UNIVERSITEIT GENT

FACULTEIT ECONOMIE
EN BEDRIJFSKUNDE

HOVENIERSBERG 24
B-9000 GENT
Tel. : 32 - (0)9 – 264.34.61
Fax. : 32 - (0)9 – 264.35.92

# WORKING PAPER

## Churn Prediction in Subscription Services:

## an Application of Support Vector Machines While Comparing

## Two Parameter-Selection Techniques

**Kristof Coussement** [1]

**Dirk Van den Poel** [2]

September 2006

2006/412

[1] PhD candidate, Department of Marketing, Ghent University.
[2] Associate Professor of Marketing, Department of Marketing, Ghent University


D/2006/7012/57

# Churn Prediction in Subscription Services: an Application of Support Vector Machines While Comparing Two Parameter-Selection Techniques


## Abstract

CRM gains increasing importance due to intensive competition and saturated markets. With the purpose of retaining customers, academics as well as practitioners find it crucial to build a churn prediction model that is as accurate as possible. This study applies support vector machines in a newspaper subscription context in order to construct a churn model with a higher predictive performance. Moreover, a comparison is made between two parameter-selection techniques, needed to implement support vector machines. Both techniques are based on grid search and cross-validation. Afterwards, the predictive performance of both kinds of support vector machine models is benchmarked to logistic regression and random forests. Our study shows that support vector machines show good generalization performance when applied to noisy marketing data. Nevertheless, the parameter optimization procedure plays an important role in the predictive performance. We show that only when the optimal parameter selection procedure is applied, support vector machines outperform traditional logistic regression, whereas random forests outperform both kinds of support vector machines. As a substantive contribution, an overview of the most important churn drivers is given. Unlike ample research, monetary value and frequency do not play an important role in explaining churn in this subscription-services application. Even though most important churn predictors belong to the category of variables describing the subscription, the influence of several client/company-interaction variables can not be neglected.

Keywords: data mining, churn prediction, subscription services, support vector machines, parameter-selection technique

## 1. Introduction

Nowadays, more and more companies start to focus on Customer Relationship Management, CRM. Indeed due to saturated markets and intensive competition, a lot of companies do realize that their existing database is their most valuable asset (Athanassopoulos, 2000; Jones et al, 2000; Thomas, 2001). This trend is also notable in subscription services. Companies start to shift away from their traditional, mass marketing strategies, in favor of targeted marketing actions (Burez et al. 2006). It is more profitable to keep and satisfy existing customers than to constantly attract new customers who are characterized by a high attrition rate (Reinartz and Kumar, 2003). The idea of identifying those customers most prone to switching carries a high priority (Keaveney and Parthasarathy, 2001). It has been shown that a small change in retention rate can result in significant changes in contribution (Van den Poel and Larivière, 2004). In order to effectively manage customer churn within a company, it is crucial to build an effective and accurate customer-churn model. To accomplish this, there are numerous predictive-modeling techniques available. These data-mining techniques can effectively assist with the selection of customers most prone to churn (Hung et al., 2005). These techniques vary in terms of statistical technique (e.g. neural nets versus logistic regression), variable-selection method (e.g. theory versus stepwise selection), number of variables included in the model, time spent to build the final model, as well as in terms of allocating the time across the different tasks in the modeling process (Neslin et al., 2004).

This study contributes to the existing literature by investigating the effectiveness of the support vector machines (SVMs) approach in detecting customer churn in subscription services. Ample research focuses on predicting customer churn in different industries, including investment products, insurance, electric utilities, health care providers, credit card providers, banking, internet service providers, telephone service providers, online services, … . Although SVMs have shown excellent generalization performance in a wide range of areas like bioinformatics (Chen et al., 2005; He et al., 2005, Zhong et al., 2006), beat recognition (Acir, 2006), automatic face authentication (Bicego et al., 2005), evaluation of consumer loans (Li et al., 2006), estimating production values (Chen and Wang, 2006; Pai and Lin, 2005), text categorization (Bratko and Filipic, 2006), medical diagnosis (Glotsos et al., 2005), image classification (Kim et al., 2005) and hand-written digit recognition (Burges and Scholkopf, 1997; Cortes and Vapnik, 1995), the applications in marketing are rather scarce (Cui and Curry, 2005).

To our knowledge only a few implementations of SVMs in a customer churn environment are published (Kim et al. 2005; Zhao et al. 2005). This study will extend the use of SVMs in a customer-churn context in two ways: (1) Unlike former studies that implemented SVMs on a very small sample, this study applies SVMs in a more realistic churn setting. Indeed, once a churn model has been built, it must be able to accurately validate a new marketing dataset which contains in practice ten thousands of records and often a lot of noise. This study contributes to the existing literature by using a sufficient sample size for training and validating the SVM models in a subscriber

churn framework. These SVMs are benchmarked to logistic regression and state-of-the-art random forests. Neslin et al. (2004) concluded that logistic modeling may even outperform the more sophisticated techniques (like neural networks), while in a marketing setting random forests already proved to be superior to other more traditional classification techniques (Buckinx and Van den Poel, 2005; Larivière and Van den Poel, 2005). (2) Before SVMs can be implemented, several parameters have to be optimized in order to construct a first-class classifier. Extracting the optimal parameters is crucial when implementing SVMs (Hsu et al. 2004; Kim et al. 2005). Consequently, a fine-tuned parameter selection procedure has to be applied. Hsu et al. (2004) proposed a grid search and a cross-validation to extract the optimal parameters for SVMs. This procedure tries different parameter pairs on the training set using a cross-validation procedure. Hsu et al. (2004) propose to select that pair of parameters with the best cross-validation accuracy - i.e. percentage of cases correctly classified (PCC). The second contribution of this study lies in extending this principle by selecting one additional parameter pair. Not only the parameters with the best cross-validation accuracy – as proposed by Hsu et al. (2004) - are selected, also the parameter pair which results in the highest cross-validation area under the receiver operating curve (AUC) is used. In contrast to PCC, AUC takes into account the individual class performance – by use of the sensitivity and specificity - for several thresholds on the classifier's posterior churn probabilities (Egan 1975; Swets and Pickett 1982; Swets 1989). In the end, it is possible to compare the predictive performance of these two parameter-selection techniques with that of logistic regression and random forests.

As a substantive contribution, an overview of the most important churn predictors is given within this subscription-services setting. As such, marketing managers gain insight into which predictors are important in identifying churn. Consequently, it may be possible to adapt their marketing strategies based on this newly obtained information.

Following an introduction of the modeling techniques (i.e. SVMs, random forests and logistic regression), Section 3 explains the evaluation measures used in this study. The model-selection procedure for SVMs is presented in Section 4. Section 5 presents the research data, while Section 6 explains the experimental results. Conclusions and directions for future research are given in Section 7.

## 2 Modeling techniques

### 2.1 Support Vector Machines

The SVM approach is a novel classification technique based on neural network technology using statistical learning theory (Vapnik 1995, 1998). In a binary classification context, SVMs try to find a linear optimal hyperplane so that the margin of separation between the positive and the negative examples is maximized. This is equivalent to solving a quadratic optimization problem in which only the support vectors, i.e. the data points closest to the optimal hyperplane, play a crucial role. However, in practice, the data is often not linearly separable. In order to enhance the

feasibility of linear separation, one may transform the input space via a non-linear mapping into a higher dimensional feature space. This transformation is done by using a kernel function. There are some advantages in using SVMs (Kim et al., 2005): (1) there are only two free parameters to be chosen, namely the upper bound and the kernel parameter, (2) the solution of SVM is unique, optimal and global since the training of a SVM is done by solving a linearly constrained quadratic problem, (3) SVMs are based on the Structural Risk Minimization (SRM) principle, which means that this type of classifier minimizes the upper bound on the actual risk, compared to other classifiers which minimize the empirical risk. This results in a very good generalization performance.

We will give a general overview of a SVM for a binary classification problem. For more details about SVMs, we refer to the tutorial of Burges (1998).

Given a set of labeled training examples $\{x_i, y_i\}$ with $i = 1, 2, 3, \ldots, N$ where $y_i \in \{-1, 1\}$ and $x_i \in R^n$, and $n$ the dimension of the input space. Suppose that the training data is linearly separable, there exists a weight vector $w$ and a bias $b$ such that the inequalities

$$w \cdot x_i + b \geq 1 \text{ when } y_i = 1, \tag{1}$$

$$w \cdot x_i + b \leq -1 \text{ when } y_i = -1, \tag{2}$$

are valid for all elements of the training set. As such, we can rewrite these inequalities in the form:

$$y_i (w \cdot x_i + b) \geq 1 \text{ with } i = 1, 2, 3, \ldots, N. \tag{3}$$

Eq (3) comes down to find two parallel boundaries,

$$\text{B1: } w \,.\, x_i + b = 1, \tag{4}$$

$$\text{B2: } w \,.\, x_i + b = -1, \tag{5}$$

at the opposite sides of the optimal separating hyperplane,

$$\text{H*: } w \,.\, x + b = 0, \tag{6}$$

with margin width between the two boundaries equal to $2/||w||$. Thus one can find the pair of boundaries which gives the maximum margin by:

minimizing

$$\frac{1}{2}.w^2 \tag{7}$$

subject to

$$y_i\,(w \,.\, x_i + b) \geq 1 \tag{8}$$

This constrained optimization problem can be solved using the characteristics of the Lagrange multipliers ( $\alpha$ ) by

maximizing

$$W(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i x_j \tag{9}$$

subject to

$$\alpha_i \geq 0 \text{ with } i = 1, 2, 3, \ldots, \text{N and} \sum_i \alpha_i y_i = 0 \tag{10}$$

The weight vector could be stated as follows:

$$w = \sum_i \alpha_i y_i x_i \tag{11}$$

INSERT FIGURE 1 ABOUT HERE

The decision function *f(x)* can be written as

$$f(x) = \text{sgn}\ (w \,.\, x + b) = \text{sgn}\left[\sum_i \alpha_i y_i (x.x_i) + b\right] \tag{12}$$

where sgn is a sign function. In practice, the input data will often not be linearly separable. However, one can still implement a linear model by introducing a higher dimensional feature space to which an input vector is mapped via a non-linear transformation:

$$\Theta : X \rightarrow X' \tag{13}$$

$$x_i \rightarrow \Theta(x_i) \tag{14}$$

where $X$ is the input space, $\Theta$ is the non-linear transformation and $\Theta(x_i)$ represents the value of $x_i$ mapped into the higher dimensional feature space $X'$.

Therefore Equation (9) can be transformed to

$$W(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \Theta(x_i) \Theta(x_j) \tag{15}$$

subject to

$$\alpha_i \geq 0 \text{ with } i = 1, 2, 3, \ldots, \text{N and } \sum_i \alpha_i y_i = 0 \tag{16}$$

By mapping the input space into a higher dimensional feature space, the problem of high dimensionality and implementation complexity occurs. One can introduce the concept of inner product kernels. Consequently, there is no more need to know the exact value of $\Theta(x_i)$, only the dot inner product is considered which facilitates the implementation.

$$K(x_i, x_j) = \Theta(x_i).\Theta(x_j) \tag{17}$$

Therefore the decision function becomes

$$f(x) = \text{sgn}\left[\sum_i \alpha_i y_i \Theta(x).\Theta(.x_i) + b\right] = \text{sgn}\left[\sum_i \alpha_i y_i K(x.x_i) + b\right] \tag{18}$$

For resolving this decision function, several types of kernel functions are available as given in table 1.

INSERT TABLE 1 ABOUT HERE

INSERT FIGURE 2 ABOUT HERE

It is possible to extend these ideas to handle non-separable data. In this case, the margin will become very small and it will be impossible to separate the data without any misclassification. To solve this problem, we relax the constraints (1) and (2) by introducing positive slack variables (ε) (Cortes et al., 1995).

Equations (1) and (2) become

$$w \,.\, x_i + b \geq 1 - \varepsilon_i \text{ when } y_i = 1, \tag{19}$$

$$w \,.\, x_i + b \leq -1 + \varepsilon_i \text{ when } y_i = -1, \tag{20}$$

with $\varepsilon_i \geq 0$.

Equations (19) and (20) can be rewritten as

$$y_i \,(w \,.\, x_i + b) \geq 1 - \varepsilon_i \text{ with } i = 1, 2, 3, \ldots, \text{N}. \tag{21}$$

The goal of the optimization process is to find the hyperplane that maximizes the margin and minimizes the probability of misclassification:

minimize

$$\frac{1}{2}.w^2 + C\sum_i \varepsilon_i \tag{22}$$

subject to

$$y_i \,(w \,.\, x_i + b) \geq 1 - \varepsilon_i \tag{23}$$

with $C$, the cost, the penalty parameter for the error term. The larger $C$, the higher the penalty to errors.

Adapting Equation (15) to the non-separable case, one receives the following optimization problem:

maximizing

$$W(\alpha) = \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{24}$$

subject to

$$0 \le \alpha_i \le C \text{ with } i = 1, 2, 3, \ldots, \text{N and} \sum_i \alpha_i y_i = 0 \quad (25)$$

More details concerning the optimization process can be found in Chang and Lin (2004).

## 2.2 Random Forests

In a binary classification context, Decision Trees (DT) became very popular because of their easiness and interpretability (Duda et al., 2001). Moreover, DTs have the ability to handle covariates measured at different measurement levels. One major problem with DTs is their high instability (Hastie et al. 2001). A small change in the data often results in very different series of splits, which is often suboptimal when validating the trained model. In the past, this problem was extensively researched.

It was Breiman (2001) who introduced a solution to the previously mentioned problem. The new classification technique is called: Random Forests. This technique uses a subset of *m* randomly chosen predictors to grow each tree on a bootstrap sample of the training data. Typically, this number of selected variables – i.e. *m* - is much lower than the total number of variables in the model. After a large number of trees is generated, each tree votes for the most popular class. By aggregating these votes over the different trees, each case is predicted a class label.

Random forests are already applied in several domains like bioinformatics, quantitative criminology, geology, pattern recognition, medicine, … . However, the applications in marketing are rare (Buckinx and Van den Poel, 2005; Larivière and Van den Poel, 2005). Random forests are used as benchmark in this study, mainly for five reasons: (1) Luo et al. (2004) stated that the predictive performance is among the best of the available techniques. (2) The outcomes of the classifier are very robust to outliers and noise (Breiman, 2001). (3) This classifier outputs useful internal estimates of error, strength, correlation and variable importance (Breiman, 2001). (4) Reasonable computation time is observed by Buckinx and Van den Poel (2005). (5) Random forests are easy to implement because there are only two free parameters to be set, namely *m*, the number of randomly chosen predictors, and the total number of trees to be grown. We follow Breiman's (2001) suggestions: *m* is set equal to the square root of the total number of variables - i.e. 9 because 82 explanatory variables are included in the model - and a large number of trees - i.e. 1000 - are chosen.

### 2.3 Logistic Regression

Logistic regression is a well-known classification technique for predicting a dichotomous dependent variable. In running a logistic regression analysis, the maximum likelihood function is produced and maximized in order to achieve an appropriate fit to the data (Allison, 1999). This technique is very popular for mainly three reasons: (1) logit modeling is conceptually simple (Bucklin and Gupta, 1992). (2) A closed-form solution for the posterior probabilities is available (in contrary to SVMs); (3) It provides quick and robust results in comparison to other classification techniques (Neslin et al. 2004).

## 3 Evaluation Criteria

After building a predictive model, marketers want to use these classification models to predict future behavior. It is essential to evaluate the classifier in terms of performance. Firstly, the predictive model is estimated on a training set. Afterwards, this model is validated on an unseen dataset, the test set. It is essential to evaluate the performance on a test set, in order to ensure that the trained model is able to generalize well. For all three modeling techniques, PCC, AUC and the top-decile lift are calculated.

PCC, also known as accuracy, is undoubtedly the most commonly used evaluation metric of a classifier. Practically, the posterior churn probabilities generated by the classifier are ranked from most likely to churn to least likely to churn. All cases above a certain threshold are classified as churners; all cases having a lower churn probability are classified as non-churners. In sum, PCC computes the ratio of correctly classified cases to the total number of cases to be classified.

It is important to notice that PCC is highly dependent on the chosen threshold because only one threshold is considered. Consequently, it does not give an indication how the performance will vary when the cut-off is varied. Moreover, PCC does not consider the individual class performance of a classifier. For example, within a skewed class distribution, wrong predictions for the underrepresented class are very costly. Nevertheless, a model that predicts always the most common class - thus neglecting the

minority class- still provides a relatively good performance when evaluated on PCC.

Unlike PCC, AUC takes into account the individual class performance for all possible thresholds. In other words, AUC will compare the predicted class of an event with the real class of that event, considering all possible cut-off values for the predicted class. The receiver operating curve (ROC) is a graphical plot of the sensitivity - i.e. the number of true positives versus the total number of events - and 1-specificity - i.e. the number of true negatives versus the total number of non-events. The ROC can also be represented by plotting the fraction of true positives versus the fraction of false positives. The area under the receiver operating curve is used to evaluate the performance of a binary classification system (Hanley and McNeil, 1982). In order to assess whether AUCs of the different classification techniques are significantly different from each other, the non-parametric test of Delong et al. (1988) is used.

In marketing applications, one is especially interested in increasing the density of the real events. The top 10% decile is an evaluation measure that only focuses on the 10% cases most likely to churn. Practically, the cases are first sorted from predicted most likely to churn to predicted least likely to churn. Afterwards, the proportion of real events in the top 10% most likely to churn is compared with the proportion of real events in the total dataset. This increase in density is called the top-decile lift. For example, a top-decile lift of two means that the density of churners in the top 10% is twice the density of churners in the total dataset. The higher the top-decile lift, the better the classifier. Potentially this top-decile lift is very interesting to target, because it contains a higher number of real events. In other words, marketing analysts are

interested in just 10% of the customer base – i.e. those who are most likely to churn - because marketing budgets are limited and actions to reduce churn would typically involve only 10% of the entire list of customers.

## 4. Model Selection for the Support Vector Machines

First, we will argue why the radial basis function (RBF) kernel is used as the default kernel function throughout this study. Secondly, the grid-search method and cross-validation procedure for choosing the optimal penalty parameter C and kernel parameter $\gamma$ is explained. In a third section, two types of parameter selection techniques are described.

### 4.1 RBF Kernel Function

The RBF kernel function is used as the default kernel function within this study, mainly for four reasons (Hsu et al., 2004): (1) this type of kernel makes it possible to map the non-linear boundaries of the input space into a higher dimensional feature space. So unlike the linear kernel, the RBF kernel can handle a non-linear relationship between the dependent and the explanatory variables. (2) In terms of performance Keerthi and Lin (2003) concluded that the linear kernel with a parameter C has the same performance as the RBF kernel with parameters $(C, \gamma)$. Lin and Lin (2003) showed that the sigmoid kernel behaves like the RBF kernel for certain parameters. (3) When looking at the number of hyperparameters, the polynomial kernel has more hyperparameters than the RBF kernel. (4) The RBF kernel has less numerical

difficulties because the kernel values lie between zero and one, while the polynomial kernel values may go to infinity or zero while the degree is large. On the basis of these arguments, the RBF kernel is used as the default kernel function.

### 4.2 Optimal Parameter Selection Using Grid Search and Cross-Validation

The RBF kernel needs two parameters to be set; C and $\gamma$, with C the penalty parameter for the error term and $\gamma$ as the kernel parameter. Both parameters play a crucial role in the performance of SVMs (Hsu et al. 2004; Kim et al. 2005). Improper selection of these parameters can be counterproductive. Beforehand it is impossible to know which combination of $(C, \gamma)$ will result in the highest performance when validating the trained SVM to unseen data. Some kind of parameter selection procedure has to be done. Hsu et al. (2004) propose a 'grid search' on C and $\gamma$ and a *v*-fold cross-validation on the training data. The goal of this procedure is to identify the optimal C and $\gamma$, so that the classifier can accurately predict unseen data. A common way to accomplish this is 2-fold cross-validation, where the training set is divided into two parts of which one is unseen in training the classifier. This performance better reflects the capabilities of the classifier in validating unknown data. More generally, in a *v*-fold cross-validation, the training data is split into *v* subsets of equal size. Iteratively, one part is left out for validation, while the other remaining (*v*-1) parts are used for training. Finally, each case in the training set is predicted once. The cross-validation performance will better reflect the 'true' performance as when validating the classifier to unseen data, while the validation set stays untouched. In order to identify which parameter pair performs best, one can repeat this procedure with several pairs of $(C, \gamma)$. As such it is

possible to calculate a cross-validated evaluation measure for every parameter pair. In the end, it is possible to select these parameters based on the best cross-validated performance.

### 4.3 Two Parameter Selection Techniques

In this study, a grid search on C and $\gamma$ is performed on the training set using a 5-fold cross-validation. The grid search is realized by evaluating exponential sequences of C and $\gamma$ (i.e. C = $2^{-5}$, $2^{-3}$,…, $2^{13}$ ; $\gamma = 2^{3}$, $2^{1}$,…, $2^{-15}$ ). Basically, all combinations of $(C, \gamma)$ are tried and two pairs of parameters are restrained: (1) the one with the best cross-validated accuracy – as proposed by Hsu et al. 2004 - and (2) the one with the biggest cross-validated area under the receiver operating curve. This additional parameter pair is selected for the reason that unlike PCC, AUC considers the sensitivity and specificity as individual class performance metrics over all possible thresholds. Once these optimal parameter pairs are obtained, the whole training set is trained again. Both classifiers will be used to validate an unseen dataset. In the end, one can compare and benchmark the performance of both kinds of SVMs.

## 5 Research Data

For the purpose of this study, data from a Belgian newspaper publishing company is used. The subscribers have to pay a fixed amount of money depending on the length of subscription and the promotional offer given. The company doesn't allow ending the subscription prior to the maturity date. The churn-prediction problem in this

subscription context comes down to predicting whether the subscription will/will not be renewed within a period of four weeks after the maturity date. During this four-week period, the company still delivers the newspapers to the subscribers. In this way, the company gives the subscribers the opportunity to renew their subscription. Figure 3 graphically traces back the time window of analysis. We use subscription data from January 2002 through September 2005. Using this time frame, it is possible to derive the dependent variable and the explanatory variables. For constructing the dependent variable, the renewal points between July 2004 and July 2005 are considered. Consequently, a customer is considered as "churner" when his/her subscription is not renewed within four weeks after the expiry date. The explanatory variables contain information covering a 30-month period returning from every individual renewal point. These variables contain information about client/company interactions, renewal-related information, socio-demographics and subscription-describing information (see Appendix A). This variety of information is gathered at two levels: subscription level and subscriber level. At the subscription level, all information from the current subscription is included, while at the subscriber level, all information related to the subscriber is covered. For instance, one can calculate the total number of complaints on the current subscription only - i.e. the subscription level - , while one can also consider the total number of complaints of a subscriber covering all his/her subscriptions - i.e. subscriber level. Finally, one ends up with an individual timeline per subscriber for every renewal point in the time interval.

INSERT FIGURE 3 ABOUT HERE

We decided to randomly select two samples of sufficient size; the training set is used to estimate the model, while the test set is used to validate the model. The training set contains as many churners as non-churners because many authors emphasize the need for a balanced training sample in order to reliably differentiate between defectors and non-defectors (Dekimpe and Degraeve, 1997; Rust and Metters, 1996; Yamaguchi, 1992). So it is not uncommon to train a model with a non-natural distribution (Chan and Stolfo, 1998; Weiss and Provost, 2001). The test set contains a proportion of churners that is representative for the true population in order to approximate the predictive performance in a real-life situation. For both datasets, all variables are constructed in the same way. The explanatory variables are compiled over a 30 month period, while the dependent variable contains information whether the subscription will/will not be renewed.

INSERT TABLE 2 ABOUT HERE

## 6 Empirical Analysis

### 6.1 SVM models

After conducting the grid search on the training data, the optimal $(C, \gamma)$ is $(2^{13},2^{-7})$ with a cross-validated accuracy of 78.089%. Table 3 summarizes the results of the grid search using the cross-validated accuracy as an evaluation criterion. Furthermore, parameter pair $(2^{7},2^{-7})$ results in the highest cross-validated AUC, 84.702. Table 4 considers the results of the grid-search procedure with the cross-validated AUC as a performance measure.

INSERT TABLE 3 ABOUT HERE

INSERT TABLE 4 ABOUT HERE

These two parameters pairs are used to train a model on the complete training set. Two SVMs are obtained, namely SVMacc[3] and SVMauc[4]. Finally, both models can be validated on a test set.

On the one hand, one can compare the performance among both SVMs, while on the other hand both SVMs can be benchmarked with the performance of the logistic regression and random forests.

### 6.2 Comparing Predictive Performance among both kinds of SVMs

In this section, a comparison is made between the predictive performance of SVMacc and SVMauc. The evaluation is performed in terms of AUC, PCC and top-decile lift. Both models are trained on a balanced training set, while in the end these classifiers have to be evaluated on a dataset which represents the actual density of churners (see Table 2). In order to assess the sensitivity of the results to the actual proportion of churners in the dataset, we will compare the performance of both SVMs on artificial test sets with different class distributions. More specifically, we compare

[3] SVMacc = SVM generated using parameters based on the model with the best cross-validated accuracy during grid search
[4] SVMauc = SVM generated using parameters based on the model with the best cross-validated AUC during grid search

the 'natural' distribution[5] (11,14% churners) with the artificial ones (50%, 40%, 30%, 20%, 18%, 16%, 14%). These artificial sets are created by randomly undersampling the real test set – i.e. the one with 11,14% churners.

Figures 4 through 6 and Table 5 depict the performance of SVMacc and SVMauc for the different class distributions. As such a comparison can be made between both SVMs. As one may observe from Figure 4, SVMauc performs better than SVMacc within all class distributions in terms of AUC performance. In order to ensure that the differences in AUC are significant, the test proposed by Delong et al. (1988) is applied. As such one can compare if the AUCs between SVMacc and SVMauc are significantly different within a certain class distribution. Table 5 reveals that on all test sets that contain 30% churners or less, SVMauc significantly outperforms SVMacc on a 90% confidence level (Delong et al. 1988). When validated on the 'natural' distribution, SVMauc significantly outperforms SVMacc at the 95% confidence level. Figure 5 shows the performance of both SVMs in terms of PCC. Despite the fact that the differences in PCC are rather small, one may observe that SVMauc does not have an inferior performance compared to SVMacc when coming closer to the 'natural' distribution. Previous findings are confirmed when evaluating both SVMs using the top-decile lift. There is a gap in top-decile lift between SVMacc and SVMauc. SVMauc has a higher top-decile lift compared to SVMacc. This gap increases when deviating from the original training distribution – i.e. the one with 50% churners. On the 'natural' distribution, SVMauc succeeds in retaining more churners within the top 10% customer most likely to churn in comparison to SVMacc.

[5] i.e. the distribution that contains the proportion of churners that is representative for the true population.

INSERT FIGURE 4 ABOUT HERE

INSERT TABLE 5 ABOUT HERE

INSERT FIGURE 5 ABOUT HERE

INSERT FIGURE 6 ABOUT HERE

Table 6 compares the predictive capabilities between SVMacc and SVMauc on the real test set (see Table 2). One can clearly see the gap in performance. SVMauc exhibits better predictive performance than SVMacc when both models are evaluated on the real test set. In terms of PCC, the increase is 0.55 percent points. There is also a significant improvement in AUC of 0.24 (Delong et al., 1988). With respect to the top-decile lift, an increase from 4.209 to 4.492 is achieved.

INSERT TABLE 6 ABOUT HERE

In sum, when a SVM is trained with a non-natural distribution, it may be better to select its parameters during the grid search based on the cross-validated AUC. The new parameter-selection technique significantly improves the AUC and the top-decile lift of the model, while accuracy is certainly not decreased.

In the following part, we compare the performance of both kinds of SVMs with logistic regression and random forests.

### 6.3 Comparing Predictive Performance of SVMs, Logit and Random Forests

The evaluation measures on the real test set (see Table 2) for all models are represented in Tables 7, 8 and 9. Table 7 compares the predictive performance of logit, random forests, SVMacc and SVMauc in terms of PCC and AUC. Table 8 shows the results from the test of Delong et al. (1988) which investigates if the AUCs of two models are significantly different. One can find the top-decile lift for all models in Table 9.

INSERT TABLE 7 ABOUT HERE

INSERT TABLE 8 ABOUT HERE

INSERT TABLE 9 ABOUT HERE

Additionally, Tables 7, 8 and 9 give information concerning the performance of SVMacc and SVMauc benchmarked to logistic regression. Only SVMauc differs significantly in terms of predictive performance when compared to logistic regression. In contrast to SVMauc, SVMacc classifies fewer cases correctly than logistic regression. Moreover the test of Delong et al. (1988) confirms that the AUC of SVMacc is not significantly different from that of the logistic regression. The need to select the right parameter-selection technique is confirmed when looking at the top-decile lift

criterion. SVMauc identifies more churners than logistic regression, while the top-decile lift of SVMacc is lower than that of logit.

From Tables 7, 8 and 9, one can also compare the performance of both SVMs with the performance of the random forests. It is clear that despite the parameter selection technique, SVMs are surpassed by random forests.

In sum, it is shown that the parameter-selection technique influences the predictive performance of SVMs. Consequently, when a SVM is trained on a balanced distribution, it may be viable and preferable to consider other than the traditional parameter-selection methods. Each improvement in predictive performance will result in a better return on investment of subscriber-retention actions based on these prediction models. In this study, SVMs are trained on a non-natural distribution; it is shown that selecting the parameters based on the best cross-validated AUC results in a better performance than when selecting them based on the highest cross-validated accuracy as was suggested in Hsu et al. (2004). In sum, one may say that choosing the right parameter-selection technique is vital for optimizing a SVM application.

In the end, it would also be counterproductive to simply rely on traditional techniques like logistic regression. SVMs - in combination with the correct parameter selection technique - and random forests, both outperform logistic regression. Nevertheless, in this study random forests are better in predicting churn in the subscription services than SVMs.

### 6.4. Variable importance

In this section, an overview of the most important variables is given. This is done based on the outcome of the random forest importance measures for mainly two reasons: (i) Random forests give the best predictive performance compared to logistic regression and SVM. (ii) Unlike random forests, the SVM software does not produce an internal ranking of variable importance. Moreover, we do not report any measures for logistic regression - e.g. standardized estimates - because most measures are prone to multicollinearity. However, this is not a problem when the focus lies mainly on prediction. In this study, we will elaborate the top-10 most important churn predictors.

It is clear from Appendix B that the length of the subscription and recency – i.e. elapsed time since last renewal – which both belong to the category of variables describing a subscription[6] are ranked on top. Furthermore, another variable from the same category – i.e. the month of contract expiration - is part of the top-10 most explaining churn variables. In contrast to extant research (e.g. Bauer 1988), monetary value and frequency – i.e. the number of renewal points – are not present within the top-10 list of most important churn predictors in this study.

Although most important churn predictors are variables that belong to the group of variables describing a subscription, the impact of some client/company-interaction variables cannot be neglected when investigating the top-10 list of most important variables: (i) Variables related to the ability of voluntarily suspending the subscription – during holiday, during a business trip, … - are present in the top-10. (ii) Recency of

[6] see Appendix A

complaining – i.e. the elapsed time since the last complaint - is also present in the top-10 most important churn predictors. Consequently, efficient-complaint handling strategies are important. Tax et al. (1998) already stated that companies do not deal successfully with service failures because most companies underestimate the impact of efficient complaint handling. (iii) Moreover, this study shows that the variable which indicates whether or not a subscription started from own initiative belongs to the top-10 list in contrast to similar variables related to other purchase motivators like direct mailing campaigns, tele-marketing actions, face-to-face promotions, … .

In spite of the importance of age, one can conclude that socio-demographics do not play an important role in explaining churn in this study which confirms the finding of Guadagni and Little (1983) and more recently, Rossi et al. (1996).

## 7. Conclusions and Future Research

In this study, we show that SVMs are able to predict churn in subscription services. By mapping non-linear inputs into a high-dimensional feature space, SVMs break down complex problems into simpler discriminant functions. Because SVMs are based on the Structural Risk Minimization principle that minimizes the upper bound on the actual risk, they show a very good performance when applied to a new, noisy marketing dataset. To validate the performance of this novel technique, we statistically compare its predictive performance with those of logistic regression and random forests. It is shown that a SVM – which is trained on a balanced distribution - outperforms a logistic regression only when the appropriate parameter selection technique is applied.

However, when comparing the predictive capabilities of these SVMs with state-of-the-art random forests, our study indicates that SVMs are surpassed by the random forests.

Particularly in this study, we implement a grid search using a 5-fold cross-validation for obtaining the optimal upper bound C and kernel parameter $\gamma$ that are the most important when implementing a SVM. This study offers an alternative parameter selection technique that outperforms the previously used technique by Hsu et al. (2004). The way in which the optimal parameters are selected, can have significant influences on the performance of a SVM. Taking into account alternative parameter-selection techniques is crucial because even the smallest change in predictive performance can have significantly increases in the return on investment of the marketing-retention actions based on these prediction models (Van den Poel and Larivière, 2004).

In addition, one can say that academics as well as practitioners don't have to simply rely on traditional techniques like logistic regression. SVMs – in combination with the right parameter-selection technique – and random forests offer some alternatives. Nevertheless, a trade-off has to be made between the time allocated to the modeling procedure and the performance achieved.

In this study, most important churn predictors are part of the group of variables describing the subscription. Unlike ample research, monetary value and frequency are not present in the top-10 most important churn drivers. On the other hand, several client/company-interaction variables play an important role in predicting churn. In spite of the importance of age, socio-demographics do not play an important role in

explaining churn in this study.

Directions for future research are given by the fact that nowadays there is no complete working meta-theory to assist with the selection of the correct kernel function and SVM parameters. Deriving a procedure to select the proper kernel function and correct parameter values according to a specific type of classification problem is an interesting topic for further research. Furthermore, applying SVMs using a sufficient sample size can be very time-consuming due to the long computational time and often requires specific software. Before SVMs can be widely adopted, easy-to-use computer software should be available in the traditional data mining packages.

**Acknowledgements**

We would like to thank the anonymous Belgian publishing company for disposing their data. Next, we also like to thank (1) Ghent University for funding the PhD project of Kristof Coussement (BOF 01D26705) and (2) the Flemish government and Ghent University (BOF equipment 011B5901) for funding our computing resources during this project. Also special thanks to L. Breiman (†) for freely distributing the random forest software, as well as C.-C. Chang and C.-J Lin for sharing their SVM-toolbox, LIBSVM.

## Appendix A: Explanatory variables included in the churn-prediction model

*Client/company-interaction variables*: variables describing the client/company relationship:

- The number of complaints,
- Elapsed time since the last complaint,
- The average cost of a complaint (in terms of compensation newspapers),
- The average positioning of the complaints in the current subscription,
- The purchase motivator of the subscription,
- How the newspaper is delivered,
- The conversions made in distribution channel, payment method & edition,
- Elapsed time since last conversion in distribution channel, payment method & edition,
- The number of responses on direct marketing actions,
- The number of suspensions,
- The average suspension length (in number of days),
- Elapsed time since last suspension,
- Elapsed time since last response on a direct marketing action,
- The number of free newspapers.

*Renewal-related variable*: variables containing renewal-specific information:

- Whether the previous subscription was renewed before the expiry date,
- How many days before the expiry date, the previous subscription was renewed,
- The average number of days the previous subscriptions are renewed before expiry date,
- The variance in the number of days the previous subscriptions are renewed before expiry date,
- Elapsed time since last step in renewal procedure,
- The number of times the churner did not renew a subscription.

*Socio–demographic variables*: variables describing the subscriber:

- Age,
- Whether the age is known,
- Gender,
- Physical person (is the subscriber a company or a physical person),
- Whether contact information (telephone, mobile number, email) is available.

*Subscription-describing variables*: group of variables describing the subscription:

- Elapsed time since last renewal,
- Monetary value,
- The number of renewal points,
- The length of the current subscription,
- The number of days a week the newspaper is delivered (intensity indication),
- What product the subscriber has,
- The month of contract expiration.

## Appendix B: Variable importance measures

| No. | AvgNormImp | Variable Name | Level[7] | Relative variable[8] |
|---|---|---|---|---|
| 1 | 73.946 | The length of the current subscription | Subscription | |
| 2 | 65.335 | Elapsed time since last renewal | Subscription | |
| 3 | 59.460 | Elapsed time since last suspension | Subscriber | |
| 4 | 54.764 | Elapsed time since last suspension | Subscription | |
| 5 | 54.035 | The month of contract expiration | Subscription | |
| 6 | 52.705 | Age | Subscriber | |
| 7 | 51.467 | Elapsed time since last complaint | Subscriber | |
| 8 | 51.056 | The average suspension length (in number of days) | Subscriber | X |
| 9 | 50.251 | The purchase motivator of the subscription: own initiative | Subscription | |
| 10 | 48.560 | The average suspension length (in number of days) | Subscriber | |
| 11 | 48.073 | Elapsed time since last complaint | Subscription | |
| 12 | 47.330 | Monetary value | Subscription | |
| 13 | 46.882 | Elapsed time since last step in renewal procedure | Subscription | |
| 14 | 46.520 | Physical person: physical personYES/NO | Subscriber | |
| 15 | 44.811 | The variance in the number of days the previous subscriptions are renewed before expiry date | Subscriber | |
| 16 | 44.357 | The average number of days the previous subscriptions are renewed before expiry date | Subscriber | |
| 17 | 43.337 | Elapsed time since last response on a direct marketing action | Subscriber | |
| 18 | 42.310 | The average number of days the previous subscriptions are renewed before expiry date | Subscription | |
| 19 | 40.011 | The number of renewal points | Subscription | |
| 20 | 38.448 | The number of suspensions | Subscriber | X |
| 21 | 37.295 | The average suspension length (in number of days) | Subscription | X |
| 22 | 37.158 | The purchase motivator of the subscription: direct marketing action | Subscription | |
| 23 | 36.536 | The number of suspensions | Subscription | X |
| 24 | 35.519 | How many days before the expiry date, the previous subscription was renewed | Subscription | |
| 25 | 35.279 | Elapsed time since last conversion in payment method | Subscriber | |
| 26 | 33.802 | Elapsed time since last conversion in payment method | Subscription | |

[7] see section 5: Research Data.
[8] correction of the variable by using the length of subscription

| | | | | |
|---|---|---|---|---|
| 27 | 33.396 | The number of complaints | Subscriber | X |
| 28 | 33.146 | The average positioning of the complaints in the current subscription | Subscription | |
| 29 | 32.520 | The conversions made in payment method | Subscription | |
| 30 | 32.481 | The average suspension length (in number of days) | Subscription | |
| 31 | 32.107 | The conversions made in payment method | Subscription | X |
| 32 | 31.637 | The number of responses on direct marketing actions | Subscriber | X |
| 33 | 31.144 | The variance in the number of days the previous subscriptions are renewed before expiry date | Subscription | |
| 34 | 29.640 | The conversions made in payment method | Subscriber | |
| 35 | 28.116 | What product the subscriber has: edition X | Subscription | |
| 36 | 28.027 | The purchase motivator of the subscription: tele marketing action | Subscription | |
| 37 | 27.860 | What product the subscriber has: edition Y | Subscription | |
| 38 | 27.584 | The conversions made in payment method | Subscriber | X |
| 39 | 26.390 | Elapsed time since last conversion in edition | Subscriber | |
| 40 | 25.442 | The number of responses on direct marketing actions | Subscriber | |
| 41 | 24.942 | Elapsed time since last conversion in distribution channel | Subscription | |
| 42 | 24.802 | The number of suspensions | Subscriber | |
| 43 | 24.237 | The number of complaints | Subscription | X |
| 44 | 24.193 | Whether the previous subscription was renewed before the expiry date | Subscription | |
| 45 | 23.993 | Elapsed time since last conversion in edition | Subscription | |
| 46 | 23.545 | The purchase motivator of the subscription: promotional offer | Subscription | |
| 47 | 23.008 | The number of suspensions | Subscription | |
| 48 | 22.991 | Elapsed time since last conversion in distribution channel | Subscriber | |
| 49 | 22.486 | The number of complaints | Subscriber | |
| 50 | 21.466 | How the newspaper is delivered: private distribution channel | Subscription | |
| 51 | 20.917 | Gender: female YES/NO | Subscriber | |
| 52 | 20.087 | The number of complaints | Subscription | |
| 53 | 19.624 | Physical person: company YES/NO | Subscriber | |
| 54 | 19.600 | How the newspaper is delivered: individual newsboy | Subscription | |
| 55 | 18.930 | Whether the age is known | Subscriber | |
| 56 | 18.906 | The number of times the subscriber did not renew a subscription | Subscriber | |
| 57 | 18.426 | The conversions made in distribution channel | Subscriber | X |
| 58 | 17.802 | The conversions made in edition | Subscriber | X |

| | | | | |
|---|---|---|---|---|
| 59 | 17.718 | The conversions made in distribution channel | Subscriber | |
| 60 | 17.289 | The purchase motivator of the subscription: direct marketing action | Subscription | |
| 61 | 17.249 | The purchase motivator of the subscription: face-to-face marketing | Subscription | |
| 62 | 16.996 | The conversions made in edition | Subscriber | |
| 63 | 16.534 | The conversions made in distribution channel | Subscription | |
| 64 | 16.095 | The conversions made in edition | Subscription | X |
| 65 | 15.446 | The conversions made in distribution channel | Subscription | X |
| 66 | 15.406 | What product the subscriber has: edition Z | Subscription | |
| 67 | 15.222 | The conversions made in edition | Subscription | |
| 68 | 14.531 | The average cost of a complaint (in terms of compensation newspapers) | Subscriber | X |
| 69 | 13.995 | The average cost of a complaint (in terms of compensation newspapers) | Subscription | X |
| 70 | 13.602 | Gender: male YES/NO | Subscriber | |
| 71 | 12.587 | The average cost of a complaint (in terms of compensation newspapers) | Subscriber | |
| 72 | 12.005 | How the newspaper is delivered: public distribution channel | Subscription | |
| 73 | 11.830 | Gender: private company YES/NO | Subscriber | |
| 74 | 11.550 | The purchase motivator of the subscription: direct marketing mailing action | Subscription | |
| 75 | 11.059 | How the newspaper is delivered: pick up newspaper at shop | Subscription | |
| 76 | 10.651 | The average cost of a complaint (in terms of compensation newspapers) | Subscription | |
| 77 | 7.601 | Gender: public company YES/NO | Subscriber | |
| 78 | 7.027 | The number of free newspapers | Subscription | |
| 79 | 5.190 | The number of days a week the newspaper is delivered (intensity indication) | Subscription | |
| 80 | 4.979 | Whether contact information (telephone, mobile number, email) is available | Subscriber | |
| 81 | 2.991 | How the newspaper is delivered: delivered abroad via courier | Subscription | |
| 82 | 2.093 | What product the subscriber has: edition W | Subscription | |

**References**

Acir, N., A support vector machine classifier algorithm based on a perturbation method and its application to ECG beat recognition systems, Expert Systems with Applications, 31 (1) (2006) pp. 150-158.

Allison P.D., Logistic regression using the SAS sytem: theory and application, Cary, NC: SAS Institute Inc. (1999).

Athanassopoulos, A.D., Customer satisfaction cues to support market segmentation and explain switching behavior, Journal of Business Research 47 (3) (2000) pp. 191-207.

Bauer, C.L., A direct mail customer purchase model, Journal of Direct Marketing, 2, (3), (1988) pp. 16-24.

Bicego M., Grosso E. and Tistarelli M., Face authentication using one-class support vector machines, Lecture Notes in Computer Science 3781 (2005) pp. 15-22.

Bratko A. and Filipic B., Exploiting structural information for semi-structured document categorization, Information Processing & Management 42 (3) (2006) 679-694.

Breiman, L., Random forests, Machine Learning 45 (1) (2001) pp. 5-32.

Buckinx W. and Van den Poel D., Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting, European Journal Of Operational Research 164 (1) (2005) pp. 252-268.

Bucklin, R.E. and Gupta, S., Brand choice, purchase incidence and segmentation: an integrated modeling approach, Journal of Marketing Research, 29 (1992) pp. 201-215.

Burez, J. and Van den Poel, D., CRM at Canal+ Belgique: reducing customer attrition through targeted marketing, forthcoming in Expert Systems with Applications.

Burges, C.J.C. and Scholkopf, B., Improving the accuracy and speed of support vector machines, in Mozer, M., Jordan, M., Petche, T., Advances in Neural Information Processing Systems, Cambridge, M.A. MIT Press (1997).

Burges, C.J.C., A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery 2 (2) (1998) pp. 121-167.

Chan P. K. and Stolfo S. J., Learning with non-uniform class and cost distributions: a case study in credit card fraud detection, Proceedings Fourth Intl. Conf. On Knowledge Discovery and Data Mining (1998), pp. 164-168.

Chang, C.-C. and Lin, C.-J., LIBSVM: a library for support vector machines, Technical Report, Department of Computer Science and Information Engineering, National Taiwan University (2004).

Chen, K.-Y. and Wang, C.-H., A hybrid SARIMA and support vector machines in forecasting the production values of the machinery industry in Taiwan, forthcoming in Expert Systems with Applications.

Chen X.J., Harrison R. and Zhang Y.Q., Multi-SVM fuzzy classification and fusion method and applications in bioinformatics, Journal of Computational and Theoretical Nanoscience 2 (4) (2005) pp. 534-542.

Cortes, C. and Vapnik, V., Support-vector networks, Machine Learning 20 (3) (1995) pp. 273-297

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

Dekimpe, M.G. and Degraeve, Z., The attrition of volunteers, European Journal of Operational Research 98 (1) (1997) pp. 37-51.

DeLong, E.R., DeLong, D.M. and Clarke-Pearson, D.L., Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, Biometrics, 44 (3) (1988) pp. 837-845.

Duda R. O., Hart, P.E. and Stork, D.G., Pattern classification, Wiley, New York (2001).

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

Glotsos D., Tohka J. and Ravazoula P., Automated diagnosis of brain tumours astrocytomas using probabilistic neural network clustering and support vector machines, International Journal of Neural Systems 15 (1-2) (2005) pp. 1-11.

Guadagni, P.M. and Little J.D.C., A logit model of brand choice calibrated on scanner data, Marketing Science 2(3) (1983) pp. 203–238.

Hanley, J.A. and McNeil, B.J., The meaning and use of the area under a receiver operating characteristic (roc) curve, Radiology, 143 (1) (1982) pp. 29-36.

Hastie, T., Tibshirani, R. and Friedman, J., The elements of statistical learning: data mining, inference and prediction, Springer-Verlag (2001).

He J.Y., Hu H.J. and Harrison R., Understanding protein structure prediction using SVM_DT, Lecture Notes in Computer Science 3759 (2005) pp. 203-212.

Hsu, C.-W., Chang, C.-C. and Lin, C.-J., A practical guide to support vector classification, Technical Report, Department of Computer Science and Information Engineering, National Taiwan University (2004).

Hung S.-Y., Yen D.C. and Wang H.-Y., Applying data mining to telecom churn management, forthcoming in Expert Systems with Applications.

Jones, M.A., Mothersbaugh, D.L. and Beatty, S.E., Switching barriers and repurchase intentions in services, Journal of Retailing 76 (2) (2000) pp. 259-374.

Keaveney, S. and Parthasarathy M., Customer switching behavior in online services: an exploratory study of the role of selected attitudinal, behavioral and demographic factors, Journal of the Academy of Marketing Science 29 (4) (2001) pp. 374-390.

Keerthi, S.S. and Lin, C.-J., Asymptotic behaviours of support vector machines with gaussian kernel, Neural Computation 15 (7) (2003) pp. 1667-1689.

Kim S., Shin K.S. and Park K., An application of support vector machines for customer churn analysis: credit card case, Lecture Notes in Computer Science 3611 (2005) pp. 636-647.

Kim S.K., Yang S. and Seo K.S., Home photo categorization based on photographic region templates, Lecture Notes In Computer Science 3689 (2005) pp. 328-338.

Larivière B. and Van den Poel D., Predicting customer retention and profitability by using random forests and regression forests techniques, Expert Systems With Applications 29 (2) (2005) pp. 472-484.

Li, S.-T., Shiue, W. and Huang M.-H., The evaluation of consumer loans using support vector machines, Expert Systems with Applications, 30 (4)(2006) pp. 772-782.

Lin, H.-T. and Lin, C.-J., A study on sigmoid kernels for SVM and the training of non-psd kernels by SMO-type methods, Technical report, Department of Computer Science and Information Engineering, National Taiwan University (2003).

Luo, T., Kramer K., Goldgof, D.B., Hall, L.O., Samson, S., Remsen, A. and Hopkins, T., Recognizing plankton images from the shadow image particle profiling evaluation recorder, IEEE Transactions on Systems Man and Cybernetics Part B – Cybernetics, 34 (4) (2004) pp. 1753-1762.

Neslin, S.A., Gupta S., Kamakura W., Lu J. and Mason C., Defection detection: improving predictive accuracy of customer churn models, Working Paper (2004).

Pai P.F. and Lin C.S., Using support vector machines to forecast the production values of the machinery industry in Taiwan, International Journal of Advanced Manufacturing Technology 27 (1-2) (2005) pp. 205-210.

Reinartz, W. and Kumar, V., The impact of customer relationship characteristics on profitable lifetime duration, Journal of Marketing 67 (1) (2003) pp. 77-99.

Rossi, P.E., McCulloch R.E. and Allenby G.M., Value of household information in target marketing, Marketing Science 15 (1996) pp. 321–340.

Rust, R.T. and Metters, R., Mathematical models of service, European Journal of Operational Research 91 (3) (1996) pp. 427-439.

Swets, J.A., Roc analysis applied to the evaluation of medical imaging techniques, Investigative Radiology, 14 (1989), pp 109-121.

Swets, J.A. and Pickett, R.M., Evaluation of diagnostic systems: methods from signal detection theory, Academic Press, New York (1982).

Tax, S.S., Brown, S.W. and Chandrashekaran, M., Customer Evaluations of Service Complaint Experiences: Implications for Relationship Marketing, Journal of Marketing 62 (April) (1998) pp. 60-76.

Thomas, J.S., A methodology for linking customer acquisition to customer retention, Journal of Marketing Research 38 (2) (2001) pp. 262-268.

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

Vapnik, V., Statistical learning theory, Wiley, New York (1998).

Vapnik, V., The nature of statistical learning theory, Springer, New York (1995).

Weiss G. and Provost F., The effect of class distribution on classifier learning, Technical Report ML-TR-43, Department of Computer Science, Rutgers University (2001).

Yamaguchi, K., Accelerated failure-time regression models with a regression model of surviving fraction: an application to the analysis of 'permanent employment' in Japan, Journal of the American Statistical Association 87 (No. 418) (1992) pp. 284-292.

Zhao Y., Li B. and Li X., Customer churn prediction using improved one-class support vector machine, Lecture Notes in Artificial Intelligence 3584 (2005) pp. 300-306.

Zhong W., He J., Harrison R., Tai P.C. and Pan Y., Clustering support vector machines for protein local structure prediction, forthcoming in Expert Systems with Applications (2006).

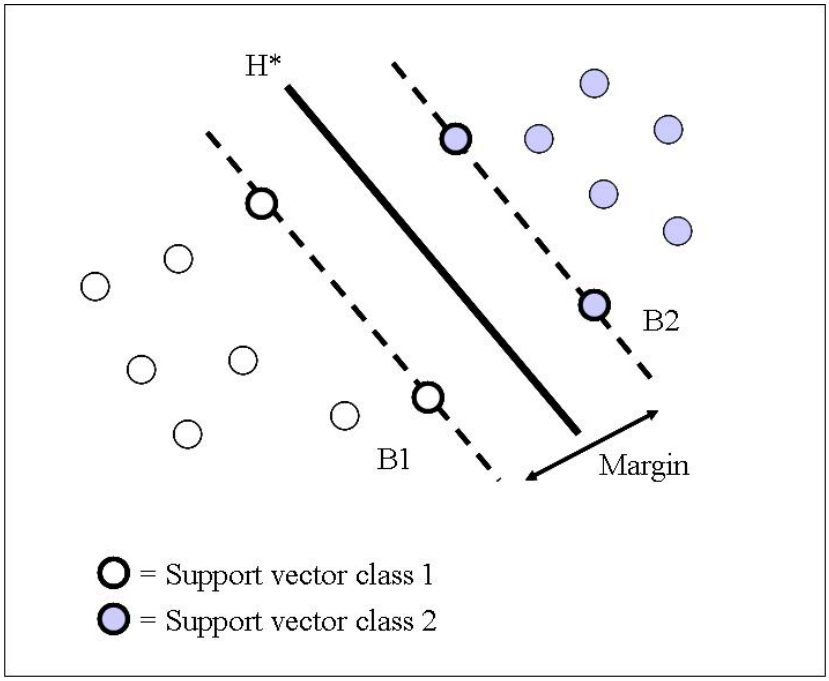


**Figure 1: This figure shows the solution for a binary linearly separable classification problem. The boundaries B1 and B2 separate the two classes. Data points on the boundaries are called support vectors. Thus one tries to find the hyperplane H* where the margin is maximal.**

| Kernel function | Mathematical form* |
|---|---|
| Linear Kernel | $K(x,x_i) = (x \cdot x_i)$ |
| Polynomial Kernel of degree d | $K(x,x_i) = (\gamma\, x \cdot x_i + r)\ \hat{}\ d$ |
| Radial Basis Function | $K(x,x_i) = \exp\{-\gamma \|x-x_i\|^2\}$ |
| Sigmoid Kernel with $r \in N$ | $K(x,x_i) = \tanh(\gamma\, x \cdot x_i + r)$ |

* $d, r \in N$; $\gamma \in R^+$

**Table 1: Overview of the different kernel functions.**

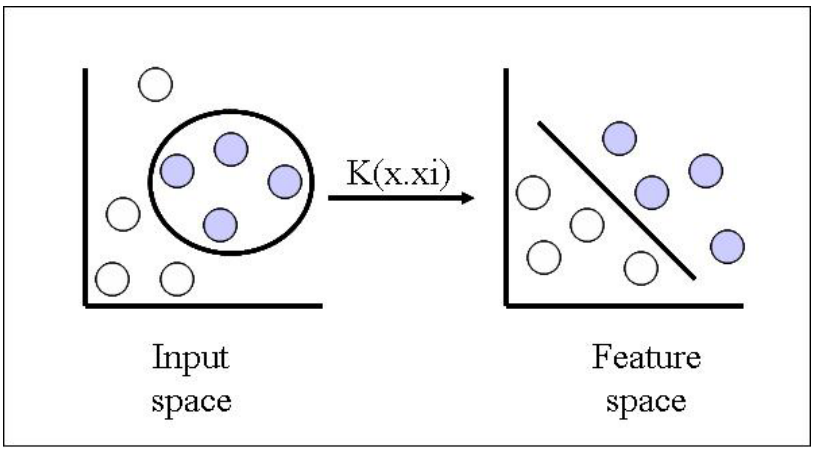


**Figure 2: The non-linear boundary in the input space is mapped via a kernel function into higher dimensional feature space. The data becomes linearly separable in the feature space.**

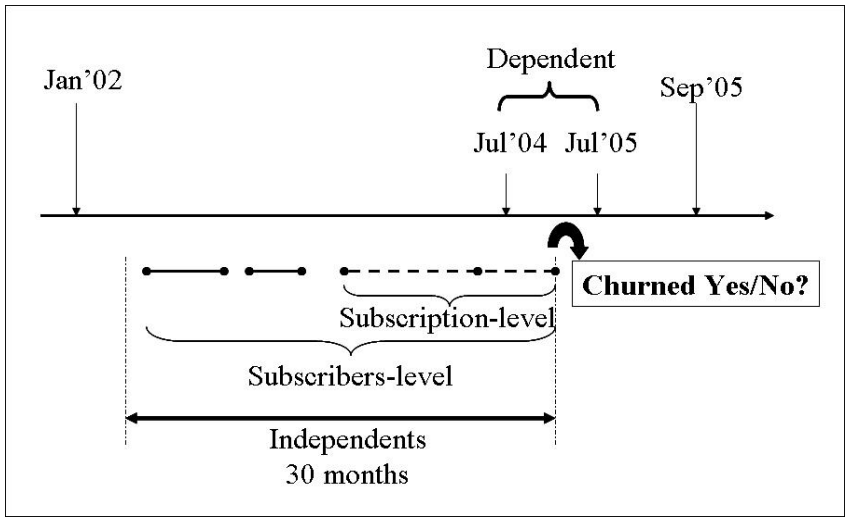


**Figure 3: Graphical display of the time window used to build the churn model.**

| | Number of observations | Relative percentage |
|---|---|---|
| **Training set** | | |
| Subscriptions not renewed | 22500 | 50% |
| Subscriptions renewed | 22500 | 50% |
| *Total* | *45000* | 100% |
| **Test set** | | |
| Subscriptions not renewed | 5014 | 11,14% |
| Subscriptions renewed | 39986 | 88,86% |
| *Total* | *45000* | 100% |

**Table 2: Distribution of the training set and test set.**

| $\gamma$ | C | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $2^{-5}$ | $2^{-3}$ | $2^{-1}$ | $2^{1}$ | $2^{3}$ | $2^{5}$ | $2^{7}$ | $2^{9}$ | $2^{13}$ |
| $2^{3}$ | 56,360 | 64,351 | 65,756 | 66,248 | 65,469 | 65,181 | 64,924 | 64,519 | 64,342 |
| $2^{1}$ | 68,147 | 70,525 | 71,733 | 71,418 | 70,453 | 69,458 | 68,836 | 68,314 | 68,087 |
| $2^{-1}$ | 75,353 | 76,582 | 77,127 | 76,262 | 74,627 | 72,958 | 71,947 | 70,859 | 70,394 |
| $2^{-3}$ | 75,959 | 77,144 | 77,649 | 77,622 | 77,558 | 76,440 | 74,918 | 73,320 | 71,842 |
| $2^{-5}$ | 74,789 | 76,164 | 76,960 | 77,471 | 78,039 | 77,996 | 77,924 | 78,056 | 76,396 |
| $2^{-7}$ | 74,367 | 74,948 | 75,975 | 76,440 | 77,118 | 77,758 | 77,719 | 78,084 | 78,089 |
| $2^{-9}$ | 75,163 | 74,349 | 74,827 | 75,907 | 76,167 | 76,693 | 76,959 | 77,722 | 77,726 |
| $2^{-11}$ | 74,240 | 75,209 | 74,344 | 74,840 | 75,856 | 76,107 | 76,271 | 76,517 | 77,144 |
| $2^{-13}$ | 54,767 | 74,213 | 75,198 | 74,403 | 74,836 | 75,860 | 76,093 | 76,202 | 76,398 |
| $2^{-15}$ | 50,000 | 64,406 | 74,103 | 75,198 | 74,406 | 74,829 | 75,872 | 76,089 | 76,182 |

**Table 3: The cross-validated accuracy per (C, $\gamma$ ).**

| $\gamma$ | C | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $2^{-5}$ | $2^{-3}$ | $2^{-1}$ | $2^{1}$ | $2^{3}$ | $2^{5}$ | $2^{7}$ | $2^{9}$ | $2^{13}$ |
| $2^{3}$ | 75,710 | 76,201 | 76,083 | 75,279 | 74,283 | 73,669 | 73,273 | 72,918 | 72,610 |
| $2^{1}$ | 80,059 | 80,221 | 80,092 | 78,600 | 77,058 | 75,878 | 75,007 | 74,441 | 74,085 |
| $2^{-1}$ | 82,703 | 83,552 | 83,722 | 82,728 | 80,951 | 79,069 | 77,616 | 76,386 | 75,442 |
| $2^{-3}$ | 83,865 | 84,296 | 84,507 | 84,472 | 83,857 | 82,406 | 80,500 | 78,402 | 76,388 |
| $2^{-5}$ | 83,373 | 83,926 | 84,172 | 84,496 | 84,691 | 84,592 | 84,212 | 83,239 | 81,745 |
| $2^{-7}$ | 82,871 | 83,188 | 83,670 | 83,896 | 84,172 | 84,514 | 84,702 | 84,699 | 84,477 |
| $2^{-9}$ | 82,232 | 82,810 | 83,087 | 83,506 | 83,625 | 83,861 | 84,173 | 84,504 | 84,674 |
| $2^{-11}$ | 80,998 | 82,229 | 82,790 | 83,059 | 83,448 | 83,462 | 83,593 | 83,869 | 84,190 |
| $2^{-13}$ | 72,936 | 80,996 | 82,228 | 82,785 | 83,052 | 83,431 | 83,393 | 83,436 | 83,601 |
| $2^{-15}$ | 50,000 | 72,987 | 80,995 | 82,228 | 82,784 | 83,051 | 83,427 | 83,377 | 83,375 |

**Table 4: The cross-validated performance (AUC) per (C, $\gamma$ )**

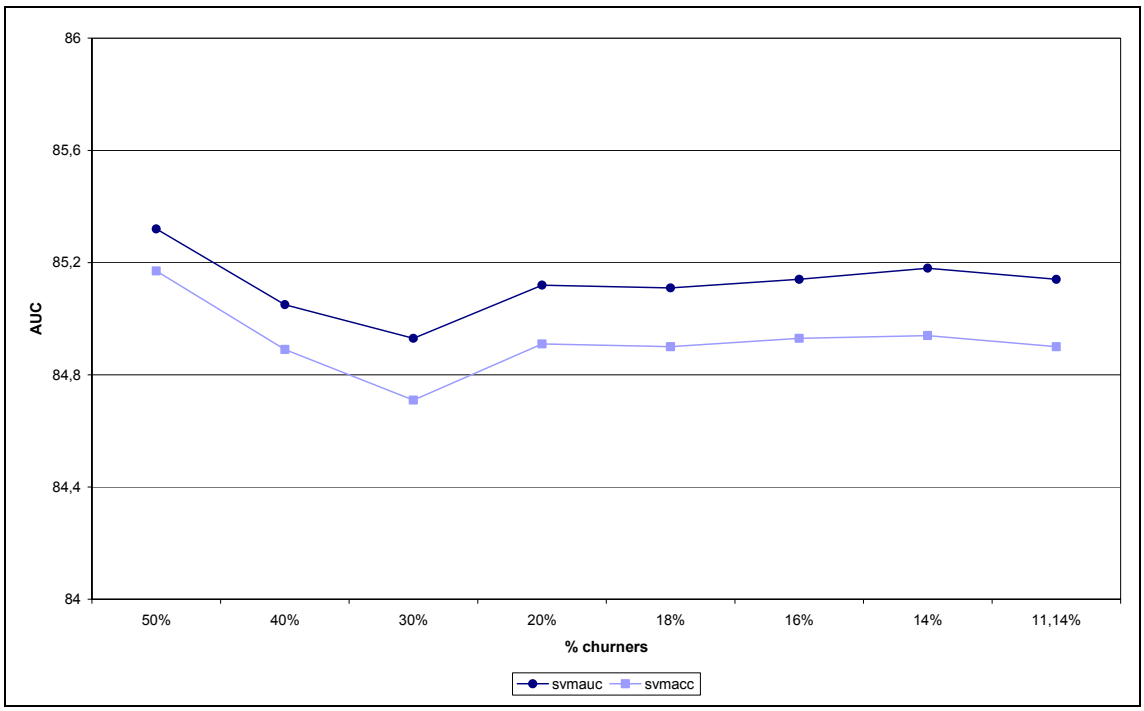


**Figure 4: Area under the receiver operating curve for SVMacc and SVMauc applied to several test sets with different class distributions**

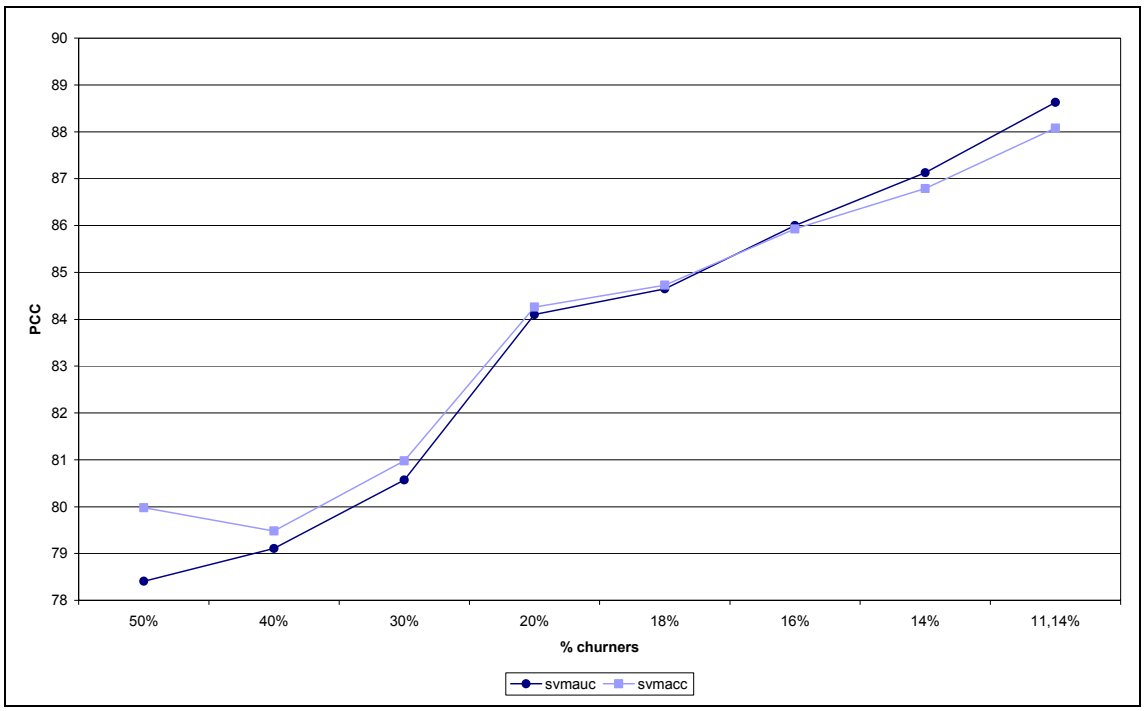


**Figure 5: Percentage correctly classified for SVMacc and SVMauc applied to several test sets with different class distributions**

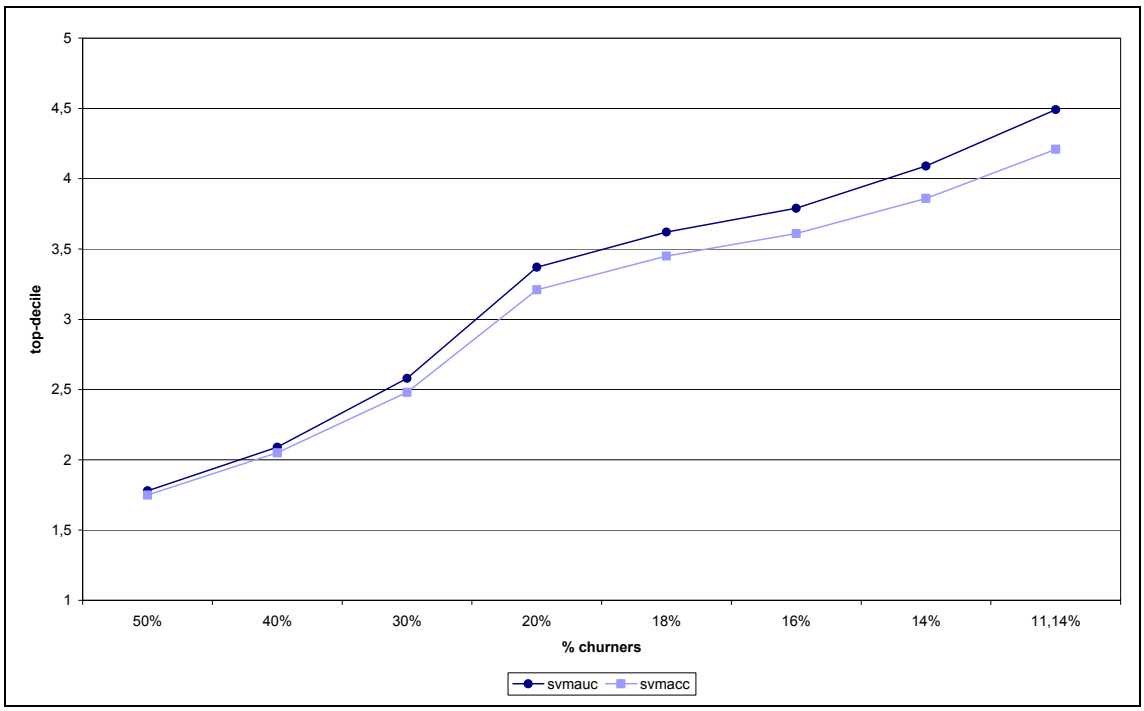


**Figure 6: Top-decile lift for SVMacc and SVMauc applied to several test sets with different class distributions**

| Number of churners | SVMacc – SVMauc |
|---|---|
| 50% | 1,16 (1); 0,281 |
| 40% | 1,57 (1); 0,210 |
| 30% | 3,65 (1); 0,056[a] |
| 20% | 3,78 (1); 0,052[a] |
| 18% | 4,35 (1); 0,037[a,b] |
| 16% | 4,27 (1); 0,039[a,b] |
| 14% | 5,96 (1); 0,014[a,b] |
| 11,14% | 6,04 (1); 0,014[a,b] |

$Chi^2$ (df); p-value
(a) significantly different on 90% confidence level
(b) significantly different on 95% confidence level

**Table 5: Pairwise comparison of performance (AUC) among several test sets using different class distributions**

| | PCC | AUC | Top-Decile lift |
|---|---|---|---|
| SVMacc | 88,08 | 84,90[a] | 4,209 |
| SVMauc | 88,63 | 85,14[a] | 4,492 |

(a) significantly different on 95% confidence level

**Table 6: The performance of SVMacc and SVMauc: PCC and AUC on the real test set**

| Model | PCC | AUC |
|---|---|---|
| Logit | 88,47 | 84,60 |
| Random Forests | 89,14 | 87,21 |
| SVMacc | 88,08 | 84,90 |
| SVMauc | 88,63 | 85,14 |

**Table 7: The performance of the different algorithms: PCC and AUC on the real test set**

| | Random forests | SVMacc | SVMauc |
|---|---|---|---|
| Logit | 219,52 (1)[a] | 2,53 (1)[b,c] | 12,56 (1)[a] |
| Random forests | | 190,44 (1)[a] | 166,58 (1)[a] |
| SVMacc | | | 6,04 (1)[a] |

$Chi^2$ (df)
(a) significantly different on 95% confidence level
(b) equal on a 95% confidence level
(c) equal on a 90% confidence level

**Table 8: Pairwise comparison of performance (AUC) on the real test set**

| Logit | Random forests | SVMacc | SVMauc |
|---|---|---|---|
| 4,478 | 4,754 | 4,209 | 4,492 |

**Table 9: The performance of the different algorithms: Top-decile lift on the real test**