Available online at www.sciencedirect.com
ScienceDirect
Journal of Interactive Marketing 24 (2010) 198–208
JOURNAL OF INTERACTIVE MARKETING
www.elsevier.com/locate/intmar

Staying Power of Churn Prediction Models

Hans Risselada,* Peter C. Verhoef & Tammo H.A. Bijmolt

University of Groningen, Faculty of Economics and Business, Department of Marketing, P.O. Box 800, NL-9700 AV Groningen, The Netherlands

Abstract

In this paper, we study the staying power of various churn prediction models. Staying power is defined as the predictive performance of a model in a number of periods after the estimation period. We examine two methods, logit models and classification trees, both with and without applying a bagging procedure. Bagging consists of averaging the results of multiple models that have each been estimated on a bootstrap sample from the original sample. We test the models using customer data of two firms from different industries, namely the internet service provider and insurance markets. The results show that the classification tree in combination with a bagging procedure outperforms the other three methods. It is shown that the ability to identify high risk customers of this model is similar for the in-period and one-period-ahead forecasts. However, for all methods the staying power is rather low, as the predictive performance deteriorates considerably within a few periods after the estimation period. This is due to the fact that both the parameter estimates change over time and the fact that the variables that are significant differ between periods. Our findings indicate that churn models should be adapted regularly. We provide a framework for database analysts to reconsider their methods used for churn modeling and to assess for how long they can use an estimated model.

Keywords: Churn prediction; Scoring models; Customer relationship management

Introduction

Churn management, being a part of Customer Relationship Management (CRM), is of utmost importance for firms, since they strive for establishing long-term relationships and maximizing the value of their customer base (Bolton, Lemon, and Verhoef 2004; Rust and Siong 2006). Losing a customer negatively affects a company in a number of ways. First, it leads to an immediate decrease in sales revenue and given that a company will have to attract more new customers when churn rates are higher, it will lead to an increase in acquisition costs (e.g. Athanassopoulos 2000; Rust and Zahorik 1993). Moreover, in the case of services that are sold on a contractual basis, losing a customer is not just a product less sold, but in fact the well-defined termination of a relationship.[1] Potential future cash flows by means of cross- or upselling are lost (Gupta, Lehmann, and Stuart 2004). Hence, accurate predictions of churn probabilities are a key element of customer lifetime value calculations and CRM in general (Blattberg, Malthouse, and Neslin 2009; Donkers, Verhoef, and De Jong 2007; Drèze and Bonfrer 2008; Fader and Hardie 2007; Gupta 2009; Pfeifer and Farris 2004). The importance of accuracy becomes even more apparent if CLV is used for marketing resource allocation (Venkatesan and Kumar 2004).

In the literature several churn model approaches have been discussed. The most commonly used methods are classification trees and logistic regression models (Neslin et al. 2006). Recently, machine learning based methodologies, such as bagging and boosting, have been applied (Ha, Cho, and MacLachlan 2005; Lemmens and Croux 2006). Bagging consists of averaging the results of multiple models that have each been estimated on a bootstrap sample from the original sample. Studies reporting the predictive performance of these models usually only consider hold-out sample or one-period-ahead validation. For example, Lemmens and Croux (2006) predict churn probabilities for one period after the estimation period. In these studies little attention is paid to the staying

* Corresponding author.
E-mail address: H.Risselada@rug.nl (H. Risselada).

[1] Although the company might provide other services also, the products under study here are central in the relationship and cancellation of a contract will most likely imply ending of the entire relationship. In addition, sometimes contracts are cancelled because two separate customers will begin to share the same address or two customers will get divorced. Unfortunately, the data do not allow us to distinguish those cases and hence, we treat all cancelled contracts as churn.

1094-9968/$ - see front matter © 2010 Direct Marketing Educational Foundation, Inc. Published by Elsevier Inc. All rights reserved.
doi:10.1016/j.intmar.2010.04.002

power. That is, how well a model predicts in a number of periods subsequent to the estimation period (Neslin et al. 2006).

Knowledge on the staying power provides database marketers with a framework to reconsider the methods used for churn modeling and to assess for how long an estimated model can be used. This in turn will help to improve CLV predictions, since they depend heavily on churn probabilities. Hence, insights on the staying power of churn prediction models can be used to determine a reliable time horizon of CLV calculations (Blattberg, Malthouse, and Neslin 2009). However, obtaining accurate predictions comes at a cost; gathering the right data, cleaning up the data sets, and estimating a model can be very time-consuming (e.g. Malthouse and Derenthal 2008). Hence, a balance between model accuracy and model building efficiency is desirable. To increase the model building efficiency we investigate in what way models need to be adapted over time. These insights can make the process of churn prediction less cumbersome, and thereby save time and money.

In this study we use two customer databases, one from a large internet service provider and one from a health insurance company, to analyze the staying power of the most commonly used churn models, namely the logit model, the classification tree, and both methods in combination with a bagging procedure. To evaluate the staying power the top-decile lift and Gini coefficient are calculated for different time periods.

The results show that the application of a bagging procedure has little effect on the predictive performance of the logit models, but that it increases the accuracy of the predictions of the classification trees. Overall, the classification tree in combination with a bagging procedure leads to the highest predictive performance over time. However, the staying power of all models is low, as the predictive performance deteriorates considerably after the estimation period. Furthermore, we find that for all models the significance and size of the parameter estimates vary over time. In sum, our results show that the optimal strategy for our data sets is to regularly estimate a new classification tree in combination with a bagging procedure and start the modeling process with selecting the appropriate variables for that particular period.

The contribution of this paper to the existing literature on churn modeling is twofold. First, this is the first paper in the marketing literature that investigates the staying power of churn prediction models over a longer time span. Previous research mainly used hold-out samples or one-period-ahead validation. Since CLV calculations depend heavily on predicted churn probabilities this study also contributes to the literature on CLV calculation and CLV-based marketing resource allocation (e.g. Blattberg, Malthouse, and Neslin 2009; Venkatesan and Kumar 2004). Second, we more specifically contribute to the churn modeling literature by testing the predictive power of prediction methods in two industries, namely the Internet service provider and insurance markets. In the extant literature, some researchers have shown a superiority of bagging in one industry (Lemmens and Croux 2006), while others suggested a better performance for the logistic regression in another industry (e.g. Donkers, Verhoef, and De Jong 2007). Hence it is important to test the predictive power of churn models across multiple industries (Verhoef et al. 2010).

The remainder of this paper is structured as follows. In the next section we present a concise overview of the literature on scoring models. In the Data section we describe the data, followed by the Methodology section. In the Empirical results section we describe the results for the two data sets separately and in the last two sections we summarize our main findings and formulate avenues for future research.

Modeling Churn

Scoring Models

In the literature various methods for churn analysis have been described. These methods are very similar to those traditionally used in the direct marketing field, since identification of customers that are likely to churn is similar to the identification of customers that are likely to respond to a mailing. Analogous to other papers in this area (e.g. Malthouse and Derenthal 2008; Verhoef et al. 2010) we will refer to these models as scoring models. Two scoring models that have extensively been studied in the marketing field are logistic regression models and classification trees (see Table 1). In the marketing literature several studies have compared the two, but did not reach a consensus on a clear winner; the observed differences in the performance of the two methods were often rather small (Hwang, Jung, and Suh 2004; Levin and Zahavi 2001; Neslin et al. 2006).

Although more sophisticated models have been studied within marketing, such as neural networks (Zahavi and Levin 1997), random forests (Buckinx and Van den Poel 2005; Coussement and Van den Poel 2008; Larivière and Van den Poel 2005), multiple adaptive regression splines (Deichmann et al. 2002), ridge regression (Malthouse 1999), and support vector machines (Coussement and Van den Poel 2008), they have not yet gained widespread popularity due to limited gains in accuracy and a substantial increase in complexity (see Table 1).[2] This is supported by Neslin et al. (2006), who found that logistic regression models and classification trees accounted for 68% of the entries of a churn modeling contest in which both practitioners and academics participated.

In sum, prior marketing literature suggests that logistic regression models and classification trees are commonly used by academics and practitioners and that both methods have good predictive performance. However, based on the aforementioned papers a superior method has not been identified.

Scoring models have been applied in many research areas other than marketing, for example the machine learning field. In that field three large-scale comparative studies have appeared, in which the performance of many different models, including logistic regression models and classification trees, has been assessed on a large number of data sets (King, Feng, and Sutherland 1995; Lim, Loh, and Shih 2000; Perlich, Provost, and Simonoff 2004). A general conclusion is that the performance of a particular method depends heavily on the characteristics of the data. King, Feng, and Sutherland (1995)

[2] For an in-depth discussion of these methods we refer to Hastie, Tibshirani, and Friedman (2008).

Table 1
Scoring model literature framework.

Study	Literature stream	Compared methods (of interest)	Number of periods	Main findings
King, Feng, and Sutherland (1995)	Machine learning	–Logistic regression –Classification trees (among many others)	1	–No single best algorithm –Performance depends on the characteristics of the data set –Overall, discriminant and regression algorithms perform well in terms of accuracy –These methods performed well on the data sets on which the tree algorithms performed worse –Tree algorithms are accurate if the data has extreme distributions –Logistic regression is a bad choice if the data is far from normal and if there are many categorical variables in the data –Tree algorithms are the easiest to use and understand
Lim, Loh, and Shih (2000)	Machine learning	–Logistic regression –Classification trees (among many others)	1	–Differences in error rates of many algorithms are statistically insignificant –Classification trees perform well and are easiest to interpret
Perlich, Provost, and Simonoff (2004)	Machine learning	–Logistic regression –Classification tree –Both+bagging	1	–Logistic regression performs better for smaller data sets –Tree induction performs better for larger data sets –Higher signal-separability situation is favorable for trees –Classification trees: bagging often improves accuracy, sometimes substantially –Logit: bagging is detrimental
Buckinx and Van den Poel 2005	Marketing	–Logistic regression –Random forest –Neural network	1	–Differences among techniques are statistically insignificant
Coussement and Van den Poel 2008	Marketing	–Logistic regression –Support vector machines (SVM) –Random forest	1	–SVM outperforms the logistic regression, only when the appropriate parameter-selection technique is used –Random forests approach outperforms SVM –The trade-off between time allocated to the modeling procedure and the performance is emphasized
Deichmann et al. 2002	Marketing	–Multiple adaptive regression splines combined with logistic regression (hybrid approach) –Logistic regression	1	–MARS+logit is slightly better than logistic regression –The use and interpretation of MARS is complicated –MARS is not widely available
Ha, Cho, and MacLachlan (2005)	Marketing	–Logistic regression –Neural network –neural network+bagging	1	–Neural network+bagging outperforms both the neural network and logit
Haughton, and Oulabi 1993	Marketing	–CART –CHAID	1	–CHAID performs slightly better for problems with many categorical variables –The best solution is to use both CART and CHAID, compare the results and choose the best
Hwang, Jung, and Suh (2004)	Marketing	–Logistic regression –Neural network –Classification tree	1	–Methods perform similarly
Kumar, Rao, and Soni (1995)	Marketing	–Logistic regression –Neural network	1	–Neural networks account for complex relationships in the data and produces better classification than the logistic regression –The logistic regression technique has a closed form solution and is easier to interpret
Larivière and Van den Poel 2005	Marketing	–Logistic regression –Random forest	1	–The random forest approach performs better than the logistic regression
Lemmens and Croux 2006	Marketing	–Classification tree –Classification tree+bagging	1	–Tree+bagging outperforms the single classification tree
Levin and Zahavi 2001	Marketing	–Classification trees –Logistic regression	1	–The logistic regression model outperforms the other models, but the differences are small –Classification trees are easier to use and interpret

Table 1 (*continued*)

Study	Literature stream	Compared methods (of interest)	Number of periods	Main findings
Neslin et al. 2006	Marketing	–Logistic regression –Classification tree –Classification tree + bagging (see Lemmens and Croux 2006) –Neural network –Discriminant analysis –Cluster analysis –Bayes	2 (one-period-ahead forecast)	–Tree + bagging performs best –Logistic regression and classification trees perform similarly and outperform the neural network approach, discriminant analysis, cluster analysis, and Bayes –The models have staying power (i.e. they perform similarly in the second period)
Xie et al. 2009	Marketing	–Improved balanced random forest (IBRF) –Neural network –Classification tree –Support vector machines	1	–IBRF outperforms the other three methods
Zahavi and Levin 1997	Marketing	–Neural network –Logistic regression	2 (one-period-ahead forecast)	–The difference in the performance of both methods is rather small –Neural network approach is complicated
This study	Marketing	–Logistic regression –Classification trees –Both + bagging	3–4	–Classification trees in combination with bagging have the strongest predictive performance. –Staying power of models is relatively small.

find that the logistic regression model is outperformed by the tree-based methods if the data is far from normal and contains many categorical variables. However, Perlich, Provost, and Simonoff (2004) emphasized that the size of the estimation sample has a major impact on the performance, and hence they argued that comparisons cannot be made on a single version of a data set. In their study, logistic regression outperformed classification trees on smaller data sets ($n \approx 1000$), but the opposite held for larger data sets. Furthermore, they found that performance is influenced by the signal-to-noise ratio; the higher this ratio, the better the classification trees perform. If signal and noise are hardly separable there is a high risk of over fitting with tree-based methods due to the "massive search" of the algorithms (Perlich, Provost, and Simonoff 2004).

Aggregation Methods

To improve the performance of the aforementioned methods predictions could be obtained by averaging the results of a large number of models. The intuition behind aggregating multiple model results is that the quality of a single predictor might depend heavily on the specific sample (Breiman 1996b) and is not known beforehand. Averaging predictors that vary substantially will result in a more stable predictor (Breiman 1996a; Malthouse and Derenthal 2008). Recently, a number of aggregation methods have been introduced in marketing. Malthouse and Derenthal (2008) aggregated predictions based on a large number of cross-sectional models, each of them estimated on a data set from a different moment in time. In their study, the aggregated models outperformed the single models. Another aggregation method originating in the machine learning field is bootstrap aggregation, or bagging, which has been applied by Lemmens and Croux (2006) to model churn of a US wireless telecommunications company. In the bagging procedure a model is estimated on a number of bootstrap samples of the original estimation sample, resulting in a number of predictions for every customer. The final prediction is obtained by taking the average of all predictions (Breiman 1996a). The bagging procedure provided classifiers that were substantially better than those obtained by a single classification tree. Although it might improve the performance of classification trees, bagging might have a negative effect on the performance of the logistic regression model (Perlich, Provost, and Simonoff 2004). Bootstrap samples are random samples of size n drawn with replacement and hence the number of original observations in the bootstrap samples is smaller than in the complete sample. Therefore the performance of the logistic regression is likely to be worse. Furthermore, the logistic regression model tends to be less sensitive to the specific sample that is used to estimate the model. Due to less variance in the estimated churn probabilities averaging them will have less effect.

To summarize, aggregating predictors is a simple way to improve the performance of commonly used models. However, bagging the logistic regression might not lead to the expected improvement due to effect of data set size and lower sample sensitivity.

Staying Power and Model Adaptation

So far, all the results we have discussed are based on in-period or one-period-ahead forecasts (see Table 1). However, building churn prediction models is a time-consuming and therefore costly operation (Malthouse and Derenthal 2008) and hence it is valuable to assess how well these models perform in the longer term.

To obtain accurate long-term churn predictions firms need a good prediction model, producing results that are reliable and

generally accepted as such. Leeflang et al. (2000) proposed five implementation criteria for the structure of good models and one of those criteria is adaptivity. In general, models can be adapted in three ways; re-estimation of the parameters, including or excluding variables from the model, or changing the entire structure. Changing the structure refers to using a different type of model, a different unit of analysis, or modeling a situation that has changed over time, e.g. a sales model of a retailer that set up a new distribution network (Leeflang et al. 2000). The models we analyze in this paper are adaptive in the sense that we can re-estimate the model parameters and add or delete variables from the models. We leave the possibility of changing the model structure aside since the aim of this paper is to compare the performance of a limited number of models with a fixed structure.

In general one would prefer a churn prediction model with large staying power. Unfortunately, this does not always occur in practice. An important factor is that changes in the market environment (i.e. the competitive setting) might affect customer behavior (Blattberg, Kim, and Neslin 2008; Malthouse and Derenthal 2008). As the market environment is typically not included in churn prediction models, these changes will lead to a decrease in staying power, since the previously estimated model no longer matches with the actual situation.

Given that models are adaptive, the question is if, when, and how models need to be adapted. The most important determinant of the need for adaptation is the difference between actual and predicted values of the dependent variable, i.e. the predictive performance over time (Leeflang et al. 2000). In particular, the staying power, defined as the predictive performance of a model in a period x months after the estimation period (Neslin et al. 2006), is an important aspect. To assess how a model needs to be adapted, it should be estimated on a number of consecutive periods using a fixed set of variables. By looking at the size, sign, and significance of the parameters one can decide whether the same parameters have to be re-estimated or different variables have to be included in the model.

There are reasons to expect that staying power differs between models in a dynamic environment. Over time, estimation samples change. Especially classification trees seem to be vulnerable to these changes (Breiman 1996a). As the bagging procedure consists of averaging the predictions based on models that have been estimated on a large number of slightly different samples, the in-period accuracy increases, and hence the staying power might increase as compared to the single model case. As we mentioned earlier, the logit model is less sensitive to minor changes in the estimation sample (Perlich, Provost, and Simonoff 2004) and hence classification trees are expected to benefit most from the bagging procedure.

In sum, based on the discussion above we expect that the classification trees will benefit most from the bagging procedure. Furthermore, we expect that the trees in combination with the bagging procedure will outperform the other three methods. However, a limited size of the data set and a low signal-to-noise ratio are in favor of the logistic regression model and might weaken the results.

Data

In the empirical study we used two data sets. The first set of data we use is part of a customer database of a large internet service provider (ISP), which is owned by a telecommunications company offering a wide range of services (e.g. fixed phone line subscriptions and digital television). The data set consists of observations for the period January–September 2006, which we divide into four periods of equal length (labeled Q1 to Q4). We include customers with an ADSL connection and exclude out-dated dial-up subscriptions since the company was actively changing these subscriptions. This forced switching behavior would disturb the analyses. Churners are those customers that have an internet subscription at the beginning of the observation period and have no subscription at the end.

The second data set comes from a health insurance company and consists of yearly churn data of the period 2004–2006. We use yearly data since customers typically switch at most once a year in this industry in the Netherlands (Dijksterhuis and Velders 2009; Donkers, Verhoef, and De Jong 2007). Churners are the customers who have an insurance at the beginning of the year but do no longer have one at the end of the year.

The variables we include in the models can be divided into two groups: customer characteristics and relationship characteristics (Prins and Verhoef 2007). The first group consists of sociodemographic variables, socioeconomic variables and commitment. The relationship characteristics consist of relationship length, breadth, and depth (Bolton, Lemon, and

Table 2
Link between included predictors and CRM literature.

Predictors ISP data	Predictors Insurance data	Theory	Studies
		Customer characteristics	
Age, household size, move	Age, family configuration	Sociodemographics	Mittal and Kamakura (2001) Verhoef et al. (2003)
Income	Income	Socioeconomics	Mittal and Kamakura (2001) Verhoef et al. (2003)
Carrier pre-select (CPS)		Commitment	Gruen et al. (2000) Verhoef (2003)
		Relationship characteristics	
Relationship age company, relationship age ISP	Relationship age company	Length	Bolton (1998)
Value added services fixed phone line		Breadth	Bolton et al. (2004)
Revenue fixed phone line, subscription type fixed phone line, connection speed	Insurance package type, Individually/collectively insured	Depth	Lemon et al. (2002) Bolton et al. (2000)

Verhoef 2004). A brief overview of the link between these predictors and the extant CRM literature is provided in Table 2. Please note that we included log-transformed versions of the revenue variable in the ISP data and of the relationship length variable in the insurance data to reduce the skewness of the distribution.

Methodology

Sampling

For the ISP data we use two different samples per period: a balanced sample (50%–50% churners–nonchurners) to estimate the models and a proportional random sample to validate the models. All random samples consist of 100K customers; the sizes of the balanced samples are 7063 (Q1), 6967 (Q2), 7146 (Q3), and 7001(Q4).

For the analysis of the insurance data we use balanced samples only; proportional random samples were not available to us. The sizes of those samples are 1789 (2004), 1294 (2005), and 1474 (2006).

We use balanced samples for estimation since the obtained classifiers outperform the ones obtained by random samples (Donkers, Franses, and Verhoef 2003; Lemmens and Croux 2006).

Models

All models are estimated using a fixed set of variables per data set as described in the Data section.[3] For a detailed description of the logistic regression model, we refer to a statistical textbook (e.g. Franses and Paap 2001). The classification trees are generated using a splitting rule based on the commonly used Gini index of diversity, suggested by Breiman et al. (1984). To avoid overfitting of the trees we use the cost-complexity pruning method (Breiman et al. 1984).

In the bagging procedure a model is estimated on B bootstrap samples of the original estimation sample, resulting in B different predictions for every customer. The final prediction is obtained by averaging all B predictions (Breiman 1996a). The top-decile lift was used to determine the optimal value of B; we set it equal to 100 in all cases[4] (Lemmens and Croux 2006).

Performance Measures

To compare the predictive performance of the various models we use two performance measures. A measure that is commonly used for these types of models is the top-decile lift (TDL; Lemmens and Croux 2006; Malthouse 1999; Neslin et al. 2006). The TDL is defined as the fraction of churners in the top-decile divided by the fraction of churners in the whole set (Blattberg, Kim, and Neslin 2008). This measure represents the ability of a model to identify those customers that have a high churn probability, the so-called high risk customers.

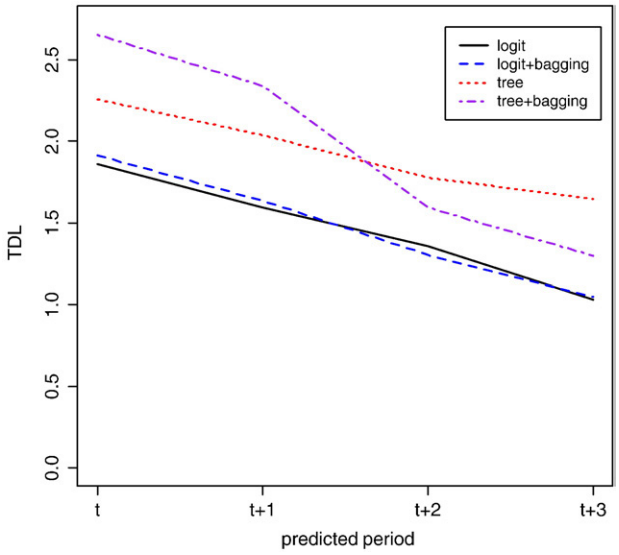

Fig. 1. Average top-decile lifts of models estimated at time t (ISP data).

The second measure we use is the Gini coefficient, which takes into account the overall performance of the model. This coefficient is frequently used to measure income inequality. Here, we use it to compare the quality of a model-based selection with a random selection of customers. We calculate the Gini coefficient by dividing the area between the cumulative lift curve and the 45-degree line by the area under 45-degree line (Blattberg, Kim, and Neslin 2008).[5]

Empirical Results

ISP Data

Staying Power: Top-decile Lift

Fig. 1 shows the average top-decile lifts of the four different models. The results have been aggregated across estimation periods for the sake of clarity. The estimation period is denoted by t. The ability of the estimated models to correctly identify high risk customers is decreasing over time, since all lines are downwards sloping. A substantial decrease in period $t+2$ can be observed. Furthermore, the figure shows that the classification trees outperform the logit models in this respect, because both the line of the tree model and the line of the tree+bagging model are above the lines of the logit model. With respect to the effect of applying a bagging procedure, the following can be observed. The logit model does not benefit from this procedure,

[3] The aim of the paper is to compare the performance of two commonly used methods and hence we used a fixed set of variables to estimate a logit model and a tree model. The resulting logit models and trees are not necessarily the same in the sense that the set of significant parameters in the logit models may differ over time and the shape of the trees may vary over time. This is a direct result of the methods used and therefore we decided to compare them this way.

[4] To find the optimal value of B we used a number of values for B (50, 100, 150) and compared the results using the top-decile lift. The results for $B=50$ and $B=100$ were different and the results for $B=100$ and $B=150$ were very similar, hence we concluded that a value larger than 100 would not add much to the analysis and decided to use the value 100.

[5] Based on comments of a reviewer we have also assessed the variance of the forecasts using a bootstrapping procedure (Blattberg et al. 2009). The variance of the forecasts is rather stable over time. Hence, we decided not to discuss this explicitly. Details on this analysis can be requested from the authors.

since both lines overlap in Fig. 1. However, the bagging procedure improves the predictive performance of the classification tree substantially. Both for the in-period and one-period-ahead predictions the TDL is higher for the tree in combination with a bagging procedure than for the single tree.

Staying Power: Gini Coefficient

In Fig. 2 the average Gini coefficients of the four models are shown. Similar to what we found for the top-decile lift, the overall performance of all models decreases over time, indicated by the downwards sloping lines. Again, the tree models outperform the logit models and the bagging procedure improves the predictions of the classification trees but has little effect on the logit model results.

Parameter Assessment

In Table 3 the parameter estimates of the single logit models are presented for each estimation period. The most important observation is that the significance and size of the parameter estimates change over time. Only 4 of the 25 variables (16%) have a significant effect on churn in all periods. None of these four effects changes in sign. Customers with a higher revenue on their fixed phone line have a higher probability to churn on their internet subscription and those with the cheapest fixed phone subscription (*type 1*) have a higher churn probability than those with a more expensive subscription. Customers that used carrier pre-select (CPS) in the past have a higher probability to churn and older people ($age \geq 65$) have a lower churn probability than young people. Three additional variables have a significant effect of the same sign in three of the four periods and four variables have a significant effect in the same direction in only two periods. There are five variables that have a significant effect only in Q1, where the sign of the effect mostly stays the same in the subsequent periods though the effect is no longer significant. Finally, the effects of two variables, relationship age ISP and connection speed medium are significant in Q1 and Q3, but the sign of the effects is opposite in the two periods; in Q1 the effect is positive, in Q3 it is negative, which clearly indicates low parameter stability.

Table 3
Parameter estimates of the single logit model (ISP data).

Variable	Period			
	Q1	Q2	Q3	Q4
Revenue fixed phone line (€)	.1038**	.1605**	.1097**	.1248**
Carrier pre-select	.1608*	.4869**	.5716**	.3507**
Relationship age company (months)	.0000	−.0007**	−.0010**	−.0012**
Relationship age ISP (months)	.0084**	.0009	−.0019*	.0001
Connection speed (ref. cat. 'slow')				
Medium	.7834**	.0199	−.5348**	−.1809*
High	.8992**	.3996**	−.1658	−.1195
Fixed phone subscription (ref. cat. 'standard')				
Type 1 (cheapest)	.7715**	1.0344**	.9207**	.4580**
Type 3	−.2412**	−.1509*	−.0198	.0340
Type 4	−.2957**	−.0438	−.1181	.0846
Type 5	−.4219**	−.2906*	.1256	.1229
Household size (ref. cat. '3')				
1	−.3177**	−.1016	−.1293	−.1103
2	−.1597*	−.0426	−.0191	−.0756
4	.0132	−.0853	−.1349	−.0796
5	.3009**	.1406	−.0562	.0100
>6	.1219	.0628	−.2390	−.3527*
Age (ref. cat. '25-35')				
<25	.0626	−.0240	−.1151	.2874*
35–45	.0909	.0048	−.1234	−.0112
45–55	.1169	.0947	−.0257	−.0088
55–65	−.2477**	−.2320*	−.1511	−.1338
≥65	−.4017**	−.2286*	−.3038**	−.3023**
Income (ref. cat. '1.5 times standard')				
<Standard income	.2101	.2119*	.3205**	.4043**
Standard income	.0757	.1762*	.1109	.3189**
2 times standard income	.0094	−.1038	−.0048	.0173
>2 times standard	−.2028**	−.2498**	−.2147**	−.0876
Value added services fixed phone line	−.0057	−.1553*	−.0476	−.0603

* $p<.05$.
** $p<.01$.

The results of the logit model in combination with a bagging procedure are very similar to those of the single logit model. The parameter estimates show very little variation over the 100 bootstrap samples. Hence, we do not present them here.

Table 4 shows the splitting variables of the single classification trees for all four periods. The results show that relationship ISP, connection speed, and age (33% of the variables) appear in all trees. Two of these variables, namely connection speed and age, appear in all logit models and all trees and can thus be considered important predictors of churn here. In contrast with the logit results, the variable value added services does not play a role in the classification tree in Q2.

A summary of the results of the classification trees in combination with a bagging procedure is provided in Fig. 3. A large diversity in the frequencies can be observed, indicating that most variables are used only in a subset of all the bootstrap samples. This corresponds to the notion of instability with respect to the estimation sample. Two variables (20%),

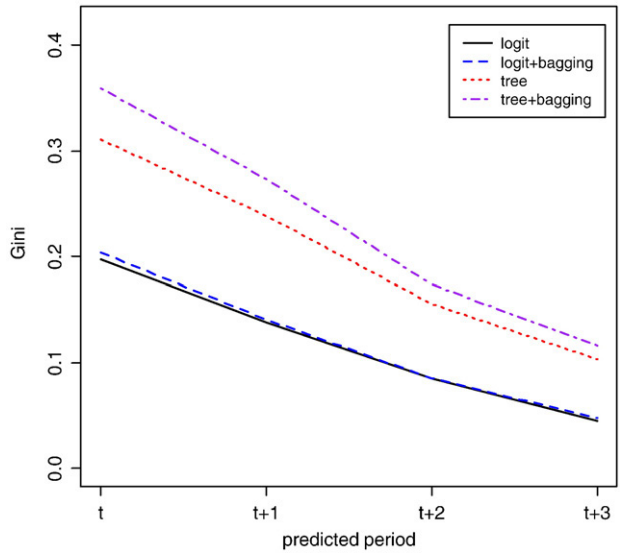

Fig. 2. Average Gini coefficients of models estimated at time *t* (ISP data).

Table 4
Splitting variables in the estimated classification trees (ISP data).

Variable	Period			
	Q1	Q2	Q3	Q4
Revenue fixed phone line		x		
Carrier pre-select (CPS)			x	
Relationship age company	x		x	x
Relationship age ISP	x	x	x	x
Connection speed	x	x	x	x
Fixed phone subscription	x	x	x	
Household size	x			
Age	x	x	x	x
Income	x			
Value added services fixed phone line				

x: variable has been used as splitting variable in the tree.

relationship age ISP and connection speed, have a stable effect on churn, since they appear in nearly all the trees in all periods.

Model Variability

There are three possible explanations for the changes in the models that we estimate: multicollinearity, omitted variables, and actual changes in the situation that we model. To check whether the data suffers from multicollinearity we calculated the condition indices. All indices are smaller than 32 (they range from 1 to 21) and hence there is no severe problem with multicollinearity (Gujarati 2003). With respect to the omitted variable problem we acknowledge that we do not take information like the market environment and customer attitudes into account. However, we argue that we included all variables that are both relevant in this situation and very common in database marketing studies, see Table 2. Therefore, the most plausible explanation is that the situation that we model changes over time due to changes in the environment (i.e. increasing price competition). These environmental changes cannot be included in standard churn models, which predict churn at a specific point in time using database data. For that purpose dynamic churn models should be developed (Leeflang et al. 2009).

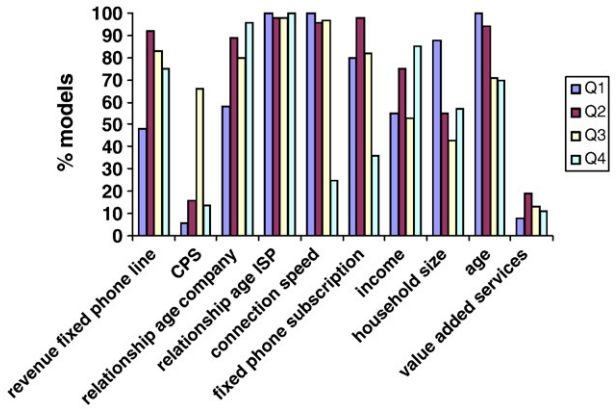

Fig. 3. Fraction of the 100 classification trees in the bagging procedure in which variables are used as a splitting variable (ISP data).

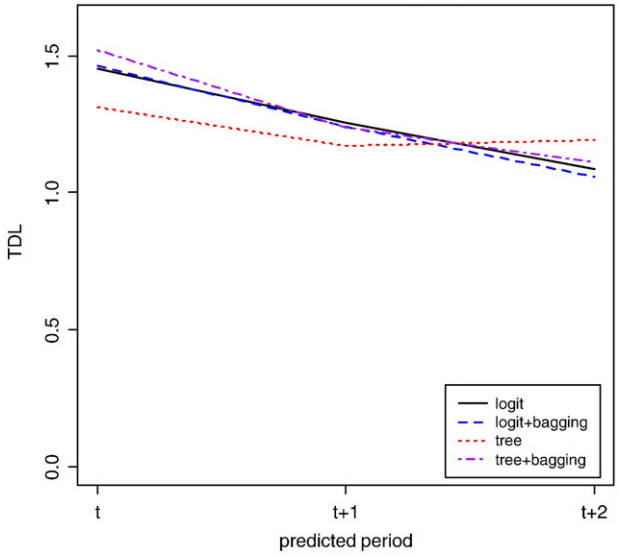

Fig. 4. Average top-decile lifts of models estimated at time t (insurance data).

Insurance Data

Staying Power: Top-decile Lift

Fig. 4 shows the average top-decile lifts of the four different models. We again aggregated the results for the sake of clarity. As was the case for the ISP data the lines are downward sloping except for the tree-line between $t+1$ and $t+2$. A possible explanation could be that the model is too simple and captures only a few main effects, since the model performs the worst in period t and $t+1$, but performs slightly better than the other models in $t+2$. With respect to applying the bagging procedure we again observe that the logit models do not benefit, since the two lines overlap in Fig. 4. However, the predictive performance of the classification trees improves due to the bagging procedure. Both in period t and $t+1$ the line of the tree in

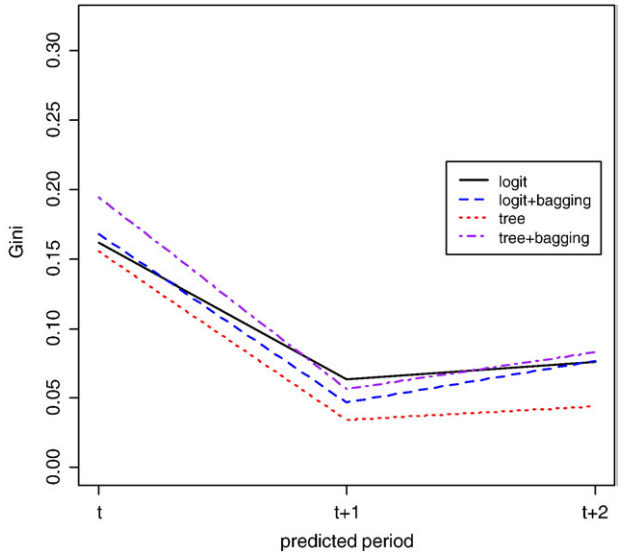

Fig. 5. Average Gini coefficients of models estimated at time t (insurance data).

combination with a bagging procedure is above the line of the single tree.

Staying Power: Gini Coefficient

In Fig. 5 the average Gini coefficients of the four models are depicted. Here, a steep decrease can be observed between period t and $t+1$, which indicates a substantial decrease in overall model performance. After reaching a rather low level of about .05 the curve flattens out between $t+1$ and $t+2$. Apart from the shape of the decrease the findings are similar to what we found for the top-decile lifts.

Parameter Assessment

The parameters of the single logit models are shown in Table 5. The sign, size, and significance of the estimates vary substantially over time. Only one of the parameters is significant in all three periods; the unknown family configuration group has a higher probability to churn than customers from the other groups. Three parameters (13%) are significant and have the same sign in two periods. Age, relationship length, and the moving indicator all have a negative effect on churn. Furthermore, five parameters have a significant effect in only one period. Finally, three of the package type dummies have a significant effect on churn in two periods. However, the effect is negative in 2004 and positive in 2006. Again, this illustrates the low parameter stability of the model.

Table 5
Parameter estimates of the single logit models (insurance data).

Variable	Period		
	2004	2005	2006
Age (years)	−.0104 **	−.0004	−.0144 **
Relationship length (years)	−.3625 **	−.2248 **	−.1116
Package type (ref. cat. '0')			
1	.0955	.8462	−.5464
2	−.0216	.4145	.5371
3	−.7110 **	−.2046	.2680
4	−.9320 **	−.1590	.2495
5	−.9182 **	−.2868	.6868 *
6	−.8702 **	−.2811	.5635 *
7	−.9738 **	−.0779	.7244 *
8	-1.1030 **	.1386	.4352
Family configuration (ref.cat. 'single')			
No kids	.1086	−.1515	.8324 **
Kids	.1337	−.0048	−.3571
Family1	−.1053	−.3456	.3325
Family2	.2331	.0938	.2395
Unknown	.6019 **	.3903 **	.9138 **
Income (ref. cat. 'unknown')			
>2 times standard	.3756	.0737	−.0448
Standard-2 times standard	−.0415	.0770	.1938
Standard income	−.0917	.0553	−.2693
Minimum–standard income	−.1248	−.0318	−.4017 *
Minimum	.0497	−.0822	−.7447 *
Variable	−.3054	−.1121	.0525
Collectively insured	−.1912	−.1767	−.4183 *
Moved	-3.7922 **	−.0730	−.5359 **

* $p<.05$.
** $p<.01$.

Table 6 shows the splitting variables of the classification trees for all three periods. The results show that only one variable (age) is used as a splitting variable in all three periods. This variable had a significant effect in the logit model for two out of the three periods and can hence be considered as a relatively important predictor of churn. Furthermore, four of the variables appear in only one of three trees.

Table 6
Splitting variables in the estimated classification trees (insurance data).

Variable	Period		
	2004	2005	2006
Age	x	x	x
Relationship length	x	x	
Moved	x		
Package type			x
Family configuration			x
Income		x	
Collectively insured			

x: variable has been used as splitting variable in the tree.

In Fig. 6 a summary of the classification trees in combination with the bagging procedure is provided. As was the case for the ISP data, a large diversity in the frequencies can be observed. Moreover, in this case none of the variables appears in a large proportion of the trees in all periods.

Model Variability

Likewise as in the ISP case, there are three possible explanations for the changes found in the estimated models. Again, we have no reason to believe that the data suffers from severe multicollinearity; all condition indices are well below 32 (range from 1 to 12). Like in the ISP case we include commonly used predictors in churn models. We suspect that the changes in the model mainly occur due to changes in the environment, which are again not captured in the currently used churn models. Again this would pledge for the inclusion of more dynamics in churn models, which is currently not done.

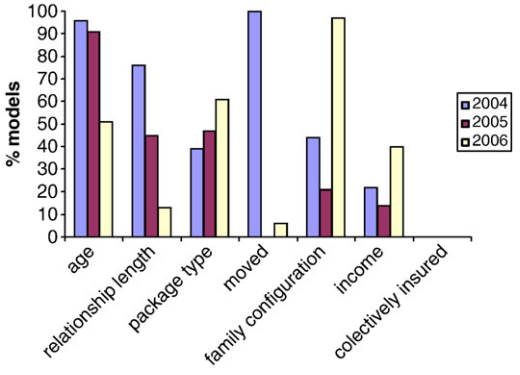

Fig. 6. Fraction of the 100 classification trees in the bagging procedure in which variables are used as a splitting variable (insurance data).

Conclusion

In this paper, we contribute to the existing literature on churn prediction models by studying the staying power of frequently used and well-performing prediction models. Furthermore, we tested the predictive performance of frequently used methods (i.e. logistic regression, trees and bagging) in two industries. Specifically we analyzed customer data of an ISP and a health insurance company. We evaluated the results of a logistic regression model, a classification tree, and of both in combination with a bagging procedure estimated on a number of consecutive periods.

The general conclusions of our model comparison are as follows:

- Confirming prior studies (i.e. Lemmens and Croux 2006) our study shows that overall classification trees combined with a bagging procedure provide the best predictive performance for all studied time periods. These findings are stronger for the ISP data which is probably due to the size of the data sets; trees tend to perform better on larger data sets (Perlich, Provost, and Simonoff 2004).
- The predictive quality of the investigated models declines over time. A substantial decrease in predictive quality is found in period $t+2$ for the ISP data and in period $t+1$ for the insurance data. This indicates that the staying power of these models is very limited.
- Although the bagging procedure improves the predictive power of classification trees, there is no strong evidence that this procedure improves the staying power. The predictive performance declines similarly for all studied models.

The limited staying power of the models implies that models cannot be used for a long time period in this specific setting. Both studies indicate that a churn model should be used for a maximum of one period subsequent to the estimation period; for the prediction of churn in period $t+2$ new models should be built. Simply updating a churn prediction model will not be sufficient to obtain reliable estimates; the model building procedure should start with selection of the important variables. The benefits of this new model development are substantial. In our empirical studies, using a more recent model leads on average to an increase of 20% in the number of churners in the predicted top-decile.

The limited staying power of the studied churn prediction models illustrates that assuming a constant churn probability for CLV calculations is a risky strategy; churn predictions and hence CLV predictions become very unreliable in the longer term. This is in line with the findings of Malthouse and Blattberg (2005). This can potentially have strong implications for CLV-based marketing resource allocation strategies (Donkers, Verhoef, and De Jong 2007; Venkatesan and Kumar 2004; Zeithaml, Rust, and Lemon 2001). Our results suggest that these allocation models should also be updated regularly.

The need for regular re-estimation illustrates the importance of automation of the modeling process; this would increase the model building efficiency and thus lower the costs. However, the estimation procedure is complicated by the required variable selection. Therefore, to automate the churn modeling process, implementation of advanced model building tools is essential.

Limitations and Suggestions for Future Research

Although we are confident with the results, it remains unclear whether they are generalizable over a broader range of services than the two we studied. Therefore, it would be interesting to validate our findings on customer databases of other services. This would reveal whether the limited staying power and the instability of the parameters is typical for the industries studied or whether these findings hold for other service sectors as well. Unfortunately, we did not have access to such data sets.

A second valuable extension of this study would be to analyze more periods than the maximum of four we used. A longitudinal data set containing a large number of periods would allow us to estimate time-varying parameters and possibly seasonality effects. It could be the case that there exists a certain pattern in the churn behavior of customers that could only be observed over a longer time period consisting of at least multiple years of data.

One additional avenue for further research is the development of more dynamic churn prediction models. The currently used models are not suited for the inclusion of dynamic changes in the customer base, market environment etc. in the model. This may explain, why the parameter estimates of our studied models are not stable over time. We therefore urge researchers to take a next step in churn modeling and to develop models that include more dynamics (see also Leeflang et al. 2009).

Finally, our results indicate that the predictive performance of models depends on the characteristics of the data. Hence, more research is needed within marketing to assess under what circumstances different types of prediction models perform best.

Acknowledgments

We thank a Dutch telecommunications company for providing the data, and Aurélie Lemmens for providing the S-code for the bagging algorithm. We thank Jenny van Doorn for her helpful comments and Jaap Wieringa for sharing data. We thank the editor Ed Malthouse and two anonymous reviewers for their helpful comments.

References

Athanassopoulos, Antreas D. (2000), "Customer Satisfaction Cues to Support Market Segmentation and Explain Switching Behavior," *Journal of Business Research*, 47, 3, 191–207.

Blattberg, Robert C., Byung-Do Kim, and Scott A. Neslin (2008), *Database Marketing: Analyzing and Managing Customers*. New York: Springer Science+Business Media.

———, Edward C. Malthouse, and Scott A. Neslin (2009), "Customer Lifetime Value: Empirical Generalizations and some Conceptual Questions," *Journal of Interactive Marketing*, 23, 2, 157–68.

Bolton, Ruth 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.

———, P.K. Kannan, and Matthew D. Bramlett (2000), "Implications of Loyalty Program Membership and Service Experiences for Customer Retention and Value," *Journal of the Academy of Marketing Science*, 28, 1, 95–108.

———, Katherine N. Lemon, and Peter C. Verhoef (2004), "The Theoretical Underpinnings of Customer Asset Management: A Framework and Propositions for Future Research," *Journal of the Academy of Marketing Science*, 32, 3, 271–92.

Breiman, Leo, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone (1984), *Classification and Regression Trees*. Belmont: Wadsworth.

——— (1996a), "Bagging Predictors," *Machine Learning*, 24, 2, 123–40.

——— (1996b), "Heuristics of Instability and Stabilization in Model Selection," *Annals of Statistics*, 24, 6, 2350–83.

Buckinx, Wouter and Dirk Van den Poel (2005), "Customer Base Analysis: Partial Defection of Behaviourally Loyal Clients in a Non-contractual FMCG Retail Setting," *European Journal of Operational Research*, 164, 1, 252–68.

Coussement, Kristof and Dirk Van den Poel (2008), "Churn Prediction in Subscription Services: An Application of Support Vector Machines while Comparing Two Parameter-selection Techniques," *Expert Systems with Applications*, 34, 1, 313–27.

Deichmann, Joel, Abdolreza Eshghi, Dominique Haughton, Selin Sayek, and Nicholas Teebagy (2002), "Application of Multiple Adaptive Regression Splines (MARS) in Direct Response Modeling," *Journal of Interactive Marketing*, 16, 4, 15–27.

Dijksterhuis, Marc and Steef Velders (2009), "Predicting Switching Behavior in a Market with Low Mobility: A Case Study," in *Developments in Market Research 2009*, A. E. Bronner, ed. Haarlem: Spaar en Hout, 167–80.

Donkers, Bas, Philip H. Franses, and Peter C. Verhoef (2003), "Selective Sampling for Binary Choice Models," *Journal of Marketing Research*, 40, 4, 492–7.

———, Peter C. Verhoef, and Martijn De Jong (2007), "Modeling CLV: A Test of Competing Models in the Insurance Industry," *Quantitative Marketing and Economics*, 5, 2, 163–90.

Drèze, Xavier and André Bonfrer (2008), "An Empirical Investigation of the Impact of Communication Timing on Customer Equity," *Journal of Interactive Marketing*, 22, 1, 36–50.

Fader, Peter S. and Bruce G.S. Hardie (2007), "How to Project Customer Retention," *Journal of Interactive Marketing*, 21, 1, 76–90.

Franses, Philip H. and Richard Paap (2001), *Quantitative Models in Marketing Research*. Cambridge: Cambridge University Press.

Gruen, Thomas W., John O. Summers, and Frank Acito (2000), "Relationship Marketing Activities, Commitment, and Membership Behaviors in Professional Associations," *Journal of Marketing*, 64, 3, 34–49.

Gujarati, Damodar N. (2003), *Basic Econometrics*. New York: McGraw-Hill/Irwin.

Gupta, Sunil, Donald R. Lehmann, and Jennifer A. Stuart (2004), "Valuing Customers," *Journal of Marketing Research*, 41, 1, 7–18.

——— (2009), "Customer-based Valuation," *Journal of Interactive Marketing*, 23, 2, 169–78.

Ha, Kyoungnam, Sungzoon Cho, and Douglas MacLachlan (2005), "Response Models Based on Bagging Neural Networks," *Journal of Interactive Marketing*, 19, 1, 17–30.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2008), *The Elements of Statistical Learning*. New York: Springer Science+Business Media.

Haughton, Dominique and Samer Oulabi (1993), "Direct Marketing Modeling with CART and CHAID," *Journal of Direct Marketing*, 11, 4, 42–52.

Hwang, Hyunseok, Taesoo Jung, and Euiho Suh (2004), "An LTV Model and Customer Segmentation Based on Customer Value: A Case Study on the Wireless Telecommunication Industry," *Expert Systems with Applications*, 26, 2, 181–8.

King, Ross D., C. Feng, and Alistair Sutherland (1995), "Statlog — Comparison of Classification Algorithms on Large Real-world Problems," *Applied Artificial Intelligence*, 9, 3, 289–333.

Kumar, Akhil, Vithala R. Rao, and Harsh Soni (1995), "An Empirical Comparison of Neural Network and Logistic Regression Models," *Marketing Letters*, 6, 4, 251–63.

Larivière, Bart and Dirk Van den Poel (2005), "Predicting Customer Retention and Profitability by Using Random Forests and Regression Forests Techniques," *Expert Systems with Applications*, 29, 2, 472–84.

Leeflang, Peter S.H., Dick R. Wittink, Michel Wedel, and Philippe A. Naert (2000), *Building Models for Marketing Decisions*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

———, Tammo H.A. Bijmolt, Jenny van Doorn, Dominique M. Hanssens, Harald J. van Heerde, Peter C. Verhoef, and Jaap E. Wieringa (2009), "Creating Lift versus Building The Base: Current Trends in Marketing Dynamics," *International Journal of Research in Marketing*, 26, 1, 13–20.

Lemmens, Aurélie and Christophe Croux (2006), "Bagging and Boosting Classification Trees to Predict Churn," *Journal of Marketing Research*, 43, 2, 276–86.

Lemon, Katherine N., Tiffany B. White, and Russell S. Winer (2002), "Dynamic Customer Relationship Management: Incorporating Future Considerations into the Service Retention Decision," *Journal of Marketing*, 66, 1, 1–14.

Levin, Nissan and Jacob Zahavi (2001), "Predictive Modeling using Segmentation," *Journal of Interactive Marketing*, 15, 2, 2–22.

Lim, Tjen-Sien, Wei-Yin Loh, and Yu-Shan Shih (2000), "A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms," *Machine Learning*, 40, 3, 203–28.

Malthouse, Edward C. (1999), "Ridge Regression and Direct Marketing Scoring Models," *Journal of Interactive Marketing*, 13, 4, 10–23.

——— and Robert C. Blattberg (2005), "Can We Predict Customer Lifetime Value?," *Journal of Interactive Marketing*, 19, 1, 2–16.

——— and Kirstin M. Derenthal (2008), "Improving Predictive Scoring Models through Model Aggregation," *Journal of Interactive Marketing*, 22, 3, 51–68.

Mittal, Vikas and Wagner A. Kamakura (2001), "Satisfaction, Repurchase Intent, and Repurchase Behavior: Investigating the Moderating Effect of Customer Characteristics," *Journal of Marketing Research*, 38, 1, 131–42.

Neslin, Scott A., Sunil Gupta, Wagner Kamakura, Lu Junxiang, and Charlotte H. Mason (2006), "Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models," *Journal of Marketing Research*, 43, 2, 204–11.

Perlich, Claudia, Foster Provost, and Jeffrey S. Simonoff (2004), "Tree Induction vs. Logistic Regression: A Learning-curve Analysis," *Journal of Machine Learning Research*, 4, 2, 211–55.

Pfeifer, Phillip E. and Paul W. Farris (2004), "The Elasticity of Customer Value to Retention: The Duration of a Customer Relationship," *Journal of Interactive Marketing*, 18, 2, 20–31.

Prins, Remco and Peter C. Verhoef (2007), "Marketing Communication Drivers of Adoption Timing of a New E-service among Existing Customers," *Journal of Marketing*, 71, 2, 169–83.

Rust, Roland T. and Anthony J. Zahorik (1993), "Customer Satisfaction, Customer Retention, and Market Share," *Journal of Retailing*, 69, 2, 193–215.

——— and Chung T. Siong (2006), "Marketing Models of Service and Relationships," *Marketing Science*, 25, 6, 560–80.

Venkatesan, Rajkumar and V. Kumar (2004), "A Customer Lifetime Value Framework for Customer Selection and Resource Allocation Strategy," *Journal of Marketing*, 68, 4, 106–25.

Verhoef, Peter C. (2003), "Understanding the Effect of Customer Relationship Management Efforts on Customer Retention and Customer Share Development," *Journal of Marketing*, 67, 4, 30–45.

———, Penny N. Spring, Janny C. Hoekstra, and Peter S.H. Leeflang (2003), "The Commercial use of Segmentation and Predictive Modeling Techniques for Database Marketing in the Netherlands," *Decision Support Systems*, 34, 4, 471–81.

———, Rajkumar Venkatesan, Leigh McAllister, Edward C. Malthouse, Manfred Krafft, and Shankar Ganesan (2010), "On CRM in Data Rich Multi-channel Retailing Environments," *Journal of Interactive Marketing*, 24, 2.

Xie, Yaya, Xiu Li, E.W.T. Ngai, and Weiyun Ying (2009), "Customer Churn Prediction using Improved Balanced Random Forests," *Expert Systems with Applications*, 36, 3, 5445–9.

Zahavi, Jacob and Nissan Levin (1997), "Applying Neural Computing to Target Marketing," *Journal of Direct Marketing*, 11, 1, 5–22.

Zeithaml, Valerie A., Roland T. Rust, and Katherine N. Lemon (2001), "The Customer Pyramid: Creating and Serving Profitable Customers," *California Management Review*, 43, 4, 118.