Available online at www.sciencedirect.com

ScienceDirect

Expert Systems with Applications 32 (2007) 125–134

Expert Systems with Applications

www.elsevier.com/locate/eswa

Predicting customer loyalty using the internal transactional database

Wouter Buckinx, Geert Verstraeten *, Dirk Van den Poel

Department of Marketing, Ghent University, Hoveniersberg 24, 9000 Ghent, Belgium

Abstract

Loyalty and targeting are central topics in Customer Relationship Management. Yet, the information that resides in customer databases only records transactions at a single company, whereby customer loyalty is generally unavailable. In this study, we enrich the customer database with a prediction of a customer's behavioral loyalty such that it can be deployed for targeted marketing actions without the necessity to measure the loyalty of every single customer. To this end, we compare multiple linear regression with two state-of-the-art machine learning techniques (random forests and automatic relevance determination neural networks), and we show that (i) a customer's behavioral loyalty can be predicted to a reasonable degree using the transactional database, (ii) given that overfitting is controlled for by the variable-selection procedure we propose in this study, a multiple linear regression model significantly outperforms the other models, (iii) the proposed variable-selection procedure has a beneficial impact on the reduction of multicollinearity, and (iv) the most important indicator of behavioral loyalty consists of the variety of products previously purchased.

Keywords: Predictive modeling; Customer relationship management; Behavioral loyalty; Overfitting; Multicollinearity; Data enrichment

1. Introduction

In the two latest decades, Customer Relationship Management (CRM) has grown to be one of the major trends in marketing, both in academia and in practice. This evolution took form in a dramatic shift in the domain, evolving from transaction-oriented marketing to relationship-oriented marketing (Grönroos, 1997), and builds strongly on the belief that it is several times less demanding – i.e. expensive – to sell an additional product to an existing customer than to sell the product to a new customer (Rosenberg & Czepiel, 1984). Hence, it has been argued that it is particularly beneficial to build solid and fruitful customer relationships, and in this discourse, customer loyalty has been introduced as one of the most important concepts in marketing (Reichheld, 1996).

From an analytical point of view, several tools have emerged in recent years that enable companies to strengthen their relationships with customers. Moreover, the rise of new media such as the World Wide Web, and the continuous technological improvements have further increased the opportunities to communicate in a more direct, one-to-one manner with customers (Van den Poel & Buckinx, 2005). Response modeling – i.e. predicting whether a customer will reply to a specific offer, leaflet or product catalog – represents the most central application in this domain, and serves as a tool to manage customer relationships. Indeed, it would be beneficial for the company–customer relationship that the latter party would receive only information that is relevant to him/her, hence allowing the company to present only those offers for which the individual customer shows a high response probability (Baesens, Viaene, Van den Poel, Vanthienen, & Dedene, 2002). Related to this, cross-selling analysis is involved with finding the optimal product to offer to a given customer (Chintagunta, 1992; Larivière & Van den Poel, 2004). Additionally, upselling analysis is focused on selling more – or a more expensive version – of the products that are currently purchased by the customer. Both techniques share a similar goal, i.e. to intensify the customer relationship by raising the share of products that is

* Corresponding author. Tel.: +32 9 264 35 24; fax: +32 9 264 42 79.
E-mail address: Geert.Verstraeten@UGent.be (G. Verstraeten).

0957-4174/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2005.11.004

purchased at the focal company, and to prevent that these products would be purchased at competitive vendors. The fear of losing sales to competitors also features in churn analysis, which is focused on detecting customers exhibiting a large potential to abandon the existing relationship. Churn analysis has received great attention in the domain ever since it has been proven that even a small improvement in customer defection can greatly affect a company's future profitability (Reichheld & Sasser, 1990; Van den Poel & Larivière, 2004). Finally, lifetime value (LTV) analysis is a widely used technique to predict the future potential of customers, in order to target only the most promising customers (Hwang, Jung, & Suh, 2004). While these techniques can each serve individually to enhance customer relationships, it should be clear that additional advantages reside in the combination of these analytic techniques. Two recent attempts to integrate such techniques can be found in Baesens et al. (2004) and Jonker, Piersma, and Van den Poel (2004).

2. The need for predicting customer loyalty

In sum, we could state that both the focus on customer loyalty and the analytic tools described above have emerged from the CRM discourse. However, it is very unusual that actual customer loyalty is used to either devise or evaluate a company's targeted marketing strategies. The major cause of this deficiency lies most likely in the unavailability of information. Currently, while companies are maintaining transactional databases that store all details on any of a given customer's contacts with the focal company, these databases cannot capture the amount of products that this customer purchases at competing stores. Indeed, a study by Verhoef, Spring, Hoekstra, and Leeflang (2002) showed that only 7.5% of companies involved in database marketing activities collect such purchase behavior. Hence, the real behavioral loyalty of a certain customer is generally unavailable in the company's records, whereby the full potential of the customer (i.e., the total needs of the customer for products in the relevant category) is unknown to any specific company. However, this information could prove to be extremely valuable in different applications.

First, the knowledge of a customer's loyalty would be useful for improving CRM. We illustrate this with an example from a banking context. It would most likely be more lucrative to offer an additional savings product to a customer who has a high balance at the focal bank and at the same time has large amounts invested at other banking institutions, than to offer the savings product to a customer that has an equally high balance, but where all his/her money is invested at the focal bank. Secondly, a notion of a customer's loyalty could be used for adapting the usefulness of the model-building process. For example, currently, cross-selling models are being built on the total customer database, whereby the users will estimate the probability of purchasing this product *at the focal company*, whereas from a cross-sales point of view, it would be more interesting to estimate whether they are interested in the product category *in general*. To overcome this, it could be interesting to build a cross-selling model on loyal customers only, because only for these customers, their total product needs are known. In this context, when attempting to model the real – and total – product needs of customers, it might seem suboptimal to include nonloyal customers into the analysis. Thirdly, the knowledge of a customer's loyalty and the evolution therein could be useful for evaluating the results of CRM-related investments, and monitoring whether certain actions lead to the desired results in the relevant customer segments.

While such loyalty information can be obtained through a questionnaire, it would prove to be financially infeasible to obtain this information for each individual customer, especially when customers would have to be surveyed regularly in order to track changes in their loyalty profile. Consequently, in this paper, we will prove that it is

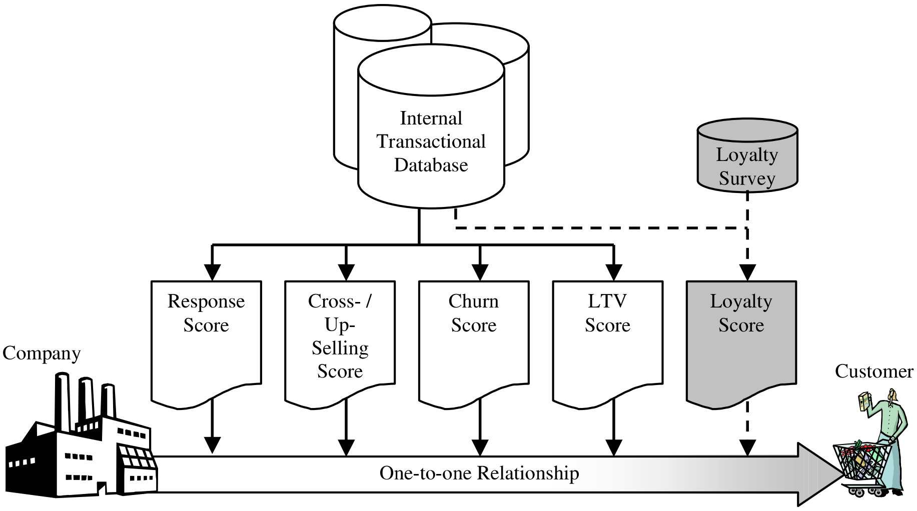


Fig. 1. Creating a loyalty score from transactional data and a loyalty survey.

sufficient to survey a sample of the company's customers, since we will combine the information stemming from the survey and the internal transactional database in order to create a loyalty score for all individual customers. Hence, as summarized in Fig. 1, this score could provide additional information to the scores based on the transactional data only, and form a valuable expert tool for managing customer relationships.

The remainder of this paper is structured as follows. The next section covers the methodology used, and focuses on a description of the applied predictive techniques, the need for adequate cross-validation, and the variable-selection procedure we propose. Next, we will describe the data used for this study. In a subsequent section, we discuss the results of the proposed predictive modeling study. Finally, we end the paper with a section covering the conclusions and directions for further research.

3. Methodology

3.1. Predictive techniques

Technically, in this study, we will predict this loyalty for customers that do not belong to the surveyed sample by use of the data that is available for all customers, i.e. the transactional data. In essence this is a problem of predictive modeling. It is not our ambition to compare all possible predictive techniques. Instead, we will compare three techniques that show interesting differences and similarities. Because of the need for an accurate prediction as well as an understanding of the model – in order to explain the findings to management – we only considered models that were expected to (i) deliver adequate predictive performance on a validation set and (ii) provide an insight into the most important variables in the model. As a benchmark predictive technique, we have used a multiple linear regression (MLR) model (Cohen & Cohen, 1983), because of the widespread usage of this statistical technique in industry and academia. We compared this benchmark with two state-of-the-art techniques from the machine learning and data mining domain. First, given the widespread use of decision trees in prediction problems where the user seeks insight into the predictive process, we have implemented Random Forests (RF, Breiman, 2001). This technique focuses on growing an ensemble of decision trees using a random selection of features to split each node (i.e. the random subspace method), where the final prediction is computed as the average output from the individual trees. RF models have been argued to possess excellent properties for feature selection, and to avoid overfitting given that the number of trees is large (Breiman, 2001). In this approach, we will grow 5000 trees, as in other applications (e.g. Geng, Cosman, Berry, Feng, & Schafer, 2004). Finally, since Artificial Neural Networks (ANNs) have often been credited for achieving higher predictive performance, we selected MacKay's Automatic Relevance Determination (ARD, MacKay, 1992) neural network because it additionally reveals a Bayesian hyperparameter per input variable, representing the importance of the variable. To this end, the relevance of the features is detected by maximizing the model's marginal likelihood. We respected the author's view that a large number of hidden units should be considered in order to build a reliable model. The use of the ARD model is made possible using Markov Chain Monte Carlo techniques, hence avoiding overfitting due to the use of a Bayesian 'Occam's razor' while allowing an interpretation of the variables' importance (MacKay, 1992).

3.2. Cross-validation

An important early topic in predictive modeling consists in validating the predictive power of a model on a sample of data that is independent of the information used to build the model. In this study, the limited number of observations in each of the two settings and the elaborate number of independent variables make it hard to split our data in an estimation and a hold-out validation set. As a consequence, we prefer a resampling method called leave-one-out cross-validation because it proves to be superior for small datasets (Goutte, 1997). Using this procedure, our data are divided into k subsets, where k is equal to the total number of observations. Next, each of the subsets is left out once from the estimation set and is then used to perform a validation score. To compute the real-life power of the model, the final validation set is built by stacking together the k resulting validations and the predictive performance is computed on this stacked set. The performance of the model – on the estimation set as well as on the validation set – is evaluated by computing (i) the correlation between surveyed loyalty and its prediction, (ii) R^2, (iii) adjusted R^2, (iv) Mean Squared Error (MSE) and (v) the Root of the MSE (RMSE).

3.3. Variable selection

In the current study, it is likely that we can compute a large number of database-related variables in comparison with the number of observations (i.e. the number of respondents of this questionnaire). While both the RF and ARD models claim to avoid overfitting, this effect does provide a reasonable threat to the multiple regression model (Cohen & Cohen, 1983). To overcome this problem, we will make use of a variable-selection technique. Thanks to this method, the dimensionality of the model can be reduced and redundant variables are removed, which is in favor of the model's performance. Additionally, a variable-selection procedure will allow us to gain insight in selecting the variables with good predictive capacities, and allows us to interpret the parameter estimates due to a plausible reduction of multicollinearity.

Fig. 2 partitions the variable-selection procedure that was used in this study into six disjoint steps. In step (i), we apply the leaps-and-bounds algorithm proposed by Furnival and Wilson (1974) on the estimation set. Their

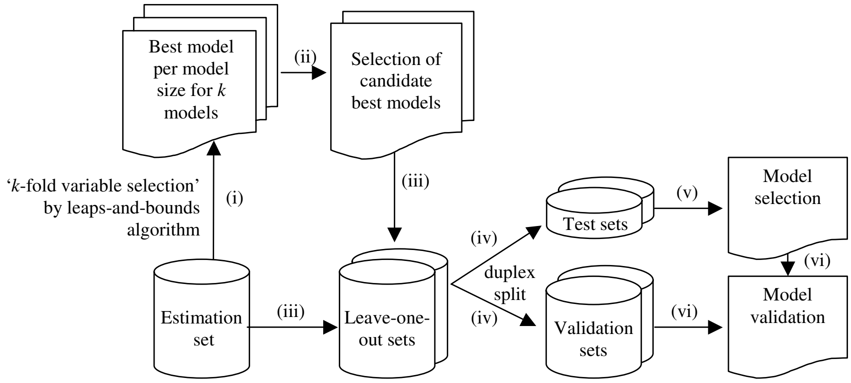


Fig. 2. Model selection and validation for the multiple linear regression model.

efficient technique identifies the model with the largest adjusted R^2 for any given model size (i.e. starting from the best model with only one variable to the full model) and at the same time avoids a full search of the variable space. However, because of the leave-one-out procedure described previously, in this case, we cannot simply perform this procedure on the total estimation set. Indeed, in order to allow for a validation of the model, the estimated models should be built when at least one observation is set aside for validation. Since it would be suboptimal to select this observation randomly, in this study we propose an iterative process in which we set aside one observation at a time, such that we create k new estimation sets, where k equals the total number of observations in the original estimation set. Hence, the outcome of this procedure – to which we refer as 'k-fold variable selection' – will consist in a list of k best models per model size. Next, in step (ii) to ensure tractability and to avoid the choice of selecting an unstable model, we reduce this list by selecting, per model size, only those models that were 'winners' in at least 5% of the occasions. In step (iii), we create the leave-one-out predictions for each candidate model using the procedure described in the previous paragraph. In the following steps, we are concerned with selecting the best models, and validating the performance of these models. Because of this dual need, in step (iv) we divide the leave-one-out dataset per candidate model into a test set containing 25% of the observations, that will be used for model selection; and a validation set consisting of the remaining 75% of the observations, that will be used for detecting the real predictive performance of the model. Considering both the importance of a good split and the low number of observations available, we do not perform a random split, but rather complete the division via the Duplex algorithm (Snee, 1977), which performs best in separating a dataset into two sets covering approximately the factor space. Concretely, here, this factor space is composed of the set of independent variables created for the study. Next, in step (v), based on the leave-one-out test set performance, we select the best-performing model per model size among the selection of candidate models. Additionally, we select the model with the highest overall performance. In the final step (vi), we validate the real predictive performance of the models selected in the previous step on the unseen data.

4. Data description

We use data from two retail stores belonging to the same large European chain which were considered, according to management, to be representative for the entire chain. The stores carried a product assortment normally associated with grocery stores (e.g., food and beverages, cosmetics, laundry detergents, household necessities). Detailed purchase records were tracked for a period of 51 months and a summarized customer table was available that tracked basic customer demographics as well as date of first purchase.

4.1. Computation of database-related variables

It is important to mention that all transactions could be linked to customers, as the store requires use of a customer identification card. In total, 35 independent variables are computed, that are related to the following topics: (i) monetary spending, (ii) frequency of purchasing, (iii) recency of last purchase, (iv) length of the customer–company relationship, (v) interpurchase time, (vi) returns of goods, (vii) purchase variety, (viii) promotion sensitivity, (ix) responsiveness on mailings and (x) distance to the store. The inclusion of these variables was mainly based on previous literature in the domain of predicting the strength of the relationship between a company and its customers (see, e.g., Baesens et al., 2004; Buckinx & Van den Poel, 2005; Bult & Wansbeek, 1995; Reinartz & Kumar, 2000; Srinivasan, Anderson, & Ponnavolu, 2002). Table 1 summarizes all these variables, together with a brief description of how they are calculated.

4.2. Loyalty survey

In addition to these transactional data, a self-administered survey was used as a complementary data collection method. Data collection took place in each of the retail

Table 1
Description and predictive performance of variables used

Variable	Description	MLR	RF	ARD
		Standardized parameter estimates	Variable importance	Alpha (importance)
Spending_1M	Spending during last month	0.3540***	0.0086	21.49
Spending_6M	Spending during last six months	0.4582***	0.1136	13.00
Spending_1Y	Spending during last year	0.4789***	0.2246	15.58
Spending_2Y	Spending during last two years	0.4742***	0.0228	23.63
Spending	Spending in total history	0.4714***	0	32.08
NumItems	Number of product items bought	0.4705***	2.3071	16.27
Spending_Fresh	Spending on fresh food products	0.4395***	0.2985	17.52
rSpend_Freq	Average Spending per visit	0.1785***	0.0055	7.11
rSpend_Lor	Spending relative to the length of the customer's relationship	0.4726***	0.4104	0.16
Frequency_1M	Number of purchases during last month	0.3477***	0	2.41
Frequency_6M	Number of purchases during last six months	0.4356***	0.035	3.76
Frequency_1Y	Number of purchases during last year	0.4455***	0.0544	3.91
Frequency_2Y	Number of purchases during last two years	0.4494***	0	2.77
Frequency	Number of purchases in total history	0.4389***	0	3.87
Recency	Number of days since last purchase	−0.2035***	0	24.44
Ipt	Average number of days between store visits	−0.2965***	0.6045	17.23
Std_Ipt	Standard deviation of the number of days between the purchases	−0.3227***	0.292	13.20
Lor	Length of customer relationship	0.0940***	0	29.92
Numcat_LY	Number of different product categories purchased from during last year	0.5221***	0.543	6.32
Numcat_2Y	Number of different product categories purchased from during last two years	0.4770***	0.2001	3.09
Numcat_3Y	Number of different product categories purchased from during last three years	0.4460***	0.1434	5.27
Numcat	Number of different product categories purchased from during the total history	0.4805***	0.2233	10.28
Neg_Inv	Dummy to indicate if the customer ever had a negative invoice (1/0)	0.2919***	0.1115	2.42
Ret_Item	Dummy to indicate if the customer ever returned an item (1/0)	0.2656***	0.0293	1.56
Returns	Total value of returned goods	0.1572***	0	11.90
NumPromItems	Number of items bought that appeared in company's promotion leaflet	0.4539***	0.9065	9.62
SpenPromItems	Money spent on products that appeared in promotion leaflet	0.4572***	0.0064	11.79
Visitspromitems	Number of visits on which a product is bought that appeared in the promotion leaflet	0.4680***	0.0342	5.35
PercNumPromItems	Percentage of products bought that appeared in leaflet	0.0139	0.06	8.48
PercResp_Leaf	Percentage of times a purchase is made given that a promotion leaflet was received	0.4792***	0	0.22
PercResp_Noleaf	Percentage of times a purchase is made given that no promotion leaflet was received	0.3098***	0	1.32
Perc_Noleaf_Freq	PercResp_Noleaf divided by shopping frequency	−0.2235***	0.1258	2.57
MoreThanOnce	Number of times that a customer visits more than once within the same promotion period	0.4308***	0	2.92
PercMoreThanOnce	MoreThanOnce divided by the number of times a customer bought in a promotion period	0.2940***	0	0.34
Distance	Distance to the store	−0.1265***	0.0457	6.07

*** $p < .01$.

stores mentioned previously. Surveys were randomly distributed to customers during their shopping trips, and customer identification numbers were recorded for all customers who received a questionnaire.

A customer's behavioral loyalty was determined as a composite measure by comparing a customer's spending at the retailer with their total spending in the relevant product category. As a first item, and similar to Macintosh and Lockshin (1997), the percentage of purchases made in the focal supermarket chain versus other stores was assessed on an 11-point scale that ranged from 0% to 100% in 10% increments (i.e., 0%, 10%, 20%, and so on). Additionally,

Table 2
Wording of the items of the loyalty scale

Item 1	Buy (much less … much more) grocery products at XYZ than at competing stores
Item 2	Visit other stores (much less frequently … much more frequently) than XYZ for your grocery shopping (–)
Item 3	Spend (0% … 100%) of your total spending in grocery shopping at XYZ

two seven-point Likert-type items assessed the shopping frequency of the customers for the focal store when compared to other stores. We pretested the questionnaire and refined it on the basis of pretest results. Table 2 gives the exact wording of the items used. After rescaling the second item (due to its expected negative correlation with both other items), we standardized the three loyalty-related questions, and averaged them to represent the behavioral loyalty construct.

5. Results

5.1. Survey response

Of the 1500 distributed questionnaires, we received 878 usable responses (i.e. a ratio of usable response of 58.33%). We successfully tested for nonresponse bias by comparing database variables such as spending, frequency of visiting the store, interpurchase time, length-of-relationship and response behavior towards companies' mailings between respondents and nonrespondents.

A usable response had all fields completed, and the respondent could be successfully linked to his or her transaction behavior in the customer database. We tested construct reliabilities of the loyalty scale by means of Cronbach's coefficient alpha. The resulting coefficient of 0.871 clearly exceeds the 0.7 level recommended by Nunnally (1978), which proves it is a reliable scale, especially given the fact that reverse coding was used to measure one item of the 3-item scale.

5.2. Predictive performance

In terms of predictive performance, in Table 3, we compare the results of the different models. Considering the MLR models, we compared the full model with the final model resulting from the variable-selection procedure described previously, which resulted in a selection of just four variables. Regarding the results from the RF model, all variables were introduced, yet only 24 variables were selected by the technique. In terms of the ARD model, after extensive trial-and-error testing, we reached an optimal performance by using 24 hidden units. No variables were selected by the latter technique so each variable contributes, to some extent, to the predictive performance.

Different interesting conclusions can be drawn from Table 3. First, it is clear that – as was expected – overfitting prevails in the MLR model, and does not appear in the RF model. This finding is in line with Breiman's (2001) initial claims as well as findings by other authors (e.g. Buckinx & Van den Poel, 2005). Indeed, the adjusted R^2 of the full MLR model drops from 0.2926 on the estimation set to 0.2301 on the validation set, which introduces skepticism on the validity of this model. Second, the variable-selection procedure we described previously succeeds in reducing the negative impact related to overfitting. Indeed, the difference between the adjusted R^2 on the estimation set (0.3032) versus the test set performance (0.2919) is sufficiently small. Thirdly, contrarily to what might have been expected using the Bayesian 'Occam's razor' (MacKay, 1992), the ARD model also proves to be sensitive to overfitting, as the performance on the estimation set is substantially higher than the performance after cross-validation. Fourth, given that an efficient variable-selection procedure is performed to the regression model, this model clearly outperforms the other models in terms of predictive performance. Fifth, in order to test whether this result is significant, we tested whether the correlations (R) differ significantly using a test of the difference of dependent samples described in Cohen and Cohen (1983, p. 57). From this test, we can conclude that the MLR model significantly outperforms the RF ($t = 2.57$, $p = 0.01022$) and ARD models ($t = 2.68$, $p = 0.00747$). However, the difference in performance between the RF and ARD models is not significant ($t = 1.39$, $p = 0.16421$).

In sum, given that the adjusted coefficient of determination of the final MLR model is fairly high (0.2919) for cross-sectional data, and given its significance ($F = 96.39$,

Table 3
Model performances

	MLR				RF		ARD	
	Full model ($v = 35$)		Final model ($v = 4$)		Full model ($v = 35$)		Full model ($v = 35$)	
	Estimation	Validation	Estimation	Validation	Estimation	Validation	Estimation	Validation
R	0.5664	0.5107	0.5535	0.5442	0.5186	0.5238	0.5714	0.4935
R^2	0.3208	0.2608	0.3064	0.2962	0.2689	0.2744	0.3265	0.2435
R^2adj	0.2926	0.2301	0.3032	0.2919	0.2385	0.2442	0.2985	0.2121
MSE	0.5586	0.6107	0.5502	0.5569	0.6023	0.5969	0.5586	0.6237
RMSE	0.7474	0.7815	0.7417	0.7463	0.7761	0.7726	0.7474	0.7898

$p = <0.0001$), we can state that it is possible to predict a customer's loyalty to a reasonable degree from the internal transactional database using a regression model – provided that an elaborate variable-selection procedure is performed. Because of the importance of the latter procedure, we discuss its implications in detail in the following paragraph.

5.3. Usefulness of the variable-selection technique

In Fig. 3, we illustrate the effect of the variable-selection technique by plotting the estimation, test and validation performance of the best-performing model per model size. While the adjusted R^2 of the estimation dataset does not decrease substantially as the number of variables increases, the validity of these models is severely hampered. However, the splitting of the leave-one-out sample into a test and validation set does clearly allow us to select the best-performing model and validate this model, while efficiently exploiting the available observations. Hence, the test set reached its highest level with the use of only four variables, whereby overfitting is reduced. Appendix A features a similar graph illustrating overfitting in terms of the RMSE.

While we have focused on the negative impact of using a large set of variables on the predictive performance of the model, an additional threat resides in the occurrence of multicollinearity. Indeed, it is likely that, when using a large number of predictors, several predictors that are jointly used might be severely correlated. Hence, the affected parameter estimates might become unstable and may exhibit high standard errors, reflecting the lack of properly conditioned data (Belsley, Kuh, & Welsch, 1980). In this section, we will illustrate the existence of multicollinearity graphically. To this goal, we follow the procedure of Belsley et al. (1980), and hence we present the evolution of the condition index of the best performing model per model size in Fig. 4. Considering the author's informal suggestion that, at an index larger than 15, weak dependencies may start to affect the regression estimates (Belsley et al., 1980, p. 153), those models incorporating more than seven variables might exhibit unstable estimates and high standard errors. In order to validate this rule of thumb we have attempted to provide a graphical representation of the stability of the estimates. To this effort, we have computed the parameter estimates of all variables when they are used separately in univariate predictive models. Next, we compared the signs of these parameters – to which we refer as the 'correct' signs – with the signs of the best multiple regression models, and we plotted the percentage of 'correct' signs in Fig. 5. The results confirm the previously offered rule-of-thumb, as at least some parameter signs differ in models that contain more than seven variables. Hence, in these models, the parameter estimates can be considered as unstable.

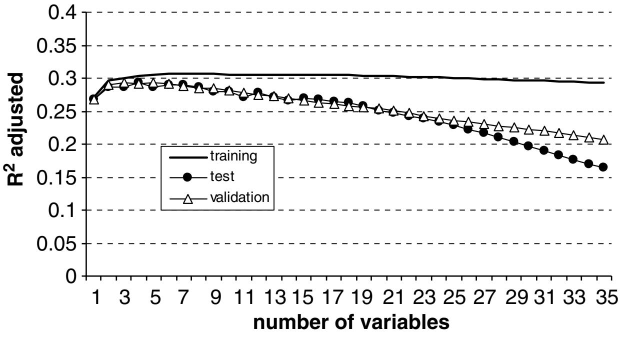


Fig. 3. Evidence of overfitting when the number of variables is increased.

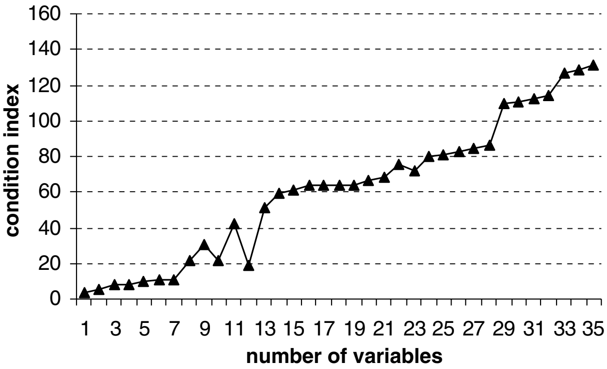


Fig. 4. Detecting multicollinearity by the condition index.

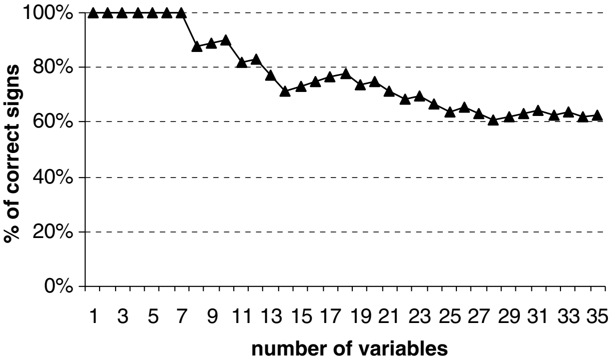


Fig. 5. An illustration of the effect of multicollinearity on the parameter signs.

To conclude this section, the full model – containing all variables – shows evidence of multicollinearity that is manifested in a condition index of 131.6 and the fact that only 63% of the parameter signs correspond to their univariate counterparts. However, these problems seem efficiently solved in the final model – containing only the four selected variables – showing a condition index of only 8.5 and a proportion of 100% 'correct' parameter signs.

5.4. Variable importance

In order to discuss the importance of the variables to predict behavioral loyalty, we will look both at the univariate performances as well as the inclusion of these variables

Table 4
Parameter estimates of the best predictive models

Number of variables	Variable	Standardized estimate	*t*-Value	Pr > $\lvert t\rvert$	R^2adj validation
1	Intercept	0	−15.69	<.0001	0.2678
	Numcat_LY	0.5221	18.12	<.0001	
2	Intercept	0	−14	<.0001	0.2905
	Spending	0.2154	5.56	<.0001	
	Numcat_LY	0.3751	9.67	<.0001	
3	Intercept	0	−14.3	<.0001	0.2934
	Numcat_LY	0.2979	6.16	<.0001	
	PercResp_Leaf	0.1240	2.64	0.0084	
	rSpend_Lor	0.1859	4.59	<.0001	
4	Intercept	0	−13.62	<.0001	0.2919
	Spending_Fresh	0.0887	2.12	0.0343	
	Numcat_LY	0.2741	5.54	<.0001	
	PercResp_Leaf	0.1145	2.43	0.0151	
	rSpend_Lor	0.1468	3.31	0.001	
5	Intercept	0	−11.91	<.0001	0.2926
	Spending_ Fresh	0.0994	2.41	0.0162	
	Numcat_LY	0.2389	4.54	<.0001	
	NumItems	0.1017	2.16	0.031	
	PercResp_Leaf	0.1651	3.07	0.0022	
	rSpend_Freq	0.0739	2.21	0.027	
6	Intercept	0	−8.46	<.0001	0.2911
	Spending_ Fresh	0.1024	2.48	0.0133	
	Numcat_LY	0.2193	4.06	<.0001	
	NumItems	0.1043	2.22	0.0269	
	PercResp_Leaf	0.1487	2.72	0.0066	
	rSpend_Freq	0.0732	2.2	0.0284	
	Std_Ipt	−0.0553	−1.64	0.1007	
7	Intercept	0	−8.53	<.0001	0.2881
	Spending_ Fresh	0.1009	2.44	0.0147	
	Neg_Inv	0.0396	1.2	0.2313	
	Numcat_LY	0.2172	4.03	<.0001	
	NumItems	0.0990	2.09	0.0365	
	PercResp_Leaf	0.1367	2.46	0.0141	
	rSpend_Freq	0.0769	2.3	0.0219	
	Std_Ipt	−0.0520	−1.54	0.1237	

into the MLR models. First, in terms of the univariate importances, Table 1 illustrates that the different models emphasize different variables. For example, in the ARD model, the length of relationship is considered as the second most important variable, while in the MLR model it features as the second least important variable, and the variable was not selected in the RF model. The difference between the models can be evaluated more formally through the computation of the correlation between the variable importances. The correlation between the MLR model and RF model is 0.08862 ($p = 0.6127$), between the MLR model and the ARD model −0.16933 ($p = 0.3308$), and between the RF model and the ARD model 0.12051 ($p = 0.4905$), so we conclude that the models really emphasize different predictors. Since the MLR model outperforms the other models, in the remainder of this paragraph, we will focus on the importance of variables according to the MLR model. From the univariate performances, we note that the purchase variety clearly forms the best predictor of loyalty. However, several groups of variables have only a slightly lower performance. Variables related to the spending, frequency, promotion behavior and response on mailings all have a good predictive performance. The other variables, such as recency, interpurchase time, length of relationship, average spending per visit, returns of goods and distance to the store clearly exhibit lower univariate predictive performance.

An additional insight can be gained from the inclusion of the variables in the best performing multivariate models. Hence, in Table 4, we present the variables of the selected models that contain up to seven variables. This confirms the fact that purchase variety, spending and a customer's response on mailing folders present the most useful information for predicting behavioral loyalty.

6. Conclusions and directions for further research

Following the prevalence of the CRM discourse, companies have started to realize the value of loyal customers, and have acquired the competences to manage customer relationships through targeted communications. Intriguingly however, these relationships are currently managed almost unanimously based on transactional data (such as recency, frequency, and monetary value of a customer) while the behavioral loyalty and hence the full potential of a customer is generally unavailable. In this study, we have constructed a reliable three-item scale to measure behavioral loyalty, and we have proven that it is possible to predict a customer's behavioral loyalty to a reasonable degree based on his/her transactional information. Hence, we have provided a viable methodology for building a loyalty score for all customers, based on a limited sample of customers for which behavioral loyalty was surveyed. This additional customer knowledge can be useful in many marketing applications within the area of customer relationship management, be it direct marketing, model building and customer evaluation.

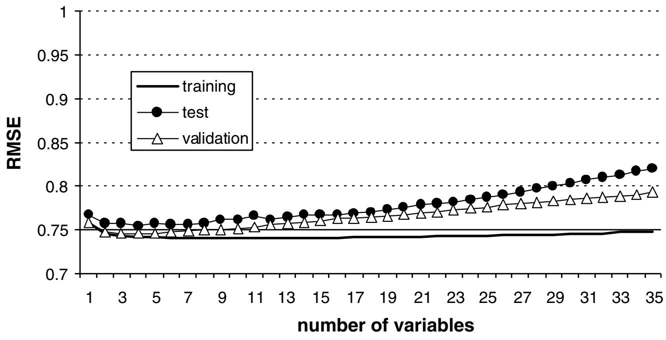
To this end, we compared three techniques that have been argued to show a good predictive performance and an interpretation of the importance of the predictors. More specifically, we compared multiple linear regression with two state-of-the-art techniques, namely Breiman's regression forests and MacKay's automatic relevance determination. The predictive modeling we propose in this study is different from the general situation of predicting transactional behavior by use of historic transactional behavior in the sense that here, the target variable is only known for a limited set of customers. Because overfitting is more likely to occur when the observations are limited compared to the number of variables, and since overfitting is a well-acknowledged problem in multiple linear regression, the major contribution of this study lies in designing an effective variable-selection procedure. Hence, considering the limited sample size, we propose a model selection and validation procedure that is based on the leaps-and-bounds algorithm using an intelligent split of a leave-one-out cross-validation sample. In a real-life study, we show that this procedure effectively increases the validation performance to an extent that the linear regression model outperforms the other models in terms of predictive accuracy, and that multicollinearity is removed to an adequate degree in the resulting model, allowing for a sound interpretation of the parameters. Hence, we show that purchase variety is the best performing predictor of behavioral loyalty, and that a customer's spending, frequency, promotion behavior, response to mailings and regularity of purchasing all provide useful information to deliver an adequate prediction of a customer's behavioral loyalty.

As any other study, this study has its limitations which may lead to further research. First of all, in this paper it was not our ambition to compare all possible predictive modeling techniques. Hence, it is not excluded that other techniques serve even better to predict behavioral loyalty. Instead, we have confirmed that a proper use of sound statistical techniques is at least able to compete with two state-of-the-art predictive techniques. Second, contrarily to what was expected, we gained evidence of overfitting in the ARD model. While again it was not the focus of this specific study, this finding seems at least intriguing. Hence, further research might focus on performing a (possibly similar) variable-selection technique for the ARD model to account for the overfitting that was detected. Thirdly, in this case, we have used a leave-one-out cross-validation sample. It is not unlikely, however, that for future usage, the procedure could be applied in a more resource-efficient way by applying a leave-*k*-out cross-validation, where *k*is increased while carefully monitoring the validity of the results. Finally, in this procedure, due to financial constraints, it was not possible to perform an out-of-sample cross-validation to account for any possible model drift. Indeed, a subsequent survey of the behavioral loyalty would prove useful in evaluating the stability of the model for future loyalty predictions.

Acknowledgments

The authors are grateful to Leo Breiman (2001) for the public availability of the random forests software, and to Ian T. Nabney (2001) for his implementation of the technique of automatic relevance determination using neural networks.

Appendix A



References

Baesens, B., Verstraeten, G., Van den Poel, D., Egmont-Petersen, M., Van Kenhove, P., & Vanthienen, J. (2004). Bayesian network classifiers for identifying the slope of the customer lifecycle of long-life customers. *European Journal of Operational Research, 156*(2), 508–523.

Baesens, B., Viaene, S., Van den Poel, D., Vanthienen, J., & Dedene, G. (2002). Bayesian neural network learning for repeat purchase modelling in direct marketing. *European Journal of Operational Research, 138*(1), 191–211.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics*. New York: John Wiley & Sons, Inc.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

Buckinx, W., & Van den Poel, D. (2005). Customer base analysis: Partial defection of behaviorally-loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research, 164*(1), 252–268.

Bult, J. R., & Wansbeek, T. (1995). Optimal selection for direct mail. *Marketing Science, 14*(4), 378–394.

Chintagunta, P. K. (1992). Estimating a multinomial probit model of brand choice using the method of simulated moments. *Marketing Science, 11*(4), 386–407.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Furnival, G. M., & Wilson, R. W. (1974). Regressions by leaps and bounds. *Technometrics, 16*, 499–511.

Geng, W., Cosman, P., Berry, C. C., Feng, Z., & Schafer, W. R. (2004). Automatic tracking, feature extraction and classification of C elegans phenotypes. *IEEE Transactions on Biomedical Engineering, 10*(51), 1811–1820.

Goutte, C. (1997). Note on free lunches and cross-validation. *Neural Computation, 9*, 1245–1249.

Grönroos, C. (1997). From marketing mix to relationship marketing – towards a paradigm shift in marketing. *Management Decision, 35*(4), 839–843.

Hwang, H., Jung, T., & Suh, E. (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), 181–188.

Jonker, J. J., Piersma, N., & Van den Poel, D. (2004). Joint optimization of customer segmentation and marketing policy to maximize long-term profitability. *Expert Systems with Applications, 27*(2), 159–168.

Larivière, B., & Van den Poel, D. (2004). Investigating the role of product features in preventing customer churn, by using survival analysis and choice modeling: The case of financial services. *Expert Systems with Applications, 27*(2), 277–285.

Macintosh, G., & Lockshin, L. S. (1997). Retail relationships and store loyalty: A multi-level perspective. *International Journal of Research in Marketing, 14*(5), 487–497.

MacKay, D. J. (1992). Bayesian interpolation. *Neural Computation, 4*, 415–447.

Nabney, I. T. (2001). *Netlab algorithm for pattern recognition*. Springer.

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.

Reichheld, F. F. (1996). *The loyalty effect*. Cambridge, MA: Harvard Business School Press.

Reichheld, F. F., & Sasser, W. E. Jr., (1990). Zero defections: Quality comes to service. *Harvard Business Review, 68*(5), 105–111.

Reinartz, W. J., & Kumar, V. (2000). On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing. *Journal of Marketing, 64*(October), 17–35.

Rosenberg, L. J., & Czepiel, J. A. (1984). A marketing approach to customer retention. *Journal of Consumer Marketing, 1*, 45–51.

Snee, R. D. (1977). Validation of regression models: Methods and examples. *Technometrics, 19*(4), 415–428.

Srinivasan, S. S., Anderson, R., & Ponnavolu, K. (2002). Customer loyalty in e-commerce: An exploration of its antecedents and consequences. *Journal of Retailing, 78*, 41–50.

Van den Poel, D., & Buckinx, W. (2005). Predicting online purchasing behaviour. *European Journal of Operational Research, 166*(2), 557–575.

Van den Poel, D., & Larivière, B. (2004). Customer attrition analysis for financial services using proportional hazard models. *European Journal of Operational Research, 157*(1), 196–217.

Verhoef, P. C., Spring, P. N., Hoekstra, J. C., & Leeflang, P. (2002). The commercial use of segmentation and predictive modeling techniques for database marketing in the Netherlands. *Decision Support Systems, 34*, 471–481.