Document de travail du LEM
2013-23

MAXIMIZE WHAT MATTERS: PREDICTING CUSTOMER CHURN WITH DECISION-CENTRIC ENSEMBLE SELECTION

Stefan LESSMANN
University of Hamburg

Kristof COUSSEMENT
IESEG School of Management (LEM-CNRS)

Koen W. DE BOCK
IESEG School of Management (LEM-CNRS)

Maximize What Matters:

Predicting Customer Churn With Decision-Centric Ensemble Selection

Stefan Lessmann[$]

$ University of Hamburg,
Institute of Information Systems, D-20146 Hamburg, Germany

Kristof Coussement[¤] - Koen W. De Bock[¤]

¤IESEG School of Management (LEM-CNRS) – Lille Catholic University, F-59000, Lille (France).

Abstract

Churn modeling is important to sustain profitable customer relationships in saturated consumer markets. A churn model predicts the likelihood of customer defection. This helps to target retention offers to the right customers and use marketing resources efficiently. Several statistical prediction methods exist in marketing, but all these suffer an important limitation: they do not allow the analyst to account for campaign planning objectives and constraints during model building. Our key proposition is that creating churn models in awareness of actual business requirements increases the performance of the final model for marketing decision support. To demonstrate this, we propose a decision-centric framework to create churn models. We test our modeling framework on eight real-life churn data sets and find that it performs significantly better than state-of-the-art churn models. We estimate that our approach increases the per customer profits of retention campaigns by $.47 on average. Further analysis confirms that this improvement comes directly from maximizing business objectives during model building. The main implication of our study is thus that companies better shift from a purely statistical to a more business-driven modeling approach when predicting customer churn.

Keywords: Churn Prediction, Marketing Decision Support, Choice Modeling, Ensemble Selection

1 Introduction

Today, managers are more than ever interested to build enduring customer relationships and to maximize the value of the customer base (e.g., Fader and Hardie 2010). Acquiring new customers in saturated markets is challenging, and more expensive than retaining existing customers (e.g., Bhattacharya 1998; Colgate and Danaher 2000). Moreover, long-term customers generate higher profits, are less sensitive to competitive actions, and may act as promoters through positive word of mouth (e.g., Ganesh et al. 2000; Reichheld 1996; Zeithaml et al. 1996). Losing customers also creates opportunity costs because of lost sales (Rust and Zahorik 1993). Finally, retention is closely coupled with important managerial metrics such as the value of the customer base (e.g., Gupta and Zeithaml 2006).

Although relationship management instruments such as loyalty programs often reduce churn (e.g., Kopalle et al. 2012; Lewis 2004; Verhoef 2003), customer attrition remains a major threat to the financial health of many companies (e.g., Risselada et al. 2010; Schweidel et al. 2008; Thomas et al. 2004). For example, T-Mobile USA lost half a million of its most lucrative customers in the first quarter of 2012 (Bensinger and Tibken 2012). It is known that contract churn rates for many types of communication services are in the three-percent-per month range (Kim 2010). This means that a provider needs to refresh nearly 100% of its customer base about every three year.

Churn management is a prominent analytical marketing tool to fight attrition, inform resource allocation decisions, and eventually increase firm performance. A churn model predicts whether a customer will defect in the near future. The firm then targets the most risky customers with retention campaigns to prevent attrition (e.g., Ganesh et al. 2000; Libai et al. 2009; Shaffer and Zhang 2002). Targeted marketing actions toward risky customers can significantly reduce churn rates and increase firm profits (Burez and Van den Poel 2007).

Marketing analysts can choose from a variety of forecasting methods to build churn models. The choice of the modeling technique is important because it has a direct impact on prediction quality and thus on the profitability of all subsequent targeted marketing efforts (e.g., Neslin et al. 2006; Risselada et al. 2010). Many studies have thus compared different methods to identify a 'best' churn modeling technique; Verbeke et al. (2012) provide a comprehensive overview. This approach seeks profit improvements through more accurate forecasts that, in turn, result from better methods. In this work, we advocate a more direct link between churn modeling and campaign profitability.

Previous churn modeling techniques are general purpose procedures. They embody the standard philosophy toward predictive learning: maximize the fit between the model and some historical data. Here, model fit refers to a statistical quality criterion such as the likelihood. We argue against the adoption of this notion of fit and performance. There is no guarantee that a maximum likelihood churn model – or a model that is optimal in terms of some other statistical indicator – is also optimal from a business viewpoint. Marketers use churn models to aid resource allocation decisions. If a marketing budget facilitates soliciting N customers with a retention program, the churn model's task is to identify the top-N customers with the highest attrition risk. Conventional churn modeling techniques are agnostic of this application context. Research on marketing decision support systems suggests that this mismatch between the actual decision task (resource allocation) and its representation in the churn model (likelihood maximization) has a negative impact on decision outcomes and performance (e.g., Lilien 2011). Therefore, our key proposition is that creating churn models in awareness of business requirements and objectives improves the quality of resource allocation decisions and thus the profitability of retention activities.

We develop a decision-centric churn modeling framework on the basis of a recent machine learning approach called ensemble selection (e.g., Partalas et al. 2010). Ensemble selection is

a two-stage modeling paradigm that can accommodate arbitrary business-driven accuracy indicators. The lift index (defined more formally below) is a well-established measure to assess campaign planning models. It embodies the constraints (retention budget) and objectives (campaign profit) that characterize a manager's decision task. This way, the lift quantifies the degree to which a model-based targeting uses marketing resources efficiently. It is also directly connected to the profitability of a retention campaign (Neslin et al. 2006). We incorporate this business-oriented notion of model performance into our modeling framework. Using the ensemble selection methodology, we create churn models that explicitly maximize lift. We call this approach decision-centric ensemble selection (DCES) because it emphasizes the ultimate decision problem during model building.

To explore the effectiveness of our approach, our first research questions compare DCES to conventional churn modeling techniques.

- RQ1: Does DCES outperform the popular logit choice model?
- RQ2: How does DCES perform in relation to advanced single classifiers?
- RQ3: Can DCES beat sophisticated ensemble learners?

These questions reflect the technical evolution in churn modeling and systematically compare DCES to what can be considered the state-of-the-art (e.g., Lemmens and Croux 2006; Risselada et al. 2010; Verbeke et al. 2012). The key difference between our approach and conventional churn models is that DCES accounts for business objectives during model building. We strive to establish a direct link between this feature and model performance in our last research question.

- R4: Does our lift-based modeling philosophy explain the performance of DCES?

We test our research questions through empirical analysis on a collection of eight real-world churn data sets. Our results indicate that DCES performs significantly better than any conventional approach, and increases the per customer profits of retention campaigns by $.47

on average. We also find evidence that the success of DCES comes largely from the particular way it accounts for business objectives during model building. A first implication of these results is that it is both feasible and effective to use business performance measures for creating churn models. As we explain below, DCES grounds on the ensemble principle to combine multiple models' forecasts. A second implication of our study is thus that the search for one 'best' forecasting method should be abandoned. Whenever a set of alternative churn models is available, e.g., because of preliminary experiments, analysts should not try to select one seemingly best model for deployment. Instead, they should appropriately combine the available models to produce a final forecast. In this sense, our results prompt a change in current churn modeling practices.

We organize the remainder of the article as follows: In the next section, we provide an overview of the related literature. We then discuss the lift index as a measure of resource allocation efficiency, before we present our DCES framework. Next, we describe the data sets employed in our study and answer our research questions. Afterwards, we conclude the paper with a discussion of findings and implications.

2 Related Literature

Modeling customer churn is an essential part of retention management and belongs to the general field of managing customer relations (e.g., Musalem and Joshi 2009). A large number of retention models have been proposed in the marketing literature. One stream of research concentrates on predicting the length of customer relationships by means of hazard or [NBD]/Pareto models (e.g., Bolton 1998; Jerath et al. 2011; Schmittlein and Peterson 1994). These models are especially important in the context of customer lifetime value calculations (e.g., Gupta et al. 2006). A second stream of research views customer switching to competitors as transient and uses migration or Markov models to estimate transition

probabilities (see Gupta and Zeithaml 2006 for an overview). Such models can support resource allocation decisions where customer state (active or inactive) is unobservable like for instance in non-contractual business settings.

In general, churn models can be grouped into two categories, explanatory and predictive. Approaches of the first category develop models to explain churn patterns on the basis of various constructs, including the firms' marketing activities (Lewis 2004), customer knowledge (Capraro et al. 2003), or attitudinal concepts such as satisfaction (Bolton 1998; Gustafsson et al. 2005) or perceived quality (Zeithaml et al. 1996). Understanding the sources of defection is important to improve customer-centric business processes and reduce attrition rates in the long run. However, these models are less suitable to support operational business decisions in the short run. This is because explanatory models sacrifice some accuracy and forecast not as good as models explicitly designed to maximize prediction performance (Shmueli and Koppius 2011). Predictive ability is, in itself, important in marketing research and practice (Cui and Curry 2005). It is especially important in churn modeling to target retention offers to the right customers and use marketing resources efficiently (e.g., Lemmens and Croux 2006; Neslin et al. 2006; Risselada et al. 2010). Company databases contain vast amounts of customer data. It is plausible that data on, e.g., product/service usage, purchase and payment behavior, etc. embodies latent patters that are somewhat indicative of defection. In the face of hundreds or thousands of customer characteristics, and possible interactions among them, developing a formal theory how these attributes influence churn is impossible; simply because cognitive limitations prohibit decision makers from processing such large amounts of information (e.g., Lilien et al. 2004). A data-driven, predictive modeling approach is thus the only way to fully exploit the available data. In addition, this approach can account for nonlinear relationships, which is useful when modeling consumer behavior (West et al.

1997). Finally, it is possible to approximate perceptual churn drivers through behavioral data (Zorn et al. 2010).

Data-driven campaign planning models use cross-sectional data to predict whether a customer will leave the firm. Likewise, the modeling goal can be to forecast whether a customer will stop using a service (i.e., while staying with the company and continuing to use other services). A multitude of prediction methods, also called scoring methods (e.g., Malthouse and Derenthal 2008; Verhoef et al. 2010), have been developed for this kind of classification task (e.g., Hastie et al. 2009). The prevailing approach in churn prediction is to use a single model. The logit choice model is often considered "the gold standard" in marketing and churn prediction in particular (Cui and Curry 2005). It is simple and easy to understand, often performs well (e.g., Neslin et al. 2006; Risselada et al. 2010), and is widely available in standard software packages such as SAS or IBM SPSS. In addition to the logit model, prior research has also examined various other (more advanced) models (see Verbeke et al. 2012 for a comprehensive overview). The main conclusion emerging from previous work is that ensemble methods predict churn most accurately (e.g., Lemmens and Croux 2006). Ensembles are more robust prediction methods that combine the forecasts of multiple member models (Malthouse and Derenthal 2008). Much theoretical and empirical research has shown that forecast combination improves prediction quality (e.g., Batchelor and Dua 1995; Gupta and Wilton 1987; Winkler and Makridakis 1983). In particular, model averaging reduces the two sources of forecast errors, bias and variance (e.g., Ha et al. 2005).

The ensemble principle constitutes the basis of our DCES framework. However, whereas previous studies use general-purpose ensemble learners such as bagging or boosting to predict churn incidents more accurately (e.g., Lemmens and Croux 2006), we focus on the model's ultimate task to support targeting decisions. In particular, given a set of candidate churn models, we pursue a combination strategy that concentrates on the selection of

customers for a retention campaign and explicitly maximizes campaign profitability. To the best of our knowledge, the marketing literature does not contain any reference to such a decision-centric modeling approach.

3 Performance Measurement

The lift measure is a well-established performance indicator for targeting models (e.g., Ling and Li 1998). The lift grounds on a list of customers ordered according to their model-estimated scores (from highest to lowest risk of attrition in churn prediction). We define the lift measure L_d for some decile *d* of the ordered customer list as:

$$L_d = \frac{\hat{\pi}_d}{\hat{\pi}}, \qquad (1)$$

where $\hat{\pi}$ and $\hat{\pi}_d$ denote the fraction of actual churners among all customers and those ranked in the top-*d* decile, respectively. If customers were solicited at random, the fraction of actual churners reached with a retention campaign would equal $\hat{\pi}$. The lift measure quantifies how much a model improves over a random targeting.

Although lift can be defined for any decile *d* of the ranked customer list, a sensible choice is to set *d* such that it reflects the available marketing budget. If the budget allows contacting a fraction of *d* customers, it is intuitive to use an equal-sized fraction of customers to assess the targeting model. In this sense, the lift measure, with suitable choice of *d*, embodies a budget constraint. Moreover, to achieve maximal lift, a churn model must maximize the number of actual churners in the top-*d* decile. These are exactly the customers one would include in the retention campaign. Consequently, the lift measure rewards an efficient resource allocation. The higher the lift the more of the marketing budget is used on actual churners. Neslin et al. (2006) further extend the business-oriented notion of the lift measure and show that it is directly connected to the profitability of a customer retention campaign. As such, the lift captures the notion of model performance that is most relevant in churn prediction and

reflects campaign planners' marketing objectives. This is why we characterize lift as a decision-centric performance measure.

4 Decision-Centric Ensemble Selection

4.1 Motivation and Overview

The prevailing approach to develop a churn model is to use some general-purpose prediction method. Such methods build a model by minimizing some statistical loss function over training data. For example, the logit choice model minimizes the negative log-likelihood, whereas decision tree–based methods use information-theoretic criteria. The analyst can select the prediction method but has no choice in the loss function. Consequently, there is some mismatch between the analyst's objective and the objective function within the prediction method. To achieve more consistency with how analysts assess performance, our DCES framework accounts for business objectives during model building.

DCES grounds on a generic modeling paradigm called ensemble selection (e.g., Partalas et al. 2010). Ensemble selection consists of three stages: (1) constructing a library of candidate models (*model library*), (2) selecting an "appropriate" subset of models for the ensemble (*candidate selection*), and (3) combining the predictions of the chosen models to produce the final (ensemble) forecast (*forecast combination*). Several alternative approaches follow these guidelines and differ mainly in how to organize candidate selection in stage two (e.g., Kuncheva 2004). The directed hill-climbing strategy (Caruana et al. 2004) is particularly well suited for our purpose because it can accommodate arbitrary accuracy indicators. The following subsection detail the stages of this approach, and our specific design decisions to develop a churn modeling framework that is driven by actual business objectives.

4.2 Model Library

In the first modeling stage, we construct a large library of candidate churn models. The success of any ensemble strategy depends on the diversity of ensemble members (e.g., Kuncheva 2004). Our approach to control the error-correlation among candidate models' prediction is twofold. First, we employ different prediction methods, including (1) the established logit model; (2) other well-known, easy-to-use algorithms, such as discriminant analysis or tree-based procedures; (3) advanced single classifiers, such as artificial neural networks or support vector machines; and (4) powerful off-the-shelf ensembles, such as bagging or boosting (e.g., Lemmens and Croux 2006). Second, we vary the metaparameter settings of individual learners. Metaparameters, such as the number of hidden nodes in a neural network (e.g., West et al. 1997), allow the analyst to adapt a prediction method to a particular modeling task (Hastie et al. 2009). This suggests that a single method will produce somewhat different (i.e., diverse) models if it is invoked with different settings for algorithmic parameters.

Table 1 summarizes the classification methods and metaparameter settings in our model library. Our particular selection of prediction methods and metaparameter settings is based on previous churn modeling studies (e.g., Verbeke et al. 2012) and literature recommendation (e.g., Caruana et al. 2004; Partalas et al. 2010), respectively. Note that some prediction methods exhibit multiple metaparameters. In this case, we define a set of candidate values for every metaparameter and create churn model for all possible value combinations. A comprehensive discussion of all techniques and their algorithmic details is available in Hastie et al. (2009).

TABLE 1: CLASSIFICATION METHODS AND METAPARAMETER SETTINGS EMPLOYED IN THIS STUDY

Classification Method	Metaparameter[a]	Candidate Settings[b]
Single Classifiers		
Classification and Regression Tree (CART) 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 (ANN) Three-layered architecture of information processing-units referred to as neurons. Each neuron receives an input in form of a weighted some 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)	1, 2, …, 20 $10^{[-4, -3.5, \ldots, 0]}$
k-Nearest-Neighbor (kNN) 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 (LDA) 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 (LogR) 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 (NB) 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 (QDA) 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 (RLR) 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 (SVM-Lin) 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 (SVM-Rbf) 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	$2^{[-12, -11, \ldots, 12]}$ $2^{[-12, -11, \ldots, -1]}$
Homogeneous Ensemble Learners		
AdaBoost (AdaB) 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 (BagDT) 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 (BagNN) 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 (RF) 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 No. of covariates randomly selected for node splitting	100, 250, 500, 750, 1000, 1500, 2000 $[0.1, 0.5, 1, 2, 4] \cdot \sqrt{M}$ [c]
LogitBoost (LoB) 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 (SGB) 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

[a] Table 1 depicts only metaparameters that have been varied to increase the number of candidate models. A classification method may offer additional parameters that have not been considered in this study. Hastie, Tibshirani, and Friedman (2009) provide a fully comprehensive description of all methods and their parameters.

[b] We consider all possible combination of metaparameter when a method exhibits more than one parameter.

[c] M represents the number of covariates.

4.3 Candidate Selection

In accordance with Caruana et al. (2004), candidate selection begins with finding the best performing individual churn model in the candidate library, and appending this model to an empty ensemble. To improve performance, we then assess all pairwise combinations of the chosen model and one other model from the library. This way, we obtain a collection of possible two member ensembles. We select the best-performing size-two ensemble if it outperforms the best individual model. Next, we examine the best-performing ensemble of size three. That is, we assess all combinations of the current size-two ensemble and one other candidate model from the library. The stepwise ensemble growing procedure stops as soon as appending additional members does not improve performance. A special characteristic of the selection strategy is that it allows a single model to enter the ensemble multiple times (Caruana et al. 2004). We illustrate this feature in the following example and discuss its implications in section 4.4.

Table 2 exemplifies candidate selection. Suppose we have five customer records, a zero-one response variable *y* (churn = no/yes), and a library of four candidate models (M1–M4). Each model estimates a churn probability for every customer, and we assess these predictions in terms of their mean squared error (MSE).

M4 is the best individual model (lowest error), so we select it as first ensemble member in iteration 1. Next, we assess all possible size-two ensembles (including M4) in the second iteration. Note that we combine multiple models through averaging their predictions (indicated by an "&" in Table 2). Given that combining M4 and M1 improves performance (MSE .26, compared to .27 when using only M4), we append M1 to the ensemble and continue the selection process. However, the third iteration achieves no further error reduction, so that we stop ensemble growing. The final ensemble is then given by the combination of M1 and M4.

TABLE 2 ENSEMBLE SELECTION ON MSE WITH A FOUR-MODEL LIBRARY

Iteration	y	M 1	M 2	M 3	M 4
	1	.89	.33	.43	.45
	0	.87	.84	.64	.55
	0	.31	.37	.90	.69
	1	.49	.83	.69	.70
	0	.65	.85	.60	.38
MSE		**.31**	**.41**	**.40**	**.27**
		M 1 & M 4	**M 2 & M 4**	**M 3 & M 4**	**M 4 & M 4**
	1	.67	.39	.44	.45
	0	.71	.69	.59	.55
	0	.50	.53	.80	.69
	1	.60	.76	.70	.70
	0	.52	.62	.49	.38
MSE		**.26**	**.32**	**.33**	**.27**
		M 1 & M 4 & M 1	**M 2 & M 4 & M 1**	**M 3 & M 4 & M 1**	**M 4 & M 4 & M 1**
	1	.74	.55	.59	.59
	0	.76	.75	.69	.66
	0	.44	.46	.64	.57
	1	.56	.67	.63	.63
	0	.56	.63	.55	.47
MSE		**.27**	**.30**	**.30**	**.26**

The above example illustrates two technical issues that deserve a mention. First, averaging model predictions is feasible only if all models produce forecasts of a common scale. This is typically not the case when working with different prediction methods. Therefore, we convert all model predictions into churn probabilities. Specifically, we project model outputs to the interval [0, 1] by means of a logistic link function (Platt 2000). Second, assessing alternative model combinations requires auxiliary validation data. That is, we need one set of data to build the candidate models M1–M4, and a second set of (validation) data to measure the performance of individual and combined models, respectively. We construct these two samples by means of cross-validation because previous studies find it superior to alternative approaches (Partalas et al. 2010).

The candidate selection strategy of Caruana et al. (2004) is generic and not restricted to any specific loss function or performance measure, respectively. For convenience, we used MSE

in the above example. However, we can gear ensemble construction toward any objective functions that depends on the estimated churn probabilities. The lift index fulfills this requirement and is directly connected to the profitability of retention campaigns (Neslin et al. 2006). By maximizing lift during candidate selection, we can thus devise ensembles that explicitly pursue actual business objectives during model building.

4.4 Forecast Combination

A combination of multiple prediction models occurs during candidate selection and also when the final ensemble is employed to generate churn scores for novel customers. We pool models by averaging over their predictions, which is the standard approach in ensemble learning and forecast combination, respectively (e.g., Armstrong 2001; Kuncheva 2004). However, it is noteworthy that we effectively compute a weighted average. This is because the candidate selection procedure allows models to enter the ensemble multiple times (Caruana et al. 2004). For example, the best-performing size-three ensemble in Table 2 (M4 & M4 & M1) is equivalent to a weighted average of the form .67 × M4 + .33 × M1. In Table 2, multiple inclusions of M4 (i.e., model weighting) do not reduce MSE compared to the simpler (equally weighted) ensemble M4 & M1. However, the opportunity to weight candidate predictions whenever the data suggest that some members deserve a greater influence on the forecast adds to the flexibility of ensemble selection and may increase performance under certain circumstances.

5 Data

We examine our research questions in an empirical study related to telecommunications churn. Customer attrition has been well addressed in the telecommunications industry (e.g., Kim 2010). This suggests that providers have developed sophisticated predictors of the churn event (i.e., the covariates used to build a churn model), which work well with standard

modeling techniques. Therefore, outperforming conventional models on real-world telecommunications data is particularly challenging.

A churn event involves a customer cancelling or not renewing a subscription in a predefined time window. To compile a churn modeling data set, the carrier monitors customers' calling patterns and other traits of behavior over some observation period (three to six month for our data sets) and determines which customers have actually abandoned their business relationship in a follow-up period (the following month for our data sets).

Eight real-life datasets are used to validate our DCES framework. Our first data set, *Operator*, contains the records of 47,761 customers of a U.S. domestic carrier (Mozer et al. 2000). Four other data sets, *Duke 1–4*, also include U.S. customers and were obtained from the Center for Customer Relationship Management at Duke University (http://www.fuqua.duke.edu/centers/ccrm/datasets/download.html). The size of these data sets ranges from 12,410 (*Duke 3*) to 93,893 (*Duke 2*) observations. *Duke 2* has been used in the Duke/NCR Teradata Churn Modeling Tournament (Neslin et al. 2006) and later churn prediction studies (e.g., Lemmens and Croux 2006; Verbeke et al. 2012). Two other data sets, *EuroOp* and *KDD09*, containing 21,143 and 50,000 customers, respectively, refer to the European telecommunications market. De Bock and Van den Poel (2011) obtained the *EuroOp* data set from a domestic carrier, and France Telecom–Orange donated the *KDD09* data set for the KDD Cup 2009 (see http://www.kddcup-orange.com/). Finally, our study incorporates *UCI*, a publicly available data set (see www.sgi.com/tech/mlc/db) that contains 5000 observations.

The number of covariates to model the binary response variable *churn = yes/no* varies from 20 (*UCI*) to 359 (*EuroOp*), and each data set contains continuous and categorical predictors. Most of the variables in all the data sets are associated with call detail records (e.g., number of calls, call duration), customer demographics (e.g., postal code, age group), contract

characteristics (e.g., presence of an international calling plan, subscription to additional services), relational information (e.g., length of overall relationship, response to mail offer), or billing data (e.g., monthly fee, total amount spent). *KDD09* is the only exception; this data set has been made anonymous for the KDD Cup, so that no variable information is available.

To prepare our data sets, we perform several standard preprocessing operations. These include removing variables that linearly depend on other variables or have zero variance, impute missing values in continuous (categorical) variables with the mean (mode) of this variable, and standardizing the values of continuous variables. We then create two versions of each data set, one for prediction methods that can process categorical data (e.g., tree-based methods) and one for methods such as neural networks that require an addition category encoding (e.g., Crone et al. 2006). In the latter case, we transform each categorical variable into a set of indicator variables to represent every possible category with one binary variable. To avoid an excessive growth of the number of variables, we reduce the number of categories to maximal five per variable before encoding using agglomerative clustering (Verbeke et al. 2012). Finally, we randomly partition all data sets into an in-sample training set (60%) and a holdout test set (40%). This is a standard approach in predictive modeling and avoids finding idiosyncratic characteristics of the training data set that do not hold up in real life (e.g., Shmueli and Koppius 2011). We use the training and testing partition to build and evaluate prediction models, respectively.

6 Results

This section addresses our four research questions. First, we compare DCES to previous churn modeling approaches. We then examine whether our lift-based modeling philosophy explains performance differences between DCES and conventional churn models.

In all experiments, we use the top-decile-lift, $L_{.1}$ as performance measure. Concentrating on the top-decile of customers is consistent with industry practices where retention campaigns target only a small fraction of customers. Top-decile lift is also a frequent choice in previous churn modeling studies (e.g., Lemmens and Croux 2006; Risselada et al. 2010).

6.1 RQ1: Does DCES Outperform the Popular Logit Choice Model?

The prevailing approach in churn prediction is the logit choice model. We compare DCES to this benchmark in Table 3 and find that DCES produces higher lift scores on all eight churn data sets (Table 3). On the basis of a Wilcoxon signed-rank test, the recommended approach for comparing two classifiers over multiple data sets (Demšar 2006), we conclude that DCES performs significantly better than the logit choice model ($S = 0$, $p = .008$). We then compute the median of the pairwise differences of the two models lift scores. This measure is a robust estimate of the expected performance difference between DCES and the logit choice model when working with other data sets (García et al. 2010). Our results suggest that this difference amounts to .185 units in lift.

TABLE 3: PERFORMANCE OF DCES VERSUS THE LOGIT CHOICE MODEL IN TERMS OF $L_{.1}$

Data Set	DCES	LogR	Percent Improvement
Duke 1	**1.471**	1.330	11%
Duke 2	**1.612**	1.419	14%
Duke 3	**2.444**	2.159	13%
Duke 4	**1.838**	1.500	23%
EuroOp	**2.622**	2.446	7%
KDD09	**1.885**	1.837	3%
Operator	**3.770**	3.673	3%
UCI	**6.821**	3.500	95%
Median difference DCES vs. LogR		.185	

The superior performance of DCES may not come as a surprise. It is an advanced modeling paradigm and can capitalize on a large library of candidate models when forming the ensemble. The logit choice model is still an important benchmark because of its popularity in

marketing (e.g., Cui and Curry 2005). Moreover, the results of Table 3 do not originate from a simple logit model. For every data set, we build 20 alternative logit models with different covariates (Table 1) and then select the best of these models (on the validation sample) for the final comparison with DCES (on the test set). A significant improvement of over this tuned logit model is a notable achievement.

6.2 RQ2: How Does DCES Perform in Comparison to Advanced Single Classifiers?

A variety of single classifiers have been considered for churn prediction (e.g., Verbeke et al. 2012). Many of these are more advanced than the logit choice model and thus represent a more challenging benchmark. We compare DCES to nine such methods in Table 4.

TABLE 4: PERFORMANCE OF DCES VERSUS SINGLE CLASSIFIERS IN TERMS OF $L_{.1}$

Data Set	DCES	RLR	ANN	SVM-Lin	SVM-Rbf	NB	kNN	QDA	LDA	CART
Duke 1	**1.471**	1.325	1.248	1.317	1.337	1.219	1.276	1.294	1.331	1.120
Duke 2	**1.612**	1.425	1.505	1.422	1.477	1.042	1.371	1.332	1.424	1.116
Duke 3	**2.444**	2.221	2.402	2.107	2.345	1.388	2.138	1.905	2.133	1.942
Duke 4	**1.838**	1.500	1.576	1.523	1.452	1.294	1.446	1.394	1.493	1.513
EuroOp	**2.622**	2.289	2.133	2.456	2.055	1.624	1.908	2.201	2.387	1.272
KDD09	**1.885**	1.823	1.748	1.851	1.213	0.932	1.542	1.707	1.775	1.200
Operator	**3.770**	1.363	3.520	1.628	3.088	1.085	3.450	3.269	3.673	2.379
UCI	**6.821**	3.143	5.893	2.786	5.857	1.000	4.321	3.643	3.179	4.429
Avg. rank	1.000	5.125	3.750	5.250	4.875	9.750	6.375	6.750	4.500	7.625
z_j		2.725	1.817	2.807	2.560	5.780	3.551	3.798	2.312	4.376
p_j *adj.*		.0125	.050	0.01	.017	.006	.008	.007	.025	.006
Contrast DCES vs. classifier j		.3278	.2270	.3177	.3331	.9028	.3786	.4047	.2835	.6127

According to Table 4, DCES produces the highest lift scores of all models on all data sets. To confirm the significance of this result, we test the null-hypothesis of equal performance using the Friedman test. This test is based on classifiers' average performance ranks and best-suited to compare multiple models over multiple data sets (Demšar 2006). For the results of Table 4, we reject the null hypothesis with high significance (Friedman's $\chi^2 = 43.47$, d.f. = 9, $p < .001$). Multiple comparison procedures provide further insight into performance differences and

enable us to identify the models that are significantly inferior to DCES. To that end, we compute the following test statistic for all $k - 1$ pairwise comparisons of DCES with one other churn model (García et al. 2010):

$$z_j = \left(R_{ES} - R_j\right) \Big/ \sqrt{\frac{k(k+1)}{6n}}, \qquad (2)$$

where R_{ES} and R_j represent the average rank of DCES and benchmark j, respectively, and n is the number of data sets. We can translate the z_j into a probability (p_j) using the standard normal distribution table. The resulting p-values require further adjustment to control the familywise error level and ensure an overall significance level of $\alpha = .05$. We use the Hommel procedure for this purpose because it is one of the most powerful approaches available (García et al. 2010). Table 4 shows the z_j and the adjusted p-values (p_j *adj.*) corresponding to all pairwise comparisons. The results indicate that DCES performs significantly better than all single classifiers.

An important managerial issue is the magnitude of performance improvements. The last row of Table 4 reports the performance contrasts between DCES and its competitors, which we compute on the basis of median performance differences with the approach of García et al. (2010). The contrasts estimate the expected performance difference between two models (i.e., when using other data sets than employed in the study). The strongest competitor consists of a neural network model (average rank of 3.750). Compared to the ANN benchmark, DCES improves $L_{.1}$ by .227 units on average. Recall that our experimental setup comprises extensive tuning of single classifiers. For example, we determine the best network architecture for each data set out of 162 alternatives (Table 1). The superiority of DCES to this method may well be above .227 units when working with less tuned neural networks. For other benchmarks, the expected advantage of DCES ranges from approximately one-quarter to a full unit in $L_{.1}$.

6.3 RQ3: Can DCES Beat Standard Ensemble Learners?

Previous studies suggest that standard ensemble algorithms represent the most challenging benchmark in churn modeling (e.g., Lemmens and Croux 2006; Risselada et al. 2010). We compare DCES with six state-of-the-art ensembles, including stochastic gradient boosting, which was the best-performing method in the Duke/NCR Teradata Churn Modeling Tournament (Neslin et al. 2006).

TABLE 5: PERFORMANCE OF DCES VERSUS STANDARD ENSEMBLES IN TERMS OF $L_{.1}$

Data Set	**DCES**	**BagDT**	**BagNN**	**RF**	**AdaB**	**SGB**	**LoB**
Duke 1	**1.471**	1.457	1.382	1.466	1.406	1.435	1.415
Duke 2	**1.612**	1.590	1.495	1.601	1.565	1.554	1.560
Duke 3	**2.444**	2.392	2.423	2.387	2.330	2.247	2.278
Duke 4	**1.838**	1.811	1.651	1.800	1.671	1.760	1.728
EuroOp	2.622	2.407	2.368	2.358	2.417	2.642	**2.661**
KDD09	1.885	1.542	1.775	1.707	1.864	1.878	**1.899**
Operator	3.770	3.172	3.812	3.575	**3.895**	3.631	3.700
UCI	**6.821**	6.750	5.964	6.786	4.214	4.214	4.571
Avg. rank	1.625	4.125	5.000	4.000	4.563	4.688	4.000
z_j		2.315	3.125	2.199	2.720	2.836	2.199
p_j *adj.*		.017	.008	.025	.013	.010	.050
Contrast DCES vs. ensemble j		.0506	.1131	.0451	.0871	.0761	.0710

Table 5 illustrates that, although not winning on all data sets, DCES still achieves a much lower (better) average rank than RF and LogitBoost (LoB), the two runners-up in the comparison (1.625 vs. 4.000). Using the Friedman test, we reject the null hypothesis of equal performance (Friedman's $\chi^2 = 12.76$, d.f. = 6, $p = .0470$). Furthermore, Hommel's procedure rejects all pairwise hypotheses of equal performance between DCES and one other standard ensemble at $\alpha = .05$ for the adjusted p-values in Table 5. Given that the ensemble benchmarks have shown excellent performance in previous research (e.g., Ha et al. 2005; Lemmens and Croux 2006; Verbeke et al. 2012), outperforming all of them with a significant margin provides strong evidence for the effectiveness of DCES. However, we note that the advantage in terms of expected gains in lift (last row of Table 5) is smaller than in previous comparisons. In this sense, Table 5 confirms the competitiveness of standard ensemble

methods. Algorithms of the boosting family perform particularly well. For example, Adaboost achieves the highest lift on the *Operator* data set, while logistic boosting performs best in the case of *KDD09* and *EuroOp.* DCES excels on all other data sets.

6.4 Does Our Lift-Based Modeling Philosophy Explain the Performance of DCES?

Previous comparisons reveal that DCES performs significantly better than many previous churn models. It is important to understand which factors govern its success. In particular, we strive to confirm that the appealing performance in terms of lift is a direct consequence of our choice to incorporate this measure into the model building process.

The main characteristics that distinguish DCES from previous churn models are threefold: (1) the availability of a large library of candidate models, (2) the practice to average multiple models' predictions, and (3) the lift-maximizing ensemble selection strategy. In the following, we explore the individual importance of these factors. This is to apportion DCES performance to its characteristic components and to obtain a clear view on their relative merits.

6.4.1 Library Size

We begin with a sensitivity analysis of the library size. The question asked here is whether DCES requires a large model library and to which extent smaller libraries are still effective. In particular, we randomly delete 2% of the candidate models from the library, create an ensemble using particular selection strategy, and assess its performance in terms of $L_{.1}$. We repeat this procedure 50 times, each time reducing the size of the library by two percent. Figure 1 depicts the corresponding development of DCES performance. It also shows the lift-scores of the logit model (LogR), ANN, and LoB ensemble for comparative purpose. We include LogR because of its popularity in churn prediction. The other two benchmarks represent the best single and ensemble classifier from previous comparisons, respectively.

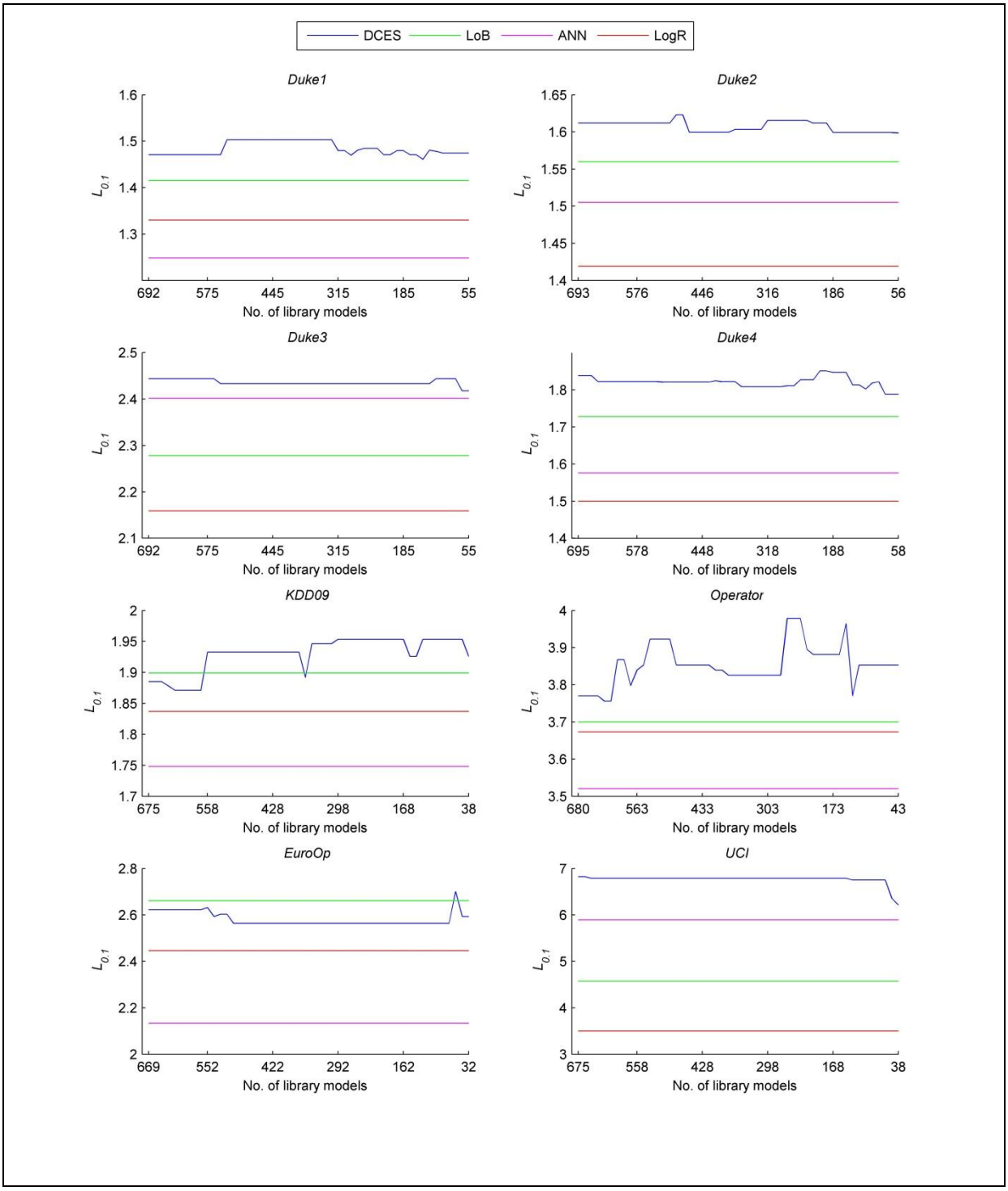

Figure 1: Development of DCES Performance When Repetitively Removing 2% (of the Original Library Size) Randomly Selected Candidate Models for 50 Iterations.

Figure 1 reveals that DCES is robust toward a random elimination of candidate models. Even small libraries of approximately 50 models suffice to perform well. In particular, Figure 1 shows that DCES is consistently better than LogR and ANN, regardless of the size of the model library. Figure 1 also shows that reducing library size never decreases the performance

of DCES below the LoB level, in settings where DCES outperformed LoB when using the full library. The two data sets where LoB was originally superior show mixed results. Whereas LoB achieves higher lift than DCES in almost all sensitivity analysis iterations on *EuroOp*, DCES actually achieves higher lift most of time on *KDD09*. In view of these results, we conclude that the success of DCES does not depend on the size of the model library.

6.4.2 Forecast Averaging

A second characteristic of the DCES framework is that it produces a composite forecast by combining member model predictions. To what extent does forecast averaging explain the success of DCES? If the main driver of its performance is indeed our lift-maximizing candidate selection strategy, we would expect approaches that average the predictions of the full library (i.e., without candidate selection) to perform as well as DCES. To confirm this, we compare DCES to four popular forecast combination approaches (e.g., Armstrong 2001): (1) a simple average (SAvg) over all candidate models' forecasts; (2) a weighted average (WAvg), in which a model's accuracy determines its influence on the composite forecast; (3) a trimmed average (TAvg) that is similar to the SAvg, except that we discard, for every customer, the *n*% most extreme churn predictions (highest and smallest) before computing the average churn score; and (4) a weighted average resulting from regressing the binary churn indicator variable on the library models' predictions (RAvg).[1] We calculate the weight of library model *j*, w_j, in WAvg as:

$$w_j = L^j_{.1} \Big/ \sum_i L^i_{.1} \qquad (3)$$

[1] In particular, we conduct a regularized logistic regression (e.g., Hastie et al. 2009). Compared to ordinary logistic regression, the regularized approach is more robust toward multicollinearity. This is important when using regression for forecast combination because the covariates are the predictions of the library models. These are inevitably correlated since they forecast the same phenomenon.

where $L_{.1}^{j}$ denotes the top-decile lift of model *j*, which we compute on the validation sample. Similarly, TAvg and RAvg employ the validation sample to select the trimming fraction *n* from the interval [.5, .1, …, .95] and build the regression model, respectively.

TABLE 6: PERFORMANCE OF DCES VERSUS AVERAGE-BASED FORECAST COMBINATION IN TERMS OF $L_{.1}$

Data Set	DCES	SAvg	WAvg	TAvg	RAvg
Duke 1	**1.471**	1.382	1.382	0.941	1.326
Duke 2	**1.612**	1.566	1.568	1.068	1.424
Duke 3	**2.444**	2.361	2.366	1.134	2.195
Duke 4	**1.838**	1.715	1.718	1.077	1.498
EuroOp	**2.622**	2.553	2.553	0.969	0.929
KDD09	**1.885**	1.871	1.878	1.254	1.158
Operator	3.770	3.965	**3.979**	0.543	1.113
UCI	**6.821**	6.143	6.357	1.464	0.179
Avg. rank	1.250	2.750	2.000	4.625	4.375
z_j		1.897	1.897	4.269	3.953
p_j *adj.*		.025	.050	.013	.017
Contrast ES vs. Avg j		.0802	.0763	.9987	.5633

Average-based combination mechanisms can capitalize on the predictions in the model library and have been shown to be highly successful in previous studies (e.g., Batchelor and Dua 1995; Gupta and Wilton 1987). However, with the exception of the *Operator* data set, they perform not as good as DCES (Table 6). Using the same statistical tests as above, we conclude that DCES performs significantly better than all average-based competitors but WAvg (Friedman's $\chi^2 = 27.7$, d.f. = 4, $p < .001$).[2] On the basis of the estimated contrasts (last line in Table 6), we expect that DCES improves lift by .08 to .99 points on average. This suggests that forecast averaging alone cannot be the reason why DCES performs well. Moreover, comparing Table 5 and Table 6, we find that the average-based combination schemes do not improve on the standard ensemble learners already well-established in churn prediction. This is noteworthy because DCES operates similar to WAvg and RAvg in that it

[2] According to Hommel's approach, an adjusted *p*-value of .05, which we show for the comparison of DCES with WAvg in Table 6, is not sufficient to reject the null hypothesis of equal performance at $\alpha = .05$.

also forms a weighted average over library model predictors. However, WAvg and RAvg retain non-zero weights for all predictions and only shrink the influence of less accurate models on the combination. DCES operates differently and identifies a subset of models before computing the final forecast.

6.4.3 Lift-Maximizing Candidate Selection

Having ruled out the influence of library size and model averaging, we conclude that the success of DCES comes largely from our lift-maximizing candidate selection strategy. It is somewhat intuitive that a churn model built to maximize lift achieves higher lift than a model built to maximize some other indicator such as likelihood. However, to support empirical evidence, we must seek a formal explanation for the efficacy of DCES and our candidate selection strategy in particular.

Theory suggests that the prosperity of any ensemble is related to the strength and the diversity of its members (e.g., Ha et al. 2005; Kuncheva 2004). These goals conflict because perfect models that discriminate between switchers and stayers with maximal accuracy must be perfectly correlated and thus lack diversity. DCES outperforms standard ensembles and this is mainly extent due to its particular strategy for candidate selection. This suggests that maximizing lift when building the ensemble model achieves in a better balance between strength and diversity than the corresponding mechanisms in standard ensembles.

To explore how DCES and competing ensemble strategies differ in terms of their emphasis on strength and diversity, we perform a kappa-lift analysis. Cohen's κ measures the degree to which two models agree in their classifications. Values of one and zero indicate that two classifiers agree on every sample, and that the agreement of two classifiers equals what would be expected by chance, respectively (Margineantu and Dietterich 1997). The κ statistic is based on discrete predictions, whereas our churn models provide continuous churn

probabilities. Therefore, to employ κ in a churn context, we convert churn probabilities into crisp classifications of the form churn = yes/no. In particular, we assign all customers whose estimated probability belongs to the top-decile of the ordered customer list to the group of churners, and all other customers to the group of nonchurners. This way, the agreement between two models captured by κ refers to the number of customers they both place in the top-decile of most-likely churners and recommend for targeting, respectively. Given an ensemble of n members, we compute κ for all $(n \times [n - 1])/2$ possible pairs of members. Similarly, we compute the mean lift score for all possible pairs of ensemble members. This allows us to depict the relationship between strength and diversity in a scatterplot (Figure 2). To put the composition of DCES ensembles into context, Figure 2 also incorporates the pairwise κ and $L_{.1}$ statistics for members of the LoB ensemble. Previous analysis suggests that LoB is the strongest competitor among the standard ensembles. It should thus have achieved a "reasonably good" balance between strength and diversity.[3]

[3] We performed the analysis for all other standard ensembles and observed similar results. Consequently, our conclusions are independent from the choice of LoB as reference ensemble.

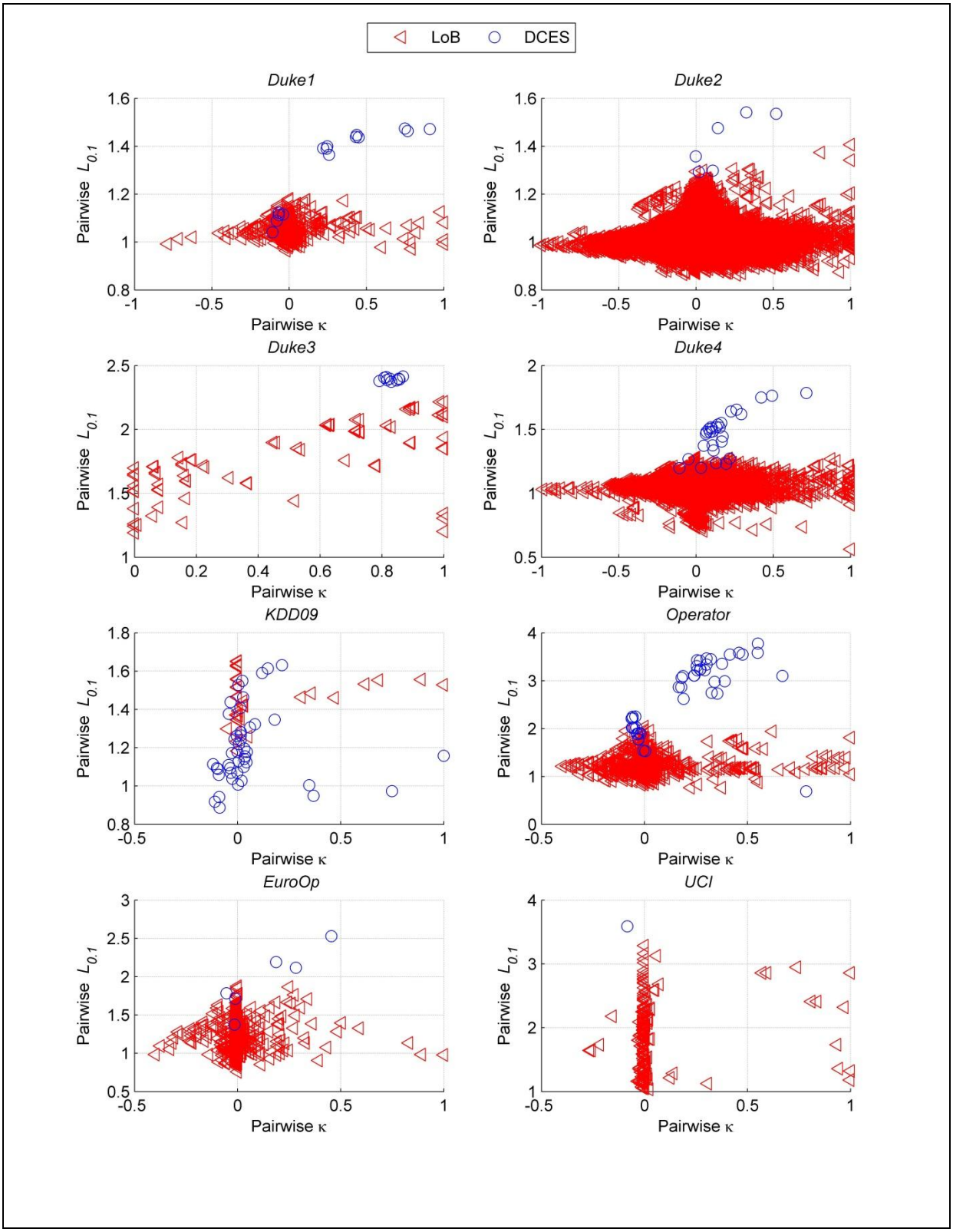

Figure 2: Kappa-Lift Analysis of the Strengh and Diversity of DCES and LoB Ensemble Member.

Note that the number of points in the scatterplot (pairs of ensemble members) is directly related to the size of an ensemble. Therefore, Figure 2 reveals that DCES leads to

parsimonious ensembles, which normally embrace substantially fewer members than the best LoB ensembles.[4] This is appealing because smaller ensembles consume less memory and predict at higher speeds, both of which are important quality criteria for any ensemble strategy (e.g., Margineantu and Dietterich 1997). However, the main result of Figure 2 is that the pairwise lift scores of ensemble members are usually higher when using DCES. This tendency persists for all data sets but *KDD09*. With respect to diversity, identifying a clear pattern is more complicated. Although the *Duke 1* and *Duke 3* data sets indicate that there may be a trend of ensemble members being less diverse (larger κ) when pursuing a decision-centric modeling strategy, we observe no such evidence in the other data sets. This suggests that differences in diversity between DCES and LoB are not systematic. In summary, Figure 2 provides evidence that DCES emphasizes strength over diversity. Compared with LoB, it leads to relatively small ensembles of strong individual members.

This kappa-lift analysis supports our proposition that the success of DCES is mainly due to maximizing lift during member selection. By choosing candidate models with high lift, the final ensemble includes only members that perform well (are strong) in terms of this metric. Figure 2 indicates that high lift can often be achieved without sacrificing diversity. Compared to LoB, ensemble members are not generally less diverse when using DCES. The reason DCES achieves a better balance between strength and diversity in our churn context is precisely that it is able to concentrate on the "right" measure of strength. Standard ensembles strategies also balance strength and diversity. However, their notion of strength is different, internally fixed by the underlying learning algorithm and agnostic of application characteristics. By illustrating the ensembles resulting from LoB in kappa-lift space, Figure 2 shows that they exhibit comparable degrees of diversity but at lower levels of strength. This does not mean that DCES is a better modeling approach in general. It is, however, a more

[4] Recall that the size of LoB ensembles ranges between 10 and 2000 (Table 1) and is tuned for every data set.

flexible approach and facilitates governing member selection toward arbitrary performance measures. This feature is particularly valuable in applications with some discrepancy between accuracy indicators that are typically incorporated in standard learners (e.g., classification error, entropy or likelihood) and performance measures that matter from a business perspective. Churn prediction is such an application and aims at models with high lift. DCES takes this objective into account, and this is why it outperforms alternative approaches.

7 Discussion

We set out to develop a framework for decision-centric churn modeling and to test its effectiveness in a large-scale empirical study. We compare our approach to several previous modeling approaches, including the popular logit model, sophisticated single classifiers, and powerful standard ensemble learners. These models are well established in academia and corporate practice. However, we find that DCES performs significantly better than any of these benchmarks.

Our results also enable us to identify the factors that explain why DCES performs well. Although it can benefit from large model libraries in our study, a sensitivity analysis reveals that the number of candidate models is not a key success factor. Thus, combining a few churn models is enough, if these are carefully chosen. The unique advantage of DCES stems from the opportunity to organize this choice process in a way that reflects actual business objectives. Building the ensemble model so as to maximize lift, DCES concentrates on the performance criterion that matters from a business standpoint. We find that this facilitates to balance strength and diversity, the key determinants of ensemble success, more appropriately than alternative ensemble regimes.

It may seem intuitive that a model optimized toward lift produces higher lift than a model optimized toward some other criteria. So why has this approach not been taken in previous

work? A possible explanation is that maximizing a discontinuous function such as lift during model fitting is highly challenging from a mathematical point of view. However, a more important reason is that model fitting is an induction problem. Even if we can overcome mathematical obstacles, approaching a statistical problem exclusively from a business angle may not be the right approach after all. A conceptual advantage of our DCES framework is that it unifies these two worlds. It leverages established statistical methods for building the candidate library and then shifts attention to the business perspective when finding the subset of models most suitable for solving the decision problem.

7.1 Implications

Our results have several implications for the science and practice of churn management. First, the finding that the new ES approach significantly outperforms what is considered the state-of-the-art emphasizes that exploring novel ways to anticipate churn and developing novel modeling frameworks is a fruitful avenue of research. Although the field has witnessed much advancement, as we show in this work, it is still possible to improve on the best models known today, identify likely churners with greater accuracy, and eventually increase the effectiveness of churn management activities.

Second, it is feasible and effective to consider business performance measures when building a churn model. Unlike previous approaches, which carry out the whole model-building process in a statistical world, DCES takes marketing objectives into account when creating the model. This is more aligned with how managers make decisions and increases the model's fit for the ultimate decision support task. In a churn context, the lift measure captures typical business objectives. Our results confirm the effectiveness to introduce this notion of performance into the model building process from an empirical and theoretical angle.

Third, our study calls for a change in standard churn modeling practices. Analysts often test alternative approaches before deploying a final churn model. Such alternatives may originate

from exploring different prediction methods and/or from experimenting with different sets of customer characteristics (i.e., covariates). The standard approach is then to pick the single "best" model and discard all the others. Our results suggest that an appropriately chosen combination of some of these alternative models will increase model performance. This selection and combination step is an excellent opportunity to introduce business objectives into the modeling process.

From a managerial perspective, a key question is to what extent better churn models add to the bottom line. Much research has shown that customer retention is an important determinant of firm performance (e.g., Gupta and Zeithaml 2006). Churn prediction aims at targeting retention programs to the 'right' customers (i.e., likely churners) and thus supports customer retention. This suggests that an indirect link between accurate churn predictions and firm performance exists. Neslin et al. (2006) examine the profit impact of churn modeling in more detail and quantify the monetary value of improvements in lift. In their most conservative scenario, per-customer profit increases by $1.71 per unit change in lift. We find in our experiments that the expected improvement of DCES over previous churn models is on average .276 lift units (computed over all benchmarks, i.e., the last rows of Tables 3−5). This suggests that a company can expect an increase in per-customer profits of $.47 ($1.71 × .276) when adopting our DCES approach. Several service markets such as the markets for communication or financial services are highly concentrated. For example, the joint market share of the two largest companies in the US market for wireless communication, AT&T and Verizon Wireless, was roughly 62%, and the market share of the four largest players together was above 85% in 2010 (Datamonitor 2011). If a company of this size contacts one percent of their customers for a churn management campaign, a $.47 increase in per-customer profits can easily amount to changes in profit in the hundreds of thousands of dollars.

An additional advantage of DCES is that it requires little human intervention. Modeling tasks typically carried out by the analyst include testing different covariates, transformations of the covariates to increase their predictive value, and alternative prediction methods. With DCES, decision makers can easily automate these tasks. They only need to incorporate the candidate models that represent the choice alternatives into the model library. The selection strategy will then pick the most beneficial model combination in a decision-centric manner. This frees marketers from laborious, repetitive modeling tasks and opens up valuable resources.

Finally, it is important to note that DCES is easy to adopt. Standard data-mining packages such as SAS Enterprise Miner or IBM SPSS Modeler provide a scripting environment to extend the base system's functionality. Implementing our DCES framework in such an environment is relatively straightforward and should not require much effort. We also show that resource-intensive computations to build up a large model library are dispensable. Marketing analysts could simply take the alternative models they routinely benchmark and reuse them as a candidate set for DCES. The bottom-line message is that additional efforts to use DCES instead of a standard modeling approach are rather small.

In summary, churn analysts have little to lose and much to gain by shifting to a decision-centric modeling approach.

7.2 Avenues for Further Research

Our study suggests several directions for further research. First, DCES works well for predictive modeling but does not allow an interpretation of how customer characteristics influence the estimated churn scores. Accuracy is crucial in churn contexts and is linked directly to company profits. However, marketers also require comprehensible models to understand which customer traits are indicative of churn. This is important to revise customer-centric business processes and increase loyalty in the long run. In addition, managers are often unwilling to follow the recommendations of a model they don't fully

understand. Therefore, it is important to develop procedures that clarify how covariates influence DCES predictions and what are the main drivers of customer churn.

A second issue pertains to the cross-sectional design of our study. All our data sets represent a snapshot, drawn from a company database at a given point in time. However, churn is a dynamic phenomenon and the causes for defection change over time. It would thus be interesting to explore the potential of DCES in a longitudinal setting.

The previous point exemplifies a third direction for research. The DCES philosophy is generic and can be applied to any discrete or continuous prediction task. It is important to validate the appropriateness of DCES in marketing applications other than churn modeling. Tasks such as scoring new product acceptance, estimating direct mail response, or predicting share of wallet are good examples. The opportunities to account for business objectives and constraints in the model-building process extend to these settings. Reproducing our results and confirming the effectiveness of a decision-centric modeling philosophy in other marketing applications would thus be a particularly fruitful research avenue.

References

J.S. Armstrong. 2001. Combining Forecasts. *Principles of Forecasting: A Handbook for Researchers and Practitioners*, J.S. Armstrong (ed.), Boston, Kluwer, 417-439.

R. Batchelor, Dua, P. 1995. Forecaster diversity and the benefits of combining forecasts. *Management Science*. **41**(1) 68-75.

G. Bensinger, Tibken, S. 2012. T-Mobile Struggles to Stem Customer Losses. *The Wall Street Journal*. Retrieved 12. July 2012.

C.B. Bhattacharya. 1998. When customers are members: Customer retention in paid membership contexts. *Journal of the Academy of Marketing Science*. **26**(1) 31-44.

R.N. Bolton. 1998. A dynamic model of the duration of the customer's relationship with a continuous service provider: The role of satisfaction. *Marketing Science*. **17**(1) 45-65.

J. Burez, Van den Poel, D. 2007. CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services. *Expert Systems with Applications*. **32**(2) 277-288.

A.J. Capraro, Broniarczyk, S., Srivastava, R.K. 2003. Factors influencing the likelihood of customer defection: The role of consumer knowledge. *Journal of the Academy of Marketing Science*. **31**(2) 164-175.

R. Caruana, Niculescu-Mizil, A., Crew, G., Ksikes, A. 2004. Ensemble Selection from Libraries of Models. C.E. Brodley (ed.), *Proc. of the 21st Intern. Conf. on Machine Learning*, New York, ACM, 18-25.

M.R. Colgate, Danaher, P.J. 2000. Implementing a customer relationship strategy: The asymmetric impact of poor versus excellent execution. *Journal of the Academy of Marketing Science*. **28**(3) 375-387.

S.F. Crone, Lessmann, S., Stahlbock, R. 2006. The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing. *European Journal of Operational Research*. **173**(3) 781-800.

D. Cui, Curry, D. 2005. Prediction in marketing using the support vector machine. *Marketing Science*. **24**(4) 595-615.

Datamonitor. 2011. Wireless Telecommunication Services in the USA. *Industry Profile 0072-2154*, Datamonitor, New York.

K.W. De Bock, Van den Poel, D. 2011. An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction. *Expert Systems with Applications*. **38**(10) 12293-12301.

J. Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*. **7** 1-30.

P.S. Fader, Hardie, B.G.S. 2010. Customer-base valuation in a contractual setting: The perils of ignoring heterogeneity. *Marketing Science*. **29**(1) 85-93.

J. Ganesh, Arnold, M.J., Reynolds, K.E. 2000. Understanding the customer base of service providers: An examination of the differences between switchers and stayers *Journal of Marketing*. **64**(3) 65-87.

S. García, 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.

S. Gupta, Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., Ravishanker, N., Sriram, S. 2006. Modeling customer lifetime value. *Journal of Service Research*. **9**(2) 139-155.

S. Gupta, Wilton, P.C. 1987. Combination of forecasts: An extension. *Management Science*. **33**(3) 356-372.

S. Gupta, Zeithaml, V. 2006. Customer metrics and their impact on financial performance. *Marketing Science*. **25**(6) 718-739.

A. Gustafsson, Johnson, M.D., Roos, I. 2005. The effects of customer satisfaction, relationship commitment dimensions, and triggers on customer retention. *Journal of Marketing*. **69**(4) 210-218.

K. Ha, Cho, S., MacLachlan, D. 2005. Response models based on bagging neural networks. *Journal of Interactive Marketing*. **19**(1) 17-30.

T. Hastie, Tibshirani, R., Friedman, J.H. 2009. *The Elements of Statistical Learning*. Springer, New York.

K. Jerath, Fader, P.S., Hardie, B.G.S. 2011. New perspectives on customer "death" using a generalization of the Pareto/NBD model. *Marketing Science*. **30**(5) 866-880.

G. Kim. 2010. AT&T Churn Rate Offers Lesson. Retrieved March 12, 2010, University.

P.K. Kopalle, Sun, Y., Neslin, S.A., Sun, B., Swaminathan, V. 2012. The joint sales impact of frequency reward and customer tier components of loyalty programs. *Marketing Science*. **31**(2) 216-235.

L.I. Kuncheva. 2004. *Combining Pattern Classifiers Methods and Algorithms*. Wiley, Hoboken.

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

M. Lewis. 2004. The influence of loyalty programs and short-term promotions on customer retention. *Journal of Marketing Research*. **41**(3) 281-292.

B. Libai, Muller, E., Peres, R. 2009. The diffusion of services. *Journal of Marketing Research*. **46**(2) 163-175.

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

G.L. Lilien, 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.

C.X. Ling, Li, C. 1998. Data Mining for Direct Marketing: Problems and Solutions. R. Agrawal, Stolorz, P.E., Piatetsky-Shapiro, G. (eds.), *Proc. of the 4th Intern. Conf. on Knowledge Discovery and Data Mining*, Menlo Park, AAAI Press, 73-79.

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

D.D. Margineantu, Dietterich, T.G. 1997. Pruning Adaptive Boosting. D.H. Fisher (ed.), *Proc. of the 14th Intern. Conf. on Machine Learning*, San Fransisco, Morgan Kaufmann, 211-218.

M.C. Mozer, Wolniewicz, R., Grimes, D.B., Johnson, E., Kaushansky, H. 2000. Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Transactions on Neural Networks*. **11**(3) 690-696.

A. Musalem, Joshi, Y.V. 2009. How much should you invest in each customer relationship? A competitive strategic approach. *Marketing Science*. **28**(3) 555-565.

S.A. Neslin, 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.

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

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

F.F. Reichheld. 1996. Learning from customer defections. *Havard Business Review*. **74**(2) 56-69.

H. Risselada, Verhoef, P.C., Bijmolt, T.H.A. 2010. Staying power of churn prediction models. *Journal of Interactive Marketing*. **24**(3) 198-208.

R.T. Rust, Zahorik, A.J. 1993. Customer satisfaction, customer retention, and market share. *Journal of Retailing*. **69**(2) 193-215.

D.C. Schmittlein, Peterson, R.A. 1994. Customer base analysis: An industrial purchase process application. *Marketing Science*. **13**(1) 41-67.

D.A. Schweidel, Fader, P.S., Bradlow, E.T. 2008. Understanding service retention within and across cohorts using limited information. *Journal of Marketing*. **72**(1) 82-94.

G. Shaffer, Zhang, Z.J. 2002. Competitive one-to-one promotions. *Management Science*. **48**(9) 1143-1160.

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

J.S. Thomas, Blattberg, R., Fox, E. 2004. Recapturing lost customers. *Journal of Marketing Research*. **41**(1) 31-56.

W. Verbeke, 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.

P.C. Verhoef. 2003. Understanding the effect of customer relationship management efforts on customer retention and customer share development. *Journal of Marketing*. **67**(4) 30-45.

P.C. Verhoef, Venkatesan, R., McAlister, L., Malthouse, E.C., Krafft, M., Ganesan, S. 2010. CRM in data-rich multichannel retailing environments: A review and future research directions. *Journal of Interactive Marketing*. **24**(2) 121-137.

P.M. West, Brockett, P.L., Golden, L.L. 1997. A comparative analysis of neural networks and statistical methods for predicting consumer choice. *Marketing Science*. **16**(4) 370-391.

R.L. Winkler, Makridakis, S. 1983. The combination of forecasts. *Journal of the Royal Statistical Society: Series A (General)*. **146**(2) 150-157.

V.A. Zeithaml, Berry, L.L., Parasuraman, A. 1996. The behavioral consequences of service quality. *Journal of Marketing*. **60**(2) 31-46.

S. Zorn, Jarvis, W., Bellman, S. 2010. Attitudinal perspectives for predicting churn. *Journal of Research in Interactive Marketing*. **4**(2) 157-169.