Predicting Customer Attrition Probability Using Cox Regression

Hyuncheol Kang
Department of Mathematics, Hoseo University
29-1, Sechul-Ri
Asan, Korea
hychkang@office.hoseo.ac.kr

Sang Tae Han
Department of Mathematics, Hoseo University
29-1, Sechul-Ri
Asan, Korea
sthan@office.hoseo.ac.kr

Jaehyung Cha
Actuarial Team, LG Insurance Co., Ltd.
85, Da-Dong, Chung-Gu
Seoul, Korea
jhcha@lginsure.com

1. Introduction

The goal of this study is to construct the model to predict attrition probability at points in time and to validate the model for applying to life insurance company. The data is collected from a life insurance for the past 5 years from January in 1995 to December in 1999 and about one million contracts data was collected. This data has about 100 variables including the variables related to time and the variables about demographic feature, socio-economic feature, policy, transactions in past, and solicitation agent, etc..

There are so many complicated data (censored data) that collected data to reach our goal like as lapse - in life insurance, termination of policy because of failure to pay a premium and lack of sufficient cash value to make premium loan. In this case, we don't know that the policy will be persisted in the future because the policy in the state of lapse has the grace period - period after the date the premium is due during which the premium can be paid with no interest charged. Hence, we apply to survival analysis especially cox regression (Cox, 1972, 1975) used proportional hazard model. Also, to fit the model, we used the tool of data mining for large sample size.

2. Modeling for Cox regression and prediction of attrition probability

Finally, we carry out following Cox regression model with fixed and time-dependent explanatory variables (Allison, 1995; Lawless, 1981)

$$\log h_i(t) = \alpha(t) + \boldsymbol{\beta}_1' \mathbf{x}_{i1} + \boldsymbol{\beta}_2' \mathbf{x}_{i2}(l), \tag{1}$$

where $h_i(t)$ is the hazard of the i-th individual at the time t, $\mathbf{x}_{i1}$ the values of fixed explanatory variables, $\mathbf{x}_{i2}(l)$ the values of time-dependent explanatory variables at the time l, l the maximum of t_k's which are less than t, and $(t_1 = 3, t_2 = 6, t_3 = 9, t_4 = 12, t_5 = 18, t_6 = 24, t_7 = 36)$.

Let T_0 denote the event time at a certain time. With estimated survival function $S(t)$, we predict a customer's attrition probability, which indicate the posterior probability that a customer are likely to cancel the insurance in the next c months, by following conditional probability

$$\begin{aligned} AP(T_0, c \mid \mathbf{x}) &= P(T_0 < t < T_0 + c)/P(T_0 < t) \\ &= 1 - S(T_0 + c)/S(T_0), \end{aligned} \tag{2}$$

where $\mathbf{x}$ is the values of explanatory variables at the time T_0.

3. Validation of the model

To validate above model, we extract the information on a hundred thousand customers who don't cancel the insurance yet at December 1999 and collect the validation data set of which variables have the values at December 1999. Also, we obtain additional information that each customer cancel the insurance or not during the time period between January 2000 and Jun 2000. The observations are grouped into deciles by order of the predicted probabilities and then the gain (percentage of the event) is calculated within each group. In this study, by means of gains table, we observe that the model well predict the likelihood of customer attrition.

REFERENCES

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

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

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

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