FACULTEIT ECONOMIE
EN BEDRIJFSKUNDE

TWEEKERKENSTRAAT 2
B-9000 GENT
Tel. : 32 - (0)9 – 264.34.61
Fax. : 32 - (0)9 – 264.35.92

WORKING PAPER

Improving Customer Attrition Prediction by Integrating Emotions from Client/Company Interaction Emails and Evaluating Multiple Classifiers.

Kristof Coussement[1]

Dirk Van den Poel [2]

July 2008

2008/527

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

For more full-paper downloads about Customer Relationship Management: visit www.crm.UGent.be

D/2008/7012/36

Abstract

Predicting customer churn with the purpose of retaining customers is a hot topic in academy as well as in today's business environment. Targeting the right customers for a specific retention campaign carries a high priority. This study focuses on two aspects in which churn prediction models could be improved by (i) relying on customer information type diversity and (ii) choosing the best performing classification technique. (i) With the upcoming interest in new media (e.g. blogs, emails, ...), client/company interactions are facilitated. Consequently, new types of information are available which generate new opportunities to increase the prediction power of a churn model. This study contributes to the literature by finding evidence that adding emotions expressed in client/company emails increases the predictive performance of an extended RFM churn model. As a substantive contribution, an in-depth study of the impact of the emotionality indicators on churn behaviour is done. (ii) This study compares three classification techniques – i.e. Logistic Regression, Support Vector Machines and Random Forests – to distinguish churners from non-churners. This paper shows that Random Forests is a viable opportunity to improve predictive performance compared to Support Vector Machines and Logistic Regression which both exhibit an equal performance.

Keywords: Churn Prediction, Subscription Services, Call Center Email, Classification, Random Forests, Support Vector Machines.

1. Introduction

As markets become increasingly saturated, academic researchers and companies have acknowledged that focussing on identifying customer most likely to churn is of crucial importance (Keaveney and Parthasarathy, 2001). Reinartz et al. (2004) remark that organizations are realizing that not all customers generate the same economic value to the company. Establishing valuable relationships with existing customers produces higher revenues and margins than attracting new customers (Reichheld and Sasser, 1990). Consequently, investments for retention strategies have a higher net return than for acquisitions. So, it is supported that companies first spend their marketing resources to keep existing customers rather than to attract new ones (Rust and Zahorik, 1993). In order to preserve the existing customer base, marketing consultants try to proactively target those customers most likely to churn. Indeed, companies are moving away from traditional mass marketing strategies in favour of a customer-focussed strategy (Burez and Van den Poel, 2007). As such, Customer Intelligence researchers are increasingly investigating the underlying break-through of improving customer attrition modelling in several research settings (e.g. Larivière and Van den Poel, 2004, Burez and Van den Poel, under review).

In recent years, data mining techniques explore and analyse huge amounts of available data in order to assist with the selection of customers most prone to switch (Hung et al., 2006). Many academics and practitioners have built different model types that attempt to predict customers' future behaviour. The research on improving the methodology of churn prediction models is still growing because of (a) it is worthwhile to target customers proactively based on a churn prediction model by sending them churn prevention actions. For instance, a field experiment by Burez and Van den Poel (2007) has shown that companies can double their

profits by sending prevention actions to their most likely churners based on a prediction model, (b) keeping existing customers is less expensive then acquiring new ones which are often characterized by a high attrition rate (Reinartz and Kumar, 2003) and (c) the churn prediction system has to be as accurate as possible, because Van den Poel and Larivière (2004) have shown that even a small change in retention rate can result in significant changes in contribution.

This study focuses on two aspects on how to optimize a classification model, i.e. (i) incorporated information and (ii) classification technique. (i) Different types of information are available for churn prediction. Fader et al. (2005) conclude that customer's past behaviour is an important predictor for one's future behaviour. In the direct marketing literature, it is common practice to summarize customers' past behaviour in terms of their Recency (i.e. the elapsed time since last purchase or renewal), Frequency (i.e. the number of prior purchases or renewals) and Monetary value (i.e. the total amount of purchases) or their RFM characteristics. However, attrition models are often more complex because several other variables, such as socio-demographics and other transactional data, are included on top of the RFM characteristics. We will refer to such a model as an extended RFM model (or eRFM model).

Nowadays, new opportunities arise to increase the prediction power of an eRFM model by incorporating information from new media (e.g. websites, blogs and emails). New valuable information is available to the data analyst because customers interact more frequently with the company through these media types. Indeed, Weng and Liu (2004) conclude that clients become more familiar with sending emails as a substitute for traditional communication. Moreover, the customer relationship literature states that interactions between client and company are potentially important (Ganesan, 1994). As a consequence, academics and

marketing consultants try to improve customer relationships by incorporating client/company information in their analysis. For instance, online feedback mechanisms experience growing popularity and have important implications for a wide range of management activities including customer acquisition and retention (Dellarocas, 2003). Moreover, call centres which handle telephone calls as well as emails are often seen as the link between customers and the organization. They are potentially important because they offer clients the possibility (a) to report service failures or (b) to ask product related information. (a) Indeed, complaint handling is an important tool to win competitive advantage (Brown, 1997) and it provides a good way to enhance the retention of customers who experience service problems (e.g. Hart et al., 1990). (b) Besides complaints, information requests are another way in which clients interact with the company. For instance, a lot of customers ask information about promotional deals and subscription related aspects. However, it is remarkable that the impact of information requests on customer churn is underinvestigated. All client/company interaction information offers the marketing manager a unique opportunity to explore the client/company relationship and to improve the performance of their churn model. More specifically, Mattsson et al. (2004) focus on the fact that client/company interactions express several emotions via the words they contain. Commonly, one argues that these interactions are homogeneous in terms of expressed emotions, but this is certainly not the case. A common typology to classify emotions is to consider positive as well as negative emotions (Chaudhuri, 1998). This study investigates whether these distinct emotions from call centre emails increase the prediction performance of a churn system. Furthermore, it deepens out the impact of these emotions expressed in client/company emails on one's attrition behaviour.

(ii) Next to the incorporation of new information types, several classification techniques are at the disposal of a data analyst. This study benchmarks the predictive performance of two state-of-the-art classifiers, Support Vector Machines and Random Forests, with the base classifier in marketing, Logistic Regression. It was Neslin et al. (2006) who conclude that Logistic Regression is a well-known and robust classification technique in marketing, while Thomas (2000) confirms that it is most often used by predictive model builders in industry. As a consequence, Logistic Regression is used to benchmark Support Vector Machines and Random Forests, two state-of-the-art classifiers within this study. Support Vector Machines have already proven his excellence performance in a wide range of industries like bioinformatics, beat recognition, image classification, ... while over the years it is gaining popularity in churn prediction too (e.g. Kim et al., 2005; Zhao et al., 2005 and Coussement and Van den Poel, 2008). Moreover, Random Forests have also shown their good predictive capabilities in a lot of industries including customer churn prediction (e.g. Larivière and Van den Poel, 2005). The purpose of this research study is to find evidence which state-of-the-art classification technique performs best in optimizing the performance of a churn system.

In conclusion, consolidating customer relationships by avoiding attrition is an important issue for marketing managers and CRM consultants. A first step in addressing this issue is finding who to target in retention actions. The choice of the most appropriate classification technique is an important issue in improving the performance of a churn model. This study compares the predictive performance of Logistic Regression, Support Vector Machines and Random Forests in distinguishing churners from non-churners. Moreover, customer's past behaviour is an important predictor for one's future behaviour by which RFM models are typically built. Such models are often extended with other transactional and socio-demographic variables.

Due to the rapid development of internet and information technology, new client/company information is available. This study investigates whether the emotions expressed in customer emails have an impact on subsequent churn behaviour. Therefore, this paper contributes to the existing literature by investigating (i) whether the predictive performance of an eRFM model increases when emotionality indicators of client/company interaction variables are included?, (ii) which classification technique – i.e. Logistic Regression, Support Vector Machines or Random Forests - performs best in distinguishing churners from loyal customers? and (iii) how do emotions expressed in written client/company interactions through call centre emails (i.e. service failures and information requests) relate to one's churn behaviour?

This paper is organised as follows. Section 2 gives an overview of the classification methods (i.e. Logistic Regression, Support Vector Machines and Random Forests) used throughout this study, while Section 3 gives an overview of the evaluation criteria for the different classification models. Section 4 describes how the emotionality indicators are extracted from the call center emails. In a next Section, the research setting is explained. The empirical results are given in Section 6, while in a last Section conclusions are presented.

2. Classification Methods

This paragraph introduces the three classification techniques used throughout this study, i.e. Logistic Regression, Support Vector Machines and Random Forests.

2.1. Logistic Regression

Logistic Regression is a well-known technique that is often used in traditional marketing applications (Neslin et al., 2006). Moreover, it is a simple technique (Bucklin and Gupta, 1992), while it provides quick and robust results. Furthermore, a closed-form solution for the 'a posteriori' probabilities is available. Logistic Regression tries to maximize the log-likelihood function in order to become an appropriate fit to the data (Allison,1999). Including all predictors into one regression model often results in overfitting and poor predictions, in settings where many variables have little to add to the prediction model. Kim (2006) states that variable selection improves the comprehensibility of the resulting model and makes the resulting models generalize better. As done by many researcher and consultants (e.g. Neslin et al., 2006), this study employs a stepwise Logistic Regression for churn prediction.

2.2 Support Vector Machines

Support Vector Machines (SVMs) were introduced by Vapnik and his colleagues for solving binary classification problems (Cortes and Vapnik, 1995 and Vapnik, 1998). The purpose of SVMs in a binary classification context comes down to finding an optimal hyperplane that maximizes the margin between positive and negative examples. We refer to the tutorial of Burges (1998) for more details about SVMs, while more information on the optimization process is provided by Chang and Lin (2004).

In order to implement SVMs, a decision on the kernel function is needed. This study uses the RBF kernel (instead of the linear kernel, the sigmoid kernel or the polynomial kernel) as the default kernel function. In contrast to the *linear kernel* function, RBF kernel makes it possible

to map non-linear boundaries of the input-space into a higher dimensional feature space (Hsu et al., 2004). Lin and Lin (2003) found in their research that the *sigmoid kernel* behaves like the RBF kernel for certain parameters. When looking at the number of hyperparameters, the *polynomial kernel* has more hyperparameters than the RBF kernel, which makes the optimization process more complex. Moreover, 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. Considering these arguments, the RBF kernel function is used as the default kernel function throughout this study.

In order to obtain an optimal performance using a RBF kernel function, one needs to optimize two parameters, namely C and γ with C the penalty parameter for the error term and γ the kernel parameter. Both parameters play a crucial role in the performance of SVMs (e.g. Hsu et al. 2004). This study applies a 'grid-search' on C and γ with a two-fold cross-validation as described by Hsu et al. (2004) for optimizing these parameters. The best parameter pair (C, γ) based on the highest cross-validated AUC is used for further analysis.

2.3. Random Forests

Random Forests is a classification technique that was introduced by Breiman (2001). It is an ensemble technique that grows many classification trees in order to overcome the instability of a traditional decision tree (Hastie et al. 2001). To classify a new object from an input vector, the input vector is run through each of the trees in the forest. Each tree gives a classification and votes for the most popular class. The forest chooses to classify the case according to the label with the most votes over all the trees in the forest.

We follow Breiman's (2001) suggestions where the number of variables for growing the tree is set equal to the square root of the total number of variables and a large number of trees - i.e. 1,000 - is chosen.

3. Evaluation Criteria

In order to evaluate the performance of the classification techniques, two criteria are used throughout this study, namely the Percentage Correctly Classified (PCC) and the Area Under the receiving operating Curve (AUC). Both measures are often used as performance criteria in different retention studies (e.g. Buckinx and Van den Poel, 2005). The PCC compares the 'a posteriori' probability of being a churner with the true status of that customer. If TP, FP, TN and FN are the True Positives, False Positives, True Negatives and False Negatives in the confusion matrix, then the PCC is defined as (TP+TN)/(TP+FP+TN+FN). The disadvantage of PCC is that it is not very robust concerning the chosen cut-off value on the 'a posteriori' churn probability (Baesens et al., 2002). In contrast to PCC, AUC performance takes into account all possible cut-off levels on the 'a posteriori' probabilities. For these cut-off points, it considers the sensitivity (i.e. the number of True Positives versus the total number of churners) and the specificity (i.e. the number of True Negatives versus the total number of non-churners) of the confusion matrix in a two-dimensional graph, resulting in the Receiving Operating Curve or ROC curve. The area under the ROC curve is used to evaluate the performance of a classification model (Hanley and McNeil, 1982). In order to compare if two classification models are significantly different in terms of AUC, the non-parametric test of Delong et al. (1988) is used.

4. Extracting Emotions from Client/Company Interaction Emails.

The ways in which individuals write provide windows into their emotive and cognitive worlds. For instance, Pennebaker and Francis (1996) investigated cognitive, emotional and language processes in disclosure, while Newman et al. (2003) predict deception from language characteristics. Text files are analyzed using the computerized text analysis program Linguistic Inquiry and Word Count, LIWC (Pennebaker, 2001 and Zijlstra et al., 2004). As such, it is possible to measure the amount of emotions in written communication.

The scientific program consists of a predefined set of words categorized by psychometric experts into positive and negative emotions. The category of positive emotions is summarized by 690 target words (e.g. happy, good,…), while 1347 target words are used to categorize negative emotions (e.g. hate, sad, ...). The externally-validated program searches the individual text files on a word-by-word basis. Each word in the text is compared against the predefined set of words. After counting the words in each category, the program computes the percentage of the total words for that text. Besides the fact that computerized word count approaches are typically blind to context in which the words are used, they have shown promising and reliable results in personality, social and clinical psychology (e.g. Mergenthaler (1996), Pennebaker et al. (2001),…). In short, the program makes it possible to compile positive as well as negative emotionality indicators from call center emails. These figures can be further employed to investigate their impact of emotions on subsequent churn behaviour.

5. Research Setting

For this study, data is collected from the largest Belgian newspaper company. The subscribers have to pay a fixed amount of money for their newspaper. Subscription data is used from January 2002 through September 2005. Figure 1 visualizes the window of analysis.

INSERT FIGURE 1 OVER HERE

Using this time frame, it is possible to define the dependent variable and the explanatory variables. For defining whether a subscriber is a churner or not, all renewal points between July 2004 and July 2005 are considered within this study. Moreover, only subscriptions having at least one email sent during the last subscription term are included in subsequent analysis. There are 11,836 subscriptions included of which 9,600 subscribers (81.11%) renew their subscription and are considered to be behavioural loyal, while 2,236 (18.89%) do not renew their product and are considered as churners. Someone is considered a churner when he/she does not renew his/her subscription within a four week period after the renewal date. During this four week period, the newspaper company still delivers the newspapers to the customers in order to give them the opportunity to renew their subscription. Besides the dependent variable, several explanatory variables need to be constructed in order to predict one's churn behaviour. The explanatory variables, such as RFM and other transactional information, contain information covering a 30-month period returning from every individual renewal point. This information is stored in a large transactional database. Moreover, subscribers have the possibility to report service failures or to ask questions via email. All emails are manually labelled by the staff of the call center by which a distinction between information requests and complaint emails is available. All emails from the last term of a

subscription are considered for further analysis. There are 18,331 emails of which 6,560 complaints (35.79%) and 11,771 information requests (64.21%). Since subscribers can send more than one email during the last term of their subscription, the emotionality indicators extracted using the methodology as explained in Section 4 are averaged per subscription type. In other words, the emotionality variables represent the general positive/negative tone in which subscribers interact with the company by means of written complaints or information requests to the call center.

In order to correctly assess the predictive capabilities of different classification models, the dataset is divided into a training and test set. The former one is composed by randomly assigning 70 percent of the subscriptions, while the other 30 percent are assigned to the test set. The training set is used to estimate the different classification models, while the test set is used for assessing the performance of the different models to an unseen data sample. The data set characteristics are given in Table 1.

INSERT TABLE 1 OVER HERE

6. Empirical Results

6.1. Churn Predictors

In the CRM literature, it is a common use to summarize customers' future behaviour based on their past behaviour. The available data are often stored in large databases and consist of RFM figures, other transactional data and socio-demographic information (see Appendix 1 for an in-depth overview of the variety of churn predictors). However, new media (e.g. email) can

assist with the improvement of client/company relations. As a consequence, additional explanatory variables are extracted from this new information type. Using the methodology as described in Section 4, several features indicating the emotionality in client/company interactions are extracted from the call center emails (see Table 2).

INSERT TABLE 2 OVER HERE

Consequently, several measurement models are built to correctly discriminate churners from loyal customers. For validating the hypothesis of additional predictive performance of the emotionality indicators on top of the churn variables (see Section 6.2.1.) and for comparing the predictive performance of the different classifiers (see Section 6.2.2.), two different churn models are built. The first one is an extended RFM model (hereafter abbreviated as eRFM) that uses the traditional RFM figures, while also adding other customer information like other transactional data and customer socio-demographics. A second prediction model adds the emotionality related variables from Table 2 to eRFM (hereafter abbreviated as eRFM-EMO). As such these models are used to assess the impact of emotionality from client/company interactions to customer's churn behaviour, while also the performance of the different classification models is compared. In order to correctly verify the impact of emotions in client/company emails on churn behaviour, univariate Logistic Regression models are built using only the emotionality indicators (see Section 6.3).

6.2. Predicting Churn Using Emotionality Indicators

6.2.1 Impact of Emotionality Indicators on Predictive Performance

This paragraph compares the predictive performance between an eRFM model and that of an eRFM-EMO model. Table 3 gives an overview of the predictive performance of all models in terms of AUC and PCC, while Table 4 contains the significance tests of Delong et al. (1988) for comparing similar classification techniques between an eRFM and eRFM-EMO context. In summary, this paragraph investigates whether adding emotionality indicators to an attrition model, results in an additional increase in predictive performance in distinguishing churners from non-churners. In other words, does an eRFM-EMO model perform better than an eRFM model for a given classifier?

INSERT TABLE 3 OVER HERE

INSERT TABLE 4 OVER HERE

Table 3 and Table 4 let us conclude that incorporating emotions from client/company interactions is a viable strategy for improving predictive performance of an eRFM churn model. The models including the additional client/company interaction variables perform always better in distinguishing churners from non-churners. Indeed, an eRFM-EMO model always has a higher predictive performance than the corresponding eRFM model in terms of PCC, while also the differences in terms of AUC are significant between similar classifiers (Delong et al., 1988).

In conclusion, incorporating emotions from client/company emails into a traditional customer attrition model is beneficial from a prediction point of view.

6.2.2. Comparing Predictive Performance of Logit, SVMs and Random Forests

This paragraph compares the predictive performance of Logit, SVMs and Random Forests within this churn context. Table 5 and Table 6 give an overview of the results from the significance test of Delong et al. (1988).

INSERT TABLE 5 OVER HERE

INSERT TABLE 6 OVER HERE

From Table 5 and Table 6, it is clear that Random Forests performs best in distinguishing churners from non-churners. Its performance is always significantly higher than the performance of Logit and SVMs in terms of PCC, as well as in terms of AUC (Delong et al., 1988).

Besides the excellent performance of Random Forests, one notices that Logit and SVMs perform equally well in this churn context. As one observes from Table 5 and Table 6, the AUC performance of Logit and SVMs are not significantly different within the different research contexts.

In summary, implementing Random Forests within this churn context is a viable opportunity to improve predictive performance in comparison to the performance of Logit and SVMs which both have an equal performance (Delong et al., 1988).

6.3. Impact of Emotionality Indicators on Churn Behaviour

This Section explores the impact of emotions in client/company emails on customer attrition. Table 7 shows the individual standardized parameter estimates, the Wald significance tests and the odds ratios as estimated during a univariate Logistic Regression using the emotionality variables. In other words, this paragraph measures the impact of every single variable from Table 2 on churn behaviour.

INSERT TABLE 7 OVER HERE

A primary finding of this research is that there is a significant relationship between positive expressed emotions in complaint emails and customer attrition (posemo_complaint:β=-0.1801,$p<0.001$). This result confirms the findings of Mattsson et al. (2004) who also found that positive emotions expressed during complaining reduce the chance of churning. The positive emotions expressed during the reporting of a service failure are assumed to counteract the eagerness to punish the company for the failure and to indicate the good prior relationship with the company. Moreover, this study shows evidence that a significant relationship exists between negative emotions expressed in complaint emails and customer churn (negemo_complaint:β=-0.1940,$p<0.001$). Contrary to the results of Mattsson et al. (2004), we find that the more negative emotional words are used in complaint emails, the lower the risk that the customer will leave the company. Considering that the company has a

good failure recovery system, this result is in line with the satisfaction framework of Oliver (1980). He states that when the gap between the outcome – in this case the service failure recovery – and the expectations – in this case the service recovery expectations – increases, the satisfaction of that particular customer increases. In other words, the more negative emotional words one uses during complaining, the more disillusioned the customer is. However, when the company decently recovers the failure, the customer will be very satisfied.

This study found no significant relationship between positive expressed emotions in information requests and someone's churn behaviour (posemo_contact:β=-0.0006,ns). Moreover, negative expressed emotions in information requests seem to have a significant influence on customers' churn behaviour (negemo_contact:β=-0.1231,$p<0.001$). One can say that the more negative emotional words are used in emails other than complaints, the lower the chance that the customer will churn. In intensive markets like the newspaper industry, customer satisfaction level becomes an important issue (Jones and Sasser, 1995). So increasing customer satisfaction is the starting point of the email handling process. For instance, customers are often comparing alternatives for their current product by the end of the subscription period. As such they often come in touch with new promotional deals, they want to capture. These information requests via email are often more negatively connoted than others because they express the disillusion of not having a promotional offer on their current subscription. However, the company tries to handle all questions as efficiently as possible by offering them a comparable subscription deal. This act increases customer satisfaction, because Oliver (1980) states that satisfaction is increased when the final outcome – i.e. here the proposal for a new subscription - exceed the rather bad expectations of fulfilling the request.

Moreover, this study found a strong relationship between the percentage of complaints of all client/company interactions and one's churn behaviour (percentage_complaint:β=-0.3561,$p<0.001$). The more complaints a customer has in his/her portfolio of emails sent to the company, the more certain he/she stays with the company. As such, complaining does not necessarily mean that the customer will leave the company. For instance, Maxham (2001) and Keaveney (1995) state that complaint behaviour results in a favourable behaviour towards the company when service failure recovery is satisfactory. Moreover, Bougie et al. (2003) state that complainers are less vulnerable to switch because they have a higher commitment and trust towards the company (e.g. Tax et al. 1998). Also Conlon and Murray (1996) state that customer who complain and receive a proper response to their service failures are more likely to stay.

7. Conclusion & Further Research

As a conclusion, one can say that the predictive performance of a churn model can be optimised by (i) exploring and adding new types of customer information into a conventional churn model and (ii) choosing the right classification technique. (i) In the last decade, there is a rapid development of the internet and information technology. As such, new information types are available to the data analyst. Nowadays, emails are seen as a valid alternative (next to letters and telephone calls) for customers to interact with the company. Most companies store these huge amounts of textual information in large databases, but hardly use them in their day-to-day analysis. As a consequence, unique opportunities arise to extract information from these emails to enrich churn models. This study shows the beneficial effect of including emotionality indicators extracted from call center emails into a customer attrition model. Indeed, the predictive performance significantly increases when these emotionality indicators

are included into a eRFM attrition model. (ii) The predictive performance of two state-of-the-art classifiers SVMs and Random Forests is benchmarked with that of a base classifier, Logistic Regression. It is shown that Random Forests significantly outperforms the other two classification techniques, while the predictive performance of Logistic Regression and Support Vector Machines is not significantly different within this research setting. Compared to Logistic Regression and SVMs, Random Forests' predictive dominance lies in the ability to discover hidden patterns in the complex data structure by (a) combining several outputs of 'weak' classifiers into a strong ensemble of classifiers and (b) by doing a random feature selection to split each node in the individual decision trees (Breiman, 2001).

Despite the fact that differences in performance may not seem large, the impact on the retention rate can be significant when targeting the right customers accordingly. Table 8 shows that increasing the retention rate with only 1% already has large implications for the long-term increase in profitability within this specific case.

INSERT TABLE 8 OVER HERE

In the current situation, about 18% of the clients defect. As such, we expect to keep each year about 82% of the subscribers. Additionally, the ideal situation of 100% retention rate is also included into the analysis for comparison reason only. This situation is utopian because there will always be people who will churn, e.g. due to natural defection. Suppose further that due to an increase in predictive performance and targeting the right customers accordingly, the company can increase the retention rate with 1% - i.e. from 82% to 83% retention rate. Table 8 indicates that an additional increase in retention rate with 1% results in a boost of total contribution over 5 year per 1,000 customers from 657,355 Euro to 669,799 Euro having an

average contribution of 200 Euro and a discount rate of 4%. If the company succeeds in increasing the retention rate with only 1%, an additional contribution of 12,445 Euro per 1,000 customers is gained.

As a substantive contribution, the impact of emotionality indicators on churn is investigated. It is shown that the impact on churn is positive when more emotional related words – i.e. positive emotions or negative emotions - are used. Customers that use more positive emotions in complaint emails are by nature more satisfied and certainly do not want to punish the company for the service failures which decreases of course the fact that one will attrite. The use of negative emotions in emails seems to have a positive influence on one's churn behaviour. This means that people who use more negative emotional connoted words tend to be more loyal. These results have large managerial implications for all managers dealing with call center management and email handling. Emotional emails require special treatment, because these people tend to more loyal than others. Even in the case of a very negative complaint email, the call center agent must be aware of the fact that this customer is of high value for the company. In fact, when customers express their dissatisfaction with the service by writing complaint emails, this does not necessarily mean that these customers will churn. Contrary, this study found that the higher the portion of complaint emails, the lower the chance that this specific customer will churn.

While we strongly believe that this research paper adds value to the current literature, there is still scope for some suggestions for further research. In this study, service recovery data was unavailable to the data analyst. As such, additional efforts could be made in collecting service recovery measures for every client/company interaction. Consequently, additional analysis using this type of data can be introduced in the current framework. Moreover, the proposed

framework of customer churn prediction is applied in a newspaper subscription business. There is a wide variety of data mining applications, e.g. cross- and up-sell applications, customer acquisition… in wide range of industries, e.g. retail, financial services, e-commerce… where the incorporation of this type of 'soft' data can improve the predictive performance. Additional analysis need to be done to validate the findings proposed in this research paper.

Acknowledgements

We would like to thank the anonymous Belgian company for providing us with data for testing our research questions and Ghent University for funding the PhD project of Kristof Coussement (BOF 01D26705). Also special thanks to L. Breiman (†) for freely distributing the Random Forests software, as well as C.-C. Chang and C.-J. Lin for sharing their SVM-toolbox, LIBSVM. Moreover, we would like to thank Bart Larivière and Ilse Bellinck for their insights during this project.

Appendix 1

	Variable type		Description
eRFM	***RFM***		Elapsed time since last renewal (Recency) The number of renewal points (Frequency) Monetary value (Monetary value)
	Other transactional information	*Client/company characteristics*	The purchase motivator of the subscription How the newspaper is delivered The number of complaints Elapsed time since the last complaint The average cost of a complaint (in terms of compensation newspapers) 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 Elapsed time since last response on a direct marketing action The number of free newspapers
		Renewal-related variables	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
		Subscription-describing variables	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
	Socio-demographics		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

References

Allison, P.D. (1999). Logistic Regression Using the SAS System: Theory and Application. SAS Institute Inc.: Cary, NC.

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

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

Brown, S.W. (1997). Service recovery through IT: complaint handling will differentiate firms in the future. Marketing Management, 6 (3), 25-27.

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

Bucklin, R.E. & Gupta, S. (1992). Brand Choice, Purchase Incidence and Segmentation: an Integrated Modeling Approach. Journal of Marketing Research, 29 (2), 201-215.

Burez, J. & Van den Poel, D. (2007). CRM at Canal+ Belgique: reducing customer attrition through targeted marketing. Expert Systems with Applications, 32 (2), 277-288.

Burez, J. & Van den Poel, D., Handling Class Imbalance in Customer Churn Prediction, Expert Systems With Applications, under review.

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

Bougie, R., Pieters, R. & Zeelenberg, M. (2003). Angry Customers Don't Come Back, They Get Back: The Experience and Behavioral Implications of Anger and Dissatisfaction in Services. Journal of the Academy of Marketing Science, 31 (4), 377-393.

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

Chaudhuri, A. (1998). Product class effects on perceived risk: the role of emotion. International Journal of Research in Marketing, 15, 157–168.

Conlon, D.E. & Murray, N.M. (1996). Customer Perceptions of Corporate Responses to Product Complaints: The Role of Explanations. Academy of Management Journal, 39 (4), 1040-1056.

Cortes, C. & Vapnik, V. (1995). Support-vector network. Machine Learning, 20, 273–297.

Coussement, K. & Van den Poel, D. (2008). Churn Prediction in Subscription Services: an Application of Support Vector Machines While Comparing Two Parameter-Selection Techniques. Expert Systems with Applications, 34 (1), 313-327.

Dellarocas, C. (2003). The digitization of word of mouth: promise and challenges for online feedback mechanisms. Management science, 49 (10), 1407-1424.

DeLong, E.R., DeLong, D.M. & Clarke-Pearson, D.L. (1988). Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: a Nonparametric Approach. Biometrics, 44 (3), 837-845.

Fader, P.S., Hardie, B.G.S. & Lee, K.L. (2005). RFM and CLV: Using Iso-Value Curves for Customer Base Analysis. Journal of Marketing Research, 62, 415-430.

Ganesan, S. (1994). Determinants of long-term orientation in buyer-seller relationships. Journal of Marketing, 58 (2), 1-19.

Hanley, J.A. & McNeil, B.J. (1982). The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve. Radiology, 143 (1), 29-36

Hart, C.W.L., Heskett, J.L. & Sasser, Jr. W.E. (1990). The profitable art of service recovery. Harvard Business Review, July-August, 146-156.

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

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

Hung, S.-Y., Yen, D.C. & Wang, H.-Y. (2006). Applying data mining to telecom churn management. Expert Systems with Applications, 31 (3), 515-524.

Jones, T.O. & Sasser, Jr. W.E. (1995). Why Satisfied Customer Defect. Harvard Business Review, 73, 88-99.

Keaveney, S. (1995). Customer Switching Behavior in Service Industries: An Exploratory Study. Journal of Marketing, 59 (April), 71-82.

Keaveney, S. & Parthasarathy, M. (2001). 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), 374-390.

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

Kim, Y.S. (2006). Toward a Successful CRM: Variable Selection, Sampling and Ensemble. Decision Support Systems, 41 (2), 542-553.

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

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

Lin, H.-T. & Lin, C.-J. (2003). 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.

Mattsson, J., Lemmink, J. & McColl, R. (2004). The Effect of Verbalized Emotions on Loyalty in Written Complaints. Total Quality Management & Business Excellence, 15 (7), 941-958.

Maxham, J.G. (2001). Service recovery's influence on consumer satisfaction, positive word-of-mouth, and purchase intentions. Journal of Business Research, 54 (1), 11-24.

Mergenthaler, E. (1996). Emotion-abstraction patterns in verbatim protocols: A new way of describing psychotherapeutic processes. Journal of Consulting and Clinical Psychology, 64, 1306-1315.

Neslin, S.A., Gupta, S., Kamakura, W., Lu, J. & Mason, C. (2006). Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models. Journal of Marketing Research, 43 (2), 204-211.

Newman, M.L., Pennebaker, J.W., Berry, D.S. & Richards, J.M. (2003). Lying words: Predicting Deception from Linguistic Style. Personality and Social Psychology Bulletin, 29, 665-675.

Oliver, R.L. (1980). A Cognitive Model of the Antecedents and Consequences of Satisfaction Decisions. Journal of Marketing Research, 17 (4), 460-469.

Pennebaker, J.W. & Francis, M.E. (1996). Cognitive, emotional and language processes in disclosure: physical health and adjustment. Cognition and Emotion, 10, 601-626.

Pennebaker, J.W., Francis, M.E. & Booth, R.J. (2001). Linguistic Inquiry and Word Count (LIWC). Erlbaum Publishers; Mahwah, NJ.

Reichheld, F.F. & Sasser, W.E. (1990). Zero Defections: Quality Comes to Services. Harvard Business Review, 68 (5), 105-111.

Reinartz, W., Krafft, M. & Hoyer, W.D. (2004). The Customer Relationship Management Process: Its Measurement and Impact on Performance. Journal of Marketing Research, 41 (3), 293-305.

Reinartz, W. & Kumar, V. (2003). The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration. Journal of Marketing, 67 (1), 77-99.

Rust, R.T. & Zahorik, A.J. (1993). Customer Satisfaction, Customer Retention, and Market Share. Journal of Retailing, 69 (2), 193-215.

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

Thomas, L.C. (2000). A Survey of Credit and Behavioral Scoring: Forecasting Financial Risk of Lending to Customers. International Journal of Forecasting, 16 (2), 149-172.

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.

Vapnik V. (1998). Statistical Learning Theory. Wiley; New York.

Weng, S.S. & Liu, C.K. (2004). Using Text Classification and Multiple Concepts to Answer Emails. Expert Systems with Applications, 26 (4), 529-543.

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

Zijlstra, H., van Meerveld, T., van Middendorp, H., Pennebaker, J.W. & Geenen R. (2004). Dutch Version of the Linguistic Inquiry and Word Count (LIWC); a Computerized Text Analysis Program. Behaviour and Health (Dutch journal), 32.

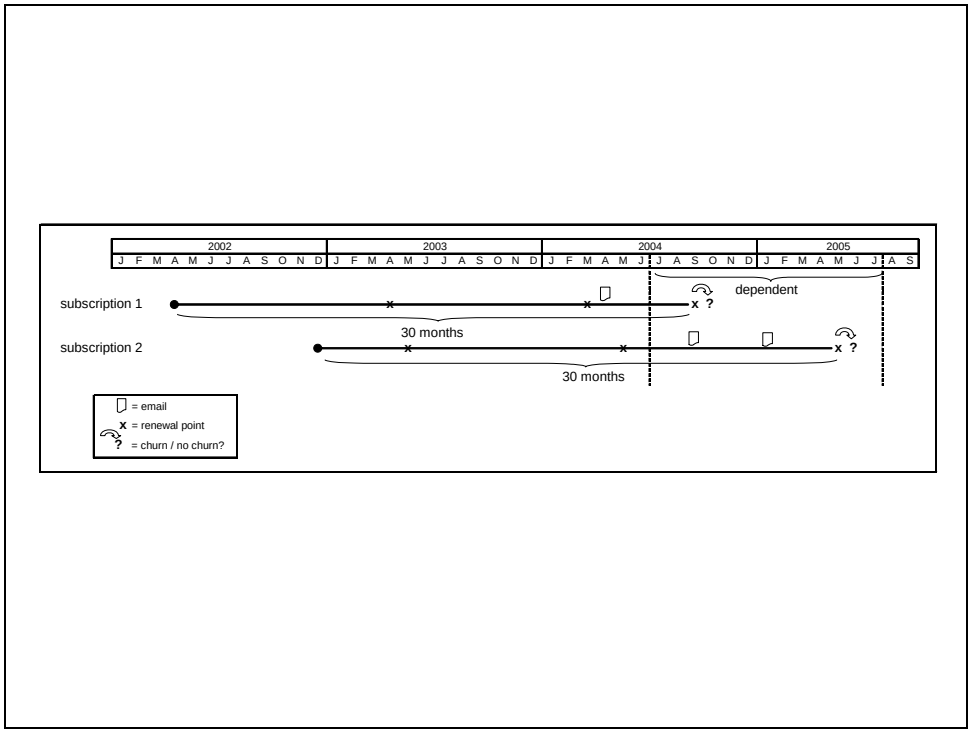

Fig. 1. Window of analysis.

	Number of subscriptions	Relative percentage
Training set		
Subscriptions not renewed	*1,792*	*18.93%*
Subscriptions renewed	*7,676*	*81.07%*
Total	*9,468*	*100%*
Test set		
Subscriptions not renewed	*444*	*18.75%*
Subscriptions renewed	*1,924*	*81.25%*
Total	*2,368*	*100%*

Table 1. Overview of the data characteristics.

Variable name	Description
Posemo_complaint	Positive emotions in complaint emails
Posemo_contact	Positive emotions in information requests
Negemo_complaint	Negative emotions in complaint emails
Negemo_contact	Negative emotions in information requests
Percentage_complaint	Percentage of complaints from total client/company interactions

Table 2. Emotionality indicators extracted from client/company interactions.

Model		eRFM		eRFM-EMO	
		AUC	**PCC**	**AUC**	**PCC**
Logit	*Training set*	73.59	77.76	74.89	77.99
	Test set	73.24	77.28	74.16	77.70
Random Forests	*Training set*	74.84	78.26	75.65	78.37
	Test set	75.12	78.29	76.02	78.97
SVMs	*Training set*	72.99	77.90	74.51	78.43
	Test set	72.53	77.53	73.83	77.87

Table 3. The predictive performance of Logit, Random Forests and SVMs.

Model	AUC		Significantly different on 95% confidence level?
	eRFM	**eRFM-EMO**	
Logit	73.24	74.16	YES
Random Forests	75.12	76.02	YES
SVMs	72.53	73.83	YES

Table 4. Pairwise comparison of AUC performance between eRFM models and eRFM-EMO models on test set.

Model 1	Model 2	AUC		Significantly different on 95% confidence level?
		Model 1	Model 2	
Logit	Random Forests	73.24	75.12	YES
SVMs	Random Forests	72.53	75.12	YES
Logit	SVMs	73.24	72.53	NO

Table 5. Pairwise comparison of AUC performance on the eRFM test set.

Model 1	Model 2	AUC		Significantly different on 95% confidence level?
		Model 1	Model 2	
Logit	Random Forests	74.16	76.02	YES
SVMs	Random Forests	73.83	76.02	YES
Logit	SVMs	74.16	73.83	NO

Table 6. Pairwise comparison of AUC performance on the eRFM-EMO test set.

Variable name	Standardized estimate	Wald significance tests	Odds ratio
Posemo_complaint	-0.1801**	98.1339	0.674
Posemo_contact	-0.0006 ns	0.00240	0.999
Negemo_complaint	-0.1940**	82.2778	0.561
Negemo_contact	-0.1231**	53.3812	0.752
Percentage_complaint	-0.3561**	255.994	0.418

**: *p < 0.001*;*ns*: *not significant*

Table 7. Univariate Standardized Parameter Estimates, Wald Significance Tests & Odds Ratio.

Retention rate (= 100%-churn%)	No of customers					Average Contribution per subscriber per year (in Euro)	Total Contribution after 5 year per 1,000 subscribers (in Euro)	Additional profit over 82% after 5 year per 1,000 subscribers (in Euro)
	Year							
	1	2	3	4	5			
82%	1,000	820	672	551	452	200	657,355	
83%	1,000	830	689	572	475	200	669,799	12,445
100%	1,000	1,000	1,000	1,000	1,000	200	925,979	256,180

Table 8. Real-life retention example.