MODELLING AND PREDICTING CUSTOMER CHURN FROM AN INSURANCE COMPANY

BY

CLARA-CECILIE GÜNTHER*, INGUNN FRIDE TVETE *, KJERSTI AAS *,

GEIR INGE SANDNES† AND ØRNULF BORGAN‡

*NORWEGIAN COMPUTING CENTER, P.O. BOX 114 BLINDERN, 0314 OSLO, NORWAY

† GJENSIDIGE, P.O. BOX 276, 1326 LYSAKER, NORWAY

‡ DEPARTMENT OF MATHEMATICS, UNIVERSITY OF OSLO, PO BOX 1053, BLINDERN, 0316 OSLO, NORWAY

Abstract

Günther, C.-C., Tvete I. F., Aas K., Sandnes G. I., Borgan Ø. Customer churn from an insurance company. Scandinavian Actuarial Journal. Within a company's customer relationship management strategy, finding the customers most likely to leave is a central aspect. We present a dynamic modelling approach for predicting individual customers' risk of leaving an insurance company. A logistic longitudinal regression model that incorporates time-dynamic explanatory variables and interactions is fitted to the data. As an intermediate step in the modelling procedure, we apply generalised additive models to identify non-linear relationships between the logit and the explanatory variables. Both out-of-sample and out-of-time prediction indicate that the model performs well

in terms of identifying customers likely to leave the company each month. Our approach is general and may be applied to other industries as well.

*K*eywords: Generalised additive models, logistic regression, longitudinal data, CRM, non-life insurance.

1. Introduction

Insurance companies can no longer rely on a steady customer base. In recent years, it has become easier for customers in many countries to change insurance provider. In Norway, the insurance regulations changed in 2006, allowing customers to cancel their policies at any time, and not only on due date. With a large number of market competitors and increasingly conscious customers, it has become more and more important for companies to retain their customers. The cost of attracting new customers can be up to 12 times the cost of retaining the existing ones (Torkzadeh et al., 2006). Having a small increase in retention rates may add millions to premium revenue, and hence customer retention is an important aspect of customer relationship management (CRM).

When a customer cancels all his policies, either to switch insurance provider or because the need of insurance is no longer present, the customer has churned. For obvious reasons, the most important customer retention strategy is to identify the customers who are likely to churn. Once they are identified, customer retention programs can be developed and actions can be taken. Customer churn has been studied in different industries, e.g. telecommunications, financial services and

insurance, using different statistical techniques. Several researchers have used logit models, e.g. Brockett et al. (2008), Ahn et al. (2006), Burez and Van den Poel (2007), Kim and Yoon (2004), Lemmens and Croux (2006), Mozer et al. (2000) and Neslin et al. (2006), whereas others have analysed customer churn in continuous time by using survival analysis techniques, e.g. Brockett et al. (2008), Bolton (1998), Burez and Van den Poel (2007), Drew et al. (2001), Jamal and Bucklin (2006), Mani et al. (1999) and Schweidel et al. (2008). Finally, some data mining approaches, such as tree-based methods and neural networks have been used, e.g. Burez and Van den Poel (2007), Drew et al. (2001), Hung et al. (2006), Lemmens and Croux (2006), Mani et al. (1999), Mozer et al. (2000), Neslin et al. (2006), Wei and Chiu (2002) and Zhang et al. (2006).

Of the above-mentioned methods, the logit-model seems to be the most popular in a churn context. This is probably because this model is relatively simple and still shows good performance. Moreover, it is robust and the parameter estimates are interpretable in terms of odds ratios. However, an important drawback with the standard logistic regression model is that it assumes linear relationships between the logit and the explanatory variables. When this is not the case, information is lost and the conclusions drawn from the analysis might not be valid. Therefore, in this paper we present an extension of the logit model that allows for more complex non-linear relationships between the response and the explanatory variables. Like Coussement et al. (2010) we use generalised additive models (GAM) (see Hastie and Tibshirani (1990) for a general introduction to

GAM), to identify the form of the functional relationship between the response and the explanatory variables. However, while they use the fitted GAM model for prediction, we use it as an intermediate step in our model building process to redefine some of the explanatory variables for a subsequent logit-model analysis. Hence, we propose an approach where we utilise the advantages of a GAM model approach in the exploratory part of the analysis. Yet we avoid an overfitted model, which may be hard to interpret due to the potential non-linear relationship between the dependent variable and the explanatory ones.

In most of the applications of the logit-model referenced above, the authors focus on whether or not customers have churned during a certain period. However, as companies typically keep monthly tracking records, information on each customer consists of a time series of observations (longitudinal data). One should therefore construct models describing customers' monthly behaviour. We predict the probability of customers' risk of leaving the company each month. We also utilise changes in customer relevant information over time. Hence, our approach is a time dynamic one. Additionally, interactions between explanatory variables may influence the churn risk, and are therefore included in our model.

Our analysis of a portfolio of private insurances from a major Scandinavian insurance company identifies some key indicators that may predict which customers are most likely to leave the company. These empirical results should be of interest to readers within as well as outside Scandinavia.

The remainder of this paper is organised as follows. In Section 2 we describe our data set, while an outline of the statistical modelling framework is given in Section 3. A summary of our statistical analysis and the model obtained for predicting customer behaviour is described in Section 4, with a study of the estimated effects and prediction performance of the model following in Section 5. Finally, in Section 6 we summarise our findings and discuss some remaining challenges.

2. Data

We consider a portfolio of private insurances from Gjensidige, the largest non-life insurance company in Norway. We define three main types of insurance coverage: (i) car, (ii) home, and (iii) health (death, disease, disablement, and accidents). In addition to the main types, a customer might have other types of policies, as for instance a motorcycle or boat insurance.

We will use monthly data for the period from November 2007 until May 2009. For each of these 19 months and each customer, the insurance company has information on several explanatory variables that may help to predict customer behaviour. A summary of these variables is given in Table 1. Most of the variables are self-explanatory. The variable `Discount` indicates whether a customer receives a discount on his total insurance premium due to membership in a specific organisation, such as a national automobile association or a federation of trade unions. The `Discount` variable is grouped into five categories, of which the last indicates that a customer is not in a discount program. The variable

`Lifetime` keeps track of the time elapsed since the earliest registration of each customer. Due to system changes in the registration procedures, the oldest lifetime registered is 12.72 years. We define a customer as being *active* if he has at least one policy in the company, as opposed to churned (that is *non-active*) if all the policies are cancelled. Note that a customer who has left the company may later obtain new insurance coverage, and hence become active again.

As the portfolio is very large, we did not use the complete portfolio, but extracted a random sample containing information on approximately 160 000 customers. Further, we constrained our study to customers between 18 and 75 years of age with a yearly premium of at most 50 000 NOK. The cut-offs were set to exclude children, elderly and highly covered customers, the latter group already being closely monitored by the company. Customers who died during our 19 month analysis period, as well as customers with no information on the `Lifetime` variable were also excluded.

With the above constraints we are left with a data set of 127 961 customers. This data set was split at random into a training set consisting of about 10% of the customers used for fitting the model (see Section 4) and a test set consisting of the remaining 90% of the customers used to evaluate the prediction performance of the model (see Section 5.2).

TABLE 1. Description of available explanatory variables.

Explanatory variable	Description
`Premium`	Yearly total premium in NOK (range [0,50000]).
`Age`	Age of customer (range 18-75).
`Gender`	Gender of customer (0=Female, 1=Male).
`Partner`	Customer's spouse or partner has also a policy in the company (0=No, 1=Yes).
`Discount`	Discount program, ({1,2,3,4,5}, 5 denotes no discount program).
`Car`	Customer has car insurance (0=No, 1=Yes).
`Home`	Customer has home insurance (0=No, 1=Yes).
`HomePolicies`	Number of home insurance policies (range 0-28).
`Health`	Customer has health insurance (0=No, 1=Yes).
`Lifetime`	Registered duration (in years) of continuous customer relationship (range [0,12.72]). If a customer exits and later returns, the value is set to 0 at the point of return.

3. Model

Our aim is to build a statistical model that for each month is able to predict which customers are most likely to leave the insurance company. In this section, we describe our modelling framework in general terms, leaving the detailed discussion of the actual model fitting procedure to Section 4.

Customers may enter or leave the company each month, and hence the number of active customers in the portfolio will change throughout our analysis period of $T = 19$ months. For customer i and month t, we introduce the following notation. The indicator $R_{i,t}$ takes the value 1 if customer i is active in month t and $R_{i,t} = 0$ otherwise. If customer i is active in month t, we let $\mathbf{Z}_{i,t}$ be the vector of the explanatory variables given in Table 1 for the customer in this month. If the customer is non-active, we do not observe $\mathbf{Z}_{i,t}$. Hence, $\mathbf{Z}_{i,t}$ is observed only for $R_{i,t} = 1$. Finally, we define

$$Y_{i,t} = \begin{cases} 1 & \text{if customer } i \text{ leaves the company in month } t \\ 0 & \text{if customer } i \text{ does not leave the company in month } t. \end{cases}$$

Note that $Y_{i,t} = 1$ if $R_{i,t-1} = 1$ and $R_{i,t} = 0$. The available data for customer i are

$$\{(R_{i,t}, R_{i,t}\mathbf{Z}_{i,t}, Y_{i,t});\, t = 1, \ldots, T\}.$$

We now introduce the history $\mathcal{H}_{i,t}$ that contains all information available on customer i by time t, i.e. by observing $(R_{i,s}, R_{i,s}\mathbf{Z}_{i,s}, Y_{i,s})$ for $s = 1, \ldots, t$. Using

a slightly informal notation, the likelihood for customer i may then be given as

$$\begin{aligned} L_i^{\text{full}} &= P(R_{i,1}, R_{i,1}\mathbf{Z}_{i,1}, Y_{i,1}, \ldots, R_{i,T}, R_{i,T}\mathbf{Z}_{i,T}, Y_{i,T}) \\ &= P(R_{i,1}, R_{i,1}\mathbf{Z}_{i,1}, Y_{i,1}) \prod_{t=2}^{T} P(Y_{i,t} \,|\, \mathcal{H}_{i,t-1}) P(R_{i,t}, R_{i,t}\mathbf{Z}_{i,t} \,|\, \mathcal{H}_{i,t-1}, Y_{i,t}). \end{aligned}$$

As we do not want to specify a model for the development of the explanatory variables, we omit the leading factor and the last factor in the product above, to obtain the partial likelihood (Cox, 1975) for customer i:

$$L_i = \prod_{t=2}^{T} P(Y_{i,t} \,|\, \mathcal{H}_{i,t-1}).$$

The conditional distribution $P(Y_{i,t} \,|\, \mathcal{H}_{i,t-1})$ is degenerate when $R_{i,t-1} = 0$. Hence we get a contribution to the partial likelihood only when customer i is active in month $t-1$. If we assume that the n customers in the training set constitute an i.i.d. sample, the partial likelihood for all the n customers may be written

$$L = \prod_{i=1}^{n} \prod_{t=2}^{T} P(Y_{i,t} \,|\, \mathcal{H}_{i,t-1}). \tag{1}$$

To make further progress, we need to specify a model for the conditional distributions of the Bernoulli variables $Y_{i,t}$ given the histories of the customers. To this end we introduce

$$p_{i,t} = Pr(Y_{i,t} = 1 \,|\, \mathcal{H}_{i,t-1}), \tag{2}$$

which is the probability that customer i will leave the company in month t given the history for this customer up to and including the previous month. We will consider a logistic model, where the $p_{i,t}$'s may depend on the explanatory variables

given in Table 1 as well as on new variables derived from these basic variables, see Section 4. To distinguish between the basic variables of Table 1 and the vector of variables actually used in the regression modelling, we denote the latter by $\mathbf{X}_{i,t}$. Our logistic regression model then takes the form

$$p_{i,t} = \frac{e^{\alpha_t + \boldsymbol{\beta}' \boldsymbol{X}_{i,t-1}}}{1 + e^{\alpha_t + \boldsymbol{\beta}' \boldsymbol{X}_{i,t-1}}}, \tag{3}$$

which alternatively may be written

$$\log \frac{p_{i,t}}{1 - p_{i,t}} = \alpha_t + \boldsymbol{\beta}' \boldsymbol{X}_{i,t-1} = \alpha_t + \sum_{j=1}^{k} \beta_j X_{ij,t-1}. \tag{4}$$

Here $\log \frac{p_{i,t}}{1-p_{i,t}}$ is called the logit and k is the number of explanatory variables included in the model. Note that we use a separate intercept term for each month to allow for a variation in the baseline risk of leaving the company.

The partial likelihood may be maximised by using standard software for logistic regression, like the `glm` command in `R` (R Development Core Team, 2008). Furthermore, given an appropriate specification of the conditional probabilities (3), the partial likelihood has similar properties to an ordinary likelihood (Cox, 1975). Hence we may use the inverse information matrix to assess estimation uncertainty and the likelihood ratio test for comparing nested models just as for standard logistic regression (McCullagh and Nelder, 1989).

4. Building the prediction model

In a logistic regression model there is a linear relationship between the explanatory variables and the logit, as seen in (4). If the relationship is not linear, then

the estimates of the parameters and the inference based on them are misleading. In this paper, we use a GAM-model (Hastie and Tibshirani, 1990) to detect any potential non-linear relationships between the logit and the explanatory variables. In a GAM-model, the linear term $\beta_j X_{ij,t-1}$ in (4) is replaced by a smooth non-parametric function $s_j(X_{ij,t-1})$.

We will transform a given explanatory variable in such a way that it resembles the curve in a plot of its smooth function s_j (a so-called GAM plot). As an intermediate step in the model building process, we fitted a model with all the variables in Table 1. For the variable `Premium`, we used the logarithm with base 10 instead of the variable itself. For the continuous variables `Age`, `Lifetime` and log(`Premium`), smoothing splines were fitted using the `gam` command in `R` with the default choice of smoothing parameters. Figure 1 shows the resulting GAM plots.

The GAM-plot for the `Age` variable indicates a fairly linear relationship downwards between age and customer churn from age 30 and onwards. Before age 30, there seems to be an increasing trend. However, as the standard errors are quite large due to few young customers, we assume the effect of age to be constant for this group. The new variable is denoted `Age.T`, see Table 2.

The effect of the `Lifetime` variable decreases linearly up until 3 years. We categorise this variable into three categories corresponding to a lifetime less than 1 year, between 1 and 3 years and longer than 3 years, respectively. This variable is denoted `Lifetime.C`, see Table 2.

In the GAM-plot for log(`Premium`), the standard errors are large for values smaller than 3 (1000 NOK). Hence, we truncate this variable at 3 and assume a linear relationship for values larger than 3. In order to achieve a useful reference point, we further divide the premium by the median before taking the logarithm. Thus the value 0 for the new variable corresponds to the median premium. This variable is denoted `log.Premium` and is defined in Table 2.

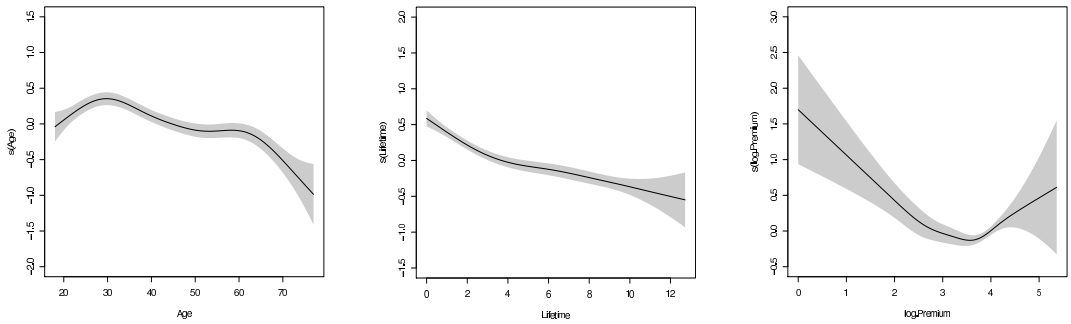

FIGURE 1. GAM-plots on logit scale of the effects of the explanatory variables `Age`, `Lifetime` and `log(Premium)` with standard errors, from the model with all variables in Table 1 included. In the first plot, the y-axis represent the non-parametric function $s_{\text{Age}}(\text{Age})$, and similarly in the two other plots for `Lifetime` and `log(Premium)`.

In addition to the transformed explanatory variables described above, we also defined some new explanatory variables based on the basic ones in Table 1. A customer may have several home insurance policies. As coverage of a house itself (exterior) and coverage of its interior define two home insurances, we assume that

two or less home insurances mean that the customer probably has one home, while more than two indicates a secondary home as well. Hence, we define the indicator variable `TwoHomes`, see Table 2.

Instead of looking at the total number of policies for each customer, we consider the number of policy types. We define the variable `MainInsurances`, which can take the values 0, 1, 2 and 3. A customer having `MainInsurances`=1, has either a car, home or health insurance policy. `MainInsurances`=0 indicates that the customer has none of the three main types of insurance, but another type of insurance, e.g. boat insurance.

Some customers leave and later rejoin the company. They might have an enhanced churn risk of leaving. We therefore define the indicator variable `ReturnedCustomer`, which indicates whether the customer has rejoined.

Changes in explanatory variables from one month to another might be of importance as these changes could indicate that the customer is phasing out his policies. Discount programs have previously been shown to be important when one estimates the risk of churn. Moreover, prior to our analysis, the insurance company believed that the car insurance policy usually was the first to be cancelled. For the `Discount` variable, we focus on whether a customer who previously had a discount (level 1 – 4), no longer has it and therefore define an indicator variable `DiscountChange` describing this. The indicator variable `CarCancelled` states whether the car insurance was cancelled last month or not. Time lags

of three and six months were also considered, but one month was found to be sufficient.

To sum up, we define a model with the following explanatory variables: `Gender`, `Partner` and `Discount` from Table 1, the transformed explanatory variables presented in Table 2, and the month specific intercept term as given in (4). Moreover, after trying various combinations, using Akaike's information criterion (AIC) and prediction ability as criteria, we allow for interaction terms between `Partner` and `log.Premium`, `Gender` and `Age.T`, `MainInsurances` and `Discount`, `log.Premium` and `Discount`, `MainInsurances` and `log.Premium`.

5. Results

5.1. **Estimated effects.** Table 3 shows the estimated main effects and standard errors for the variables included in the final model. The only effects that are not part of an interaction term, and are hence easily interpreted, are those of `Lifetime.C`, `ReturnedCustomer`, `TwoHomes`, `CarCancelled` and `DiscountChange`. The estimated effect of `Lifetime.C` is relative to the reference level `Lifetime.C`=3. We see that shorter lifetimes yield an increased churn risk. Further, if a customer has rejoined the company (`ReturnedCustomer`=1), the churn probability is increased compared to a customer who has not previously left. Both these effects seem reasonable, as long-term customers with no history of cancelling policies are loyal and hence less likely to churn. A customer with more than two home

TABLE 2. Description of derived explanatory variables. m denotes the median of the yearly premium.

Variable	Values	Condition
`Age.T`	30	if `Age`≤ 30
	`Age`	if `Age`>30
`Lifetime.C`	1	if `Lifetime` ≤ 1
	2	if $1 <$ `Lifetime` ≤ 3
	3	if `Lifetime` > 3
`log.Premium`	$\log(1000/m)$	if $\log($`Premium`$/m) \leq \log(1000/m)$
	$\log($`Premium`$/m)$	if $\log($`Premium`$/m) > \log(1000/m)$
`TwoHomes`	0	if `HomePolicies` ≤ 2
	1	if `HomePolicies` > 2
`MainInsurances`	0	if `Car` + `Home` + `Health` $= 0$
	1	if `Car` + `Home` + `Health` $= 1$
	2	if `Car` + `Home` + `Health` $= 2$
	3	if `Car` + `Home` + `Health` $= 3$
`ReturnedCustomer`	1	if the customer has rejoined the company
	0	else
`CarCancelled`	1	if $\texttt{Car}_t = 0$ and $\texttt{Car}_{t-1} = 1$
	0	else
`DiscountChange`	1	if $\texttt{Discount}_t = 5$ and $\texttt{Discount}_{t-1} \neq 5$
	0	else

insurances (`TwoHomes`=1) has a lower churn probability than the remaining customers. Further, a customer who had a discount last month, but not in the current (`DiscountChange`=1), is much less loyal than a customer who still has one. Finally, cancellation of car insurance during one month (`CarCancelled`=1) yields an increased churn probability the next month, as anticipated by the insurance company. However, this effect is far from being significant.

The variables `Partner`, `Gender`, `MainInsurances`, `log.Premium`, `Age.T` and `Discount` are all included in one or more interaction terms. The estimated main effects for these variables (shown in Table 3) apply when the other explanatory variables equal their reference values. For other values of the explanatory variables we also have to take the estimated interaction effects of Table 4 into account. Although the interaction terms are more difficult to interpret, we can observe the following. `Age.T` interacts only with `Gender`, and the negative interaction between the two means that whatever the values of the other explanatory variables, the churn risk for males is more reduced by increasing age than is the case for females. In a similar manner `Partner` interacts only with `log.Premium`, and from the estimates of Tables 3 and 4 we find that for a high yearly premium, a customer is more loyal if his partner is also a customer of the insurance company, while the opposite is the case for a low yearly premium. All pairs of the variables `MainInsurances`, `log.Premium`, and `Discount` interact, and this makes the interpretation of the effect of these three variables quite involved. But we note that customers who are in a discount program have substantially lower

churn risk than customers who are not, and that the reduction in churn risk is largest for those who have three main insurances and a large yearly premium.

5.2. **Prediction performance.** Our model was fitted to the training set described in Section 2. To evaluate the model on an independent data set, we will first do a so-called out-of-sample prediction. We then predict the probability for the customers in our test set to leave the company each month in the period November 2007 – May 2009. For the estimated probabilities, a cut-off is chosen so that the customers with churn probability higher than this cut-off will be classified as churned, and customers with churn probability lower than the cut-off will be classified as not churned. In this way, we obtain a classification rule. One way to evaluate the prediction performance of a model, is to calculate the true positive rate (TP), also known as sensitivity, and the false positive rate (FP), also known as one minus the specificity. The true positive rate is the proportion of churned customers that are correctly classified as churned, whereas the false positive rate is the proportion of customers incorrectly classified as churned among the non-churned customers. However, these rates depend on the specific cut-off chosen. To obtain a clearer view of the overall prediction performance, the receiver operating characteristic (ROC) curve (Fawcett, 2006) can be plotted. This curve shows the true positive rate plotted against the false positive rate for all possible cut-offs. If simply guessing at random which customers will churn, the ROC curve would be the diagonal line in the plot. The larger the area under

TABLE 3. Estimated main effects. Significant effects, using a 5% significance level, are shown by *. The estimated effect of the time variable α_t is not given.

Variable	Estimated effect	Standard error
`Partner =1`	0.11*	0.05
`Gender =1`	0.25	0.16
`Lifetime.C=1`	0.61*	0.06
`Lifetime.C=2`	0.35*	0.05
`MainInsurances=0`	0.50	0.35
`MainInsurances=1`	0.47	0.32
`MainInsurances=2`	0.28	0.33
`ReturnedCustomer=1`	0.59*	0.12
`TwoHomes=1`	-0.46*	0.23
`log.Premium`	-0.19	0.45
`Age.T`	-0.02*	0.003
`Discount=1`	-1.02*	0.32
`Discount=2`	-1.46*	0.42
`Discount=3`	-0.99*	0.40
`Discount=4`	-1.27*	0.57
`CarCancelled=1`	0.15	0.19
`DiscountChange=1`	1.80*	0.13

TABLE 4. Estimated interactions effects. Significant effects, using a 5% significance level, are shown by *.

Variable	log.Premium	MainInsurances			Age.T
		0	1	2	
Partner=1	-0.26*				
MainInsurances=0	0.28				
MainInsurances=1	0.93*				
MainInsurances=2	1.59*				
Discount=1	-0.22	0.68	0.53	-0.06	
Discount=2	-0.60*	0.87	0.22	-0.28	
Discount=3	-0.75*	-0.05	0.22	-0.21	
Discount=4	-1.40*	-0.34	-0.17	-0.31	
Gender=1					-0.02

the curve is, the better the model performs in terms of prediction. The solid line in Figure 2 shows the ROC curve for our test set. We see that our model performs much better than simply guessing at random. In practice, this implies that using our model will result in more customers being correctly classified as churned compared to a random selection of customers.

The churn rate, i.e. the proportion of customers actually leaving the company each month is confidential, and is thus not given here. While the ROC-curve displays the model's performance for all possible cut-offs, it could also be of interest to consider the customers corresponding to e.g. the 1000 highest predicted churn probabilities. A company might identify such a customer group for a personal follow up. Among the customers with the 1000 highest predicted probabilities, our model is able to predict the number who actually churned 15 times better than guessing at random, which we find to be a great improvement.

To validate the fitted model from Section 4 out-of-time, we use another data set consisting of the same customers as in our original data set, but during the period of June 2009 – January 2010. We divide this data set into two different test sets. Test set A consists of the customers in our original test set, whereas test set B consists of the customers in our original training set. We use the monthly covariate information for June 2009 to January 2010 in our predictions, with the estimated regression coefficients presented in Section 5.1. When we do out-of-time predictions, we need estimates of the baseline α_t for the period from June 2009 to January 2010. As there is no apparent time-trend for the estimated α_ts during the months prior to June 2009, we set all the α_ts in the test period to the mean of the estimated αs for the months prior to June 2009. As we are only interested in the ranking of the estimated churn probabilities, and not their actual size, another value of α_t would not alter the ranking.

The dashed and dotted lines in Figure 2 show the ROC curves for test sets A and B, respectively. The model performs similarly for both test sets, and only slightly worse out-of-time compared to in-time (solid line). With a true positive rate less than about 0.25, the model performs better out-of-time than in-time. This is confirmed by counting the number of customers who actually churned among the largest predicted probabilities. Our model performs 16 times better than random guessing for test set A (considering the 1000 highest predicted probabilities) and 18 times better for test set B (considering the 100 highest predicted probabilities). Since test set A is both out-of-time and out-of-sample compared to the training data set, it is to be expected that the performance for test set B is slightly better. These numbers depend on the chosen cut-off and the results would differ with another cut-off, which can be seen by considering the ROC curves.

The prediction ability was also evaluated separately for each month in the period from June 2009 to January 2010, to see whether it decreased over time. Perhaps surprisingly, the prediction ability remained fairly constant for these eight months. This indicates a stable environment with respect to effects from company or competitor strategies and campaigns.

6. Discussion

Within a company's CRM strategy, identifying the customers most likely to churn is central. In this paper, we have presented a dynamic modelling approach for

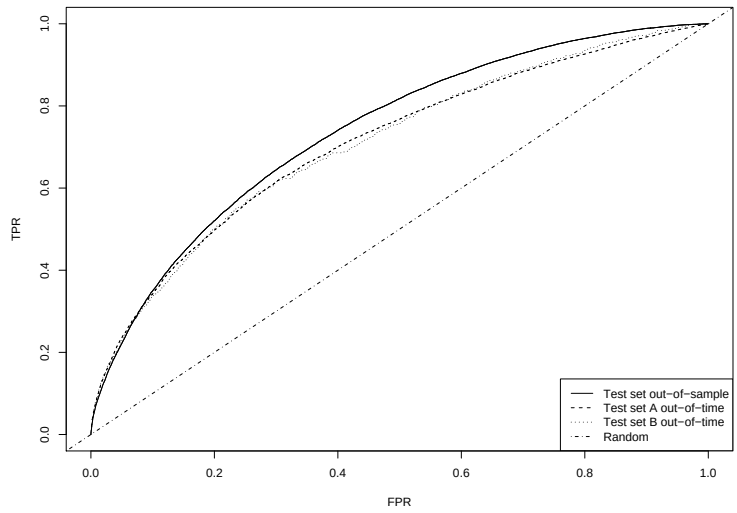

FIGURE 2. ROC-curve for in sample predictions in the test set and out-of-time predictions for test set A and B.

predicting individual customers' monthly churn risk. Our data set is from an insurance company, but the approach is general, and may be applied to other industries as well.

In our logistic regression model we have included explanatory variables describing partner, gender, lifetime, age, yearly premium, discount, number of main types of insurances, number of home insurances and changes in car insurance and discount the last month. We find that the most important factors are whether a customer has a discount or not, and changes over time in this variable. For the insurance company it should be valuable to monitor such a change, as this could be a sign of the customer being in a process of cancelling his policies. Unlike

others who have applied the logit model to customer churn, we also include interactions between some of the variables. Several of these interactions were shown to be valuable for prediction.

We strongly agree with Coussement et al. (2010) that a GAM approach can give a more realistic description of the relationship between dependent and independent variables. There are however several reasons for not using the GAM-model for prediction. First, the results of a fitted GAM-model are not easily interpreted or communicated. Moreover, when using GAM there is always a danger of overfitting the model. We therefore suggest to apply GAM as a valuable tool in the model building process, rather than as the final model approach for predicting customer churn.

Some variables that might have a great influence on a customer's decision to leave the insurance company were not available for this study. For instance, the last price the customer was offered from the company before he left, was probably very important for his decision. If this information had been available, the prediction results most likely would have been improved. In addition, other external factors like competitors' campaigns and focus in media on the benefits of switching insurance provider may have an increasing effect on churn.

Like us, Brockett et al. (2008) consider customers having multiple policies. They find that the time between the cancellation of the first and the remaining policies depends on the type of the policy first cancelled. This implies that a customer with a car and a home insurance might behave differently from a customer

with a home and a health insurance, even though the number of insurances is the same in both cases. We partly take this into account by including a variable that indicates whether the car insurance policy was cancelled previous month.

If our modelling approach is included in a CRM strategy, customers with a high churn probability can be identified early, and individual customer retention procedures can be carried out. The high probability churn customers are likely to be a diverse group, consisting both of valuable customers and customers who might not be very profitable to the company. When examining the individuals in a high risk group a sensible strategy could be to retain those with a slightly lower churn probability, but with high expected profitability. We believe our approach to be a useful part of the CRM routine for reducing the costs of marketing and client service.

Acknowledgements

The work was financed by the Norwegian Center for Research-based Innovation 'Statistics for innovation', project number 970141669. The authors thank Gjensidige, for kindly supporting the data and in particular Katrine Linnerud for valuable comments.

References

Ahn, J.-H., Han, S.-P., and Lee, Y.-S. (2006). Customer churn analysis: churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. *Telecommunications Policy*, 30(10-11):552–568.

Bolton, R. N. (1998). A dynamic model of the duration of the customer's relationship with a continuous service provider: the role of satisfaction. *Marketing Science*, 17(1):45–65.

Brockett, P. L., Golden, L. L., Guillen, M., Nielsen, J. P., Parner, J., and Perez-Marin, A. M. (2008). Survival analysis of a household portfolio of insurance policies: How much time do you have to stop total customer defection? *The Journal of Risk and Insurance*, 75(3):713–737.

Burez, J. and Van den Poel, D. (2007). CRM at a pay-TV company: using analytical models to reduce customer attrition by targeted marketing for subscription services. *Expert Systems with Applications*, 32(2):277–288.

Coussement, C., Benoit, D. F., and Van den Poel, D. (2010). Improved marketing decision making in a customer churn prediction context using generalized additive models. *Expert Systems with Applications*, 37:2132–2143.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–276.

Drew, J. H., Mani, D. R., Betz, A. L., and Datta, P. (2001). Targeting customers with statistical and data-mining techniques. *Journal of Service Research*, 3(3):205–219.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 19:861–874.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, London.

Hung, S.-Y., Yen, D. C., and Wang, H.-Y. (2006). Applying data mining to telecom churn management. *Expert Systems with Applications*, 31(3):514–524.

Jamal, Z. and Bucklin, R. E. (2006). Improving the diagnosis and prediction of customer churn: a heterogeneous hazard modeling approach. *Journal of Interactive Marketing*, 20(3-4):16–29.

Kim, H.-S. and Yoon, C.-H. (2004). Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. *Telecommunications Policy*, 28(9-10):751–765.

Lemmens, A. and Croux, C. (2006). Bagging and boosting classification trees to predict churn. *Journal of Marketing Research*, 43(2):276–286.

Mani, D. R., Drew, J., Betz, A., and Datta, P. (1999). Statistics and data mining techniques for lifetime value modeling. *Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 94–103.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall/CRC, London, 2nd edition.

Mozer, M. C., Wolniewicz, R., Grimes, D. B., Johnson, E., and Kaushansky, H. (2000). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Transactions on Neural Networks - Special Issue on Data Mining and Knowledge Representation*, 11(3):690–696.

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., and Mason, C. H. (2006). Defection detection: measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research*, 43(2):204–211.

R Development Core Team (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Schweidel, D. A., Fader, P. S., and Bradlow, E. T. (2008). Modeling retention within and across cohorts. *Journal of Marketing*, 72(1):82–94.

Torkzadeh, G., Chang, J. C. J., and Hansen, G. W. (2006). Identifying issues in customer relationship management at Merck-Medco. *Decision Support Systems*, **42**(2):1116–1130.

Wei, C.-P. and Chiu, I. (2002). Turning telecommunications call details to churn prediction: a data mining approach. *Expert Systems with Applications*, 23(2):103–112.

Zhang, Y., Qi, J., Shu, H., and Li, Y. (2006). Case study on CRM: detecting likely churners with limited information of fixed-line subscriber. *Proceedings of the International Conference on Service Systems and Service Management*,

2:1495–1500.

Corresponding author:

Clara-Cecilie Günther

Norwegian Computing Center

P.O. Box 114 Blindern

NO-0314 Oslo

Norway