The Korean Communications
in Statistics Vol. 11 No. 2, 2004
pp. 227-233

Prediction of the Probability of Customer Attrition by Using Cox Regression[1]

Hyuncheol Kang[2] and Sang-Tae Han[3]

Abstract

This paper presents our work on constructing a model that is intended to predict the probability of attrition at specified points in time among customers of an insurance company. There are some difficulties in building a data-based model because a data set may contain possibly censored observations. In an effort to avoid such kind of problem, we performed logistic regression over specified time intervals while using explanatory variables to construct the proposed model. Then, we developed a Cox-type regression model for estimating the probability of attrition over a specified period of time using time-dependent explanatory variables subject to changes in value over the course of the observations.

Keywords : Proportional Hazards Model, Cox Regression, Probability of Attrition

1. Introduction

The objective of this study is to construct a model for predicting the probability of attrition over specified points in time among customers of an insurance company. To begin with, we collected data on some one million customers who have taken out insurance for the past five years. However, there is a critical challenge of constructing a data-based model because censored observations may be contained in the data to be used. It results from the fact that, for customers who have yet to cancel their insurance policy, we know only the minimum length of time during which they are insured, not the exact duration of their insurance contract. This led to the failure to derive the proposed model by means of standard statistical approaches such as logistic regression which is often used to investigate the relationship between binary response (for example, success and failure) and a set of explanatory variables.

[1] This work was supported by grant No. R05-2003-000-11281-0 from the Basic Research Program of the Korea Science & Engineering Foundation.

[2] Senior Lecturer, Department of Informational Statistics, Hoseo University, Asan 336-795, Korea.
E-mail: hychkang@office.hoseo.ac.kr

[3] Associate Professor, Department of Informational Statistics, Hoseo University, Asan 336-795, Korea.
E-mail: sthan@office.hoseo.ac.kr

There are many types of models that have been used for censored data (or survival data). Among them, two types of models are most widely used; One is the accelerated failure time model (Kalbfleisch and Prentice, 1980) and the other the proportional hazards model, also known as Cox regression (Cox, 1972). Both models are based on the assumptions about the underlying distribution of survival times. The accelerated failure time model assumes a parametric form not only for the effects of the explanatory variables but also for the underlying survival function as a rule. Cox regression also assumes a parametric form for the effects of the explanatory variables, but it allows an unspecified form for the underlying survival function. Moreover, Cox regression makes it relatively easy to incorporate time-dependent explanatory variables which are subject to changes in value during any observation period. Consequently, Cox regression has been used in this study as a means of constructing the proposed model from a practical perspective.

Meanwhile, due to as many as about 100 explanatory variables collected (including time-dependent variables), it has not been possible, in a practical way, as it were, computationally, to develop the proposed model, taking all the variables into account at one time. Therefore, variable selection is needed as a pre-processing. In order to construct the proposed model, we used explanatory variables as well as performed logistic regression at several time intervals.

The rest of the paper is organized as follows. In Section 2, we formally introduce the basic concepts of survival analysis as well as Cox regression with time-dependent explanatory variables. In Section 3, we present the framework for variable selection as a preprocessing. Section 4 provides a description of the Cox-type regression model for predicting the probability of attrition at specified points in time while assessing the proposed model using a validation data set.

2. Basic Concepts of Cox Regression

Survival analysis is a loosely defined statistical term that encompasses a variety of statistical techniques for analyzing positive-valued random variables and finds its widest use in both reliability and clinical trial studies (Kalbfleisch and Prentice, 1980; Miller, 1981; Lawless, 1981). Survival data consist of a response variable that measures the duration of time until a specified event occurs (event time, failure time, or survival time) and possibly a set of explanatory variables thought to be associated with the survival time variable. The purpose of survival analysis is to model the underlying distribution of the survival time variable and to assess the dependence of the survival time variable on the explanatory variables.

Let the survival time variable, $T(\geq 0)$, have density $f(t)$ and distribution function $F(t)$. The survival function $S(t)$ is

$$S(t) = 1 - F(t) = P\{T > t\}, \tag{2.1}$$

and the hazard function $h(t)$ is

$$h(t) = \frac{f(t)}{S(t)}. \tag{2.2}$$

The survival function and the hazard function are interpreted as follows: {an individual survives longer than t} and $h(t) = \lim_{\delta t \to 0} P${an individual fails between t and $t + \delta t$ given it survived past time t}/ δt. The survival function can be expressed in terms of the hazard function

$$S(t) = \exp\left\{ - \int_0^t h(u) du \right\}. \tag{2.3}$$

An intrinsic characteristic of survival data is that a data set may contain possibly censored observations. That's because the event of interest is very rarely observed by the end of the study. Such censoring occurs when an entry is not observed until an event of interest (such as death or other type of failure) occurs. As it takes a long-period observation to get a response, the event of interest may not occur over the entire period of or at the termination time for data collection.

With focus more on the effect of explanatory variables on survival times, Cox regression (i.e. the proportional hazards model) is most widely used as a method for the analysis of censored data (Garfield, 1990), mainly because it eliminates the necessity of a particular probability distribution of survival times. The model is usually written as

$$h_i(t) = h_0(t) \ \exp\{ \boldsymbol{\beta}' \boldsymbol{x}_i \}, \tag{2.4}$$

where $h_0(t)$ is an arbitrary and unspecified baseline hazard function, $\boldsymbol{x}_i = (x_{i1} \cdots x_{ik})'$ is a vector of measured explanatory variables for the i-th individual, and is a vector of $\boldsymbol{\beta} = (\beta_1 \cdots \beta_k)'$ unknown regression parameters associated with the explanatory variables. Taking the logarithm of both sides, the model can be rewritten as

$$\log h_i(t) = \alpha(t) + \boldsymbol{\beta}' \boldsymbol{x}_i, \tag{2.5}$$

where $\alpha(t) = \log h_0(t)$. To estimate $\boldsymbol{\beta}$, Cox (1972, 1975) introduced the partial likelihood function, which eliminates the unknown baseline hazard $h_0(t)$ and accounts for censored survival times.

The partial likelihood of Cox also allows time-dependent explanatory variables. An explanatory variable is time-dependent if its value for any given individual can change over time and time-dependent variables have many useful applications in survival analysis. For the model with fixed and time-dependent explanatory variables, we have

$$\log h_i(t) = \alpha(t) + \beta_1' x_{i1} + \beta_2' x_{i2}(t). \tag{2.6}$$

This indicates that the hazard at time t depends on the values of fixed explanatory variables x_{i1}, and on the values of time-dependent explanatory variables x_{i2} at time t. What may not be clear is that $x_{i2}(t)$ can be defined using any information about the individual prior to time t, thereby allowing for lagged or cumulative values of some variables.

For the Cox regression model with time-dependent explanatory variables, the survival function can be expressed as

$$S(t; x_1, x_2) = [S_0(t)]^{\exp\{\beta_1' x_1 + \beta_2' x_{2(t)}\}}, \tag{2.7}$$

where $S_0(t) = \exp\left\{-\int_0^t h_0(u)\,du\right\}$ is the baseline survival function, that is, the survival function for an individual whose explanatory variables all have values of 0. After estimating β_1 and β_2 by partial likelihood, an estimate of $S_0(t)$ is made by a nonparametric maximum likelihood method (Allison, 1995).

3. Pre-Processing of Variable Selection

To construct the proposed model, we have collected data on some one million customers who have taken out insurance for the past five years (length of time between January 1995 and December 1999). The data consists of event times containing the length of time during which an insurance contract of interest is cancelled by their customers, the entire month of December 1999 during which the customers have yet to cancel the insurance contract, and a set of explanatory variables.

The set of non-time-dependent explanatory variables, x_1, is created based on the information at time $t=0$ (i.e. the time period during which a customer is insured). It also contains demographic variables (i.e. gender, age, marital status, etc.), socio-economic features (i.e. occupation, housing, etc.), policy-related information (i.e. policy categories, payment/collection methods, sum of premiums, etc.), transaction history (i.e. the date when an insurance policy of interest is signed by the customer, amount of loan, etc.), and solicitation agents (i.e. position, age, etc.).

Since it is not actually possible to create time-dependent variables at all points in time ($t \leq 60$), the set of time-dependent explanatory variables, $x_2(t_l)$, is created at seven points chosen for the time $(t_1=3, t_2=6, t_3=9, t_4=12, t_5=18, t_6=24, t_7=36)$ in which the insurance company in question experiences some of the highest attrition rates. The earlier a customer cancels the insurance policy, the higher the company suffers customer attrition. Therefore, we have selected some more points falling within earlier time periods in order to ensure accuracy in building the proposed model. The set of time-dependent explanatory variables contains the variables with lagged or cumulative values that encompass disabled status, amount of policy loan, changes/non-change made to payment method, question of whether or not a specific solicitation agent still works for the company at time interval $(0, t_l)$.

Since dummy variables are needed for a categorical variable like policy category and payment method, we have a number of parameters available for an estimate of the proposed model. For a model that allows time-dependent explanatory variables, computing the resulting partial likelihood is generally too time-consuming. A few variable selection methods for Cox regression have been suggested by SAS Institute (1999), Faraggi & Simon (1998), Ibrahim et al. (1999), and Fan & Li (2002) etc. However, those methods are not applicable to this case that allows numerous parameters as well as time-dependent explanatory variables. In actuality, it is not so easy to develop a model, taking all variables into account at one time. For this reason, we select the explanatory variables that will be used to construct the proposed model in the following exploratory manner.

Firstly, we create variables $y_l(l=1,\cdots,7)$ with the value of 1 if event time is less than or equal to t_l or 0 otherwise, and y_8 with the value of 1 if event time is not censored or 0 otherwise. Secondly, using the observations of which event time is greater than t_{l-1}(where $t_0 = 0$), we carry out usual logistic regression

$$\text{Logit}\,(y_l) = \alpha_l + \boldsymbol{\beta}_{1l}'\,\boldsymbol{x}_1 + \boldsymbol{\beta}_{2l}'\,\boldsymbol{x}_2(t_{l-1}). \tag{3.1}$$

for each response variable $y_l(l=1,\cdots,8)$ with a stepwise variable selection method. Finally, for our Cox regression model, we select the explanatory variables x_{1j}'s, of $\boldsymbol{x}_1$, in which more than one of parameters $\beta_{1l}(l=1,\cdots,8)$ are statistically significant. Additionally, we select the time-dependent explanatory variables $x_{2k}(\cdot)$'s, of $\boldsymbol{x}_2(\cdot)$, in which more than one of parameters $\beta_{2l}(l=1,\cdots,8)$ are statistically significant.

4. Prediction of the Probability of Customer Attrition and Validation of the Model

Using the selected explanatory variables as well as a pre-processing method of variable selection, we apply the following Cox regression model to the same data described in previous Section

$$\log h_i(t) = \alpha(t) + \beta_1' x_{i1} + \beta_2' x_{i2}(l), \tag{4.1}$$

where l is the maximum value of t_l's which are less than t. Once again, we select statistically significant variables from both x_{i1} and x_{i2}. Finally, we use 11 non-time-dependent explanatory variables (i.e. gender, age, policy category, payment method, amount of loan, solicitation agent's position, etc.) and 5 time-dependent explanatory variables (i.e. disabled status, changes/non-change made to payment method, the question of whether or not a specific solicitation agent still works for the company, etc.) as a means of constructing the proposed model.

Let T_0 denote the event time at a certain time. In order to curb attrition rates, the insurance company attempts to contact customers who are likely to cancel the insurance over the next six months. Thus, we can predict the likelihood of a customer's cancellation using an estimated survival function as well as the following conditional probability which is denoted by $AP6$,

$$AP6(x) = P(T_0 < t < T_0 + 6)/P(T_0 < T) = 1 - S(T_0 + 6)/S(T_0), \tag{4.2}$$

where x is the value of explanatory variables at a certain time.

To validate the model (4.2), we extracted the information about one hundred thousand customers who haven't cancelled the insurance policy yet as of December 1999, while collecting the validation data set of which variables have values as of December 1999. Furthermore, we obtained additional information about individual customers who cancelled the insurance policy during the time period between January 2000 and June 2000.

Table I shows the validation accuracy of the model in terms of rank ordering accounts on the basis of the predicted probability of attrition. This gains table, generally used as an assessment tool for evaluating the performance of a model in data mining (Rud, 2001), is constructed in the following manner: (1) A data set is sorted by the predicted probabilities of the event in descending order, (2) which is followed by grouping the observations into deciles, and (3) the gain (percent of the event) is calculated within each group. Thus, the gains of a good predictive model including its cumulative gains are dropped off abruptly at some cutoff values. Likewise, the gains of a mediocre model decrease gradually.

The lift is calculated by dividing the decile percent of the event by the overall percent of the event 7.5%. The value of 2.11 in decile 0-10% means that the customers in this decile are 211% more likely to be motivated than the overall average. At each decile, lift measurements demonstrate the model's power to beat the random approach or average performance. Table I indicates that the decile 0-10% is 2.11 times greater than the average. For the proposed model, up to and including the decile 30-40% outperforms the average. As illustrated by Table I, the proposed model has provided a high degree of predictability for customer attrition.

Table I. Gains and Lifts for Predicted Probability of Attrition

	Non-Cumulative				Cumulative			
Decile	# of Obs.	# of the events	Gain	Lift	# of Obs.	# of the events	Gain	Lift
0- 10%	10000	1583	15.8%	2.11	10000	1583	15.8%	2.11
10- 20%	10000	1035	10.4%	1.39	20000	2618	13.1%	1.75
20- 30%	10000	1019	10.2%	1.36	30000	3637	12.1%	1.61
30- 40%	10000	853	8.5%	1.13	40000	4490	11.2%	1.49
40- 50%	10000	715	7.2%	0.96	50000	5205	10.4%	1.38
50- 60%	10000	661	6.6%	0.88	60000	5866	9.8%	1.31
60- 70%	10000	637	6.4%	0.85	70000	6503	9.3%	1.24
70- 80%	10000	502	5.0%	0.67	80000	7005	8.8%	1.17
80- 90%	10000	338	3.4%	0.45	90000	7343	8.2%	1.09
90-100%	10000	166	1.7%	0.23	100000	7509	7.5%	1.00

References

[1] Allison, P.D. (1995). *Survival Analysis Using the SAS System : A Practical Guide*, SAS Institute, Cary, NC.

[2] Cox, D.R. (1972). Regression Models and Life Tables (with discussion), *Journal of the Royal Statistical Society*, Vol. B34, 187-220.

[3] Cox, D.R. (1975). Partial Likelihood, *Biometrika*, Vol. 62, 269-276.

[4] Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model, *The Annals of Statistics*, Vol. 30, 74-99.

[5] Faraggi, D. and Simon, R. (1998). Bayesian Variable Selection Method for Censored Survival Data, *Biometrics*, Vol. 54, 1475-1485.

[6] Garfield, E. (1990). 100 Most Cited Papers of All Time, *Current Contents*, Februry 12.

[7] Ibrahim, J., Chen M., and MacEachern, S. (1999). Bayesian variable selection for proportional hazards models, *The Canadian Journal of Statistics*, Vol 27, 701-718.

[8] Kalbfleisch, J.D. and Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*, John Wiley & Sons, New york.

[9] Lawless, J.F. (1981). *Statistical Models and Methods for Lifetime Data*, John Wiley & Sons, New york.

[10] Miller, R.J. (1981). *Survival Analysis*, John Wiley & Sons, New york.

[11] Rud, O. P. (2001). *Data Mining Cookbook : Modeling Data for Marketing, Risk, and Customer Relationship Management*, John Wiley & Sons, New york.

[12] SAS Institute (1999). *SAS/STAT User's Guide in SAS Online Document*, SAS Institute Inc.

[Received August 2003, Accepted April 2004]