

Statistical models to analyse the customer life cycle

Modelli statistici per l'analisi del ciclo di vita dei clienti

Silvia Figini*, Paolo Giudici

Università di Pavia

e-mail: silvia.figini@eco.unipv.it

Riassunto: Il lavoro prevede l'approfondimento metodologico dell'analisi statistica della sopravvivenza come strumento per la mitigazione del rischio di abbandono dei clienti. Nella parte applicativa si confrontano i metodi classici previsivi di data mining (alberi di classificazione e modelli lineari generalizzati) con modelli temporali semi-parametrici classici dell'analisi della sopravvivenza.

Keywords: Decision tree, Generalised linear models, Survival function, Cox model

1. Introduction

We consider the problem of estimating customer lifetime value of customers, when a large number of features are present in the data. In order to measure lifetime value we use survival analysis models to estimate customer tenure. One of the most popular methods in the analysis of survival data is the comparison of hazard and survival functions, for different classes of patients and / or treatments. Such comparison, along with Cox proportional hazard model, is well suited to infer the effect of covariates on survival time. In the paper we show how these ideas and methods can be adapted to a different environment, the estimation of customers life cycles. In such a context, a number of data mining modelling challenges arise. For example, differently from a patient, a client that is lost can come back. We will show how our approach performs, and compare it with classical churn models on a real case study, based on data from a media service company that aims to predict churn behaviours, in order to entertain appropriate retention actions. For non disclosure reasons in this paper we cannot give accurate statements and information about the company whose data we have analysed; we shall instead use general statements and normalised figures.

2. Weaknesses of traditional data mining models

Statistical models typically used to predict churn are based on logistic regression or classification tree (Giudici 2003). The use of new methods is necessary to obtain a predictive tool which is able to consider the fact that churn data is ordered in calendar time. To summarise, we can sum up at least two main weaknesses of traditional models in our set-up, which are all related to time-dependence:

1. Excessive influence of the contract deadline date, also shown by a high association of some variables in the database with the month of deadline;

2. Redundance of information: the database contains variables which gives redundant information. As these variables are time dependent, they may induce biased effects in the final estimates;

3. Survival analysis models to estimate churn

The previous points explain why we decided to look for a novel and different methodology to predict churn. A variety of statistical techniques arising from medical survival analysis (Singer and Willet, 2003) can be applied to tenure modeling. The tenure prediction models we have developed generate, for a given customer i , a hazard curve or a hazard function, that indicates the probability $h_i(t)$ of cancellation at a given time t in the future. A hazard curve (Hougaard, 1995) can be converted to a survival curve or to a survival function which plots the probability $S_i(t)$ of 'survival' (non-cancellation) at any time t , given that customer i was 'alive' (active) at time $t=1$, i.e.,

$$S_i(t) = S_i(t-1) \times [1 - h_i(t)]$$

with $S_i(1) = 1$. Once a survival curve for a customer is available, LTV for that specific customer i can be computed as:

$$LTV = \sum_{t=1}^T S_i(t) \times v_i(t)$$

where $v_i(t)$ is the expected business value of customer i at time t and T is the maximum time period under consideration. Survival analysis is concerned with studying the time between entry to a study and a subsequent event (churn). All of the standard approaches to survival analysis are probabilistic or stochastic. That is, the times at which events occur are assumed to be realizations of some random processes (Klein and Moeschberger, 1997). It follows that T , the event time for some particular individual, is a random variable having a probability distribution. A useful, model-free approach for all random variables is nonparametric, that is, using the cumulative distribution function. The cumulative distribution function of a variable T , denoted by $F(t)$, is a function that tell us the probability that the variable will be less than or equal to any value t that we choose. Thus, $F(t) = P(T \leq t)$. If we know the value of F for every value of t , then we know all there is to know about the distribution of T . In survival analysis it is more common to work with a closely related function called the survivor function defined as $S(t) = P(T > t) = 1 - F(t)$. If the event of interest is a churn the survivor function gives the probability of surviving beyond t . Because S is a probability we know that it is bounded by 0 and 1 and because T cannot be negative, we know that $S(0) = 1$. Finally, as t gets larger, S never increases (Kaplan and Meier, 1958). Often the objective is to compare survivor functions for different subgroups in a sample (clusters, regions). If the survivor function for one group is always higher than the survivor function for another group, then the first group clearly lives longer than the second group. When variables are continuous, another common way of describing their probability distributions is the probability density function. This function is defined as:

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}$$

That is, the probability density function is just the derivative or slope of the cumulative distribution function. For continuous survival data, the hazard function is actually more popular than the probability density function as a way of describing distributions. The hazard function is defined as :

$$h(t) = \lim_{\epsilon_t \rightarrow 0} \frac{Pr(t \leq t + \epsilon_t | T \geq t)}{\epsilon_t}$$

In our case, to explain $h(t)$ we have chosen to implement Cox's model. Cox made two significant innovations. First he proposed a model that is a proportional hazards model (Anderson, 1991). Second he proposed a new estimation method that was later named partial likelihood or more accurately, maximum partial likelihood. We will start with the basic model that does not include time-dependent covariate or non proportional hazards (Cox, 1972). The model is usually written as:

$$h(t) = h_0(t) \times \exp(\beta_1 X_1 + \dots + \beta_p X_p)$$

Cox model assumes that the hazard for individual i at time t is the product of two factors: a baseline hazard function that is left unspecified, and a linear combination of a set of p fixed covariates, which is then exponentiated. The baseline function can be regarded as the hazard function for an individual whose covariates all have values 0. The model is called proportional hazard model because the hazard for any individual is a fixed proportion of the hazard for any other individual. In Cox model building the objective is to identify the variables that are more associated with the churn event. This implies that a model selection exercise, aimed at choosing the statistical model that best fits the data, is to be carried out. The statistical literature presents many references for model selection (Giudici, 2003). Most models are based on the comparison of model scores. The main score functions to evaluate models are related to the Kullback-Leibler principle. This occurs for criteria that penalize for model complexity, such as AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion).

4. Application

Concerning explanatory variables, the available variables employed were taken from different databases used inside the company, which contained, respectively: socio-demographic information about the customers (hobbies, pc possession at home, address changes); information about their contractual situation and about its changes in time; information about contacting the customers (through the call centre, promotion campaigns, etc) and, finally, geo-marketing information (divided into census, municipalities and larger geographical sections). The variables regarding payment conditions include information about the type of payment of the monthly subscription (postal bulletin, account charge, credit card), with other info about the changes of the type of payment. The variables regarding contacts also include information about the number of questions made by every customer and temporal information, such as the number of calls made during the last month, the last two months and so on. Regardless of their provenience, all variables have gone through a pre-processing feature selection step aimed at reducing their very large number (equal to 606). Such step has been performed using a combination of wrapping and filter techniques, going from dimensionality reduction to

association measure ranking. The result of the procedure in our case is a set of about twenty explanatory variables. Such variables can be grouped in three main categories, according to the sign of their association with the churn rate, represented by the hazard ratio: variables that show a positive association (e.g. related to the wealth of the geographic region, the quality of the call center service, the sales channel), variables that show a negative association (e.g. number of technical problems, cost of service bought, status of payment method) and variables that have no association (e.g. equipment rental cost, age of customer, number of family components). More precisely, to calculate the previous associations we have considered the values of the hazard ratio under different covariate values. For example, for the variables indicating number of technical problems we compare the hazard function for those that have called at least once with those that have not made such calls. As the resulting ratio turns out to be equal to 0.849, the risk of becoming churning is lower for the callers than for the non callers. A very important remark is that Cox model generates survival functions that are adjusted for covariate values, computed as $S(t, X) = S_0(t) \times \exp(\sum_{i=1}^p \beta_i X_i)$. Model Fit Statistics presents an AIC value of 943788.62 (1124678.6, without covariates) and SBC value of 944347.95 (1124678.6, without covariates).

5. Conclusions

In the paper we have presented a novel data mining techniques to predict rates of churn of customers. Our conclusions show that the approach we have proposed, based on survival analysis modelling, leads to more robust conclusions. In particular, survival analysis modelling gives more valuable information, such as a whole predicted survival function, rather than a single predicted survival probability. Finally, our results show that survival analysis modelling is a much powerful tool for lifetime value analysis and, consequently, for the actual planning of a range of marketing actions that impact on both perspective and actual customers.

This work has been supported by MIUR PRIN FUNDS 'Data Mining for e-business applications', 2004-2006.

References

- Anderson K.M. (1991) A non proportional hazards Weibull accelerated failure time regression model, *Biometrics*, 47, 281-288.
- Cox D.R. (1972) Regression Models and Life Tables, *Journal of the Royal Statistical Society, Series B*, 34, 187-220.
- Giudici P. (2003) *Applied data mining : statistical methods for business and industry*, Wiley, London.
- Hougaard P. (1995) *Analysis of Multivariate Survival Data*, Springer.
- Kaplan E.L., and Meier R. (1958) Nonparametric Estimation from Incomplete Observations, *Journal of the American Statistical Association*, 53, 457-481 .
- Klein J.P., and Moeschberger M.L. (1997) *Survival analysis: Techniques for censored and truncated data*, Springer.
- Singer J.D., Willett J.B. (2003) *Applied Longitudinal Data Analysis : Modeling Change and Event Occurrence*, Oxford University Press.