Data Mining for Lifetime Value Estimation

Silvia Figini
University of Pavia, Italy

INTRODUCTION

Customer lifetime value (LTV, see e.g. Bauer et al. 2005 and Rosset et al. 2003), which measures the profit generating potential, or value, of a customer, is increasingly being considered a touchstone for administering the CRM (Customer relationship management) process. This in order to provide attractive benefits and retain high-value customers, while maximizing profits from a business standpoint. Robust and accurate techniques for modelling LTV are essential in order to facilitate CRM via LTV. A customer LTV model needs to be explained and understood to a large degree before it can be adopted to facilitate CRM. LTV is usually considered to be composed of two independent components: tenure and value. Though modelling the value (or equivalently, profit) component of LTV, (which takes into account revenue, fixed and variable costs), is a challenge in itself, our experience has revealed that finance departments, to a large degree, well manage this aspect. Therefore, in this paper, our focus will mainly be on modelling tenure rather than value.

BACKGROUND

A variety of statistical techniques arising from medical survival analysis can be applied to tenure modelling (i.e. semi-parametric predictive models, proportional hazard models, see e.g. Cox 1972). We look at tenure prediction using classical survival analysis and compare it with data mining techniques that use decision tree and logistic regression. In our business problem the survival analysis approach performs better with respect to a classical data mining predictive model for churn reduction (e.g. based on regression or tree models). In fact, the key challenge of LTV prediction is the production of segment-specific estimated tenures, for each customer with a given service supplier, based on the usage, revenue, and sales profiles contained in company databases. 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 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:

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

where $v_i(t)$ is the expected value of customer i at time t and T is the maximum time period under consideration. The approach to LTV (see e.g. Berger et. Al. 1998) computation provides customer specific estimates (as opposed to average estimates) of the total expected future (as opposed to past) profit based on customer behaviour and usage patterns. In the realm of CRM, modelling customer LTV has a wide range of applications including:

- Evaluating the returns of the investments in special offers and services.
- Targeting and managing unprofitable customers.
- Designing marketing campaigns and promotional efforts
- Sizing and planning for future market opportunities

Some of these applications would use a single LTV score computed for every customer. Other applications require a separation of the tenure and value component for effective implementation, while even others would use either the tenure or value and ignore the other component. In almost all cases, business analysts who use LTV are most comfortable when the predicted LTV score and/or hazard can be explained in intuitive terms.

Our case study concerns a media service company. The main objective of such a company is to maintain

its customers, in an increasingly competitive market; and to evaluate the lifetime value of such customers, to carefully design appropriate marketing actions. Currently the company uses a data mining model that gives, for each customer, a probability of churn (score).

The churn model used in the company to predict churn is currently a classification tree (see e.g. Giudici 2003). Tree models can be defined as a recursive procedure, through which a set of n statistical units is progressively divided in groups, according to a divisive rule which aims to maximize a homogeneity or purity measure of the response variable in each of the obtained groups. Tree models may show problems in time-dependent applications, such as churn applications.

MAIN FOCUS

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 summarize, we can sum up at least four main weaknesses of traditional models in our set-up, which are all related to time-dependence:

- excessive influence of the contract deadline date
- redundancies of information
- presence of fragmentary information, depending on the measurement time
- excessive weight of the different temporal perspectives

The previous points explain why we decided to look for a novel and different methodology to predict churn.

Future Survival Analysis Models to Estimate Churn

We now turn our attention towards the application of methodologies aimed at modelling survival risks (see e.g. Klein and Moeschberger 1997). In our case study the risk concerns the value that derives from the loss of a customer. The objective is to determine which combination of covariates affect the risk function, studying specifically the characteristics and the relation with the probability of survival for every customer.

Survival analysis (see e.g. Singer and Willet 2003) 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. 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 (see e.g. Hougaard 1995), 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 death (or, equivalently, 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. 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:

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}, \qquad (2)$$

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 (see e.g. Allison 1995) is defined as:

$$h(t) = \lim \varepsilon t \to 0 \frac{\Pr\{t \leq T < t + \varepsilon t \mid T \geq t\}}{\varepsilon t}, \qquad (3)$$

The aim of the definition is to quantify the instantaneous risk that an event will occur at time t. Since time