Targeting customers for profit:

An ensemble learning framework to support marketing decision making

Stefan Lessmann[*], Kristof Coussement[2], Koen W. De Bock[3] & Johannes Haupt[1]

[1]School of Business and Economics, Humboldt-University of Berlin,

Unter den Linden 6, D-10099 Berlin, Germany

[2]IÉSEG Center for Marketing Analytics (ICMA), IÉSEG School of Management –

Université Catholique de Lille (LEM, UMR CNRS 9221),

Department of Marketing, 3 Rue de la Digue, F-59000, Lille, France

[3]Audencia Business School, 8 Route de la Jonelière, F-44312 Nantes, France

Abstract

Marketing messages are most effective if they reach the right customers. Deciding which customers to contact is thus 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. The results of a comprehensive empirical study confirm the business value of the proposed approach in that it recommends substantially more profitable target groups than several benchmarks.

Keywords: Marketing Decision Support, Business Value, Profit-Analytics, Machine Learning

[*] Corresponding author; e-mail: stefan.lessmann@hu-berlin.de, tel.:+49.30.2093.5742

1 Introduction

Big data analytics revolutionizes the face of decision support (e.g., Gupta & George, 2016). Skepticism toward formal decision aids used to be widespread among managers (Lilien, 2011). Today, however, we witness an unprecedented interest in quantitative decision support models. Vast amounts of data, powerful pattern extraction algorithms, and easy to use software systems fuel this development and promise to improve management support. For example, based on a survey among firm executives, Germann et al. (2013) estimate that increasing marketing analytics deployment is associated with an average eight percent increase in return on assets. In a similar way, Tambe (2014) finds the use of big data technologies to be associated with an average one to three percent increase of firm productivity.

The paper concentrates on marketing decisions in campaign planning. Campaign planners need to answer three questions (Elsner et al., 2004): 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 (e.g., Phan & Vogel, 2010) and churn management (e.g., Coussement & Van den Poel, 2008) literature. To target marketing offers, companies use response models, which estimate acceptance probabilities for individual customers. This facilitates soliciting the most likely responders. Response models use a variety of prediction methods including, artificial neural networks (e.g., Olson & Chae, 2012), support vector machines (e.g., Chen et al., 2015), or tree-based approaches (e.g., Lemmens & Croux, 2006).

Prediction methods are designed for generality and solve modeling problems in various domains (e.g., Bose & Mahapatra, 2001). We argue that using an off-the-shelf method for customer targeting suffers a limitation in that 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 little effect on the estimation of the targeting model. Therefore, the objective of the paper is to develop and test a modeling framework that accounts for business objectives during the development of a targeting model. Current trends in marketing support this objective. In particular, marketing communication is increasingly personalized (e.g., Golrezaei et al., 2014) and distributed through digital channels (e.g., Ding et al., 2015). Personalization amplifies the scale of targeting decisions while digitalization often requires real-time decision making. In this regard, both trends

illustrate the need to automate customer targeting. A high recognition of business goals during model development seems especially important when targeting models operate in a self-governed manner.

The paper contributes to the literature in three ways. First, we make a methodological contribution. Relying on the principles of ensemble learning, we propose a paradigm to develop predictive marketing support models, which we call profit-conscious ensemble selection (PCES). PCES differs from previous approaches in that it integrates established principles of statistical inference with business objectives in customer targeting. We hypothesize that the explicit consideration of marketing goals at an early stage in the modeling process improves the quality of targeting decisions. Second, we perform a comprehensive empirical study including twenty-five real-world marketing data sets from different industries to test the effectiveness of PCES. In addition to comparing several targeting models, an important feature of the experiment is that it contrasts paradigms toward model development; namely: i) "profit-agnostic" models derived from minimizing statistical loss, ii) "profit-centered" models derived from maximizing business performance, and iii) an integrated approach in the form of PCES that balances statistical and economic considerations. This setup provides novel insight concerning the relative merits of fundamentally different approaches toward predictive modeling. Third, we clarify the degree to which introducing profit considerations into model development improves business performance and decision quality. We achieve this through estimating the campaign profit that emerges from model-based targeting and the marginal profit of PCES-based targeting, respectively. This provides a clear, managerially meaningful measure of the value of PCES.

The reminder of the paper is organized as follows: Section 2 reviews related literature. The proposed targeting methodology is developed in Section 3. Section 4 and 5 elaborate on the design and results of the empirical evaluation of PCES, respectively. Section 6 concludes the paper.

2 Background and related work

A large body of literature examines the antecedents of (model-based) decision support system (DSS) effectiveness (e.g., Lilien et al., 2004). Several studies highlight the importance of the DSS exhibiting high fit for the decision task (e.g., Dennis et al., 2001). However, the effect of fit depends on the (post)processing of DSS recommendations. More specifically, Fuller and Dennis (2009) demonstrate how managers learn to mitigate a lack of DSS fit and achieve performance similar to managers who

have access to better technology (i.e., higher fit). This is reasonable since managers' decision-making is guided by a mental model that enables them to appraise DSS outputs in awareness of a specific problem context, connect this output to decision quality, and, in this way, correct for misleading information from a poor decision support model (Fuller & Dennis, 2009; Lilien, 2011). This theory indicates the value of human supervision in model-based decision support. However, disadvantages of such "model-manager-tandem" include high labor costs, a possible lack of expertize, especially related to big data technologies (e.g., Manyika et al., 2011), and high latency in decision-making. PCES strives to combine the efficiency of fully automated, model-based decision-making and the ability of managers to use contextual, task-specific information to improve decision quality in targeting applications.

Data-driven prediction models are widely used to forecast customer responses to marketing campaigns (e.g., Bose & Mahapatra, 2001; Chen et al., 2015; Olson & Chae, 2012). Requiring little human intervention, they also appear well prepared to automate decision-making in real-time targeting applications such as online advertising or social media (e.g., Ballings & Van den Poel, 2015; Fan & Yan, 2015; Perlich et al., 2014). Prior work also studies the question whether the development of predictive decision support models should account for business objectives. In the forecasting literature, Granger (1969) was the first to criticize the use of quadratic loss functions for model estimation. Arguing that real-world applications rarely exhibit symmetric error costs, he proposed loss functions that penalize positive and negative residuals differently. Subsequent studies further elaborate on Granger's work and contribute theoretical as well as empirical insights (e.g., Christoffersen & Diebold, 1997; Leitch & Tanner, 1991). PCES also employs non-standard loss functions for the development of predictive models and assesses models in terms of business performance. The main differences lie in the methodology and application. We focus on multivariate machine learning models as opposed to univariate time series forecasting models and examine decision problems in marketing campaign planning. This also implies that we study a different business objective (i.e., campaign profit).

The cost-sensitive learning literature also studies asymmetric error costs. In general, cost-sensitive learning encompasses methods that operate at the data level, for example by altering the distribution between classes with higher/lower misclassification costs (e.g., Domingos, 1999) and algorithmic adaptations to make standard learners cost-aware (e.g., Žliobaitė et al., 2015). This paper also considers

class-dependent misclassification costs. 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. However, studies in cost-sensitive learning aim at generality and strive to develop modeling approaches that perform well across a variety of applications where misclassification costs differ. While generality is a goal worth pursuing, a DSS approach that focuses on a concrete application has the potential to better reflect its specific requirements. PCES is such an approach for decisions in the scope of targeted marketing. Marketing campaigns typically target 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.

There is also a large body of literature on predictive models for customer targeting. In general, previous work has studied all steps of the predictive modeling process (see Figure 1) from building an analytic database through gathering data from past campaigns and test mailings (e.g., Rokach et al., 2008) over data preparation including target variable definition (Bodapati & Gupta, 2004; Glady et al., 2009), independent variable development, encoding, and selection (e.g., Coussement et al., 2017), model estimation and tuning (e.g. Chen et al., 2015) to prediction post-processing (e.g., Coussement & Buckinx, 2011), performance evaluation (e.g., Verbraken et al., 2012) and decision-making (e.g., Schröder & Hruschka, 2016). However, the vast majority of previous studies estimate the targeting model using standard prediction methods (neural networks, support vector machines, random forest, etc.). We call this approach profit-agnostic because it does not take account of the actual business problem – campaign profit maximization – during model development.

Some studies emphasize the inability of statistical accuracy indicators (NLL, percentage correctly classified, etc.) to reflect marketing objectives and propose alternatives for specific applications such as the (expected) maximum profit criterion for churn modeling (Verbeke et al., 2012; Verbraken et al., 2012). We further extend this research in two ways. First, using a more general profit function, we consider not only churn modeling but a broad range of targeting applications. Second, focusing on profit-oriented model development, we introduce the business goal in an earlier modeling step where corresponding information can exert more influence on the eventual model. To confirm this, we empirically compare PCES to the approach of Verbeke et al. (2012).

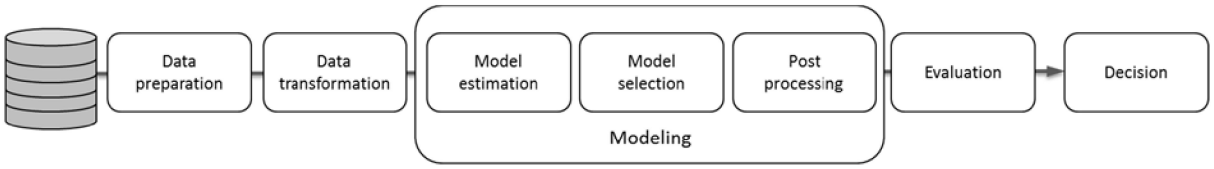

Figure 1: Predictive modeling process[2]

To our knowledge, only two studies consider a profit-oriented model development. Using a genetic algorithm (GA), Bhattacharyya (1999) estimates the parameters of a linear model so as to maximize profit. Cui et al. (2015) 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 approach of Bhattacharyya (1999).

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., Hastie et al., 2009). Assume a marketer wishes to predict the behavior of customer i, characterized by vector $\boldsymbol{x}_i = (x_{1i}, x_{2i}, \dots, x_{Mi}) \in \mathbb{R}^M$, 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). In the former case, the prediction task is a regression problem; and a classification problem otherwise. In direct marketing, modeling discrete responses decreases bias due to incorrect model specification and may thus increase prediction accuracy (Bodapati & Gupta, 2004). Accordingly, we focus on binary classification where $y_i \in \{0,1\}$ with a value of y_i=1 (y_i=0) indicating that customer i accepted (rejected) a marketing offer. A

[2] Figure 1 grounds on process models for data analysis (e.g., Li et al., 2016) and magnifies the modeling step so as to highlight tasks in predictive analytics, including the estimation of a model, the tuning of its meta-parameters, and potentially a post-processing of predictions (e.g., Coussement & Buckinx, 2011). Although not explicitly highlighted, we acknowledge that a modeling process may exhibit feedback loops.

targeting model, $f(\boldsymbol{x})$, represents a functional mapping from customer records to responses: $f_{\boldsymbol{\Lambda}}(\boldsymbol{x}): \mathbb{R}^M \longmapsto \{0,1\}$, where $\boldsymbol{\Lambda}$ denotes a vector of model parameters. Model estimation involves fitting model parameters to data. Afterwards, the specified model facilitates predicting $\boldsymbol{y}$ given $\boldsymbol{x}$. In other words, the model allows the marketer to predict customer response (and more generally behavior) from observable customer data (summarized in $\boldsymbol{x}$).

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: $\boldsymbol{\Lambda}' \leftarrow \min_{\boldsymbol{\Lambda}} Q\big(y_i, f_{\boldsymbol{\Lambda}}(\boldsymbol{x}_i)\big) \quad \forall\, i = 1, \ldots, N$, 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. To illustrate this, consider a marker who wishes to predict customers' responses to a marketing message using the well-known logit model. She estimates the model through minimizing Q, which in the case of logit is the negative log-likelihood (NLL), over a sample of observations from previous campaigns (i.e., D).

A loss function represents a model-internal notion of fit. In the previous example, a lower NLL indicates that a model fits the data more accurately. 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 (e.g., Vapnik & Kotz, 2006). 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 performance should not be taken for granted. For example, maximizing fit using some statistical loss function during model development may lead to a different model compared to maximizing campaign profit. On the other hand, statistical loss functions others have strong theoretical underpinnings and exhibit desirable properties related to generalization and the accuracy of model prediction in particular (e.g., Hastie et al., 2009). It is imperative to build on this theory when developing a prediction model to prevent overfitting (e.g., Vapnik & Kotz, 2006). This motivates our PCES approach to integrate statistical considerations (in the

form of established loss functions and estimation principles) and business objectives in campaign planning (in the form of campaign profit) during the development of a targeting model.

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 are manifold and include the mail-order industry, churn management, and cross-selling (e.g., Blattberg et al., 2008). Recently, targeting models are increasingly used in real-time settings, for example to increase purchase probabilities in e-shops through personalization (e.g., Golrezaei et al., 2014) or to guide decisions in online marketing (e.g., Xu et al., 2014).

From a managerial point of view, the business value of a targeting model depends on the degree to which it increases the profitability of a marketing campaign. More specifically, campaign profitability represents a short-term business goal. A short-term perspective may be considered problematic in that it disregards the interdependencies of different campaigns (e.g., Bleier & Eisenbeiss, 2015; Schröder & Hruschka, 2016). However, a short term perspective that concentrates on campaign profit is suitable in this paper, which concerns operational decision support in reoccurring, routine tasks and real-time applications with potential/need for decision automation. Therefore, we appraise the business value of a targeting model in terms of the overall revenue from the specific target group that the model recommends minus the total cost of solicitation. More formally, we model campaign profit, Ω, as (Martens & Provost, 2011; Piatetsky-Shapiro & Masand, 1999):

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

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:

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

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. For example, Neslin et al. (2006) devise a profit function for models that target retention actions to customers with high churn probability. The expected maximum profit criterion further refines this approach (Verbraken et al., 2012). 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., Hernández-Orallo et al., 2011; Rokach et al., 2008; Verbeke et al., 2012). Second, calculating campaign profit using the mean revenue per accepted offer may be more suitable for predictive modeling, for example because information to reliably estimate revenues at the customer level is lacking. Last, some applications do not require distinguishing revenues across customers, for example when targeting services like study programs that entail a fixed fee or 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 Figure 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 (Fuller & Dennis, 2009).

The proposed framework is based on a machine learning paradigm called ensemble selection (e.g., Caruana et al., 2006; Partalas et al., 2010; Woźniak et al., 2014). An ensemble is a collection of (base) models, all of which predict the same target. Combining multiple models in an ensemble is useful to increase predictive accuracy (e.g., Malthouse & Derenthal, 2008). 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 is similar to Caruana's et al. (2006) 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. The following subsections elaborate on this design.

3.3.1 Model library

The success of an ensemble depends on the diversity of its members (e.g., Partalas et al., 2010). To obtain a library of diverse models, we use different learning algorithms. In addition, we consider multiple settings for the meta-parameters of individual algorithms. Meta-parameters such as the number of hidden nodes in a neural network (e.g., Fletcher & Goss, 1993) facilitate adapting a learning algorithm to a particular task (Hastie et al., 2009). This suggests that prediction models from the same algorithm vary with meta-parameters and thus display diversity.

Table 1 summarizes the learning algorithms and meta-parameter settings in the model library. The selection is based upon previous literature on customer targeting and ensemble modeling (Caruana et al., 2006; Lessmann et al., 2015; Verbeke et al., 2012). Some methods have been chosen due to their popularity in academia and industry (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 Hastie et al. (2009). In total, we consider 15 learning algorithms from which we derive 877 different models.

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	$2^{[-14, -13, \ldots, 14]}$
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 Width of Rbf kernel function	$2^{[-12, -11, \ldots, 12]}$ $2^{[-12, -11, \ldots, -1]}$
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	$[0.1, 0.5, 1, 2, 4] \cdot \sqrt{M}$ ***
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.

3.3.2 Candidate selection

Given the model library, we select candidate models using directed hill-climbing (Caruana et al., 2006). 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 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[3]).

It is common practice to use statistical loss functions for ensemble member selection (e.g., Caruana et al., 2006; Partalas et al., 2010; Woźniak et al., 2014). 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 particular, we propose to maximize (1) instead of a statistical loss function during candidate selection. This way, we devise an ensemble that incorporates business objectives during model development. Specifically, PCES refines the first-stage predictions, which stem from well-established prediction models and embody the principles of statistical learning, through selective combination so as to better represent the actual decision problem. This ex-post revision of (individual model) predictions mimics the way in which managers use DSS recommendations and possibly correct for misleading advice (Fuller & Dennis, 2009).

3.3.3 Model aggregation

Model aggregation refers to a combination of models' predictions. This occurs during candidate selection and when computing the final ensemble prediction. We pool models by averaging over their predictions. Effectively, we compute a weighted average. This is because the candidate selection procedure of Caruana et al. (2006) 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

[3] Available online at https://bit.ly/pces_appendix.

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 (Platt, 2000).

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 (Asuncion & Newman, 2010), 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.

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 (Caruana et al., 2006).

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	(Verbeke et al., 2012)	40,000	70	0.50
D5	Churn prediction	Telco	(Verbeke et al., 2012)	93,893	196	0.50
D6	Churn prediction	Telco	(Verbeke et al., 2012)	12,410	18	0.39
D7	Churn prediction	Telco	(Verbeke et al., 2012)	69,309	67	0.29
D8	Churn prediction	Telco	(Verbeke et al., 2012)	21,143	384	0.12
D9	Churn prediction	Telco	KDD09	50,000	301	0.07
D10	Churn prediction	Telco	(Verbeke et al., 2012)	47,761	41	0.04
D11	Churn prediction	Telco	(Verbeke et al., 2012)	5,000	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	9,822	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[4] (the number gives the year of the competition); KDD = ACM KDD Cup[5] (the number gives the year of the competition); UCI-xxx = UCI Machine Learning Repository[6] (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).

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 (e.g., Cui et al., 2006), ii) random forest, due to its success in previous benchmarking studies (e.g., Lessmann et al., 2015; Verbeke et al., 2012), 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

[4] http://www.data-mining-cup.com
[5] http://www.sigkdd.org/kddcup/index.php
[6] http://archive.ics.uci.edu/ml/

campaign profit (1). Prior work finds a selection of prediction models using business performance measures to substantially improve decision quality (e.g., Glady et al., 2009; Verbeke et al., 2012; Verbraken et al., 2014). Therefore, we expect the benchmarks to be challenging.

The ensemble selection approach of Caruana et al. (2006) 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 the 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 (1999). 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. To implement the GA benchmark, we reuse the settings of Bhattacharyya (1999) and set the population size, crossover rate, and mutation rate to 50, 0.7, 0.2, respectively.

4.3 Configuration of ensemble selection

Caruana et al. (2006) propose some modifications of basic ensemble selection. One extension consists of an additional bagging step. More specifically, instead of selecting a single set of base models from the full model library, Caruana et al. (2006) subsample the library, select one ensemble from each subsample, and average over the resulting ensembles. 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 percent 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 (e.g., Leitch & Tanner, 1991) 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. (2012) recommend this approach and 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. Given that larger values of r give an incentive to increase campaign size, we constrain the optimization of (1) such that $\tau \leq 0.5$. Given that marketing campaigns typically target a small fraction of responsive customers (e.g., Blattberg et al., 2008), 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 (García et al., 2010). The last row of Table 3 reports the adjusted p-values.

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 (García et al., 2010). 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).

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., Glady et al., 2009; Verbeke et al., 2012; Verbraken et al., 2012). 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 percent). Before 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. (2010) 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: Comparison of Campaign Profit at Model-Optimized Campaign Sizes

	Campaign profit [\$]					
Data	**Logit**	**RF**	**BBM**	**GA**	**PAES**	**PCES**
D1	1,660	1,596	1,764	1,532	**1,874**	1,846
D2	61,612	75,816	75,989	62,953	75,725	**76,001**
D3	-2	-83	88	-104	76	**137**
D4	-2,992	-2,832	-2,832	-3,052	-2,852	**26**
D5	-7,096	-6,766	-6,766	-7,096	-6,666	**25**
D6	-1,017	-997	-977	-1,027	-997	**159**
D7	35,578	39,598	39,778	35,098	40,408	**40,618**
D8	2,966	2,926	3,270	2,756	**3,404**	3,121
D9	699	469	862	509	999	**1,139**
D10	442	876	839	590	901	**984**

D11	1,491	2,000	2,022	1,534	2,020	**2,058**
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	9,121	9,283	9,690	9,568	**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	9,954	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., 2010). 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.*

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., Lessmann et al., 2015; Verbeke et al., 2012). Profit increases of five percent above are managerially meaningful, especially for larger companies and run many campaigns (Neslin et al., 2006). 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 percent 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 selects τ close to zero. Finally, Table 5 evidences a trend of PCES to recommend smaller campaigns. The median value τ=16.66 for PCES is much less than the second-smallest value of τ=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.

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

The results of Table 4 and Table 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 (García et al., 2010). Figure 2 depicts the corresponding results. Given that smaller settings of r lead to large improvements over weaker benchmarks, we split Figure 2 into two panels which show results for all settings of r and those above five, respectively.

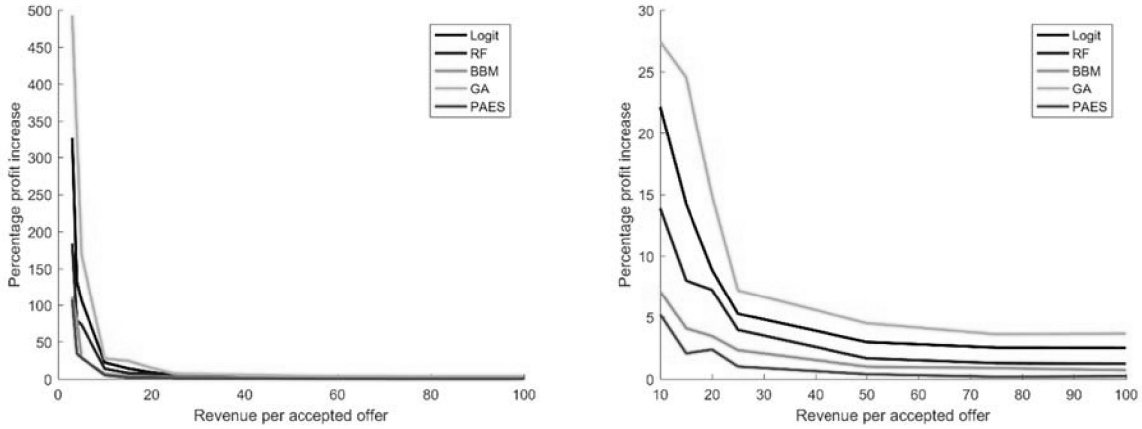

Figure 2: Expected percentage improvement in campaign profit due to using PCES for target group selection. We estimate profit contrasts in the same way as in Table 4. Panel a) shows all settings of r, whereas panel b) focuses on settings of r>5 for better readability.

Figure 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 Figure 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. In addition, the study sheds some light on the divergence between the optimization of statistical loss functions and business objectives for prediction model development in targeting applications. The experimental design includes three philosophies toward model development: a direct maximization of business performance (GA), a model selection approach, which introduces business objectives ex-post and develops models using statistical learning (Logit, RF, BBM and PAES), and PCES that shifts the consideration of the actual business objective to a previous modeling stage so as to gear model development toward the ultimate goal of the marketing campaign and achieve higher fit between the final model and the business task which it is meant to support.

Observed results suggest the direct approach to be least effective. In fact, a simple logit model consistently outperforms GA. The logit and GA model both construct a linear classifier. Better performance of the former evidence that model development through minimizing a statistical loss function is preferable to a direct maximization of business performance. Well-known estimation problems such as overfitting (e.g., Hastie et al., 2009) are a likely cause of this result. Remedies to such problems are readily available in statistical learning. However, developing predictive decision support 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 indicate 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 another model tuned to minimize NLL or some other 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 modelling, 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., Shmueli & Koppius, 2011). 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., Vapnik & Kotz, 2006). Consequently, the experimental design ensures 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[7]) 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 (Bequé et al., **2017**). TDL is related to AUC but focuses on ranking performance among of subset of customers (e.g., Neslin et al., 2006). 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

[7] Available online at https://bit.ly/pces_appendix.

estimation as in Table 4 (García et al., 2010). The e-companion provides a more detailed analysis of AUC, TDL performance in Appendix II[8], and campaign profit with budget constraint in Appendix III[8].

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. (2010).

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 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

[8] Available online at https://bit.ly/pces_appendix.

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 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.

References

Asuncion, A., & Newman, D. J. (2010). UCI Machine Learning Repository Retrieved 2009-09-02, from http://archive.ics.uci.edu/ml/

Ballings, M., & Van den Poel, D. (2015). CRM in social media: Predicting increases in Facebook usage frequency. *European Journal of Operational Research,* 244(1), 248-260.

Bequé, A., Coussement, K., Gayler, R., & Lessmann, S. (**2017**). Approaches for Credit Scorecard Calibration: An Empirical Analysis. *Knowledge-Based Systems,* 134(15), 213-227.
Bhattacharyya, S. (1999). Direct marketing performance modeling using genetic algorithms. *INFORMS Journal on Computing,* 11(3), 248-257.
Blattberg, R. C., Neslin, S. A., & Kim, B.-D. (2008). *Database Marketing: Analyzing and Managing Customers*. New York: Springer.
Bleier, A., & Eisenbeiss, M. (2015). Personalized online advertising effectiveness: The interplay of what, when, and where. *Marketing Science,* 34(5), 669-688.
Bodapati, A., & Gupta, S. (2004). A direct approach to predicting discretized response in target marketing. *Journal of Marketing Research,* 41(1), 73-85.
Bose, I., & Mahapatra, R. K. (2001). Business data mining - a machine learning perspective. *Information & Management,* 39(3), 211-225.
Caruana, R., Munson, A., & Niculescu-Mizil, A. (2006). *Getting the Most Out of Ensemble Selection*. Proc. of the 6th Intern. Conf. on Data Mining, IEEE Computer Society: Los Alamitos. pp. 828-833.
Chen, Z.-Y., Fan, Z.-P., & Sun, M. (2015). Behavior-aware user response modeling in social media: Learning from diverse heterogeneous data. *European Journal of Operational Research,* 241(2), 422-434.
Christoffersen, P. F., & Diebold, F. X. (1997). Optimal prediction under asymmetric loss. *Econometric Theory,* 13(6), 808-817.
Coussement, K., & Buckinx, W. (2011). A probability-mapping algorithm for calibrating the posterior probabilities: A direct marketing application. *European Journal of Operational Research,* 214(3), 732-738.
Coussement, K., Lessmann, S., & Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. *Decision Support Systems,* 95, 27-36.
Coussement, K., & Van den Poel, D. (2008). Integrating the voice of customers through call center emails into a decision support system for churn prediction. *Information & Management,* 45(3), 164-174.
Cui, G., Wong, M. L., & Lui, H.-K. (2006). Machine learning for direct marketing response models: Bayesian networks with evolutionary programming. *Management Sciences,* 52(4), 597-612.
Cui, G., Wong, M. L., & Wan, X. (2015). Targeting High Value Customers While Under Resource Constraint: Partial Order Constrained Optimization with Genetic Algorithm. *Journal of Interactive Marketing,* 29, 27-37.
Dennis, A. R., Wixom, B. H., & Vandenberg, R. J. (2001). Understanding fit and appropriation effects in group support systems via meta-analysis. *MIS Quarterly,* 25(2), 167-193.
Ding, A. W., Li, S., & Chatterjee, P. (2015). Learning user real-time intent for optimal dynamic web page transformation. *Information Systems Research,* 26(2), 339-359.
Domingos, P. (1999). *MetaCost: A General Method for Making Classifiers Cost-Sensitive.* In U. M. Fayyad, S. Chaudhuri & D. Madigan (Eds.). Proc. of the 5th Intern. Conf. on Knowledge Discovery and Data Mining, ACM Press. pp. 155-164.
Elsner, R., Krafft, M., & Huchzermeier, A. (2004). Optimizing Rhenania's direct marketing business through dynamic multilevel modeling (DMLM) in a multicatalog-brand environment. *Marketing Science,* 23(2), 192 - 206.
Fan, W., & Yan, X. (2015). Novel applications of social media analytics. *Information & Management,* 52(7), 761-763.
Fletcher, D., & Goss, E. (1993). Forecasting with neural networks. *Information & Management,* 24(3), 159-167.
Fuller, R. M., & Dennis, A. R. (2009). Does fit matter? The impact of task-technology fit and appropriation on team performance in repeated tasks. *Information Systems Research,* 20(1), 2-17.
García, S., Fernández, A., Luengo, J., & Herrera, F. (2010). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences,* 180(10), 2044-2064.
Germann, F., Lilien, G. L., & Rangaswamy, A. (2013). Performance implications of deploying marketing analytics. *International Journal of Research in Marketing,* 30(2), 114-128.
Glady, N., Baesens, B., & Croux, C. (2009). Modeling churn using customer lifetime value. *European Journal of Operational Research,* 197(1), 402-411.
Golrezaei, N., Nazerzadeh, H., & Rusmevichientong, P. (2014). Real-time optimization of personalized assortments. *Management Science,* 60(6), 1532-1551.
Granger, C. W. J. (1969). Prediction with a generalized cost of error function. *Operational Research Quarterly,* 20(2), 199-207.
Gupta, M., & George, J. F. (2016). Toward the development of a big data analytics capability. *Information & Management,* 53(8), 1049-1064.
Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The Elements of Statistical Learning* (2nd ed.). New York: Springer.

Hernández-Orallo, J., Flach, P. A., & Ramirez, C. F. (2011). *Brier Curves: A New Cost-Based Visualisation of Classifier Performance.* In L. Getoor & T. Scheffer (Eds.). Proc. of the 28th Intern. Conf. on Machine Learning, Omnipress: Madison. pp. 585-592.

Leitch, G., & Tanner, J. E. (1991). Economic forecast evaluation: Profits versus the conventional error measures. *The American Economic Review,* 81(3), 580-590.

Lemmens, A., & Croux, C. (2006). Bagging and boosting classification trees to predict churn. *Journal of Marketing Research,* 43(2), 276-286.

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research,* 247(1), 124-136.

Li, Y., Thomas, M. A., & Osei-Bryson, K.-M. (2016). A snail shell process model for knowledge discovery via data analytics. *Decision Support Systems,* 91, 1-12.

Lilien, G. L. (2011). Bridging the academic–practitioner divide in marketing decision models. *Journal of Marketing,* 75(4), 196-210.

Lilien, G. L., Rangaswamy, A., Van Bruggen, G. H., & Starke, K. (2004). DSS effectiveness in marketing resource allocation decisions: Reality vs. perception. *Information Systems Research,* 15(3), 216-235.

Malthouse, E. C., & Derenthal, K. M. (2008). Improving predictive scoring models through model aggregation. *Journal of Interactive Marketing,* 22(3), 51-68.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big Data: The Next Frontier for Innovation, Competition, and Productivity: McKinsey Global Institute.

Martens, D., & Provost, F. (2011). Pseudo-Social Network Targeting from Consumer Transaction Data: Faculty of Applied Economics, University of Antwerp.

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research,* 43(2), 204-211.

Olson, D. L., & Chae, B. (2012). Direct marketing decision support through predictive customer response modeling. *Decision Support Systems,* 54(1), 443-451.

Partalas, I., Tsoumakas, G., & Vlahavas, I. (2010). An ensemble uncertainty aware measure for directed hill climbing ensemble pruning. *Machine Learning,* 81(3), 257-282.

Perlich, C., Dalessandro, B., Raeder, T., Stitelman, O., & Provost, F. (2014). Machine learning for targeted display advertising: transfer learning in action. *Machine Learning,* 95(1), 103-127.

Phan, D. D., & Vogel, D. R. (2010). A model of customer relationship management and business intelligence systems for catalogue and online retailers. *Information & Management,* 47(2), 69-77.

Piatetsky-Shapiro, G., & Masand, B. (1999). *Estimating Campaign Benefits and Modeling Lift.* In S. Chaudhuri & D. Madigan (Eds.). Proc. of the 5th Intern. Conf. on Knowledge Discovery and Data Mining, ACM Press. pp. 185-193.

Platt, J. C. (2000). Probabilities for Support Vector Machines. In A. Smola, P. Bartlett, B. Schölkopf & D. Schuurmans (Eds.), *Advances in Large Margin Classifiers* (pp. 61-74). Cambridge: MIT Press.

Rokach, L., Naamani, L., & Shmilovici, A. (2008). Pessimistic cost-sensitive active learning of decision trees for profit maximizing targeting campaigns. *Data Mining and Knowledge Discovery,* 17(2), 283-316.

Schröder, N., & Hruschka, H. (2016). Investigating the effects of mailing variables and endogeneity on mailing decisions. *European Journal of Operational Research,* 250(2), 579-589.

Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS Quarterly,* 35(3), 553-572.

Tambe, P. (2014). Big data investment, skills, and firm value. *Management Science,* 60(6), 1452-1469.

Vapnik, V., & Kotz, S. (2006). *Estimation of Dependences Based on Empirical Data* (2 ed.). New York: Springer.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research,* 218(1), 211-229.

Verbraken, T., Bravo, C., Weber, R., & Baesens, B. (2014). Development and application of consumer credit scoring models using profit-based classification measures. *European Journal of Operational Research,* 238(2), 505-513.

Verbraken, T., Verbeke, W., & Baesens, B. (2012). A novel profit maximizing metric for measuring classification performance of customer churn prediction models. *IEEE Transactions on Knowledge and Data Engineering,* 25(5), 961-973.

Woźniak, M., Graña, M., & Corchado, E. (2014). A survey of multiple classifier systems as hybrid systems. *Information Fusion,* 16, 3-17.

Xu, L., Duan, J. A., & Whinston, A. (2014). Path to purchase: A mutually exciting point process model for online advertising and conversion. *Management Science,* 60(6), 1392-1412.

Žliobaitė, I., Budka, M., & Stahl, F. (2015). Towards cost-sensitive adaptation: When is it worth updating your predictive model? *Neurocomputing,* 150, Part A, 240-249.