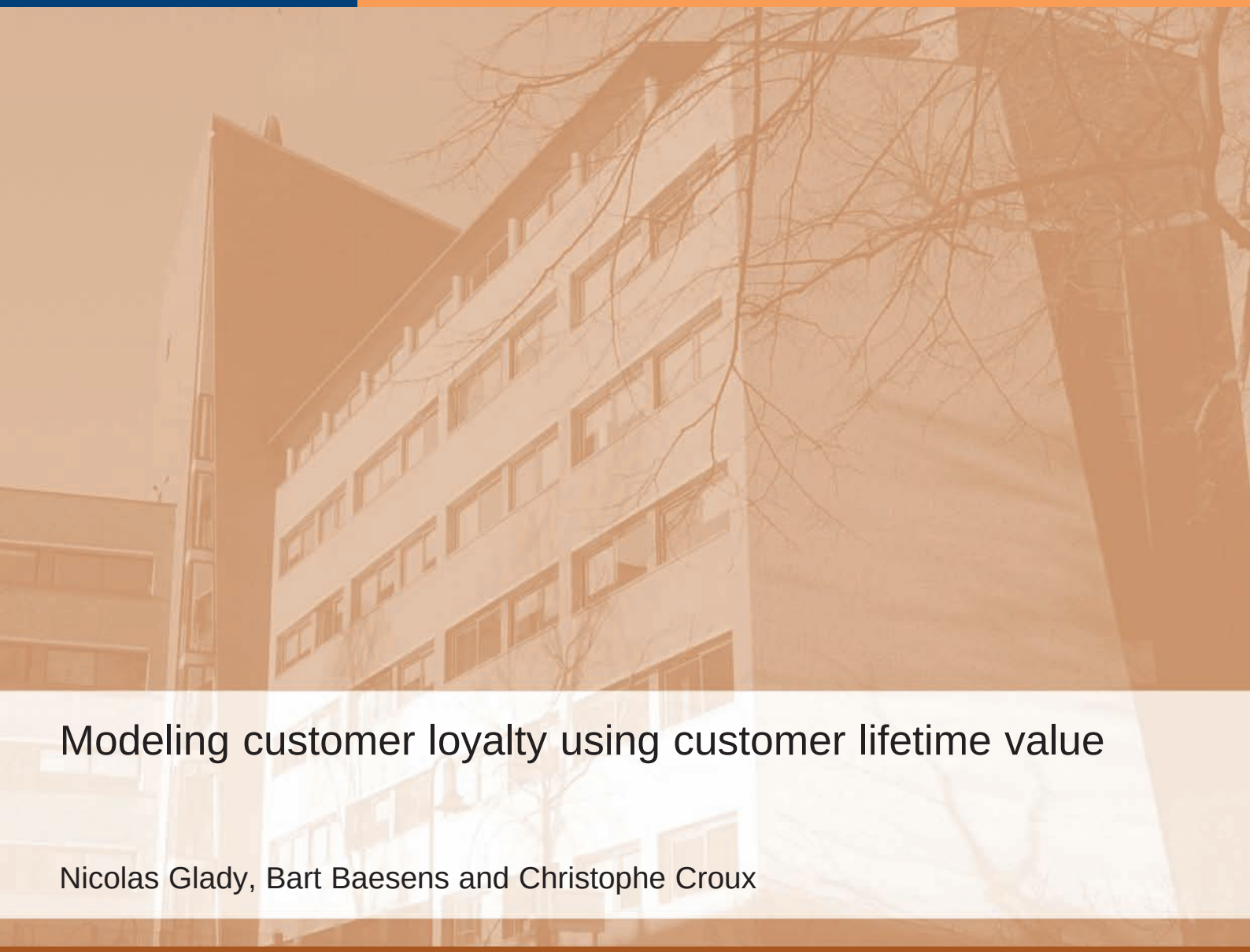
KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Economics and Applied Economics



Modeling customer loyalty using customer lifetime value

Nicolas Glady, Bart Baesens and Christophe Croux

DEPARTMENT OF DECISION SCIENCES AND INFORMATION MANAGEMENT (KBI)

KBI 0618

Modeling Customer Loyalty Using Customer Lifetime value

Nicolas Glady[a] Bart Baesens[a,b] Christophe Croux[a] *

[a] Faculty of Economics and Applied Economics, Naamsestraat 69, B-3000 Leuven, Belgium
[b] School of Management, University of Southampton, SO17 1BJ, UK.

Abstract

The definition and modeling of customer loyalty have been central issues in customer relationship management since many years. Recent papers propose solutions to detect customers that are becoming less loyal, also called churners. The churner status is then defined as a function of the volume of commercial transactions. In the context of a Belgian retail financial service company, our first contribution will be to redefine the notion of customer's loyalty by considering it from a customer-centric point-of-view instead of a product-centric point-of-view. We will hereby use the customer lifetime value (CLV) defined as the discounted value of future marginal earnings, based on the customer's activity. Hence, a churner will be defined as someone whose CLV, thus the related marginal profit, is decreasing. As a second contribution, the loss incurred by the CLV decrease will be used to appraise the cost to misclassify a customer by introducing a new loss function. In the empirical study, we will compare the accuracy of various classification techniques commonly used in the domain of churn prediction, including two cost-sensitive classifiers. Our final conclusion is that since profit is what really matters in a commercial environment, standard statistical accuracy measures for prediction need to be revised and a more profit oriented focus may be desirable.

Keywords: Churn Prediction, Classification, Customer Lifetime Value, Prediction Models.

*We thank ING Belgium for their support and useful information, especially Martine George head of the customer intelligence department. All correspondence should be sent to the first author Nicolas Glady: Naamsestraat 69, B-3000 Leuven; Nicolas.Glady@econ.kuleuven.ac.be

1 Introduction

In a time of cost-cutting and intensive competitive pressure, it becomes of crucial importance for retailers to capitalize their existing customer base. Consequently, customer retention campaigns are implemented. This requires to detect the customers decreasing their loyalty to the company, also called churners. This paper proposes a new framework for the churner detection process, using the earnings a customer yields to the company.

A churner has long been defined with regard to the longevity of his/her historical monetary value. However, Reinartz and Kumar (2000) criticized this method, since they demonstrated that a long life-cycle and profit were not necessarily related. On the opposite, Rust et al. (2004) emphasized that marketing strategy should focus on projected future financial return using the customer equity defined as the total value of the customer base. In order to predict this value, Dwyer (1997) and Berger and Nasr (1998) have provided a framework using the lifetime value of a customer.

Supporting this idea, Gupta et al. (2004) showed that the profit, and hence the firm's value, is a function of the total *Customer Lifetime Value* (CLV). Concurrently, Venkatesan and Kumar (2004) demonstrated the usefulness of CLV as a metric for customer selection, since "customers who are selected on the basis of their lifetime value provide higher profits in future periods than do customers selected on the basis of several other customer-based metrics". Finally, in a recent paper, Neslin et al. (2006) compare several churn classifiers with regard to the CLV change they incur.

This paper contributes to the existing literature by using the customer lifetime value as a basis concept for a churner classifier implementation. First, in order to define the value of a customer, we will define the CLV as the present value of future cash flows yielded by the customer's product usage, without taking into account previously spent costs. Subsequently, to detect churning behavior, we considered Baesens et al. (2003) who proposed solutions to estimate the slope of the customer life-cycle, giving an insight on future spending evolutions. Combining these two ideas, we will predict churning behavior on the basis of the slope of the customer lifetime value in time, hereby moving from a product-centric point-of-view to a customer centric point-of-view. A churner will then be defined as someone with a value decreasing over time.

Consequently, we will be able to compute the actual loss caused by a bad prediction (with no or inefficient action) by defining a new type of profit-sensitive loss function. Our key point is that in any business activity, to loose only a few profitable customers is much

more worse than to loose many non-profitable ones. That is why usual statistical accuracy measures may not be most ideal in this context.

Next, we will use and contrast several classifiers for the churning prediction. A decision tree and a neural network will be compared to a baseline logistic regression model. Moreover, a cost-sensitive design has been proposed by Turney (1995) and Fan et al. (1999). These papers provide tools to optimize classifiers using boosting with regard to a cost function. Such algorithms are called meta-classifiers, since they only optimize other "base" classifiers. Applying this idea, we will also implement a decision tree generated on a cost-sensitive training sample and AdaCost, a variant of the well-known AdaBoost algorithm. For simplicity purpose, the only predictor variables in these models will be of the RFM (recency, frequency and monetary) type: Buckinx and Van den Poel (2005) and Fader et al. (2005) proved that RFM variables can predict accurately the CLV.

In our empirical study, using data provided by a retail banker, the loss function presented will be applied to assess various common classification techniques for the detection of churning behavior. The purpose of this paper is not to provide a new classification technique, but instead, under some assumptions defined later, to construct a framework using a profit-sensitive loss function for the selection of the best classification techniques with regard to the estimated profit.

Our paper is organized as follows: in Section 2, we discuss the general definition of churning in order to propose a new one using the CLV. Likewise, in Section 3, we will discuss the usual loss functions for churn prediction and we will provide a new one using the CLV. In Section 4.1, we will describe the data set used in Section 4.2 in order to compare in Section 5 usual classification techniques used in churn behavior prediction. In the last section, we will discuss the assumptions made and the results obtained. Finally, we will propose issues for further research.

2 Definitions of churning

First, we have to define the condition under which a customer has to be considered as being decreasing his/her loyalty, and hence as churning. The issue in a competitive environment is that most people have more than only one supplier. For instance, in retail banking, a customer could have a current account in a first bank and a mortgage loan in another. Most people have several current accounts even if they do not use it (so-called "sleeping" accounts). As a matter of fact, we need to find a definition of a churner applicable to

non-contractual products, as opposed to contractual products. Contractual products are for instance insurance, mortgage, cellular phone (if high entry or exit barriers and fixed price), in other words all products with "contractual" cash-flows. On the other hand, non-contractual products could be catalog sales, cellular phones (if low entry and exit barriers and marginal price), etc.

In the empirical study, we will focus on the private person checking account of a Belgian financial institution. It corresponds to non-contractual products because even if the general relationship is long and contractual, the price for the customer to stop using the account is low and the product usage is at the customer's discretion. Analysis of churning behavior for financial services has been studied before (e.g. by Van den Poel and Larivière (2004)), but using a static definition of churning, for example, defining a churner as a customer who closed all his/her accounts.

2.1 Previous Definitions of Churners

Most definitions of the customer status are using the product activity and a threshold fixed by a business rule. If the activity of the customer has fallen below the threshold, (or equal to zero), this customer is considered as a churner. We claim that this is not always relevant, one should observe the evolution in the customer activity instead.

As an example, consider a business rule labeling all customers with a product activity below 5 transactions per year as churners. If a customer has made 4 transactions in the current year, he/she will be considered as a churner, even though during past years 5 transactions were made annually. On the other hand, if another customer had an activity of 100 transactions per year for 10 years, but has made 6 transactions only this year, he/she will not be considered as a churner. This is problematic since it is not sure that the first customer has decreasing loyalty, whereas the last customer has obviously changed product usage. A churner status function based on a major change in the activity would be more appropriate.

Furthermore, if one has to wait until the customer has ended his/her relationship with the company, it's too late to take any preemptive action. The ultimate purpose is to increase the earnings yielded by the customers, by detecting churning behavior at the very beginning. Moreover, the idea to define a churner for a non-contractual product based on life-cycle duration only, has been challenged by Reinartz and Kumar (2000). Consequently, as noted by Rust et al. (2004), only future earnings (that is what we will later define as the CLV) are relevant to take any potential preemptive action, even though assumptions for the future are

obviously made considering the past.

2.2 Churner Status Indicator Based on the Slope of the Product Usage

In a more dynamic approach, Baesens et al. (2003) describe methods to estimate the slope of future spending for long-life customers, hereby providing qualitative information for marketeers. Our contribution is to propose a framework to resolve the heterogeneity in the customer population by identifying the more profitable customers such that they can be carefully approached using future actions. Instead of looking in the past to observe whether the customer has churned, we will focus on the future in order to estimate whether the relationship will remain profitable.

Consequently, as a first definition for the churner status, we could consider that if the slope of the product usage in time is below a certain value (let us say 1, when the products usage is decreasing), then the customer should be considered as churning. With $x_{i,j,t}$ being the product j usage, during period t, of customer i, then we define $\alpha_{i,j,t}$ as the slope of the product usage:

$$x_{i,j,t+1} = \alpha_{i,j,t} \times x_{i,j,t}. \tag{1}$$

The slope of the product usage $\alpha_{i,j,t}$ could then be interpreted as a growth rate for $\alpha_{i,j,t} \gg 1$, a retention rate for $\alpha_{i,j,t} \simeq 1$ and a churning rate for $\alpha_{i,j,t} \ll 1$. The purpose of this paper is to focus on the third case, when the customer is churning. Baesens et al. (2003) defined the indicator function of the churner status $y_{i,j,t}$ for the customer i during period t for product j as,

$$y^{(1)}_{i,j,t} = I(\alpha_{i,j,t} < 1). \tag{2}$$

In other words, a customer i is then considered as a churner for product j during period t if his/her product usage will be decreasing in the near future $(t+1)$.

Although the definition of Baesens et al. (2003) is simple and easy to understand, it has the disadvantage to be product-centric. The products are considered separately, whereas a customer could have several products. The same customer could then be considered as a churner for one product but loyal for another. On the opposite, according to many authors such as Dwyer (1997), Rust et al. (2004) and Gupta et al. (2004), all marketing campaigns should be customer-centric. The churner status should ideally be defined based on the entire customer activity. That is the issue we will try to address in the next section.

2.3 A New Definition of Churner Using the Customer Lifetime Value

Our first goal is to detect the customers decreasing their loyalty, now defined as those decreasing their future customer lifetime value. Secondly, we need to identify those for which a retention action will be profitable.

2.3.1 Definition of Customer Lifetime Value

Customer valuation is a major topic since many years and has been discussed by several papers in the customer relationship management literature, see Dwyer (1997), Berger and Nasr (1998), Rust et al. (2004) and Malthouse and Blattberg (2005). Nowadays one can see a proliferation of valuation methods using both terms of "Customer Lifetime Value" or "Customer Equity", for an overview, see Pfeifer et al. (2005). This paper follows Gupta et al. (2004), defining the value of a customer as "the expected sum of discounted future earnings [...] where a customer generates a margin [...] for each period [...]."

The CLV is function of all the transactions a customer will make, for all the q products the company is selling, but it does not take into account cross-individual (word-to-mouth) effects. Consequently, the customer lifetime value of the customer i, for the horizon h from the period t is the sum of the net cash flows $CF_{i,j,t+k}$, yielded by the transaction on product j, discounted at the rate r (assumed constant)[1] and defined as

$$CLV_{i,j,t} = \sum_{k=1}^{h} \sum_{j=1}^{q} \frac{1}{(1+r)^k} \times CF_{i,j,t+k}. \tag{3}$$

Since we are focussing on retention and not acquisition, all customers were acquired in the past and only marginal earnings are to be accounted, disregarding acquisition cost, any sunk costs or fixed costs[2]. Hence if we denote the product marginal yield by unit of product usage for product j as π_j, assumed fixed by product[3], we can define the net cash flow (product profit) $CF_{i,j,t}$ generated by a product j sold to a customer i during period t as a function of the product usage $x_{i,j,t}$,

$$CF_{i,j,t} = \pi_j \times x_{i,j,t}. \tag{4}$$

[1]For simplicity purposes, we will consider the discount as if all cash flows were obtained end-of-month.

[2]In the banking case, the profit considered is nearly equal to the transaction price paid by the customer since in retail banking, the marginal transaction costs are negligible.

[3]It may depend on the type of customer, thus on i. Customers may have preferential conditions according to their status. For simplicity reasons, we will consider an average product yield.

Using (3), this is giving us the CLV for the customer i at t for all the q products,

$$CLV_{i,t} = \sum_{k=1}^{h} \sum_{j=1}^{q} \frac{1}{(1+r)^k} \times \pi_j \times x_{i,j,t+k}. \tag{5}$$

As observed in Reinartz and Kumar (2000), the CLV could be high not only if the product usage remains positive for longer horizons, but also if the product usage $x_{i,j,t}$ itself is high as well. That is our main argument to say that one should focus on profitability instead of longevity only.

2.3.2 Churner Status Indicator Based on Marginal Action Profit

Improving the churner status definition, we could use the decrease of the CLV instead of the slope of the product usage $x_{i,j,t}$ to identify the churners. First, using (1) and (4), we could re-state the product profit (net cash flow) as follows:

$$CF_{i,j,t+1} = \pi_j \times \alpha_{i,j,t} \times x_{i,j,t}.$$

Next, we reformulate the present value of future earnings for the customer i during period t for the product j (that is the CLV),

$$CLV_{i,j,t} = \sum_{k=1}^{h} \frac{\prod_{v=0}^{k-1} \alpha_{i,j,t+v}}{(1+r)^k} \times \pi_j \times x_{i,j,t}.$$

The gain in CLV due to a retention action is an opportunity gain. It is the difference between the CLV, after the retention action (e.g. $\alpha_{i,j,t}$ is kept equal to one)[4], and the CLV without action. We will call it the marginal action profit ($MAP_{i,j,t}$) and it will be denoted as

$$\begin{aligned} MAP_{i,j,t} &= \Delta CLV_{i,j,t} \\ &= CLV_{i,j,t}(\text{with action}) - CLV_{i,j,t}(\text{without action}) \\ &= \sum_{k=1}^{h} \frac{1}{(1+r)^k} \times \pi_j \times x_{i,j,t} - \sum_{k=1}^{h} \frac{\prod_{v=0}^{k-1} \alpha_{i,j,t+v}}{(1+r)^k} \times \pi_j \times x_{i,j,t}. \end{aligned} \tag{6}$$

However, equation (6) is not implementable in practice. Indeed, we would need to know all the information for h periods in advance in order to have all the $\alpha_{i,j,t+v}$ values, before being able to compute the CLV and knowing whether a customer is a churner or not. Instead, we

[4]That formula could be modified with any other value than $\alpha = 1$, with the assumption that a customer retention campaign should at least not decrease the CLV or even, increase it.

will consider that $\alpha_{i,j,t}$ is constant during h periods without action[5]. This number of periods h will obviously be finite and constant for convenience purpose. The equation (6) becomes[6]

$$\begin{aligned} MAP_{i,j,t} &= \sum_{k=1}^{h} \frac{1}{(1+r)^k} \pi_j x_{i,j,t} - \sum_{k=1}^{h} \frac{\alpha_{i,j,t}^k}{(1+r)^k} \pi_j x_{i,j,t} \\ &= \pi_j x_{i,j,t} \left(\frac{1}{r}(1 - \frac{1}{(1+r)^h}) - \frac{\alpha_{i,j,t}}{1+r-\alpha_{i,j,t}}(1 - (\frac{\alpha_{i,j,t}}{1+r})^h) \right). \end{aligned} \tag{7}$$

We will use this value as a lower bound of the profit for a customer who has in mind to churn but has been stopped to do so by a retention action. When the customer was not intending to churn, the action does not have any effect. Then the lower bound of the marginal action profit is the action effect on the customer cash flows for all the products q,

$$MAP_{i,t} = \sum_{j=1}^{q} MAP_{i,j,t}, \tag{8}$$

$$\text{with } \begin{cases} MAP_{i,j,t} = 0 & \text{for } \alpha_{i,j,t} \geq 1 \\ MAP_{i,j,t} = MAP_{i,j,t} & \text{for } \alpha_{i,j,t} < 1. \end{cases} \tag{9}$$

Finally, if our purpose is to have an efficient action and if the marginal action cost (MAC) is assumed fixed but not negligible, we arrive at the following customer-centric churner definition:

$$y_{i,t} = I(MAP_{i,t} > MAC). \tag{10}$$

In other words, a churner is defined as someone for whom a retention action is profitable.

This new indicator function offers three major advantages compared with (2) that defines a churner as someone who is decreasing product usage. First, churners not worthy to deal

[5]A constant retention rate for customer valuation was also accepted by Gupta et al. (2004). Therefore, for simplification purposes and under smoothing conditions described below, we will assume the constant character of $\alpha_{i,j,t}$ in order to have a minimum delay when wanting to assess the model.

[6]In order to have the total present value of the possible future loss for the churning behavior of customer i during period t for product j, one could use the convergence of (7) in h,

$$\lim_{h\to\infty} MAP_{i,j,t} = \pi_j \times x_{i,j,t} \times \left(\frac{1}{r} - \frac{\alpha_{i,j,t}}{1+r-\alpha_{i,j,t}} \right),$$

and passing from a single product view to a customer view (all products), we have,

$$\lim_{h\to\infty} MAP_{i,t} = \sum_{j=1}^{q} \pi_j \times x_{i,j,t} \times \left(\frac{1}{r} - \frac{\alpha_{i,j,t}}{1+r-\alpha_{i,j,t}} \right).$$

But since it may be unlikely that α remains constant, this value should be used as an informal indication only.

with will be neglected. The second advantage is a cross-product, customer-centric definition of a churner instead of a product-oriented definition. Finally, the last advantage is that, once the parameters (action cost, product profit, etc.) have been defined, this definition is applicable to every type of business.

In reality, it is always laborious to find the exact unitary action marginal cost (MAC), the exact marginal product revenue (π_j) and the exact effect of the action on the product usage (the value of $\alpha_{i,j,t}$ if the action is taken). However, if the scale of these parameters is approximately correct, this valuation gives an insight on the profit of a retention action. Moreover, that will enable us to compare the financial value of various churner detection techniques.

3 Loss Function Definition

During the empirical study, several classifiers will be compared. In order to assess the accuracy of each classifier, the loss incurred by wrong predictions needs to be quantified. A loss function needs to be defined. The most common measure of loss (or gain), is the *Percentage of Correctly Classified* (PCC) observations. This measure implicitly assumes equal misclassification costs, which is most often not the case. Moreover, this measure is very sensitive to the choice of the cut-off value used to map the classifier output to classes, as we will see below.

Another well-known classification performance metric is the *Receiver Operating Characteristic* curve (ROC), described in Egan (1975). A ROC curve is a graphical plot of the sensitivity (percentage of true positive) versus 1-specificity (percentage of false positive), letting the classification cut-off vary between its extremes. The AUROC, the *Area Under the Receiver Operating Characteristic* curve, is then a summary measure of classification performance. This second measure provides a better evaluation criterion, since it is independent of any cut-off.

Nevertheless, all misclassifications are not always causing the same loss. In a business context, a very profitable customer has to be monitored very closely, whereas churners that are not yielding any profit may be less interesting to consider. In the next subsection, we will use the CLV in order to define a new loss function proportional to the decrease in earnings generated by a bad prediction.

3.1 A New Loss Function Using the Customer Lifetime Value

In what follows, two kinds of errors are distinguished. The first one is the false positive type, when a customer is classified as a churner whereas he/she is not decreasing loyalty. In this case, an action is taken that was not necessary. The loss is the action cost, which is assumed to be the same for every customer. The second one is the false negative type, when a churner is not detected by the classifier. Here, the loss function is the difference between the earnings generated without action, and the earnings that would have been generated if the customer would have been stopped from churning (i.e. with $\alpha_{i,j,t} = 1$).

We define the loss function for a customer i during period t using (8) as follows[7]

$$L(x_{i,j,t}, \alpha_{i,j,t}, y_{i,t}, \hat{y}_{i,t}) = \begin{cases} 0 & \text{for } y_{i,t} = \hat{y}_{i,t} \\ MAC & \text{for } y_{i,t} = 0 \text{ and } \hat{y}_{i,t} = 1 \\ MAP_{i,t}(x_{i,j,t}, \alpha_{i,j,t}) - MAC & \text{for } y_{i,t} = 1 \text{ and } \hat{y}_{i,t} = 0. \end{cases} \tag{11}$$

Here, the churning status $y_{i,t}$ is defined in (10), and $\hat{y}_{i,t}$ is its prediction using a certain classification method (see Section 4.3). More profitable customers that are churning will cause a bigger loss (if misclassified) than those who are less profitable.[8]

In order to be able to compare our loss function with the PCC, we first compute the ratio between the losses incurred by the classification model, and the worst case scenario, yielding a number between 0 and 1. The worst case scenario assumes that every customer is misclassified. We denote this ratio as the cumulative loss percentage,

$$L_{tot} = \frac{\sum L(x_{i,j,t}, \alpha_{i,j,t}, y_{i,t}, \hat{y}_{i,t})}{\sum L(x_{i,j,t}, \alpha_{i,j,t}, y_{i,t}, 1 - y_{i,t})}, \tag{12}$$

where the sum is over all indices i, t and j. Finally, we define the cumulative profit percentage as the opposite of the cumulative loss percentage

$$\bar{L}_{tot} = 1 - L_{tot}. \tag{13}$$

[7]The reader has to keep in mind that we are doing an incremental analysis: what are the incremental consequences on the CLV of a retention action? Similarly, we are assessing a classifier with regard to the CLV change it will yield. In spirit of this opportunity cost or opportunity gain approach, we can state that the cost incurred by a good classification is zero.

[8]The reader should not forget that $MAP_{i,t}$, thus the loss function defined in (11), is only a lower bound of the opportunity cost of a misclassification, since it is most likely that the action effect will be more than only prevent the customer from churning, but may also increase product consumption, hence the product profit.

3.1.1 Cut-off value: Definition and Application

Most classifiers are giving a probability to belong to one of the two classes instead of giving a binary outcome. We need a threshold (or cut-off value, denoted by τ) to distinguish one class from another. Let $p_{i,t}$ be the posterior churning probability estimated by the classifier for customer i during period t. The cut-off value τ is the value between 0 and 1, such that, if $p_{i,t} \geq \tau$, then the customer is classified as a churner. Accordingly, the profit curve (PROC) $f(\tau)$ becomes:

$$f(\tau) = \bar{L}_{tot}(\tau), \tag{14}$$

for $0 < \tau < 1$. We can then define the area under the profit curve (AUPROC) as a profit based measure of classification performance which is independent of the cut-off. This curve may then also be used to set the cut-off in a profit optimal way.

To compute the AUPROC, one could use a discrete integration under the curve with an arbitrary precision parameter pr. Consider the set of $\lfloor \frac{1}{pr} \rfloor$ cut-off values $pr, 2pr, \ldots, 1$, then the approximation of the AUPROC is computed as follows

$$AUPROC = pr \times \sum_{l=1}^{\lfloor \frac{1}{pr} \rfloor} \bar{L}_{tot}(l \times pr). \tag{15}$$

4 The Empirical Study

4.1 Description of the Data Set

We studied the current account transactions (number of invoicing last month, amount invoiced last month, number of withdrawals, etc.) provided by a Belgian financial service company for a sample of $n = 10,000$ customers and $s = 9$ months (from January 2004 till September 2004). The population consisted out of new, old and sleeping (without any activities since many months) customers. All transactions were aggregated at the customer level. We considered two different product usages, the total number of debit transactions and the total amount debited in every month.

Before estimating and assessing the classification models, we separated the sample into a training set (66% of the observations) to design the classifiers and a test set (33% of the observations) for the performance assessment. The training set was composed of the product transactions from January 2004 till June 2004 (6 months). The test set contained the products transactions for the same customers but from July 2004 till September 2004 (3 months).

4.2 Implementation Details

Since the action profit (7) is very sensitive to the value of $\alpha_{i,j,t}$, we first smooth the values of both $x_{i,j,t}$ and $\alpha_{i,j,t}$ in order to remove the noise and possible instabilities in the churner status. Indeed, it could happen that the slope of the product usage goes slightly up and down from one month to another. Since we are studying the trend of the product usage, we need to have a smoothed value of this slope. Rearranging (1), we applied a standard exponential smoothing scheme. If we denote $\widetilde{x}_{i,j,t}$ the smoothed value of $x_{i,j,t}$ and $\widetilde{\alpha}_{i,j,t}$ the smoothed value of $\alpha_{i,j,t}$ then

$$\widetilde{x}_{i,j,t} = a \times x_{i,j,t} + (1-a) \times \widetilde{x}_{i,j,t-1}, \tag{16}$$

$$\widetilde{\alpha}_{i,j,t} = \frac{\widetilde{x}_{i,j,t+1}}{\widetilde{x}_{i,j,t}}. \tag{17}$$

The smoothing parameter a was set at 0.8, as determined using experimental evaluation. Next, each observation was rearranged as follows

$$\mathbf{x}_{i,t} = [\widetilde{x}_{i,1,t}, \ldots, \widetilde{x}_{i,1,t-m}, \ldots, \widetilde{x}_{i,q,t}, \ldots, \widetilde{x}_{i,q,t-m}], \tag{18}$$

whereby $\widetilde{x}_{i,j,t}$ represents the smoothed value of explanatory variable j for customer i observed during time period t. The maximum number of lags considered was $m = 3$. The vector $\mathbf{x}_{i,t}$ contains then the values of the predictor variables to be used in the classification procedures (to be discussed in Section 4.3). Note that the variables $x_{i,1,t}$ and $x_{i,2,t}$, i.e. the number of debit transactions and the total amount debited in month t for customer i, are function of the recency, frequency and monetary value of the customer. The vector $\mathbf{x}_{i,t}$ is completely observed for the training sample for $i = 1 \ldots n$ and $t = 4, 5, 6$. The corresponding $y_{i,t}$ is then computed according to (10). In the following, we denote an observation i as a couple $(\mathbf{x}_i, y_i)$, with $i = 1 \ldots N$ for the training set, dropping the dependency on time. Note that $N = 3n = 30,000$, yielding a very huge training sample size.

For the computation of the CLV, the product yield considered was directly proportional to the transaction volume (product usage 1), $\pi_1 = 0.1\%$. There was no fixed contribution by transaction (product usage 2), $\pi_2 = 0\%$. The discount rate applied was the weighted average cost of capital disclosed in the 2004 financial statement of the financial service provider, $r = 8.92\%$ yearly, giving a monthly discount rate of 0.7146%.

In order to compare short-term and long-term CLV, the study was made for two distinct values of the time horizon (h). The first measures will be made by quarter, $h = 3$. The longer-term view will be computed for a semester, $h = 6$. Finally, the churner status has

been defined using (10) with marginal action cost (MAC) fixed at 2 EUR, which is our best guess for an upper bound of the marginal average cost of a mailing retention campaign.

We will denote the AUPROC computed in (14) as AUPROC_3 for the quarterly view and AUPROC_6 for the semester view. These values will have to be compared with the non cost-sensitive AUROC values. We will denote $\bar{L}_3 = 1 - L_3$ the cumulative profit percentage for the quarterly view and $\bar{L}_6 = 1 - L_6$ the cumulative profit percentage for the semester view (see 12). Both measures will be compared with the non cost-sensitive percentage of correctly classified observations (PCC). These performance measures are computed over the test set, where the indices in (12) range from $i = 1, \ldots, n$, $j = 1, 2$ and $t = 7, 8$. Note that we cannot include the last month, $t = 9$, in the test set since $\alpha_{i,j,t}$ is not computable for it. This yields $2n = 20,000$ observations $(\mathbf{x}_i, y_i)$ in the test sample. Such a large testing sample size guarantees precise estimation of the performance measures.

4.3 Description of the Classifiers

4.3.1 Logistic Regression, Decision Trees and Neural Networks

The first classifiers applied are a selection of well-known data mining algorithms: a logistic regression, a decision tree and a neural network.

The famous logistic regression classifier results for a standard statistical binary regression model, see e.g. Agresti (2002). Decision trees are recursive partitioning algorithms, which are estimated using e.g. information theoretic concepts so as to arrive at a comprehensible tree-based decision model, that is evaluated in a top-down way as discussed in Quinlan (1992). A multi-layer perceptron neural network is a non-linear predictive model whereby inputs are transformed to outputs by using weights, bias terms, and activation functions. These last two models have been included in our study, because non-linear relationships were found in Fader et al. (2005) between CLV and RFM explanatory variables.

For all classification models, we set the cut-off value $\tau = 0.5$. The performances will be quantified using the PCC, the AUROC and the cumulative profit percentage ($\bar{L}_h$) and the AUPROC_h at horizons $h = 3$ and $h = 6$. The software used for the implementation was Matlab 6.1 using the PRtools toolbox of Duin et al. (2004).

- Given the training sample $S = \{(\mathbf{x}_1, y_1, c_1), \ldots, (\mathbf{x}_N, y_N, c_N)\}$, with $\mathbf{x}_i \in I\!R^{m \times q}$, y_i recorded such that $y_i \in \{-1, 1\}$ and $c_i > 0$
- Initialize $c_1(i) = c_i$. according to (19) for $1 \leq i \leq n$
- For $l = 1 \ldots L$
  1. Create bootstrap sample B_l using bootstrap weights $c_l(i)$.
  2. Train base learner h_l for bootstrap sample B_l.
  3. Compute the classifier $h_l \colon I\!R^{m \times q} \to [-1, 1]$ on the set S.
  4. Compute $w_l = \frac{1}{2} \times \ln(\frac{1+r}{1-r})$ where $r_l = \sum_{i=1}^{n} c_l(i) h_l(x_i) \beta_l(i) y_i$, and $\beta_l(i) = 0.5 + 0.5 \times c_l(i)$ for misclassified observations and $\beta_l(i) = 0.5 - 0.5 \times c_l(i)$ for correctly classified observations.
  5. Update the costs according to $c_{l+1}(i) = c_l(i) \exp(-w_l h_l(\mathbf{x}_i) \beta_l(i) y_i)$ and rescale them such that they sum to one.
- Output the final AdaCost classifier $\hat{f}(x_i) = \sum_l^L w_l \times h_l(\mathbf{x}_i)$

Figure 1: General AdaCost algorithm.

4.3.2 Description of the Cost-Sensitive Classifiers

AdaCost

This paper implements a version of AdaCost algorithm as proposed by Fan et al. (1999). AdaCost is basically an extension of AdaBoost (Freund and Schapire (1997)), giving better performance with regard to the cumulative loss percentage (12). It selects several times a random sample (bootstrap) of the original training set, each time estimating a classifier, $h(\mathbf{x}_i)$. Whereas AdaBoost gives the same probability of selection for every observation, in AdaCost the probability for an observation i to be selected in the bootstrap is proportional to its misclassification cost, c_i, here defined as

$$c_i = \frac{L(\mathbf{x}_i, y_i)}{\sum_{i=1}^{N} L(\mathbf{x}_i, y_i)}, \tag{19}$$

where $\mathbf{x}_i$ has been defined in (18) and, $L(\mathbf{x}_i, y_i) = L(\widetilde{x}_{i,j,t}, \widetilde{\alpha}_{i,j,t}, y_{i,t}, 1 - y_{i,t})$ as defined in (11). The algorithm is outlined in Figure 1. We used decision trees as base classifiers $h(\mathbf{x}_i)$.
The choices for w_l, $r_l$and $\beta_l(i)$ in step 4 are the same as in Fan et al. (1999). The number of iterations in the AdaCost algorithm was the usual number of iterations in the AdaBoost-like

algorithm, $L = 50$.

Cost-Sensitive Decision Tree

The last classifier we will study is a special version of AdaCost. If there is only one iteration (without re-weighting), the classifier becomes a decision tree trained on a cost-weighted bootstrap. Since such a technique is very fast, straightforward, and more readable, it may be an interesting alternative to consider.

5 Empirical Results

In this section, we will describe our empirical results. First, some descriptive statistics will be presented, showing that churners are strongly more expensive to misclassify than non-churners. Next, the accuracy of the classifiers previously described will be compared. Two points will be made. First, the new loss function is providing different results than the standard measures of accuracy. Secondly, cost-sensitive classifiers will be presented as an interesting alternative to the usual techniques.

5.1 Frequency of Churners

The churners and non-churners, defined according to (10), are distributed as indicated in Tables 1 and 2. The first line contains statistics for the total data set (training set and test set) and the second line only for the test set. In the first two columns, one can see the relative frequencies of non-churners and churners, assuming each observation has the same weight. The next two columns contain relative frequencies expressed in a cost-weighted way. For non-churners this is

$$\frac{\sum_i I(y_i = 0) \times c_i}{\sum_i c_i},$$

and for churners

$$\frac{\sum_i I(y_i = 1) \times c_i}{\sum_i c_i}.$$

Obviously, to misclassify a churner is, on average, far more expensive than to misclassify a non-churner. For a longer horizon ($h = 6$, see Table 2), we have evidently more churners. For a longer period of CLV computation, the retention action profit increases and thus, is more likely to be greater than the action cost.

The reader has to keep in mind that the reported frequencies depend on the product yield π_j and the marginal action cost. First, all other parameters being equal, the greater the

Table 1: Frequency of churners and non-churners, for $h = 3$

	Relative frequency		Cost-adjusted frequency		Total
Data Set	Non-Churners	Churners	Non-Churners	Churners	Number
Total	87.49%	12.51%	40.32%	59.68%	50,000
Test set	86.96%	13.04%	38.19%	61.81%	20,000

Table 2: Frequency of churners and non-churners, for $h = 6$

	Relative frequency		Cost-adjusted frequency		Total
Data Set	Non-Churners	Churners	Non-Churners	Churners	Number
Total	78.44%	21.56%	19.72%	80.28%	50,000
Test set	77.00%	23.00%	18.27%	81.73%	20,000

marginal action cost, the less it is cost-effective to target the customers with only moderate churning behavior ($\alpha_{i,j,t}$ close to 1). On the contrary, if the product yield was greater, these customers would be considered as worthy to start an action.

From Tables 1 and 2, one could observe that there are proportionally less churners in the total data set than in the test set. This is due to the way the data sets have been constructed. In the long-run, everybody dies, or, in our case, churns. Since the test set consisted of customers sampled during the first month and observed six months later, churning behavior is of course going to increase when customers are observed in later time periods.

5.2 Comparison of Classifiers

The classification results on the test set of the various techniques are depicted in Tables 3 and 4, for $h = 3$ and 6, respectively. Five classifiers are compared: a logistic regression, a multi-layer perceptron neural network, a decision tree, a cost-sensitive decision tree and the AdaCost boosting method previously described. Their performance is measured by the newly proposed cumulative profit percentage $\bar{L}_h$, as defined in (13), and the area under the profit curve, AUPROC, defined in (15). We also assess the classifiers by computing, as is usually done, the percentage of correct classifications (PCC) and the area under the receiver operating curve (AUROC). The last column contain the percentage of churners predicted as churners, also called the true positives.

Table 3: Performance of classifiers with $h = 3$, as measured by the cumulative profit percentage $\bar{L}_3$, and the area under the profit curve AUPROC_3, together with the percentage of correctly classified observations (PCC), the AUROC, and the percentage of true positives.

Models	$\bar{L}_3$	AUPROC_3	PCC	AUROC	True Pos.
Regression	90.64 %	88.58 %	89.38 %	92.55 %	25.84 %
Neural Network	89.55 %	85.47 %	91.43 %	**96.09 %**	47.01 %
Decision Tree	93.39 %	87.31 %	91.54 %	**94.72 %**	58.97 %
AdaCost	**96.28 %**	**96.13 %**	92.41 %	87.46 %	64.65 %
Cost-Sensitive Tree	**96.16 %**	**95.64 %**	90.13 %	94.45 %	80.60 %

Table 4: As in Table 3, but now for $h = 6$.

Models	$\bar{L}_6$	AUPROC_6	PCC	AUROC	True Pos.
Regression	85.21 %	83.74 %	80.61 %	88.62 %	26.22 %
Neural Network	93.61 %	77.34 %	84.82 %	**91.91 %**	53.43 %
Decision Tree	90.47 %	80.92 %	84.37 %	**91.14 %**	58.65 %
AdaCost	**95.62 %**	**95.42 %**	85.44 %	89.94 %	75.52 %
Cost-Sensitive Tree	**94.64 %**	**94.41 %**	77.06 %	85.80 %	96.13 %

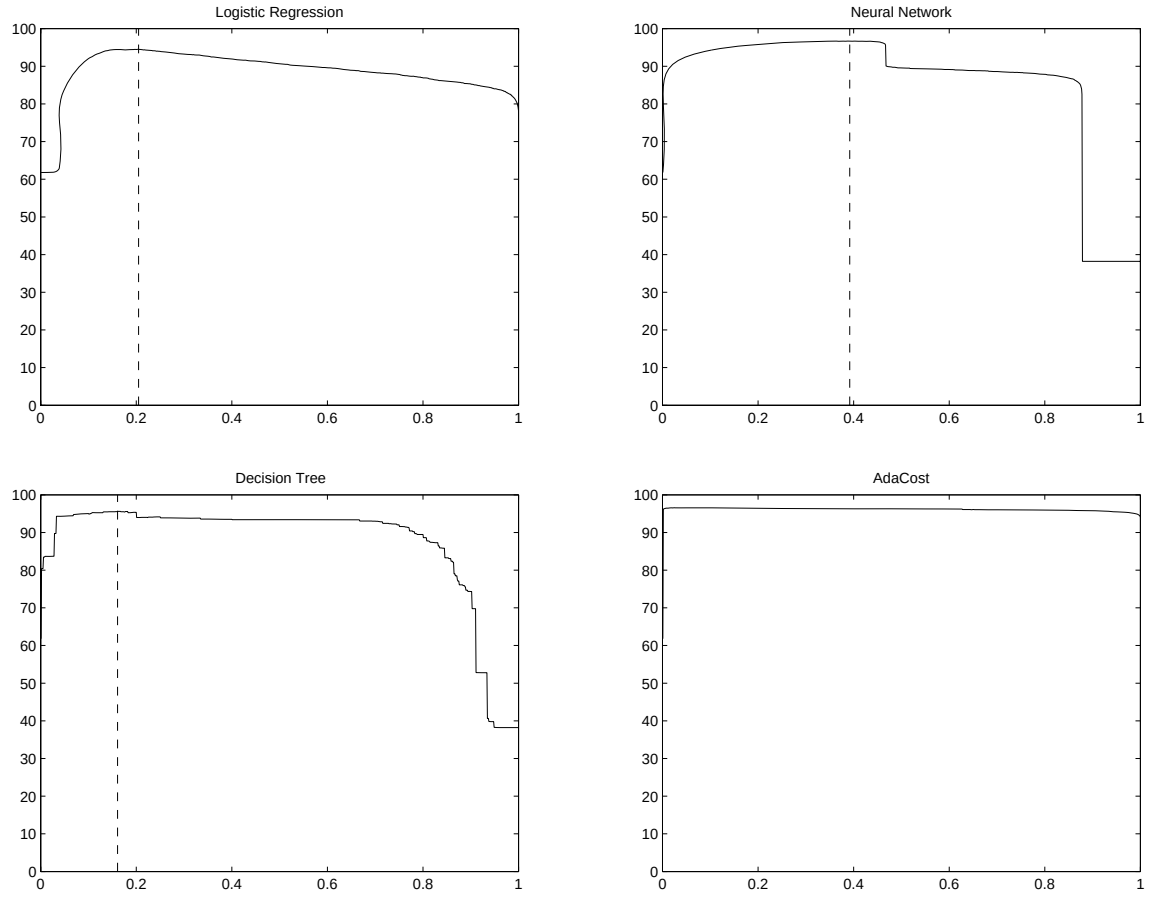


Figure 2: Profit Curves for $h = 3$, for the logistic regression, the neural network, the decision tree, and the AdaCost classifier. The dashed line indicates the maximum of the profit curve.

The profit curves, being defined in (14), are plotted in Figure 2, for the logistic regression, neural network and decision tree classifiers, together with the cost-sensitive AdaCost classifier (the profit curve for the cost-sensitive tree is similar to the latter one). The profit curve plots the cumulative profit percentage as a function of the cut-off value being used for classifying the observations as being churners or not. The plots for $h = 3$ are presented, the results for $h = 6$ being similar.

These profit curves are useful in deciding on the optimal cut-off value τ. The cut-off can be set at the maximum of the profit curve, hereby correcting for the asymmetry in the misclassification costs and the class distributions. Note that for AdaCost and the cost-sensitive decision tree the induced asymmetry is already taken into account in the construction of the classifier, hence for these methods we can still use the standard cut-off

value $\tau = 0.5$. For the non cost-sensitive classifiers, since false negatives are more expensive than false positives, all maxima are situated in the left halve of the plots. Hence, when setting the cut-off using the profit curve, more customers are classified as churners.

The area under the profit curve, summarizing the profit curve in a single number, provides an insight regarding the performance of the classifier predictions. The closer the predicted probabilities are to the extremes (0 for assumed perfect non-churners or 1 for assumed perfect churners), the higher will be the value of the area under the profit curve (AUPROC). For the same value of L_{tot}, different values of AUPROC can be obtained.

5.3 Discussion

From Tables 3 and 4, it follows that the classifiers achieving the best results in our empirical application are the AdaCost classifier and the cost-sensitive tree. They attain the highest values for the cumulative profit percentage and the AUPROC at both horizons. Since these classifiers directly include cost information in designing the classification models, it comes as no surprise that both give the best results in terms of profit. The other three classifiers are yielding a lower profit. Whereas, using the traditional non-profit based performance measures, the neural network and the decision tree give the best results. The latter observation adds to the mention of Fader et al. (2005) that there is a highly non-linear relationship between the RFM variables and the CLV. We see that non-linear model (neural network and decision tree) have a better classification accuracy than the linear one (logistic regression). Nevertheless, we do claim that in our application the relevant performance measures should be profit-sensitive.

One can see that it is well possible that two classifications methods have similar values for the PCC (or the AUROC), but perform very differently according to the profit-sensitive measures. As a matter of fact, if one would select a classifier on the basis of a standard measure of accuracy (e.g. AUROC), one would choose the neural network. Whereas, the neural network has the lowest AUPROC value. This difference is mainly explained by the fact that the misclassification cost is, on average, greater for churners than for non-churners. Consequently, the total profit for the classifiers that manage to correctly classify the churners (e.g. the cost-sensitive classifiers) is better than those that do not (e.g. the logistic regression). Nevertheless, even though the empirical study shows that, for a selected cut-off value, the proportion of true positives is crucial with respect to the profit generated, one cannot only consider the true positive accuracy. For example, as one can see from Tables

3 and 4, the cost-sensitive decision tree identifies the highest percentage of churners, whereas the Adacost classifier still has the highest values for the cumulative profit percentage and the AUPROC. The latter criteria also depend on the identification of non-churners, since profit can be made by not making a retention action for a non-churning customer.

Finally, the cost-sensitive decision tree achieved very good empirical results, in a computationally efficient way. It provides a good trade-off between the classifier construction simplicity and the profit maximization.

6 Conclusion

In this paper, we provide a framework for evaluating churner classification techniques based on a financial measure of accuracy, i.e. the profit loss incurred by a misclassification, considered from a customer lifetime value perspective. First, using a customer-centric approach, we define a churner as someone whose CLV is decreasing in time. Second, we emphasize the fact that not all customers are equal, neither are all misclassifications. Therefore, we propose a CLV-sensitive loss function and area based measure to evaluate the classifiers for several possible cut-off values. In our empirical setting, we use both traditional as well as cost-sensitive classifiers. We show that the cost-sensitive approaches achieve very good results in terms of the defined profit measure, emphasizing the point that, even though it is important to achieve a good average performance, it is at least as important to correctly classify potentially profitable churners.

In an ideal world, where every management control parameter would be known, it would be possible to estimate the exact monetary value of the retention campaign: that would be the sum of changes in the customer lifetime values generated by the model implementation. In reality, thanks to our CLV-sensitive loss function, one could have at least an insight on what is the approximate value of a customer retention campaign.

We can identify different topics for further research. As we have seen, the product usage growth rate α has a large impact on the CLV. In this paper, we assumed α to be constant. It would be interesting to allow varying α and investigate the impact on our findings. Further developments could focus on a more accurate prediction of this value or a more accurate prediction of the CLV. Also, the model we used to define the CLV has some limitations: we study only non-contractual product types without taking into account neither cross-product effects (cross-selling), nor cross-individual effects (word-to-mouth).

References

Agresti, A., 2002. Categorical data analysis. Wiley, Hoboken, New Jersey.

Baesens, B., Verstraeten, G., Van den Poel, D., Egmont-Petersen, M., Van Kenhove, P., Vanthienen, J., 2003. Bayesian network classifiers for identifying the slope of the customer lifecycle of long-life customers. European Journal of Operational Research 156 (2), 508–523.

Berger, P. D., Nasr, N. I., 1998. Customer lifetime value: Marketing models and applications. Journal of Interactive Marketing 12 (1), 17–30.

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.

Duin, R. P. W., Juszczak, P., Paclik, P., Pekalska, E., de Ridder, D., Tax, D. M. J., 2004. PRTools4, A Matlab Toolbox for Pattern Recognition. Delft University of Technology.

Dwyer, F. R., 1997. Customer lifetime valuation to support marketing decision making. Journal of Direct Marketing 11 (4), 6–13.

Egan, J. P., 1975. Signal detection theory and roc analysis. In: Series in Cognition and Perception. Academic Press, New York.

Fader, P. S., Hardie, B. G. S., Ka Lok Lee, 2005. RFM and CLV: Using iso-value curves for customer base analysis. Journal of Marketing Research 42 (4), 415–430.

Fan, W., Stolfo, S. J., Zhang, J., Chan, P. K., 1999. Adacost: misclassification cost-sensitive boosting. In: In Proc. 16th International Conf. on Machine Learning. Morgan Kaufmann, San Francisco, CA, pp. 97–105.

Freund, Y., Schapire, R. E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55 (1), 119–139.

Gupta, S., Lehmann, D. R., Stuart, J. A., 2004. Valuing customer. Journal of Marketing Research 41 (1), 7–18.

Malthouse, E. C., Blattberg, R. C., 2005. Can we predict customer lifetime value? Journal of Interactive Marketing 19 (1), 2–16.

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

Pfeifer, P. E., Haskins, M. R., Conroy, R. M., 2005. Customer lifetime value, customer profitability, and the treatment of acquisition spending. Journal of Managerial Issues 17 (1), 11–25.

Quinlan, J. R., 1992. C4.5: Programs for Machine Learning. Morgan Kaufmann Series in Machine Learning.

Reinartz, W. J., Kumar, V., 2000. On the profitability of long-life customers in a non contractual setting: An empirical investigation and implications for marketing. Journal of Marketing 64 (4), 17–35.

Rust, R. T., Lemon, K. N., Zeithaml, V. A., 2004. Return on marketing: Using customer equity to focus marketing strategy. Journal of Marketing 68 (1), 109–127.

Turney, P. D., 1995. Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. Journal of Artificial Intelligence Research 2, 369–409.

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.

Venkatesan, R., Kumar, V., 2004. A customer lifetime value framework for customer selection and resource allocation strategy. Journal of Marketing 68 (4), 106–125.