Customer relationship: a survival analysis approach

Silvia Figini[1]

University of Pavia
Via S.Felice 5
silvia.figini@eco.unipv.it

Summary. We consider the problem of estimating customer lifetime value. In order to measure lifetime value we use survival analysis models to estimate customer tenure. In the paper we show how these ideas and methods can be adapted to a different environment, the estimation of customer's life cycles. In such a context, a number of data mining modelling challenges arise. 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.

Key words: Data Mining, Predictive models, Survival analysis, Cox model.

1 Traditional data mining churn models

The company is such that most of its sales of services are arranged through a yearly contract, that allows buying different 'packages' of services at different costs. The contract of each customer with the company is thus renewed yearly. If the client does not withdraw, the contract is renewed automatically. Otherwise the client churns. In the company there are three types of churn events: people that withdraw from their contract in due time (i.e. less than 60 days before the due date); people that withdraw from their contracts overtime (i.e. more than 60 days before the due date); people that withdraw without giving notice, as is the case of bad payers. Correspondingly, the company assigns two different churn states: an 'EXIT' state to the first two classes of cutomers; and a 'SUSPENSION' state to the third. Concerning the causes of churn, it is possible to identify a number of components that can generate such a behaviour:

- A static component, determined by the characteristics of the customers and the type/subject of contracts;
- A dynamic component, that encloses trend and the contacts of the clients with the call center of the company;
- A seasonal part, tied to the period of subscription of the contract;

- External factors, that include the course of the markets and of the competitors

Statistical models typically used to predict churn are based on logistic regression or classification tree, see e.g. [GIU03].
The churn model used in the company to predict churn is currently a classification tree. We have compared the predictive performance of the best classification tree with logistic regression models. More generally, all created models were evaluated on the basis of a test sample (by definition not included in the training phase) and classified in terms of predictive accuracy with respect to the actual response values in it. In business terms, predictive accuracy means being able to identify correctly those individuals which will become really churner during the valuation phase (correct identification). Evaluation was made using a confusion, or cross validation matrix. However, there is the problem of the excessive influence of the contract deadline. The two types of trees (CART and Chaid) predict that 0.90 of customers whose deadline is in April is at risk. If we consider that the variable target was built gathering data of February, the customers whose term is in April and have to regularly unsubscribe within the 60 days allowed, must become EXIT in February. Therefore, despite their good predictive capability, these models are useless for marketing actions, as a very simple model based on customer's deadlines will perform as well.

2 Weaknesses of traditional data mining models

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:

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

- 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 Our proposal: Survival analysis models to estimate churn

The previous points explain why we decided to look for a novel and different methodology to predict churn. 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 see e.g.[AND91] . 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

$$S(t) = P(T \succ t) = 1 - P(T \leq t) = 1 - F(t) \tag{1}$$

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 Meier estimation, see e.g. [KAM58] is a very good method when the number of cases is small but representative and the exact survival times are known. In terms Kaplan-Meier Product-Limit Estimation, the estimated survival function at time t, *S(t)* has the following form:

$$S(t) = \prod_{i=1}^{t} [(n-i)/(n-i+1)]^{C_i} \tag{2}$$

where $\prod_{i=1}^{t}$ denotes the multiplication of the survival times across all cases less than or equal to t (the geometric mean), t is the time, (e.g. days, weeks, months, etc), n is the total number of cases in the sample, i is the number of cases surviving up to time t and C_i is a constant of censoring. 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:

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt} \tag{3}$$

That is, the probability density function is just the derivative or slope of the cumulative distribution function, see e.g. [HOU72]. 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 :

$$h(t) = lim_{\epsilon_t \to 0} \frac{Pr(t \leq T \leq t + \epsilon_t | T \geq t)}{\epsilon_t} \tag{4}$$

In our case, to explain *h(t)* we have chosen to implement Cox's model, see e.g. [COX72]. Cox made two significant innovations.
First he proposed a model that is a proportional hazards model. 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. The model is usually written as:

$$h(t) = h_0(t) \times exp(\beta_1 X_1 \ + ... + \beta_p X_p) \tag{5}$$

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 buiding 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 . 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).
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 was 'alive' (active) at time *t-1*, i.e.,

$$S_i(t) = S_i(t-1) \times [1 - h_i(t)] \tag{6}$$

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

$$LTV = \sum_{t=1}^{T} S_i(t) \times v_i(t) \tag{7}$$

where $v_i(t)$ is the expected business value of customer i at time t and T is the maximum time period under consideration.

4 The data available

The database available for our analysis contains information that can affect the distribution of the event time, as the demographic variables, variables about the contract, the payment, the contacts and geomarketing. We remind the readers that the target variable has a temporal nature and, for this reason, it is preferable to build predictive models through survival analysis, see e.g. [KLM97]. 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 perfomed 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.

5 Survival analysis models to estimate churn

In order to build a survival analysis model, we have constructed two variables: one variable of status (distinguish between active and non active customers) and one

of duration (indicator of customer seniority) . The survival function is estimated through the methodology of Kaplan Meier. The Kaplan Meier estimator is the most widely used method for estimating a survival function and it is based on a nonparametric maximum likelihood estimator. When there are non censored data the KM estimator is just the sample proportion of observations with event times greater than t. Suppose there are K distinct event times, $t_1 < t_2 < ... < t_k$. At each time time t_j there are n_j individuals who are said to be at risk of an event. At risk means they have not experienced an event not have they been censored prior to time t_j. If any cases are censored at exactly t_j, there are also considered to be at risk at t_j. Let d_j be the number of individuals who die at time t_j. The KM estimator is defined as

$$\hat{S}(t) = \prod_{j:t_j \leq t} [1 - \frac{d_j}{n_j}] for t_1 \leq t \leq t_k, \tag{8}$$

This formula says that, for a given time t, take all the event times that are less than or equal to t. For each of those event times, compute the quantity in brackets, which can be interpreted as the conditional probability of surviving to time t_{j+1}, given that one has survived to time t_j. Then multiply all of these survival probability together. The first step in the analysis of survival data (for the descriptive study) consists in a plot of the survival function and the risk. Figure 1 shows the survival function for the whole customer database. From Figure 1 note that the survival function has

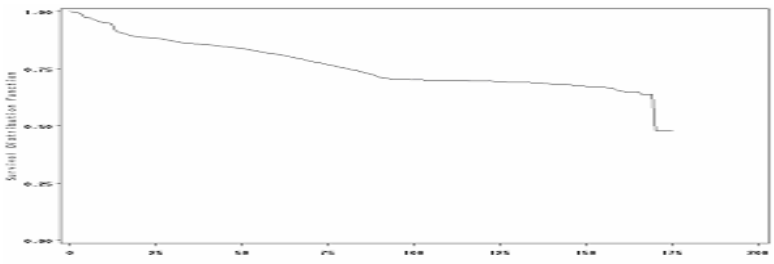

Fig. 1. Descriptive Survival function

varying slopes, corresponding to different periods. When the curve decreases rapidly we have time periods with high churn rates; when the curve decreases softly we have periods of 'loyalty'. We remark that the final jump is due to a distorsion caused by a few data, in the tail of the lifecycle distribution. A very useful information, in business terms, is the calculation of the life expectancy of the customers. This can be obtained as a sum over all observed event times:

$$\hat{S}(t_j) \times (t_j - t_{j-1}), \tag{9}$$

where $\hat{S}(t_j)$ is the estimate of the survival function at the j-th event time, obtained using Kaplan Meier method, and t is a duration indicator. We remark that life expectancy tends to be underestimated if most observed event types are censored (i.e., no more observable). All this methodological steps are implemented in the R

software. We use in particular the survival package and we have write some news function to compute for each customer the life expectancy. We now move to the building of a full predictive model. We have chosen to implement Cox's model. In our case we have compared all models and chose the one with the lowest value of AIC and BIC. The obtained model presents a good fit with AIC=943788.62 and BIC=944347.95 with the inclusion of covariates: this means that adding covariates lead to a better fit. As well known, the BIC presents higher values due to its penalty term. Log likelihood comparison can be formally embedded into an overall statistical test, such as Score, Wald or the likelihood ratio test, see e.g. [ALL95]. The result of the procedure 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)
- Variables that have no association (e.g. equipment rental cost, age of customer, number of family components).

More precisely, to calcolate the previous associations we have considered the values of the hazard ratio under different covariate values. A very important remark is that Cox model generates survival functions that are adjusted for covariate values. Figure 2 below shows a comparison between the survival curve obtained without covariates (the baseline, as in Figure 1) and the same curve adjusted for the presence of covariates. Figure 2 shows that covariates affect considerably survival times: up

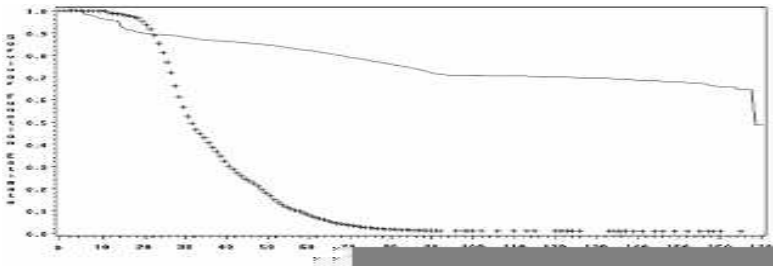

Fig. 2. Comparison between survival functions

to two years of lifetime, the Cox survival curve (described by the symbols '+') is greater with respect to the baseline (described by the continuous curve). After such period the survival probability declines abruptly and turns out to be much lower for the remaining lifespan. Once a Cox model has been fitted, it is advisable to produce diagnostic statistics, based on the analysis of residuals, to verify if the hypotheses underlying the model are correct. In our case they were found to be correct, so we could proceed with the predictive stage. This part is a new approch to evaluate

the performance of the Cox Model implemented. We have write a function with R that is able to compute the lift value, the confusion matrix and some new index. In particular, in the prediction step the goodness of the model will be evaluated in terms of predictive accuracy. In order to evaluate the predictive performance of the model, and compare it with classical data mining models, we have focused our attention to the 3 months ahead prediction. Once survival probabilities have been calculated, we have deviced and implemented a procedure to build the confusion matrix and, correspondingly, the percentage of captured true churners of the model. We remark that this is indeed not a fair comparison as survival models predict more than a point; however the company wanted this comparison as well. In correspondence of each estimated probability decile, we report the percentage true churners in it (first decile, captured 0.0504, second decile, captured 0.0345). While in the first decile (that is, among the customers with the highest estimated churn probability) 0.05 of the clients are effective churners, the same percentage lowers down in susequent deciles, thus giving an overall picture of good performance of the model. Indeed the lift of the model, as measured by the ratio between the captured true responses (model vs random) does not turn out to be substantially better with respect to what obtained with tree models. However, we remark that, differently from what occurred with classical models, the customers with the highest estimated churn rate are now not necessarily those whose contract is close to the deadline. This is the most beneficial advantage of the survival analysis approach, that, in turn, leads to substantial gains in campaign costs. A further advantage of the survival analysis approach lies in its immediate translation in terms of lifetime value analysis.

6 Conclusions

In the paper we have presented a comparison between classical and novel data mining techniques to predict rates of churn of customers. Our results show that survival analysis modelling is a much powerful tool, see e.g. [SIW03], for lifetime value analysis and, consequently, for the actual planning of a range of marketing actions that impact on both perspective and actual customers. Our feature research includes the development of bayesian models to improve data description and prediction.

7 Acknowledgements

I am must grateful to my supervisor, Prof. Paolo Giudici (University of Pavia), for his academic guidance and research assistance. This work has been supported by MIUR PRIN FUNDS 'Data Mining for e-business applications', 2004-2006.

References

[ALL95] Allison, P. D.: Survival Analysis Using the SAS Sysrem. SAS Institute, (1995)

[GIU03] Giudici, P. : Applied Data Mining. Wiley, (2003)

[KLM97] Klein, J.P. and Moeschberger, M.L.: Survival analysis: Techniques for censored and truncated data. Springer, (1997)

[SIW03] Singer,P. Willet, D.: Applied Longitudinal Data Analysis. Oxford University Press, (2003)

[AND91] Anderson, K.M.: A non proportional hazards Weibull accelerated failure time regression model. Biometrics, **47**, 281–288, (1991)

[COX72] Cox, D. R.: Regression Models and Life Tables. Journal of the Royal Statistical Society, **B34**, 187–220, (1972).

[HOU72] Hougaard, P.: Frailty models for survival data. Lifetime Data Analysis, **1**, 255–273, (1995).

[KAM58] Kaplan, E. L. and Meier, R.: Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association, **53**, 457–481, (1958).