantage of the campaign profit function (1) over subsequent advancements is generality. Connecting customer revenues, direct costs, and model accuracy through model lift, (1) can represent a variety of targeting applications including churn management, direct mail, e-couponing, etc. Therefore, we use (1) in this paper and leave the evaluation of the proposed PCES approach for specific targeting tasks such as churn modeling to future work.

An assumption of (1) and its extensions is that costs and returns are homogeneous across customers. In campaign planning, assuming constant offer costs is plausible for most marketing channels. However, disregarding variability in customer spending ($r =$ const.) is a strong simplification. Typically, the returns from accepted marketing offers differ across customers. Our justification for using (1) despite this assumption is threefold. First, it is common practice to work with class as opposed to case depending costs/returns in the marketing and cost-sensitive learning literature [e.g., [32,37]. Second, calculating campaign profit using the mean revenue per accepted offer may be more suitable for predictive modeling, for example because information to estimate revenues at the customer level reliably is lacking. Last, some applications do not require distinguishing revenues across customers, for example when targeting services entail a fixed fee or when running lead generation campaigns.

3.3. Profit-conscious ensemble selection

The proposed modeling framework is based on the view that the development of predictive decision support models should pay attention to both statistical and business considerations. Therefore, we strive to incorporate campaign profit (1) as marketing objective into model development (see Fig. 1). To achieve this, we decompose model development into two sub-steps. The first stage leverages statistical learning principles. In step two, model predictions are refined to maximize campaign profit. Recall that such multi-stage approach mimics the way in which managers use decision support models: they re-appraise and possibly correct DSS outputs in the context of their decision task [10].

The proposed framework is based on a machine learning paradigm called ensemble selection [3]. An ensemble is a collection of (base) models, all of which predict the same target. Much research confirm that combining multiple models in an ensemble is useful to increase predictive accuracy [e.g., [37]. Ensemble selection involves three steps: (i) constructing a library of candidate models (*model library*), (ii) selecting an “appropriate” subset of models for the ensemble (*candidate selection*), and (iii) integrating the predictions of the chosen models into a composite forecast (*model aggregation*). From an algorithmic point of view, PCES follows Caruana’s et al. [3] approach. Its distinctive feature is that it integrates statistical and economic objectives. This way, PCES embodies a different paradigm toward developing predictive decision support models.

3.3.1. Model library

The success of an ensemble depends on the diversity of its members. To obtain a library of diverse models, we use different learning algorithms. We also consider multiple settings for algorithmic meta-parameters. Meta-parameters such as the regularization parameter in support vector machines facilitate adapting a learning algorithm to a task, which suggests that prediction models from the same algorithm vary with meta-parameters and display diversity. Table 1 summarizes the learning algorithms and meta-parameter settings in the model library.

It is common practice to select a specific, ‘best’ set of meta-parameters for an individual learning algorithm in a model selection step. As we detail below (see Section 4.2), we also adopt this practice to obtain benchmark models against which we compare PCES. However, for PCES itself, we do not perform model selection a priori but keep all candidate models in the library. The selection of algorithms and meta-parameters is based upon previous literature on customer targeting and ensemble modeling [20,37]. Some methods have been chosen due to their popularity (e.g., logistic regression, decision trees, discriminant analysis) and others because of high performance in previous studies (e.g., random forest, support vector machines, gradient boosting). Interested readers can find a comprehensive discussion of the algorithms in [17]. In total, we consider 15 learning algorithms from which we derive 877 different models. We acknowledge that several extensions of popular machine learning algorithms have been proposed in the literature. Innovative learners like, for example, the fuzzy support vector machine [41] may give better results than the original version of the algorithm. Our reason to not include corresponding techniques comes from the design goal of PCES to be easy to implement in practice. Standard algorithms as those forming our model library are available in contemporary business analytics software such as, e.g., SAS, Microsoft Azure ML, and many others as well as popular data science programming languages such as R, Python, Scala, etc. or high-performance computing infrastructures like Apache Spark. Leveraging corresponding standards is beneficial because it ensures that companies could deploy PCES at low cost and without a need to re-implement algorithms that have mainly been used in research. The same consideration discourages an application of deep learning in this paper.

3.3.2. Candidate selection

Given the model library, we select candidate models using directed hill-climbing [3]. In particular, we first select the single best candidate model from the library. To improve this model’s performance, we next assess all pairwise combinations of the chosen model and one other base model from the library. This way, we obtain a collection of possible two-member ensembles, out of which we select the best performing candidate ensemble. We then continue with examining the set of all three-member ensembles that include the models chosen in the previous iteration. Incremental ensemble growing

Table 1
Classification methods and meta-parameter settings.

Learning algorithm	Meta-parameter*	Candidate settings**
Classification and regression tree Recursively partitions a training data set by inducing binary splitting rules so as to minimize the impurity of child nodes in terms of the *Gini* coefficient. Terminal nodes are assigned a posterior class-membership probability according to the distribution of the classes of the training instances contained in this node. To classify novel instances, the splitting rules learned during model building are employed to determine an appropriate terminal node. *Overall number of models: 6*	Min. size of nonterminal nodes Pruning of fully grown tree	10, 100, 1000 Yes, No
Artificial neural network Three-layered architecture of information processing-units referred to as neurons. Each neuron receives an input signal in the form of a weighted sum over the outputs of the preceding layer's neurons. This input is transformed by means of a logistic function to compute the neuron's output, which is passed to the next layer. The neurons of the first layer are simply the covariates of a classification task. The output layer consists of a single neuron, whose output can be interpreted as a class-membership probability. Building a neural network models involves determining connection weights by minimizing a regularized loss-function over training data. *Overall number of models: 162*	No. of neurons in hidden layer Regularization factor (weight decay)	3, 4, …, 20 $10^{[-4,\ -3.5,\ \ldots,\ 0]}$
k-nearest-neighbor Decision objects are assigned a class-membership probability according to the class distribution prevailing among its k nearest (in terms of Euclidian distance) neighbors. *Overall number of models: 18*	Number of nearest neighbors	10, 100, 150, 200, …, 500, 1000, 1500, …4000
Linear discriminant analysis Approximates class-specific probabilities by means of multivariate normal distributions assuming identical covariance matrices. This assumption yields a linear classification model, whose parameters are estimated by means of maximum likelihood procedures from training data. *Overall number of models: 20*	Covariates considered in the model	Full model, stepwise variable selection with p-values in the range 0.05, 0.1,…, 0.95
Logistic regression Approximates class membership probabilities (i.e., a posteriori probabilities) by means of a logistic function, whose parameters are estimated from training data by maximum likelihood procedures. *Overall number of models: 20*	Covariates considered in the model	Full model, stepwise variable selection with p-values in the range 0.05, 0.1,…, 0.95
Naive bayes Approximates class-specific probabilities under the assumption that all covariates are statistically independent. *Overall number of models: 9*	Histogram bin size	2, 3, …, 10
Quadratic discriminant analysis Differs from LDA only in terms of the assumption about the structure of the covariance matrix. Relaxing the assumption of identical covariance leads to a quadratic discriminant function. *Overall number of models: 20*	Covariates considered in the model	Full model, stepwise variable selection with p-values in the range 0.05, 0.1,…, 0.95
Regularized logistic regression Differs from ordinary LogR in the objective function optimized during model building. A complexity penalty given by the L1-norm of model parameters (Lasso-penalty) is introduced to obtain a "simpler" model. *Overall number of models: 29*	Regularization factor	$2^{[-14,\ -13,\ \ldots,\ 14]}$
Support vector machine with linear kernel Constructs a linear boundary between training instances of adjacent classes so as to maximize the distance between the closest examples of opposite classes and achieve a pure separation of the two groups. *Overall number of models: 29*	Regularization factor	$2^{[-14,\ -13,\ \ldots,\ 14]}$

(*continued on next page*)

Table 1 (*continued*)

Learning algorithm	Meta-parameter*	Candidate settings**
Support vector machine with radial basis function kernel Extends SVM-lin by implicitly projecting training instances to a higher dimensional space by means of a kernel function. The linear decision boundary is constructed in this transformed space, which results in a nonlinear classification model. *Overall number of models: 300*	Regularization factor	$2^{[-12,\,-11,\,\ldots,\,12]}$
	Width of Rbf kernel function	$2^{[-12,\,-11,\,\ldots,\,-1]}$
AdaBoost Constructs an ensemble of decision trees in an incremental manner. The new members to be appended to the collection are built in a way to avoid the classification errors of the current ensemble. The ensemble prediction is computed as a weighted sum over the member classifiers' predictions, whereby member weights follow directly from the iterative ensemble building mechanism. *Overall number of models: 11*	No. of member classifiers	10, 20, 30, 40, 50, 100, 250, 500, 1000, 1500, 2000
Bagged decision trees Constructs multiple CART trees on bootstrap samples of the original training data. The predictions of individual members are aggregated by means of average aggregation. *Overall number of models: 11*	No. of member classifiers	10, 20, 30, 40, 50, 100, 250, 500, 1000, 1500, 2000
Bagged neural networks Equivalent to BagDT but using ANN instead of CART to construct member classifiers. The ensemble prediction is computed as a simple average over member predictions. *Overall number of models: 5*	No. of member classifiers	5, 10, 25, 50, 100
Random forest The ensemble consists of fully grown CART classifiers derived from bootstrap samples of the training data. In contrast with standard CART classifiers that determine splitting rules over all covariates, a subset of covariates is randomly drawn whenever a node is branched and the optimal split is determined only for these preselected variables. The additional randomization increases diversity among member classifiers. The ensemble prediction follows from average aggregation. *Overall number of models: 35*	No. of member classifiers	100, 250, 500, 750, 1000, 1500, 2000***
	No. of covariates randomly selected for node splitting	
LogitBoost Modification of the AdaB algorithm which considers a logistic loss function during the incremental member construction. We employ tree-based models as member classifiers. *Overall number of models: 11*	No. of member classifiers	10, 20, 30, 40, 50, 100, 250, 500, 1000, 1500, 2000
Stochastic gradient boosting Modification of the AdaB algorithm, which incorporates bootstrap sampling and organizes the incremental ensemble construction in a way to optimize the gradient of some differential loss function with respect to the present ensemble composition. We employ tree-based models as member classifiers. *Overall number of models: 11*	No. of member classifiers	10, 20, 30, 40, 50, 100, 250, 500, 1000, 1500, 2000

* Note that Table 1 depicts only those meta-parameters for which we consider multiple settings. A classification method may offer additional meta-parameters.

** We consider all possible combination of meta-parameter settings for learners such as Random Forest that exhibit multiple meta-parameters.

*** *M* represents the number of explanatory variables (i.e., covariates) in a data set.

terminates when adding novel members stops improving performance. Interested readers find a working example of the algorithm in the e-companion (see online Appendix I).

We propose to reserve the selection step for business objectives. Using heuristic search, it is possible to gear ensemble selection toward any objective function that depends on the model-estimated probabilities. In this paper, we devise an ensemble that incorporates business objectives through maximizing (1) in the selection stage. This way, PCES refines the first-stage predictions, which stem from well-established prediction models and embody the principles of statistical learning, by means of a combination of predictions to better represent the actual decision problem.

From a mathematical point of view, configuring the hill-climbing heuristic to maximize (1) appears a minor modification. However, this modest modification leads to a fundamentally different paradigm toward prediction model development. The campaign profit function (1) captures the business value of a decision support model and characteristics of the decision context such as a budget constraint (in the form of τ. Consequently, maximizing (1) leads to a contextualized model that is aware of the environment to which it will be deployed and the decisions it is meant to support. Furthermore, an ex-post revision of (individual model) predictions as done by PCES mimics the way in which managers use DSS recommendations and possibly correct for misleading advice [10]. These features represent the real value of PCES and, as we suggest, warrant a comprehensive empirical evaluation how much a contextualized modeling paradigm improves over standard supervised learning.

3.3.3. Model aggregation

Model aggregation refers to a combination of models' predictions. PCES combines predictions in the candidate selection step (see 3.3.2). A candidate ensemble consists of a subset of base models. To assess a candidate ensemble, we compute the simple average over the predictions of the selected base models. We detail this approach in the online appendix, which provides a numerical example of candidate selection and PCES (see Appendix I in the online appendix). PCES performs the same model aggregation when computing the predictions of the final ensemble, which is the specific selection of base models that gives the best results during candidate selection.

Although we pool models by averaging over their predictions, PCES effectively computes a weighted average. This is because the candidate selection procedure of Caruana et al. [3] allows the same model to enter the ensemble multiple times. The opportunity to weight predictions whenever the data suggest that a strong model deserves greater influence on the ensemble prediction adds to the flexibility of ensemble selection. Note that averaging model predictions requires all models to produce forecasts of a common scale. To ensure this, we calibrate base model predictions using a logistic link function prior to model averaging [31].

4. Empirical design

We examine the effectiveness of PCES in the scope of an empirical benchmark. Such experiment requires suitable data, which represents the characteristics of customer targeting applications, and benchmark models to put the performance of PCES into context.

4.1. Marketing data sets

The empirical study considers 25 cross-sectional marketing data sets. The data sets stem from different industries and represent different prediction tasks, each of which requires selecting customers for targeted marketing actions. The main sources from which we gather the data sets are: (i) data mining competitions, (ii) previous modeling studies, (iii) the UCI machine learning repository [22], and (iv) projects with industry partners. Given the large number of data sets, it is prohibitive to discuss every data set in detail. Table 2 summarizes data set characteristics and identifies sources where more information is available. Every data set has been recorded at a given point in time. Accordingly, variable values give a snapshot of the state of a customer but provide no information how a variable, say customer spending, has evolved over time. For this reason, we do not consider sequence learning algorithms such as recurrent neural networks in this paper.

To simulate a real-world campaign planning setting, we randomly split data sets into two samples using a ratio of 60:40. We refer to the two samples as the training set and the test set, respectively. We develop targeting models using the training set and assess fully specified models on the test set. Certain modeling choices within PCES and the benchmark models (see below) require auxiliary validation data. Examples include the identification of the best base model in the library (as benchmark to PCES) and the heuristic search for ensemble members in the second stage of PCES. We obtain such validation data by means of five-fold cross validation on the training set [3].

4.2. Benchmark models

Alternative targeting models represent a natural benchmark to the proposed PCES approach. We consider (i) the well-known logit model, due to its popularity in marketing, (ii) random forest, due to its success in previous benchmarking studies [e.g., [20,37], and (iii) a best base model (BBM) benchmark, which is given by the strongest individual targeting model from the model library. A common denominator among these benchmarks is that they account for the problem context during *model selection*. For each marketing data set, we select among the 20 / 35 / 877 candidate logit / random forest / base models (see Table 1) the one giving maximal campaign profit (1). Prior work finds a selection of prediction models using business performance measures to substantially improve decision quality [e.g., [14,37,38]. Therefore, we expect the benchmarks to be challenging. To further elaborate on our approach toward benchmark selection, recall that our model library includes multiple models for each learning algorithm, which we derive from executing the algorithm with different settings for algorithmic meta-parameters (see Table 1). We select the logit and random forest benchmarks among all logit and random forest models in the model library for each data set and for each experimental setting. For example, we consider multiple cost-to-benefit ratios and examine model performance across these ratios on each data set. We also consider different mailing depths. In the interest of obtaining a challenging benchmark, we select the strongest logit/random forest model for each setting and data set individually. We proceed in the same way to select the BBM, this time, however not selecting the benchmark model only among candidate logit / random forest models but all models in the library.

Table 2
Data sets characteristics.

Data set	Marketing objective	Industry	Source*	Observations	Variables	P(+1)**
D1	Churn prediction	Energy	DMC02	20,000	32	0.10
D2	Churn prediction	Finance	CP	155,056	23	0.14
D3	Churn prediction	Finance	CP	30,104	47	0.04
D4	Churn prediction	Telco	[37]	40,000	70	0.50
D5	Churn prediction	Telco	[37]	93,893	196	0.50
D6	Churn prediction	Telco	[37]	12,410	18	0.39
D7	Churn prediction	Telco	[37]	69,309	67	0.29
D8	Churn prediction	Telco	[37]	21,143	384	0.12
D9	Churn prediction	Telco	KDD09	50,000	301	0.07
D10	Churn prediction	Telco	[37]	47,761	41	0.04
D11	Churn prediction	Telco	[37]	5000	18	0.14
D12	Profitability scoring	E-Commerce	DMC05	50,000	119	0.06
D13	Profitability scoring	E-Commerce	DMC06	16,000	24	0.49
D14	Profitability scoring	Mail-order	UCI-Adult	48,842	17	0.24
D15	Profitability scoring	Mail-order	DMC04	40,292	107	0.21
D16	Response modeling	Charity	KDD98	191,779	43	0.05
D17	Response modeling	E-Commerce	CP	121,511	82	0.06
D18	Response modeling	E-Commerce	CP	214,709	77	0.13
D19	Response modeling	E-Commerce	CP	382,697	76	0.09
D20	Response modeling	E-Commerce	DMC10	32,428	40	0.19
D21	Response modeling	Finance	CP	45,211	16	0.12
D22	Response modeling	Finance	UCI-Coil	9822	13	0.06
D23	Response modeling	Mail-order	DMC01	28,128	106	0.50
D24	Response modeling	Publishing	CP	300,000	30	0.01
D25	Response modeling	Retail	DMC07	100,000	17	0.24

* CP = consultancy project with industry; DMC = Data Mining Cup (http://www.data-mining-cup.com) (the number gives the year of the competition); KDD = ACM KDD Cup(http://www.sigkdd.org/kddcup/index.php) (the number gives the year of the competition); UCI-xxx = UCI Machine Learning Repository(http://archive.ics.uci.edu/ml/) (with xxx being the name of the data set in the repository).

** P(+1) denotes the prior probability of response (e.g., the fraction of customers who accept an offer).

The ensemble selection approach of Caruana et al. [3] contributes a fourth benchmark. Here, we call it profit-agnostic ensemble selection (PAES) and employ a statistical loss function (i.e., NNL) for base model selection. Therefore, PAES and PCES differ in their approach to select base models for the final ensemble in a profit-agnostic as opposed to a profit-conscious manner. This configuration allows us to attribute performance differences between PAES and PCES to the fact that the latter accounts for business performance during model development.

The last benchmark draws inspiration from Bhattacharyya [2]. It optimizes the coefficients of a linear regression function, which discriminates between responsive and non-responsive customers, using a genetic algorithm (GA). We use (1) as fitness function implying that the GA maximizes campaign profit. Focusing exclusively on business goals during model development, GA is a useful benchmark to support the design of PCES as an integrated modeling framework that balances statistical and economic considerations. GAs exhibit meta-parameters such as the size of the population, the specific type of crossover operator or the mutation rate. In configuring the GA benchmark, we rely on Bhattacharyya [2] and use their settings of population size = 50, crossover rate = 0.7, and mutation rate = 0.2.

4.3. Configuration of ensemble selection

Caruana et al. [3] propose some modifications of basic ensemble selection. One extension consists of an additional bagging step. Instead of selecting a single set of base models from the full model library, they subsample the library, select one ensemble from each subsample, and average over the resulting ensembles [3]. The basic and bagged ensemble selection algorithms represent alternative strategies to develop a model. We consider both strategies and determine the superior approach for each data set by means of model selection. For bagged ensemble selection, we consider subsample sizes of 5%, 10%, and 20% of the model library and 5, 10, and 25 bagging iterations. Importantly, PAES and PCES are treated in the same way to avoid bias.

5. Empirical results

The experimental design provides test set predictions from PCES and benchmark models across the marketing data sets. Many indicators are available to assess predictive accuracy. We suggest that a comparison in terms of business performance is most meaningful from a managerial point of view and thus assess targeting models in terms of campaign profit (1).

Recall that (1) is a function of campaign size, τ. In the following, we consider τ a decision variable and let a targeting model find the profit maximal solution to (1) over $l(\tau)$ and τ. This implies that the model determines which and how many customers to target and thus how much to spend on the campaign. Verbeke et al. [37] recommend this approach and

Table 3
Win-tie-loss statistics of PCES versus benchmarks in the flexible budget case.

	PCES vs. Logit			PCES vs. RF			PCES vs. BBM			PCES vs. GA			PCES vs. PAES		
Return (r)	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss
$2	24	1	0	21	2	2	22	1	2	25	0	0	19	3	3
$3	24	0	1	21	1	3	22	1	2	25	0	0	22	0	3
$4	25	0	0	24	0	1	21	1	3	25	0	0	20	0	5
$5	25	0	0	23	1	1	23	1	1	24	1	0	20	0	5
$10	24	0	1	24	0	1	22	0	3	24	0	1	18	0	7
$15	24	0	1	23	0	2	18	0	7	24	0	1	12	0	13
$20	24	0	1	23	0	2	22	0	3	24	0	1	17	0	8
$25	24	0	1	24	0	1	23	0	2	23	0	2	16	1	8
$50	23	0	2	23	0	2	22	0	3	24	0	1	16	0	9
$75	23	0	2	21	1	3	21	0	4	24	0	1	13	0	12
$100	23	0	2	19	1	5	20	0	5	23	1	1	11	1	13
Total	263	1	11	246	6	23	236	4	35	265	2	8	184	5	86
	96%	0%	4%	89%	2%	8%	86%	1%	13%	96%	1%	3%	67%	2%	31%
p-value*	0.000			0.000			0.000			0.000			0.000		

* The p-values correspond to pairwise comparisons of PCES and one benchmark, using Rom's procedure to protect against an elevation of alpha values in multiple pairwise comparisons [11]. Multiple pairwise comparisons are feasible since a X^2 value of 823.5 suggest that we can reject the null hypothesis of equal performance among models (Friedman test) with high confidence (p-value $<$ 0.000).

proof its effectiveness. We follow their advice but consider a different profit function to cover a larger scope of marketing applications.

To cover a broad range of application scenarios, we consider multiple settings for the monetary campaign parameters offer cost (c) and return per accepted offer (r). More specifically, it is sufficient to vary r because the profit function (1) is invariant to a linear scaling. Rescaling (1) such that $c = 1$ and $r' = r/c$ does not change the profit maximal solution. We thus fix c at $1 and consider settings of r = $2, $3, $5, $10, $15, $25, $50, $75, and $100. These values capture a range of targeting applications. Smaller values represent settings where the ratio between offer cost and return per accept is moderately skewed. Such scenario might occur when companies contact customers through a call-center or when selling products by means of printed catalogs in the mail-order industry. Both channels involve considerable offer costs (e.g., to produce a premium catalog), which could explain moderate imbalance between r and c. High skewness between these parameters arises in online marketing where digital channels facilitate reaching customers at very low costs. Larger values of r capture such applications. Overall, considering 25 marketing data sets with 11 settings for the cost-to-benefit ratio, r/c, we obtain 275 experimental settings. To carry out profit optimization, we run PCES as well as the PAES and GA benchmark for each of these settings individually. For the logit, random forest and BBM benchmark, we use the models stored in the model library and respectively select the best logit, random forest, and base model for each experimental setting. Given that larger values of r give an incentive to increase campaign size, we constrain the optimization of (1) such that $\tau \leq 0.5$. Since marketing campaigns typically target a small fraction of customers, contacting more than half of the customer base seems unrealistic.

Table 3 reports the win-tie-loss statistics of PCES vs. benchmark models for the 11 (return to cost ratios) * 25 (data sets) = 275 comparisons. Consider, for example, the comparison of PCES versus BBM at r = $2. A value of 22 suggests that PCES achieves higher campaign profit than BBM on 22 out of 25 data sets. BBM outperforms PCES on two data sets and both models tie on one data set. We also compare the statistical significance of profit differences using the Friedman test (see bottom of Table 3). For the results of Table 3, a X^2 value of 823.5 indicates that we can reject the null hypothesis of equal performance (p-value <0.000). This allows us to proceed with a set of pairwise comparisons of PCES against one benchmark to detect significant differences among individual targeting models. To protect against an elevation of alpha values in multiple pairwise comparisons, we adjust p-values using Rom's procedure [11]. The last row of Table 3 reports the adjusted p-values.

Table 3 reveals evidence that PCES produces significantly higher campaign profits than any of the benchmark models (p-values of pairwise comparisons consistently less than 0.000). Recall that the purpose of the logit, RF, and BBM benchmark is to reflect common marketing practices where a set of candidate models is developed and the strongest candidate (in terms of (1)) is selected. This is exactly the modeling paradigm advocated in previous studies [e.g., [14,37,39]]. Accordingly, the results of Table 3 indicate that introducing the relevant notion of model performance during model development (as opposed to model selection) further increases performance. However, this interpretation requires further qualification since the superiority of PCES may also come from the ability of ensemble selection to create powerful prediction models. Indeed, the PAES benchmark, an ordinary ensemble selection method, turns out to be the strongest benchmark. However, although benefitting from the same large base model library as PCES, a PAES-based customer targeting gives significantly less profit compared to using PCES. In particular, we find the latter to produces higher profits in 184 out of 275 comparisons (67%). Be-

Please cite this article as: S. Lessmann, J. Haupt and K. Coussement et al., Targeting customers for profit: An ensemble learning framework to support marketing decision-making, Information Sciences, https://doi.org/10.1016/j.ins.2019.05.027

Table 4
Comparison of campaign profit at model-optimized campaign sizes.

	Campaign profit [$]					
Data	Logit	RF	BBM	GA	PAES	PCES
D1	1660	1596	1764	1532	**1874**	1846
D2	61,612	75,816	75,989	62,953	75,725	**76,001**
D3	−2	−83	88	−104	76	**137**
D4	−2992	−2832	−2832	−3052	−2852	**26**
D5	−7096	−6766	−6766	−7096	−6666	**25**
D6	−1017	−997	−977	−1027	−997	**159**
D7	35,578	39,598	39,778	35,098	40,408	**40,618**
D8	2966	2926	3270	2756	**3404**	3121
D9	699	469	862	509	999	**1139**
D10	442	876	839	590	901	**984**
D11	1491	2000	2022	1534	2020	**2058**
D12	−8	17	−33	−310	84	**428**
D13	14,700	18,270	18,270	15,110	18,390	**18,810**
D14	34,421	34,755	35,067	34,385	35,107	**35,185**
D15	21,642	21,842	**22,012**	21,353	21,982	21,073
D16	572	6	572	208	527	**726**
D17	9121	9283	9690	9568	**10,690**	10,087
D18	64,096	101,186	105,824	63,438	105,649	**106,418**
D19	85,123	119,158	122,949	91,387	**123,881**	123,804
D20	10,424	10,614	10,564	9954	10,654	**10,884**
D21	12,877	14,534	14,632	12,708	14,498	**14,725**
D22	210	323	325	242	305	**357**
D23	29,044	29,544	**30,154**	28,454	30,074	30,004
D24	−1	−2	14	1	13	**27**
D25	47,440	53,210	53,210	50,380	**53,770**	53,660
Estimated profit increase (in percent)*	657 (22%)	407 (14%)	233 (7%)	756 (27%)	178 (5%)	

* The estimation is based on García et al. [11]. We first use their contrast estimation approach to calculate the expected profit improvement of PCES over a benchmark, and then convert this contrast to a percentage through dividing by the benchmark's median (across data sets) campaign profit.

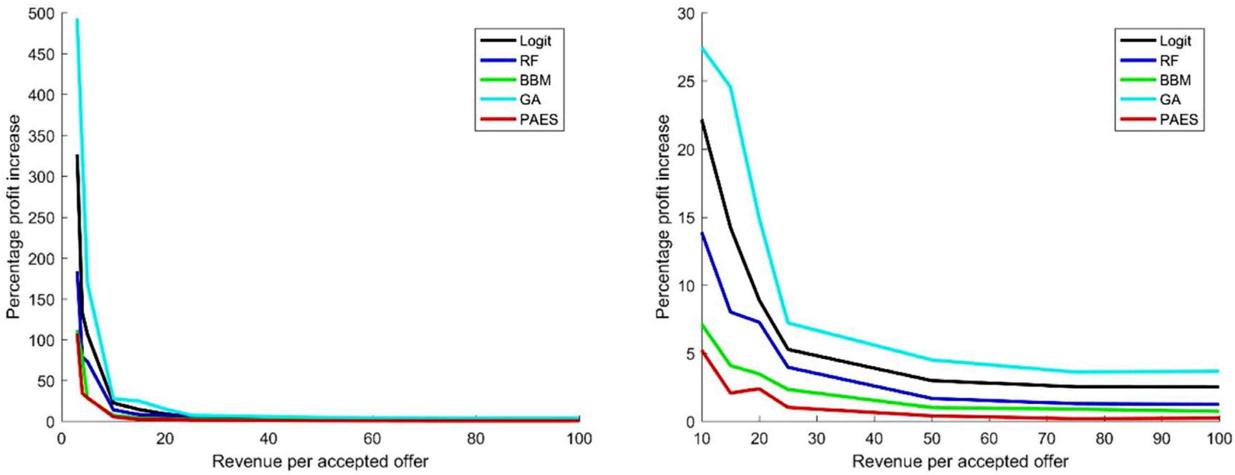
fore examining the relative performance of alternative targeting models in more detail, we note that PCES also outperforms the GA benchmark (i.e., a direct profit maximization) with substantial margin.

To obtain a clearer view on the degree to which PCES increases business performance, we calculate the profit implication resulting from using PCES or a benchmark model for campaign targeting. In particular, we consider a fictitious company with a customer base of $N = 100{,}000$ customers; and let the per-customer return from accepted offers, r, and offer costs to contact customers, c, be $10 and $1, respectively. Table 4 depicts the campaign profits emerging from a model-based targeting per marketing data set. Given that we consider campaign size a decision variable, we let every targeting model select its individually best setting τ. This way, Table 4 compares targeting models in terms of the maximal campaign profit they can produce for given r and c. Bold face highlights the best result per data set. The optimized campaign sizes corresponding to the results of Table 4 are available in Table 5. The last row of Table 4 summarizes the observed results in the form of an estimate of the expected profit increase of PCES over a benchmark. The estimation procedure comes from García et al. [11] and is based on the median profit difference between PCES and a benchmark model across the data sets. Given the scope of the empirical study (e.g., 25 real-world data sets from different industries), we consider the resulting value a reliable estimate of the profit that a targeting model achieves on unseen data.

Table 4 reemphasizes that PCES typically produces higher profits than benchmark models. This is especially apparent when examining the performance contrast shown in the last row of Table 4. Based on the observed results, we expect PCES to increase campaign profit by five percent compared to the most challenging benchmark and up to fourteen percent compared to random forest, a state-of-the-art classifier much credited for high accuracy [e.g., [20]. Profit increases of five percent and above are managerially meaningful, especially for larger companies and companies that run many campaigns [29]. It is also noteworthy that using the logit model for targeting, an approach still popular in industry, entails substantial opportunity costs. Compared to this benchmark, PCES produces higher campaign profits across all data sets and can be expected to increase profits by 22% on average. With respect to a direct optimization of campaign profit during model development, which the GA benchmark embodies, Table 4 reveals that corresponding results are the weakest in the comparison. Last, PCES is the only approach that avoids losses. For some data sets (e.g., D4–D6) the optimization of τ on validation data gives a poor result for the hold-out test data on which we calculate campaign profit. In particular, Table 5 reveals that all benchmarks select τ equal to its upper bound of 0.5 on D4–D6. This leads to large campaigns that result in a loss for the given setting of $r{:}c = 10{:}1$. PCES, on the other hand, benefits from its ability to adapt the ensemble forecast when optimizing τ, because it employs (1) during model development. This allows PCES to recognize that the level of predictive accuracy vis-à-vis the return to cost ratio might not facilitate profitable targeting. Thus, PCES selectsτ close to zero. Finally, Table 5 evidences a trend of PCES to recommend smaller campaigns. The median value $\tau = 16.66$ for PCES is much less than the second-smallest

Table 5
Model-optimized campaign sizes.

	Model-optimized campaign sizes [%]					
Data	Logit	RF	BBM	GA	PAES	PCES
D1	41.12	49.68	35.58	40.09	38.20	43.18
D2	25.78	16.21	15.67	26.15	15.49	15.86
D3	0.35	6.67	4.33	4.01	7.25	4.76
D4	50.00	50.00	50.00	50.00	50.00	0.17
D5	50.00	50.00	50.00	50.00	50.00	0.34
D6	50.00	50.00	50.00	50.00	50.00	1.97
D7	50.00	50.00	50.00	50.00	50.00	50.00
D8	46.16	47.70	46.34	46.87	49.26	50.00
D9	7.70	12.70	16.04	13.20	23.10	16.96
D10	5.07	6.56	5.81	5.44	7.69	5.74
D11	38.43	15.47	14.40	39.77	14.00	15.10
D12	14.14	15.26	17.36	12.35	16.18	7.86
D13	50.00	50.00	50.00	50.00	50.00	50.00
D14	48.52	49.62	48.59	49.83	48.85	47.68
D15	50.00	50.00	50.00	49.93	50.00	45.34
D16	3.83	0.03	3.83	0.71	2.57	4.27
D17	22.04	17.39	17.61	15.44	19.52	16.66
D18	36.83	20.09	17.74	35.45	17.56	17.03
D19	19.52	13.03	12.14	18.99	12.55	12.04
D20	50.00	50.00	50.00	50.00	50.00	50.00
D21	28.99	25.47	26.97	30.64	25.78	27.95
D22	23.65	15.44	14.63	18.51	23.02	10.77
D23	50.00	50.00	50.00	50.00	50.00	50.00
D24	0.00	0.01	0.04	0.06	0.04	0.04
D25	50.00	50.00	50.00	50.00	50.00	50.00
Median	38.43	25.47	26.97	39.77	25.78	16.66



(a)

(b)

Fig. 2. Expected percentage improvement in campaign profit due to using PCES for target group selection. We estimate profit contrasts in the same way as in Table 4. Panel (a) shows all settings of r, whereas panel (b) focuses on settings of $r > 5$ for better readability.

value of $\tau = 25.47$ for RF. Smaller campaigns are appealing since they require less resources and might be better targeted to customer interests. For example, despite recommending smaller campaigns, PCES produces higher profits than RF on all data sets, which signals higher predictive accuracy and, in turn, better targeting.

The results of Tables 4 and 5 stem from a campaign with specific setting of returns and offer costs. To confirm generalizability of results to other campaign settings, we next examine the magnitude of PCES-induced profit improvements across the full range of campaign parameters $r =$ \$2, \$3, \$5, \$10, \$15, \$25, \$50, \$75, and \$100 (with $c =$ \$1). To that end, we rerun model development (for PCES and GA) and model selection (logit, RF, BBM, PAES) for all data sets and settings of r. We then use the same contrast estimation approach (see last row Table 4) to calculate percentage profit improvements of PCES over its benchmarks [11]. Fig. 2 depicts the corresponding results. Given that smaller settings of r lead to large improvements over weaker benchmarks, we split Fig. 2 into two panels which show results for all settings of r and those above five, respectively.

Please cite this article as: S. Lessmann, J. Haupt and K. Coussement et al., Targeting customers for profit: An ensemble learning framework to support marketing decision-making, Information Sciences, https://doi.org/10.1016/j.ins.2019.05.027

Fig. 2 confirms that superior performance of PCES generalizes to other settings of campaign parameters. Above zero improvements demonstrate that PCES consistently produces higher profits than the benchmarks. GA is again the weakest benchmark in the comparison. Even in the scenario r:$c=100$:1, where high imbalance between marketing returns and costs renders the targeting task relatively easy, PCES increases campaign profits by more than five percent compared to GA. This confirms that direct maximization of campaign profits is not a suitable approach to develop targeting models. The other models ground on statistical learning. From Fig. 2, we conclude that following corresponding principles is essential when developing a targeting model. However, the specific adaptation that we propose, namely to introduce campaign profits into model development, succeeds in improving the business performance of the resulting model. Random forest, for example, recommends campaigns that are roughly 3–15% less profitable compared to PCES.

6. Discussion

The empirical analysis evidences the effectiveness of the proposed approach toward model development. Our study also sheds light on the divergence between the optimization of statistical loss and business objectives for prediction model development in targeting applications. The experimental design includes three philosophies toward model development: (i) a direct maximization of business performance (GA), (ii) a model selection approach, which introduces business objectives ex-post and develops models using statistical learning (Logit, RF, BBM and PAES), and (iii) PCES that shifts the consideration of the actual business objective to a previous modeling stage to gear model development toward the ultimate goal of the marketing campaign.

We find the direct approach to be least effective. Even a simple logit model outperforms GA. The logit and GA model both construct a linear classifier. Better performance of the former evidences that model development through minimizing statistical loss is preferable to a direct maximization of business performance. Well-known estimation problems such as overfitting [e.g., [17] are a likely cause of this result. Remedies to such problems are available in statistical learning. However, developing predictive models through profit maximization, the direct approach is unable to capitalize on this knowledge.

Considering the model selection approach, logistic regression, random forest, and BBM perform better than GA but inferior to PCES. Profit improvements over these benchmarks are often substantial. On average, PCES also recommends smaller campaigns, which indicates better targeting of PCES campaigns. Overall, these results suggest that incorporating business goals early in the modeling process has a sizeable positive effect on the quality of the prediction model and decision support, respectively.

One might object that a targeting model that is tuned to maximize profits will naturally give higher profits than a model that minimizes NLL or another loss function. Following this line of reasoning, one might question the fairness of the comparison in terms of campaign profit (1). However, it is important to recall that targeting is a prediction problem. We aim at predicting customer responses to marketing messages. In predictive modeling, it is crucial to develop a model on one set of (training) data and test it on a different, 'fresh' set of (test) data [e.g., [33]. Given disjoint data sets for model training and evaluation, it is wrong to assume that maximizing profit on the training set will naturally give higher profit on the test set. This is apparent from the poor results of the GA benchmark and, more importantly, statistical learning theory [e.g., [36]. Consequently, the experimental design facilitates a fair comparison.

However, it is still interesting to examine the performance of PCES across different evaluation measures to shed lights on the antecedents of its success in the above comparison. In particular, maximizing campaign profit (1) over $l(\tau)$ and τ, our evaluation criterion differs notably from typical accuracy indicators and statistical loss functions. We hypothesize that the advantage of PCES over benchmark models decreases when the ensemble selection criterion (i.e., business performance measure) is more similar to the loss functions that standard targeting models embody. To test this, the paper is accompanied by an e-companion, which provides results for additional performance measures; namely AUC and TDL (online Appendix II) and campaign profit under a budget constraint (online Appendix III[7]). With respect to the similarity of these measures to standard indicators of predictive accuracy and statistical loss, we suggest an ordering of the form $AUC \prec TDL \prec \Omega(l(\tau), \tau = const.) \prec \Omega(l(\tau), \tau)$. AUC captures a classifier's ranking performance. It is a standard accuracy indicator, which we consider relatively closest to standard loss functions like NLL [1]. TDL is related to AUC but focuses on ranking performance among of subset of customers [e.g., [29]. Thus, we consider it more distinct from model-internal loss functions. The same logic applies to campaign profit under a budget constrain ($\Omega(l(\tau), \tau = const.)$), just that this measure, in addition, depends on cost and benefit parameters which introduce further differences. Last, the evaluation measure we consider above, campaign profit with flexible marketing budget, $\Omega(l(\tau), \tau)$, includes the additional decision variable τ and is therefore most distinct from NLL or other standard loss functions.

Below, we summarize results from the e-companion and illustrate how the relative performance advantage of PCES develops across different performance measures. In particular, Table 6 reports the estimated performance improvement over a benchmark model across AUC, TDL, and campaign profit with fixed and flexible budget, whereby we use the same approach toward performance contrast estimation as in Table 4 [11]. The e-companion provides a more detailed analysis of AUC, TDL performance in Appendix II, and campaign profit with budget constraint in Appendix III.

Table 6 supports the view that PCES is most effective if an application specific (business) performance measure embodies a different notion of model performance than a model-internal loss function. Performance improvements are especially pronounced when assessing model performance in terms of campaign profit with flexible budget. On the other hand, improvements over the strongest competitor, PAES, vanish when using the AUC for performance evaluation, and are marginal

Table 6
Comparison of PCES and benchmarks across statistical and monetary performance measures .

	AUC	TDL	$\Omega(l(\tau), \tau = const.)$	$\Omega(l(\tau), \tau)$
Logit	7.31%	25.79%	18.10%	22.00%
RF	1.39%	3.58%	2.30%	14.00%
BBM	0.28%	3.10%	1.00%	7.00%
GA	6.23%	21.91%	15.60%	27.00%
PAES	0.00%	0.14%	0.30%	5.00%

We compute the relative performance improvements of PCES over benchmarks in the same way as in Table 4 using the contrast estimation approach of García, et al. [11].

for TDL and campaign profit under a budget constraint. The results for other benchmarks follow a similar trend, whereby PCES still provides a sizeable advantage in most cases. Overall, we take Table 6 as further evidence that incorporating profit consideration into model development is valuable. More specifically, the efficacy of PCES increases with decreasing similarity between a targeting model's internal loss function and a relevant measure of business performance.

7. Summary

We set out to develop a modeling approach that integrates principles of statistical learning with business objectives in customer targeting. To achieve this, we propose PCES, which first estimates a set of statistical prediction models and then selects from this library a subset of models so as to maximize campaign profit. The results that we obtain from a comprehensive empirical study confirm the effectiveness of this approach. We observe PCES to predict customer responsiveness more accurately than benchmarks and show that the profit of a marketing campaign increases when using PCES for target group selection. We also find this advantage over competitors to increase with decreasing correlation between a model-internal loss function and a relevant measure of business performance.

7.1. Implications

The results of our study have several implications. First, integrating business goals into the modeling process is interesting from a theoretical point of view. A large number of prediction methods have been developed in the literature. Well-grounded in the theory of statistical learning, such methods facilitate the development of empirical prediction models in diverse application settings. Generality, however, has a cost. General purpose methods disregard the characteristic properties of specific applications such as profit in campaign planning. On the other end, a common approach toward decision support in the literature involves the development of tailor-made models that fully reflect the requirements of a given application. However, tailor-made models also suffer limitations. In the case of predictive modeling, a possible shortcoming may be that they are less accurate, for example because they fail to automatically account for nonlinear patterns. We consider our results a stimulus to rethink approaches to develop prediction models. In particular, we call for the development of modeling methodologies that are both widely applicable and aware of characteristic application requirements. To some extent, the proposed PCES framework is such an approach. For example, to adapt PCES to a decision problem other than targeting, we can replace the campaign profit function (1), which guides ensemble member selection, with an objective function that captures the peculiarities of the novel business application.

Second, from a managerial perspective, the key question is to what extent novel targeting models add to the bottom line. In this sense, an implication of our study is that it is feasible and effective to develop targeting models in a profit-conscious manner. Improvements of campaign profit of several percent, which we observe in many experimental settings, are managerially meaningful and indicate that PCES is a useful addition to campaign planners' toolset. Its application seems especially rewarding in settings where companies contact a large number of customers, conduct many campaigns, and/or run campaigns with high frequency, all of which is common in digital marketing and e-commerce.

A third implication of the study is related to the way in which targeting models are commonly employed in academia and industry. In particular, a model selection approach, which involves developing a set of candidate models and selecting *one* best model for deployment should be avoided. Our study suggests that an appropriately chosen combination of (some of these) alternative models using ensemble selection is likely to increase predictive accuracy and, more generally, model performance. Furthermore, introducing an additional selection and combination step into the modeling process provides an excellent opportunity to account for business objectives during model development.

Finally, a fourth implication is that the development of targeting models requires little human intervention. Typical modeling tasks include, for example, testing different variables, transformations of variables to increase their predictive value, and testing alternative prediction methods. Using an ensemble selection framework, campaign managers can easily automate these tasks. They only need to incorporate the candidate models that represent choice alternatives into the model library. The selection strategy will then pick the most beneficial model combination in a profit-conscious manner. This frees campaign planners from laborious, repetitive modeling tasks and unlocks valuable resources, which can be spend on tasks that truly require creativity and domain knowledge. In the case of predictive modeling, engineering informative features is a good example for such task.

7.2. Limitations and future research

Clearly, the study exhibits limitations that open up avenues for further research. Most importantly, we do not account for heterogeneity among customer values. We examine a range of settings in which the return per accepted offer differ. However, the return is always the same across customers. Given that customer spending differs in many practical applications, it is important to examine customer-dependent returns in future research. Future research could also extend the proposed modeling framework. In particular, PCES is a black-box approach that does not reveal how customer characteristics influence predictions. Such insight is important to understand which factors determine customers' reactions toward marketing offers. Therefore, developing approaches that unlock the PCES black-box and clarify how variables influence predictions seems to be a fruitful avenue for future research.

Finally, our study does not consider deep learning. This may seem surprising because deep learning methods have achieved excellent results, especially when processing unstructured data in computer vision and text analysis [19]. While some corresponding studies display much relevance for marketing, for example approaches to recommend tags for images [13] or to extract consumer sentiment from written text [18], it seems fair to conclude that research on the suitability of deep learning for core marketing areas is yet scarce. In this regard, examining the suitability of deep learning for customer targeting appears an interesting avenue for future research. However, the popular deep learning architectures convolutional and recurrent neural networks are particularly suitable for processing multi-dimensional data structures such as images (multiple images each of which consists of multiple pixels each of which has multiple color channels) or texts (multiple documents each of which consists of multiple words, each of which is projected to a multi-dimensional embedding space), and appear less appropriate for the cross-sectional data we employ here and that prevails in the literature on customer targeting [e.g., [27]. For example, tabular data with two dimensions, observations and features, does not exhibit a sequential structure, which discourages application of recurrent networks. Similarly, the filtering operation in convolutional networks is not readily applicable when working with “flat tables”. In view of this, future research on deep learning-based targeting would benefit from multi-dimensional input data where the values of individual features are available over time. Until corresponding results become available, interested readers find a preliminary analysis of a deep network with our data sets in the online appendix that accompanies this paper. For these data sets, we find PCES to perform significantly and substantially better than a deep learning benchmark.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ins.2019.05.027.

References

[1] A. Bequé, K. Coussement, R. Gayler, S. Lessmann, Approaches for credit scorecard calibration: an empirical analysis, Knowl. Based Syst. 134 (2017) 213–227.
[2] S. Bhattacharyya, Direct marketing performance modeling using genetic algorithms, INFORMS J. Comput. 11 (1999) 248–257.
[3] R. Caruana, A. Munson, A. Niculescu-Mizil, Getting the most out of ensemble selection, in: Proceedings of the 6th International Conference on Data Mining (ICDM'06), Hong Kong, China, IEEE Computer Society, 2006, pp. 828–833.
[4] P.F. Christoffersen, F.X. Diebold, Optimal prediction under asymmetric loss, Econom. Theory 13 (1997) 808–817.
[5] K. Coussement, W. Buckinx, A probability-mapping algorithm for calibrating the posterior probabilities: a direct marketing application, Eur. J. Oper. Res. 214 (2011) 732–738.
[6] K. Coussement, S. Lessmann, G. Verstraeten, A comparative analysis of data preparation algorithms for customer churn prediction: a case study in the telecommunication industry, Decis. Support Syst. 95 (2017) 27–36.
[7] G. Cui, M.L. Wong, X. Wan, Targeting high value customers while under resource constraint: partial order constrained optimization with Genetic Algorithm, J. Interact. Mark. 29 (2015) 27–37.
[8] A.W. Ding, S. Li, P. Chatterjee, Learning user real-time intent for optimal dynamic web page transformation, Inf. Syst. Res. 26 (2015) 339–359.
[9] R. Elsner, M. Krafft, A. Huchzermeier, Optimizing Rhenania's direct marketing business through dynamic multilevel modeling (DMLM) in a multicatalog-brand environment, Mark. Sci. 23 (2004) 192–206.
[10] R.M. Fuller, A.R. Dennis, Does fit matter? The impact of task-technology fit and appropriation on team performance in repeated tasks, Inf. Syst. Res. 20 (2009) 2–17.
[11] S. García, A. Fernández, J. Luengo, F. Herrera, Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: experimental analysis of power, Inf. Sci. 180 (2010) 2044–2064.
[12] Y. Geng, R. Liang, W. Li, J. Wang, G. Liang, C. Xu, J. Wang, Learning convolutional neural network to maximize Pos@Top performance measure, in: Proceedings of the European Symposium on Artificial Neural Networks (ESANN'2016), 2016, pp. 589–594. 2016.
[13] Y. Geng, G. Zhang, W. Li, Y. Gu, R.-Z. Liang, G. Liang, J. Wang, Y. Wu, N. Patil, J.-Y. Wang, A novel image tag completion method based on convolutional neural transformation, in: A. Lintas, S. Rovetta, P.F.M.J. Verschure, A.E.P. Villa (Eds.), Proceedings of the International Conference on Artificial Neural Networks (ICANN'2017), Springer, Alghero, Italy, 2017, pp. 539–554.
[14] N. Glady, B. Baesens, C. Croux, Modeling churn using customer lifetime value, Eur. J. Oper. Res. 197 (2009) 402–411.

[15] N. Golrezaei, H. Nazerzadeh, P. Rusmevichientong, Real-time optimization of personalized assortments, Manag. Sci. 60 (2014) 1532–1551.
[16] C.W.J. Granger, Prediction with a generalized cost of error function, Oper. Res. Q. 20 (1969) 199–207.
[17] T. Hastie, R. Tibshirani, J.H. Friedman, The Elements of Statistical Learning, 2nd ed, Springer, New York, 2009.
[18] S. Kudugunta, E. Ferrara, Deep neural networks for bot detection, Inf. Sci. 467 (2018) 312–322.
[19] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.
[20] S. Lessmann, B. Baesens, H.-V. Seow, L.C. Thomas, Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research, Eur. J. Oper. Res. 247 (2015) 124–136.
[21] C.-T. Li, Y.-J. Lin, M.-Y. Yeh, Forecasting participants of information diffusion on social networks with its applications, Inf. Sci. 422 (2018) 432–446.
[22] M. Lichman, UCI Machine Learning Repository, University of California, School of Information and Computer Science, Irvine, CA, 2013 http://archive.ics.uci.edu/ml/.
[23] G.L. Lilien, Bridging the academic–practitioner divide in marketing decision models, J. Mark. 75 (2011) 196–210.
[24] S. Maldonado, C. Bravo, J. López, J. Pérez, Integrated framework for profit-based feature selection and SVM classification in credit scoring, Decis. Support Syst. 104 (2017) 113–121.
[25] S. Maldonado, J. Pérez, C. Bravo, Cost-based feature selection for Support Vector Machines: an application in credit scoring, Eur. J. Oper. Res. 261 (2017) 656–665.
[26] D. Martens, F. Provost, Pseudo-social network targeting from consumer transaction data, in, NYU Working Paper No. CEDER-11-05, 2011.
[27] D. Martens, F. Provost, J. Clark, E.J.d. Fortuny, Mining massive fine-grained behavior data to improve predictive analytics, MIS Q. 40 (2016) 869–888.
[28] S. Mitrović, B. Baesens, W. Lemahieu, J. De Weerdt, On the operational efficiency of different feature types for telco Churn prediction, Eur. J. Oper. Res. 267 (2018) 1141–1155.
[29] S.A. Neslin, S. Gupta, W. Kamakura, J. Lu, C.H. Mason, Defection detection: measuring and understanding the predictive accuracy of customer churn models, J. Mark. Res. 43 (2006) 204–211.
[30] C. Perlich, B. Dalessandro, T. Raeder, O. Stitelman, F. Provost, Machine learning for targeted display advertising: transfer learning in action, Mach. Learn. 95 (2014) 103–127.
[31] J.C. Platt, Probabilities for support vector machines, in: A. Smola, P. Bartlett, B. Schölkopf, D. Schuurmans (Eds.), Advances in Large Margin Classifiers, MIT Press, Cambridge, 2000, pp. 61–74.
[32] L. Rokach, L. Naamani, A. Shmilovici, Pessimistic cost-sensitive active learning of decision trees for profit maximizing targeting campaigns, Data Min. Knowl. Discov. 17 (2008) 283–316.
[33] G. Shmueli, O.R. Koppius, Predictive analytics in information systems research, MIS Q. 35 (2011) 553–572.
[34] E. Stripling, S. vanden Broucke, K. Antonio, B. Baesens, M. Snoeck, Profit maximizing logistic model for customer churn prediction using genetic algorithms, Swarm Evol. Comput. 40 (2018) 116–130.
[35] P. Tambe, Big data investment, skills, and firm value, Manag. Sci. 60 (2014) 1452–1469.
[36] V. Vapnik, S. Kotz, Estimation of Dependences Based on Empirical Data, 2 ed, Springer, New York, 2006.
[37] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, B. Baesens, New insights into churn prediction in the telecommunication sector: a profit driven data mining approach, Eur. J. Oper. Res. 218 (2012) 211–229.
[38] T. Verbraken, C. Bravo, R. Weber, B. Baesens, Development and application of consumer credit scoring models using profit-based classification measures, Eur. J. Oper. Res. 238 (2014) 505–513.
[39] T. Verbraken, W. Verbeke, B. Baesens, A novel profit maximizing metric for measuring classification performance of customer churn prediction models, IEEE Trans. Knowl. Data Eng. 25 (2012) 961–973.
[40] V.V. Vlasselaer, T. Eliassi-Rad, L. Akoglu, M. Snoeck, B. Baesens, GOTCHA! Network-based fraud detection for social security fraud, Manag. Sci. 63 (2017) 3090–3110.
[41] S. Wang, Y. Li, Y. Shao, C. Cattani, Y. Zhang, S. Du, Detection of dendritic spines using wavelet packet entropy and fuzzy support vector machine, CNS Neurol. Disord. Drug Targets 16 (2017) 116–121.
[42] H. Zhao, X. Li, A cost sensitive decision tree algorithm based on weighted class distribution with batch deleting attribute mechanism, Inf. Sci. 378 (2017) 303–316.
[43] B. Zhu, B. Baesens, S.K.L.M. vanden Broucke, An empirical comparison of techniques for the class imbalance problem in churn prediction, Inf. Sci. 408 (2017) 84–99.

Please cite this article as: S. Lessmann, J. Haupt and K. Coussement et al., Targeting customers for profit: An ensemble learning framework to support marketing decision-making, Information Sciences, https://doi.org/10.1016/j.ins.2019.05.027