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

Complete Research

Baumann, Annika, Humboldt University, Berlin, Germany, annika.baumann@wiwi.hu-berlin.de

Lessmann, Stefan, Humboldt University, Berlin, Germany, stefan.lessmann@hu-berlin.de

Coussement, Kristof, IÉSEG School of Management, Lille, France, k.coussement@ieseg.fr

De Bock, Koen, IÉSEG School of Management, Lille, France, k.debock@ieseg.fr

Abstract

Churn modeling is important to sustain profitable customer relationships in saturated consumer markets. A churn model predicts the likelihood of customer defection. This is important to target retention offers to the right customers and to use marketing resources efficiently. The prevailing approach toward churn model development, supervised learning, suffers an important limitation: it does not allow the marketing analyst to account for campaign planning objectives and constraints during model building. Our key proposition is that creating a churn model 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. Further analysis suggests that this improvement comes directly from incorporating business objectives into model building, which confirms the effectiveness of the proposed framework. In particular, we estimate that our approach increases the per customer profits of retention campaigns by $.47 on average.

Keywords: Predictive Analytics, Churn Modelling, Marketing Decision Support, Ensemble Selection.

1 Introduction

Today, managers are more than ever interested to build enduring customer relationships (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 word of mouth (e.g., Ganesh et al., 2000; Reichheld, 1996; Zeithaml et al., 1996). 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). Targeted marketing actions using retention campaigns toward risky customers can significantly reduce churn rates and increase firm profits (Burez and Van den Poel, 2007).

To do so, marketing analysts can choose from a variety of approaches to build predictive models that estimate the probability of a customer to become a churner. 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 (e.g., Verbeke et al., 2012).

Previous churn modeling techniques embody the standard philosophy toward predictive learning: maximize the fit between the model and historical data using statistical quality criterion such as the likelihood. We argue that focusing only on statistical accuracy may be misleading. 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 context. They ignore the budget constraint and tend to assess 'fit' across all customers, rather than emphasizing accuracy among recipients of the campaign. However, research on marketing decision support systems suggests that a mismatch between the actual decision task (resource allocation) and its representation in a decision support model (e.g., likelihood maximization to build a churn model) 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. To test our proposition, 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). Using this methodology, we create churn models that explicitly maximize the lift index, which is a well-established measure to assess campaign planning models. We call this approach decision-centric ensemble selection (DCES) because it emphasizes the ultimate decision problem during model building. To explore the effectiveness of DCES in a systematic way, we compare it to alternative churn modeling approaches and analyze the following research questions.

- 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?
- RQ4: Does our lift-based modeling philosophy explain the performance of DCES?

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 part of retention management (e.g., Musalem and Joshi, 2009). In general, we distinguish two groups of churn models, explanatory and predictive models. 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). Approaches that model churn probabilities in a time-dependent manner using [NBD]/Pareto or Markov models (e.g., Gupta and Zeithaml, 2006) also belong to this category. Prediction models follow a data-driven modeling paradigm and are often opaque. Their advantage is that they are explicitly designed for forecasting purposes and typically predict more accurately than explanatory models (Shmueli and Koppius, 2011). Predictive accuracy is 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 to use marketing resources efficiently (e.g., Lemmens and Croux, 2006; Neslin et al., 2006).

Several methods and algorithms are available to develop predictive churn models such as partial least squares (Kim et al., 2013) and those based on supervised learning (e.g., Hastie et al., 2009). For example,

the logit choice model is widely used in industry (Cui and Curry, 2005) and has been shown to perform relatively well when compared to more advanced techniques (e.g., Neslin et al., 2006; Risselada et al., 2010). However, large-scale benchmarking studies provide evidence that ensemble models, which combine the forecasts of multiple base models, predict customer churn most accurately (e.g., Lemmens and Croux, 2006; Verbeke et al., 2012). Empirical evidence confirms the efficacy of the ensemble paradigm and suggests that combining the predictions from multiple alternative (base) models is a powerful modeling approach in general (e.g., Bhattacharya et al., 2011; Lessmann and Voß, 2010; Loterman et al., 2012). For this reason, we chose the ensemble principle as basis for our DCES framework. In particular, to develop decision-centric churn models, we propose a base model combination strategy that maximizes the profitability of a customer retention campaign (see chapter 4 for details).

3 Performance Measurement

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

$$L_d = \frac{\hat{\pi}_d}{\hat{\pi}}, \tag{1}$$

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. Note that a campaign that targets d percent of the customers at random will, on average, reach $\hat{\pi}$ actual churners. Therefore, the lift quantifies how much a model improves over a random targeting. In addition, there is a direct link between the lift of a churn model and the profitability of a retention campaign (Neslin et al., 2006). To see this, note that selecting a value for d is equivalent to imposing a budget constraint in that it implies a specific campaign size. Furthermore, only actual churners that receive and accept the retention program creates value. This indicates that $\hat{\pi}_d$ is the key driver of campaign profitability. Interested readers are referred to Verbraken et al. (2012) for a comprehensive discussion of the economics of churn prediction and the lift measure, respectively.

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 little choice in the loss function. Consequently, there is some mismatch between the analyst's objective and the objective function within the prediction method. To better align these two objectives, our DCES framework accounts for business objectives during model building. DCES grounds on a modeling paradigm called ensemble selection. 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., Partalas et al., 2010). 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 subsections explain the stages of this approach in more detail, and our specific design decisions to develop a churn modeling framework that is driven by actual business objectives.

4.2 Model Library

At first, 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 meta-parameter settings of individual learners. Meta-parameters 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 models if it is invoked with different settings for algorithmic parameters. Table 1 summarizes the classification methods and meta-parameter settings in our model library. Our particular selection of methods and meta-parameters is based on previous churn modeling studies (e.g., Verbeke et al., 2012) and literature recommendations (e.g., Caruana et al., 2004; Partalas et al., 2010).

4.3 Candidate Selection

Following Caruana et al. (2004), we initialize candidate selection with finding the best performing individual churn model in our library. To improve performance, we then assess all pairwise combinations of this model and one other model from the library. 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. The candidate selection strategy of Caruana et al. (2004) is able to accommodate any objective function that depends on the estimated churn probabilities. We exploit this feature for our DCES approach. In particular, we organize candidate selection in such a way that it maximizes the lift index. Recall that the lift is directly connected to the profitability of retention campaigns (Neslin et al., 2006). Therefore, by maximizing lift during candidate selection, we devise ensembles that explicitly pursue actual business objectives (i.e., campaign profits) during model building.

Finally, note that assessing alternative model combinations requires auxiliary validation data. That is, we need one set of data to build the candidate models in the library, and a second set of (validation) data to calculate the lift of individual and combined models during candidate selection. We construct these two samples by means of cross-validation because previous research find it superior to alternative regimes (Partalas et al., 2010).

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. More specifically, given that the candidate selection procedure allows models to enter the ensemble multiple times, we effectively compute a weighted average (Caruana et al., 2004). The opportunity to weight base model predictions in the ensemble whenever the data suggest that some members deserve a greater influence on the composite forecast adds to the flexibility of ensemble selection and may increase performance under certain circumstances.

Finally, note that averaging model predictions is feasible only if all models produce forecasts of a common scale. To achieve this, 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).

Classification Method	Number of models	Meta-parameter	Candidate Settings
Single Classifiers			
Classification and Regression Tree (CART)	6	Min. size of nonterminal nodes Pruning of fully grown tree	10, 100, 1000 Yes, No
Artificial Neural Network (ANN)	162	No. of neurons in hidden layer Regularization factor (weight decay)	1, 2, …, 20 $10^{[-4, -3.5, \ldots, 0]}$
k-Nearest-Neighbor (kNN)	18	Number of nearest neighbors	10, 100, 150, 200, …, 500, 1000, 1500, …4000
Linear Discriminant Analysis (LDA)	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)	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)	9	Histogram bin size	2, 3, …, 10
Quadratic Discriminant Analysis (QDA)	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)	29	Regularization factor	$2^{[-14, -13, \ldots, 14]}$
Support Vector Machine with linear kernel (SVM-Lin)	29	Regularization factor	$2^{[-14, -13, \ldots, 14]}$
Support Vector Machine with Radial Basis Function Kernel (SVM-Rbf)	300	Regularization factor Width of Rbf kernel function	$2^{[-12, -11, \ldots, 12]}$ $2^{[-12, -11, \ldots, -1]}$
Ensemble Learners			
AdaBoost (AdaB)	11	No. of member classifiers	10, 20, 30, 40, 50, 100, 250, 500, 1000, 1500, 2000
Bagged Decision Trees (BagDT)	11	No. of member classifiers	10, 20, 30, 40, 50, 100, 250, 500, 1000, 1500, 2000
Bagged Neural Networks (BagNN)	5	No. of member classifiers	5, 10, 25, 50, 100
Random Forest (RF)	35	No. of member classifiers No. of covariates randomly selected for node splitting	100, 250, 500, 750, 1000, 1500, 2000 $[0.1, 0.5, 1, 2, 4] \cdot \sqrt{M}$
LogitBoost (LoB)	11	No. of member classifiers	10, 20, 30, 40, 50, 100, 250, 500, 1000, 1500, 2000
Stochastic Gradient Boosting (SGB)	11	No. of member classifiers	10, 20, 30, 40, 50, 100, 250, 500, 1000, 1500, 2000

Table 1. Classification Methods and Meta-parameter Settings Employed in the Study.

5 Data

We examine our research questions in an empirical study related to telecommunications churn. Customer attrition has been well addressed in this industry so that sophisticated variables to predict churn are available (e.g., Kim, 2010). Consequently, it is particularly challenging to outperform conventional churn models on real-world telecommunications data. Table 2 provides a summary of the eight real-life churn datasets used in our study.

Data Set	Customer Records	Description / Source
Duke 1- 4	12,410 - 93,893	U.S. customers (http://www.fuqua.duke.edu/centers/ccrm/datasets/download.html)
EuroOp	21,143	European telecommunications carrier
KDD09	50,000	European telecommunications carrier (http://www.sigkdd.org/kdd-cup-2009-customer-relationship-prediction)
Operator	47,761	U.S. domestic carrier (Mozer et al., 2000)
UCI	5000	publicly available data set (www.sgi.com/tech/mlc/db)

Table 2. Telecommunication datasets used for the validation of our DCES framework.

The number of covariates to model the binary response variable *churn = yes/no* varies from 20 (*UCI*) to 359 (*EuroOp*). Each data set contains continuous and categorical predictors. Most of the variables in all the data sets are associated with call detail records, customer demographics, contract characteristics, relational information, or billing data. For each data set, we perform several preprocessing operations (e.g. elimination of linear dependency and missing values). 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 additional 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. Finally, we randomly partition the data sets into an in-sample training set (60%) and a holdout test set (40%). We use the training and testing partition to build and evaluate prediction models, respectively (e.g., Shmueli and Koppius, 2011).

6 Results

This section reports our results. First, we compare DCES to previous churn models. Next, we examine whether our specific candidate selection strategy explains the observed performance differences. In accordance with previous literature, we use the top-decile-lift, $L_{.1}$, as performance measure (e.g., Lemmens and Croux, 2006; Risselada et al., 2010).

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

We compare DCES to the best of 20 alternative logit choice models and find that DCES produces higher lift scores on all eight churn data sets (Table 3). Next to this performance indicator we also state the size of the final ensemble as well as the model composition of each ensemble. On the basis of a Wilcoxon signed-rank test, the recommended approach for comparing two classifiers (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 the difference amounts to .185 units in lift. Additionally we state the size and synthesis of the final ensemble in Table 3.

The superior performance of DCES may seem trivial. It is an advanced modeling paradigm and can capitalize on a large library of candidate models when forming the ensemble. However, the logit choice model is still an important benchmark because of its popularity in marketing (e.g., Cui and Curry, 2005).

Data Set	DCES	LogR	Percent Improvement	Size Final Ensemble	Final Ensemble Composition
Duke 1	**1.471**	1.330	11%	10	ANN, BagDT, CART, kNN, SGB, SVM-Lin, SVM-Rbf
Duke 2	**1.612**	1.419	14%	4	BagDT, CART, SVM-Rbf
Duke 3	**2.444**	2.159	13%	8	ANN, SVM-Lin
Duke 4	**1.838**	1.500	23%	8	AdaB, ANN, kNN, RF, RLR, SVM-Rbf
EuroOp	**2.622**	2.446	7%	7	AdaB, ANN, CART, RF, SVM-Lin
KDD09	**1.885**	1.837	3%	10	AdaB, ANN, LoB, RF, SVM-Rbf
Operator	**3.770**	3.673	3%	10	ANN, BagNN, LoB, SVM-Rbf
UCI	**6.821**	3.500	95%	2	RF, RLR
		.185	*= Median difference DCES vs. LogR*		

Table 3. Performance of DCES versus the logit choice model in terms of $L_{.1}$ as well as the size of the final ensemble and model composition.

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, and find that DCES gives the highest lift scores in all comparisons. To confirm the significance of this result, we test the null-hypothesis of equal performance using the Friedman test (Demšar, 2006), and reject it with high confidence (Friedman's $\chi^2 = 43.47$, d.f. = 9, $p < .001$). We then 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 family-wise 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). The adjusted p-values (p_j *adj.*) corresponding to the pairwise comparisons indicate that DCES performs significantly better than the single classifiers (Table 4). The last row of Table 4 depicts the improvement of DCES over a benchmark churn model that can be expected when using other data than used in the study. We develop this statistic using the contrast estimation approach of García et al. (2010). The expected differences range from approximately one-quarter to a full unit in $L_{.1}$.

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
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

Table 4. Performance of DCES versus single classifiers in terms of $L_{.1}$.

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).

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
p_j *adj.*		.017	.008	.025	.013	.010	.050
Contrast DCES vs. ensemble j		.0506	.1131	.0451	.0871	.0761	.0710

Table 5. Performance of DCES versus standard ensembles in terms of $L_{.1}$.

Table 5 illustrates that DCES achieves a much better performance (e.g., lower average rank) than RF and LogitBoost (LoB), the two second best models 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 these methods with significant margin provides strong evidence for the effectiveness of DCES. However, 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.

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

It is important to understand which factors explain the success of DCES, and to confirm that its appealing performance is a consequence of our choice to incorporate the lift measure into the model building process in particular. Three main characteristics distinguish DCES from previous churn models: (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 to obtain a clear view on their relative merits.

6.4.1 Library Size

DCES has access to a library of candidate models. To test whether DCES requires a large model library and to which extent smaller libraries are still effective, we randomly delete 2% of the candidate models from the library, create an ensemble using DCES, 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 as well as the lift-scores of the logit model (LogR), ANN, and the LoB ensemble for comparative purpose. In general, 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, DCES is consistently better than LogR and ANN. Furthermore, reducing the library size never decreases the performance of DCES below the LoB level in settings where DCES has originally outperformed LoB. In view of these results, we conclude that the size of the model library does not explain the success of DCES.

6.4.2 Forecast Averaging

A second characteristic of DCES is that it develops a composite forecasting. To clarify the degree to which this feature explains the success of DCES, we compare DCES to four popular forecast combination approaches (e.g., Armstrong, 2001): (1) a simple average (SAvg); (2) a weighted average (WAvg); (3) a trimmed average (TAvg) which discards the *n*% most extreme churn predictions and (4) a weighted average resulting from regressing the binary churn indicator variable on the library models' predictions (RAvg). We calculate the weight of library model on the basis of its performance 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 to build the regression model, respectively.

With the exception of the *Operator* data set, we find that the average-based combination mechanisms perform not as well as DCES (Table 6). In particular, using the statistical testing framework elaborated above, we may conclude that DCES performs significantly better than all average-based competitors but WAvg. Given the estimated contrasts (last row of Table 6), we expect that DCES improves lift by .08 to .99 points on average. Moreover, comparing Table 5 and Table 6, we find that the average-based combination schemes do not improve over the standard ensemble learners which are already well-established in churn prediction. This is noteworthy because DCES operates similar to WAvg and RAvg in that it also forms a weighted average over library model predictors. Despite this similarity, DCES outperforms these competitors. Overall, these results suggest that forecast averaging alone cannot be the reason why DCES performs well.

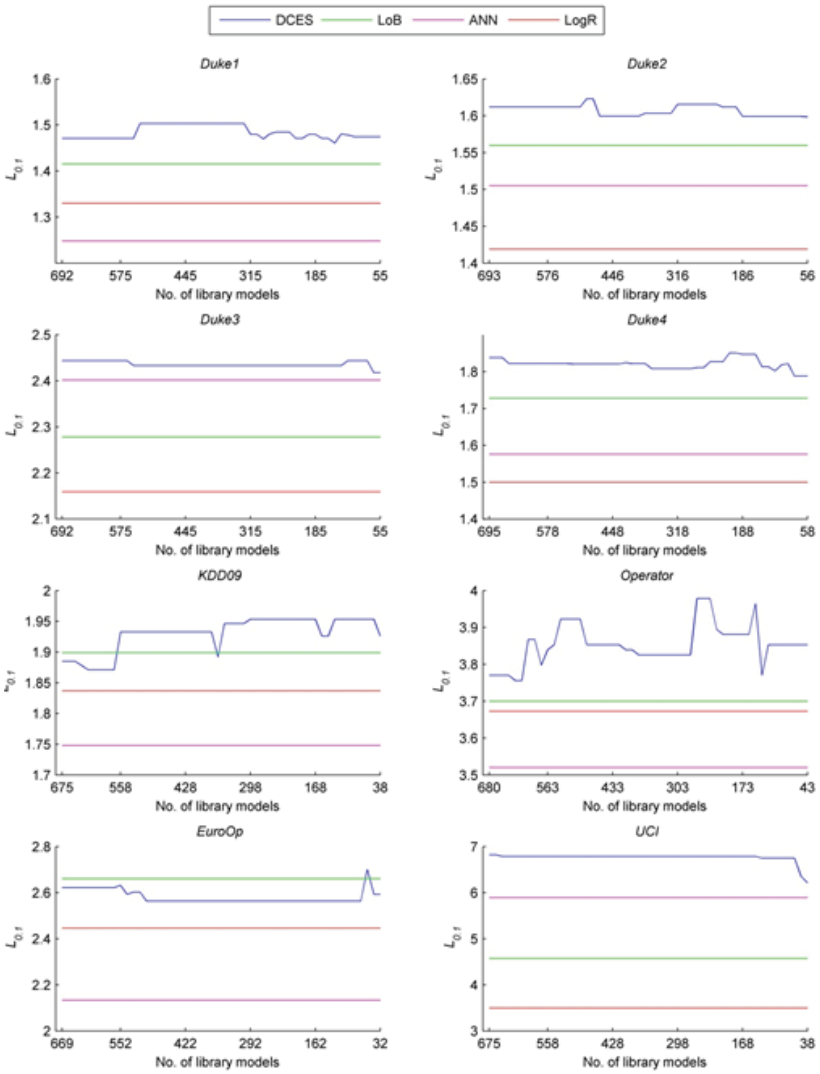

Figure 1. Development of DCES performance when repetitively removing 2% (of the original library size) randomly selected candidate models for 50 iterations.

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
p_j *adj.*		.025	.050	.013	.017
Contrast ES vs. Avg j		.0802	.0763	.9987	.5633

Table 6. Performance of DCES versus average-based forecast combination in terms of $L_{.1}$.

6.4.3 Lift-Maximizing Candidate Selection

Having ruled out the influence of the library size and the model averaging, we hypothesize that the success of DCES comes mainly from our lift-maximizing candidate selection strategy. Theory suggests that the prosperity of any ensemble is related to the strength and diversity of its members (e.g., 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. In view of the observed results, we suspect that the appealing performance of DCES stems from its specific candidate selection strategy achieving a better balance between strength and diversity.

To test this, we perform a kappa-lift analysis (Margineantu and Dietterich, 1997). In particular, given an ensemble of n members, we first compute kappa for all $(n \times [n-1])/2$ possible pairs of members and the mean lift score for all possible pairs of ensemble members. This allows us to depict the relationship between strength (i.e., lift) and diversity (i.e., kappa) in a scatterplot (see Figure 2).

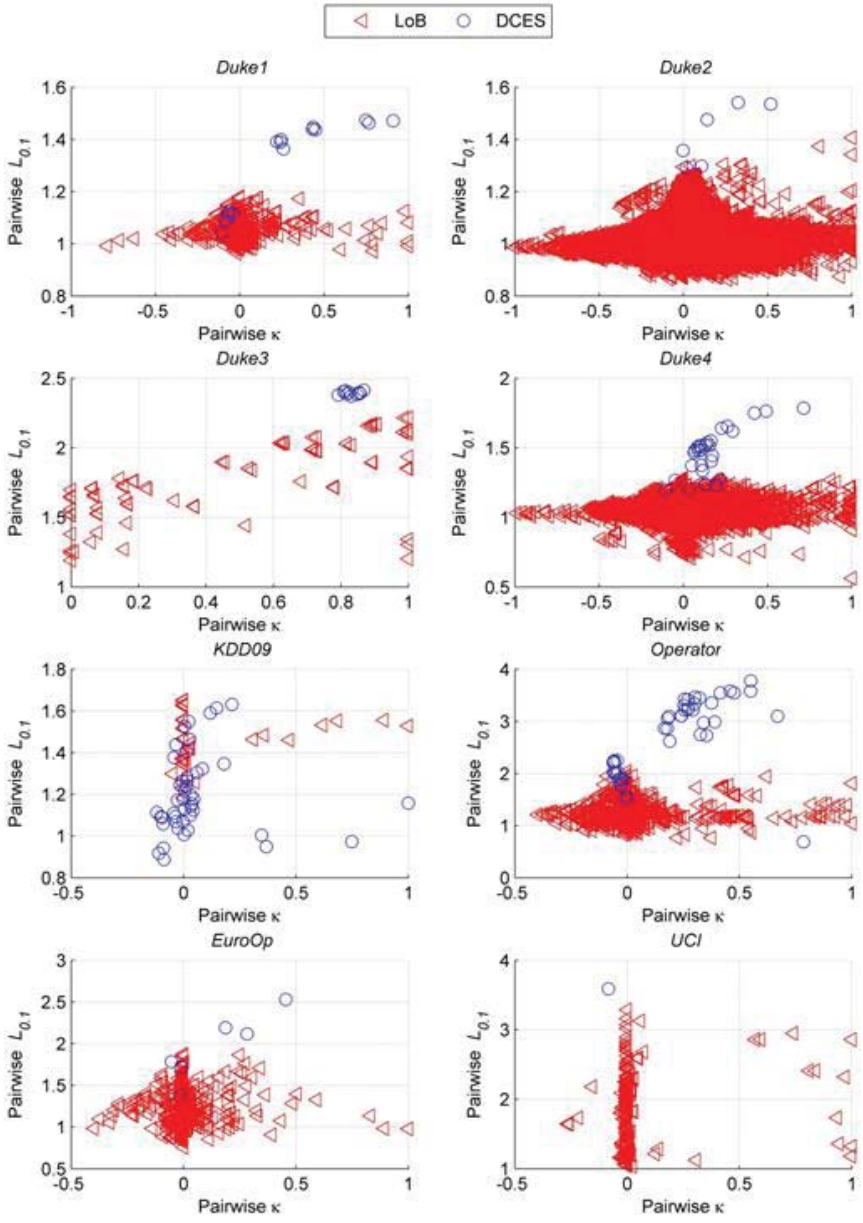

Figure 2. Kappa-Lift Analysis of the Strengh and Diversity of DCES and LoB Ensemble Member.

By this it becomes obvious that DCES leads to parsimonious ensembles, which normally embrace substantially fewer members than the best LoB ensembles. This is appealing because smaller ensembles consume less memory and predict at higher speeds (e.g., Margineantu and Dietterich, 1997). Furthermore, except for *KDD09* the pairwise lift scores of ensemble members are higher when using DCES. With respect to diversity, except for *Duke 1* and *Duke 3* we observe no trend of ensemble members being less diverse when pursuing a decision-centric modeling strategy. This suggests that differences in diversity between DCES and LoB are not systematic. Therefore, 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 in terms of lift. 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. The ensembles resulting from LoB in kappa-lift space exhibit comparable degrees of diversity but at lower levels of strength. This does not mean that DCES is a better modeling approach in general but a more flexible approach that 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 prediction methods 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 by comparing our approach to several previous well established modeling approaches (e.g., logit choice model). We find that DCES performs significantly better than any of these benchmarks. Although DCES 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. The unique advantage of DCES stems from the opportunity to organize the model 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.

Possible explanations why the DCES approach has not been considered in previous work 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 ensemble selection 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. 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, DCES takes marketing objectives into account. 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.

Third, 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 customers. 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. 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 possible 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 a per-customer profit increase of $1.71 per unit change in lift. We find that the expected improvement of DCES over previous churn models is .276 lift units. This suggests that a company can expect an increase in per-customer profits of $.47 ($1.71 × .276) when adopting our DCES approach. Depending on the size of the company a $.47 increase in per-customer profits can easily amount to changes in profit in the hundreds of thousands of dollars.

Another advantage of DCES is that it requires little human intervention. Modeling tasks typically carried out by the analyst include testing and transformation of covariates and prediction methods. With DCES, it is only necessary to incorporate the candidate models that represent the choice alternatives into the 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.

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. Since marketers and managers require comprehensible and understandable models it is important to develop procedures that clarify how covariates influence DCES predictions and what are the main drivers of customer churn.

Second, all our data sets represent a snapshot drawn at a given point of 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.

Third, it is important to validate the appropriateness of DCES in marketing applications other than churn modeling such as, e.g. scoring new product acceptance or estimating direct mail response. 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

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

Bensinger, G. and S. Tibken (2012). *T-Mobile Struggles to Stem Customer Losses*. URL: http://online.wsj.com/articles/SB10001424052702304203604577396393221489630 (visited 11/26/2014)..

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

Bolton, R. N. (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.

Burez, J. and D. Van den Poel (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.

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

Caruana, R., Niculescu-Mizil, A., Crew, G. and A. Ksikes (2004). "Ensemble Selection from Libraries of Models." In: *Proceedings of the 21st International Conference On Machine Learning*. Ed. by C. E. Brodley. ACM. New York, pp. 18–25.

classification performance of customer churn prediction models." *IEEE Transactions on Knowledge and Data Engineering* 25, 961–973.

Colgate, M. R. and P. J. Danaher (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.

Crone, S. F., Lessmann, S. and R. Stahlbock (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.

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

Demšar, J. (2006). "Statistical comparisons of classifiers over multiple data sets." *Journal of Machine Learning Research* 7, 1–30.

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

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

García, S., Fernández, A., Luengo, J. and F. Herrera (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.

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

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

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

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

Kim, G. (2010). *AT&T Churn Rate Offers Lesson*. URL: http://technews.tmcnet.com/voice-quality/topics/phone-service/articles/93062-att-churn-rate-offers-lesson.htm (visited 11/26/2014).

Kim, Y. S., Lee, H., and J. D. Johnson (2013). "Churn management optimization with controllable marketing variables and associated management costs." *Expert Systems with Applications* 40 (6), 2198–2207.

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

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

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

Lessmann, S., and S. Voß (2010). "Customer-centric decision support: A benchmarking study of novel versus established classification models." *Business & Information Systems Engineering* 2, 79–93.

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

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

Ling, C. X. and C. Li (1998). "Data Mining for Direct Marketing: Problems and Solutions." In: *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*. Eds. by R. Agrawal, P. E. Stolorz, G. Piatetsky-Shapiro. AAAI Press. Menlo Park, pp. 73–79.

Loterman, G., Brown, I., Martens, D., Mues, C., and B. Baesens (2012). "Benchmarking regression algorithms for loss given default modeling." *International Journal of Forecasting* 28, 161–170.

Margineantu, D. D. and T. G. Dietterich (1997). "Pruning Adaptive Boosting." In: Proceedings of the 14th International Conference on Machine Learning. Ed. by D. H. Fisher. Morgan Kaufmann. San Fransisco, pp. 211–218.

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

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

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

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

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

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

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

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

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

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

Verbeke, W., Dejaeger, K., Martens, D., Hur, J. and B. Baesens (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., Verbeke, W., and B. Baesens (2012). "A novel profit maximizing metric for measuring classification performance of customer churn prediction models." In: *IEEE Transactions on Knowledge and Data Engineering* 25 (5), 961–973.

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

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