Survival data mining in the telecommunications industries: usefulness and complications

Z. Mohammed & D. Kotze
Department of Statistics, University of the Western Cape, South Africa

Abstract

Decision makers in business industries have seen the change from the old economy to a new economy. The old economy is goods-centred, transaction-based in nature, focused on customer attraction, and product-based thinking whereas the new economy is service-centred, subscription-based in its nature, focused on customer retention, and customer-based thinking. The firm of today has to evaluate its performance by taking the customers into consideration. It is important to study customer lifetime value, which is the net present value of customers' profit over a time period. The two main components used to build the customer lifetime value model are customer length of service (which is represented by customers' survival curve) and customer monthly margin. While the customer monthly margin can be obtained from an accounting model, the major problem is customer survival time. In this study we take into consideration the telecommunications industry which represents a good example of subscription-based business where customer and customer relation is the vital factor in success. The nature of customer survival time in such an industry has brought many complications in customer survival modelling. Some of these complications are non-smooth survival functions, non-smooth spiky hazard functions, the possibility that both customer and company can initiate the churn, multi-churns, and multi-reactivations. In this study we review survival data mining and we discuss how survival data mining approaches are eligible to represent complications involved and how they are beneficial from both a practical and methodological point of view. A model for customer survival time is suggested and discussed and challenges to apply this model in practice are raised.

Keywords: telecommunications industry, customer equity, customer lifetime value, survival analysis, churn analysis, hazard probability.



1 Introduction: customer and measuring the firm's performance

The business world is increasingly organizing itself around customers rather than products. The importance of the customer has resulted from a series of changes that led to the transformation from an old economy to a new economy. The old economy was goods-centred, transaction-based in nature, focused on customer attraction, and product-based thinking while the new economy is service-centred, subscription-based in its nature, focused on customer retention, and customer-based thinking. Customer focus requires a new approach: managing according to customer equity (the value of a firm's customers), rather than brand equity (the value of a firm's brands), and focusing on customer profitability rather than product profitability (Rust et al. [1]). Robert Blattberg and John Deighton introduced the term customer equity (Blattberg and Deighton, [2]). This means: The total discounted future net revenue that a firm expects from its relationship with its customers today.

Although the concept is very clear and straightforward, management and application of a customer equity approach, analytically (using operational customer databases), seems to have some barriers (Bell et al. [3]). These barriers are: First, to find an effective way to assemble individual level, industry-wide customer data. Although technology makes data collection and storage easy and cheap, we still meet with some difficulties like the determination of the dimensions of a specific driver or metric, the selection of appropriate variables among a massive number of variables to answer specific business questions, the gap between statisticians (more generally modellers) and business managers in understanding the phenomena under investigation, and the sensitivity of customer information. In relation to the data, we have the issue of measuring customer equity drivers (value equity, brand equity, and retention equity) numerically. Most of these drivers depend on customer perceptions. The second challenge in using the customer as a measure of a firm's performance, is that customer asset metrics (customer lifetime value) depend on assumptions about the future stream of income of a customer, the appropriate allocation of costs to customers, the discount factor, the expected lifetime of a customer and the probability that a customer is alive at a point in time. Thirdly, modelling future revenues appropriately, maximizes not only the measurement of customer equity but also aligns the organization with customer management activities. From this list of issues that are encountered, the consideration of the customer as an asset and using an appropriate metric to evaluate them, we are going to investigate customer lifetime, or as we call it here, customer survival time: the expected duration that the customer is going to actively engage in business with the firm. Taking into consideration the telecommunication business class, customer survival time is the main component that we need in order to estimate the customer lifetime value (CLV). A similar approach to customer survival time is customer churn: the probability that a customer will stop doing business.

In the following section we will discuss customer survival time and we will review what researchers, in the data-mining field, have done so far.



2 Customer survival time and survival data mining

Survival analysis concerns time-to-event data analysis (Oakes, [4]). It was originally applied in the medical field. Examples of such problems are death, heart attack of a patient, and machine failure. Sometimes we need to follow a group of individuals over a certain period to record the occurrence of the event under study, for instance a heart attack of a patient. By the end of the follow-up period not all individuals will have experienced the event. Survival analysis techniques more specifically, deal with the analysis and modelling of time-to-event data with censored cases.

Subscription-based business such as the telecommunication industry, banking service, and insurance are good examples where you can find a well-defined start and end point of the customer's relation with the firm. It is convincing to say that the time from subscription to churn is equivalent to and can be modelled using survival analysis techniques. To study the lifetime of a group of customers (who are active at a certain point of time) by recording the date of churn, we may have censoring or truncation (in truncated data an observation is made by the investigator only when it experiences a certain event). Not all cases may have the event date when the study is terminated.

Suppose that the event time of the *i*th customer is T_i, the event time distribution is described by the survival function $S(t)$, *where*

$$S(t) = \Pr ob(T_i \geq t). \tag{1}$$

This function represents the probability that a customer will continue his/her relation with the firm up to time *t* or the proportion of customers who are actively engaged in business from the total number of customers who were followed from a fixed point in time till *t.* Another very important function is the hazard function, *h(t),* where

$$h(t_j) = \Pr ob(T = t_j \mid T \geq t_j) = 1 - \frac{S(t_{j+1})}{S(t_j)}. \tag{2}$$

The continuous version of the hazard rate is:

$$h(t) = \lim_{\Delta \to 0} \left(\frac{1}{\Delta} \Pr ob(t \leq T < t + \Delta \mid T \geq t) \right) = -\frac{d}{dt} \ln(S(t)). \tag{3}$$

The hazard rate represents the chance that a customer is going to churn at t, which is relevant and valuable information to study customer churn and survival. A more common model used is the proportional hazard model, which was introduced by David Cox [5]. The hazard function is defined as follows:

$$h_i(t) = \left(h_0(t)\right)\left(\exp\left(\beta^T z_i(t_i)\right)\right). \quad (4)$$

where $h_0(t)$ is a baseline hazard function common to all individuals, β is a vector of regression parameters of the dependence of the survival time, T_i, distribution on the covariates vector z_i. The Cox model is one which has been used most often and is highly appreciated. Different parametric and nonparametric settings have been used to represent the survival time distribution.

Recently researchers have tried to apply survival techniques in a data-mining framework and approach. Now, we can define survival data mining as the application of survival analysis, a traditional statistical technique, to data mining problems with censoring. The aim is to use historical data to build predictive models for various problems such as cancellation of service, acquiring a new product or upgrading and downgrading.

Survival data mining was discussed by a number of researchers. Will Potts (Potts [6]) outlined the application of survival analysis to predictive modelling and scoring. He included a discussion on discrete-time logistic models and piece-wise hazard models. The discussion involved the applicability and parameter estimation in each of these models in addition to Weibull and piece-wise exponential models.

Junxiang Lu (Lu [7]) discussed the estimation of customer lifetime value using survival analysis. He tried to build a model for CLV in the telecommunication industry. He emphasized that the customer survival curve and the customer monthly margin are the most important components in modelling CLV. He was interested to apply the model in terms of sampling, variable reduction, model estimation, and model validation.

Gordon Linoff (Linoff [8]) pointed out that the nature of the survival curve in a business application might have special properties that come from phenomena involved in business. Termination of contracts, non-payment, and end promotion can lead to a sharp drop in the survival curve or a non-smooth spiky hazard function.

In the following sections we will discuss customer survival time modelling that can be used to estimate customer lifetime value taking into account the telecommunication industry. Moreover we will investigate the methodological representation of these complications. The business consequences of these complications will also be taken into account.

3 A customer survival time model for the telecommunication industry

While the main role of survival analysis in medical applications is to analyse influential factors that can affect the life of the patients, it has two important functions in the business field. Firstly, we need to use survival analysis techniques to study factors that can prolong the customer's relation with the firm, which is similar in approach to the one in the medical field application. In business terminology these techniques can effectively be used to test the impact of marketing campaigns (e.g. assessing the effectiveness of different retention

programs, different levels of a campaign, different incentives being used either to acquire new customers or upgrade old customers to be more profitable and to add on new products or services). Secondly, customer survival time can be used to predict the expected future revenue from the customer, i.e. customer lifetime value. This is going to be helpful to identify and target the most profitable customers and to evaluate the firm's value by taking into consideration a customer approach. But it requires more precision otherwise we are going to lose reality along a series of estimations which involve approximations and assumptions. Existence of a massive number of records may give an advantage to business application versus medical application where a limited number of records are present. Data mining is designed to handle the situation in which we have a large data set but the methodological approaches borrowed mainly from statistics may not be able to accomplish this job in a satisfactory way. The nature of customer survival time in the telecommunication industry (and similar industries like insurance companies, internet service providers, banking, etc) and objectives (roles) of survival analysis in business require a careful handling of the problem. The problem here involves several serious issues. Firstly, both customer and company can initiate the termination of the business relation. The company can initiate the termination of the relation because, for example the customer is not able to pay or it can even try to get rid of the customer because he is not profitable. A customer can also leave the company due to unsatisfactory service he receives. Secondly, multi-churns and multi-reactivations have a great impact on both methodological bases used to estimate the survival curve and the conceptual business bases in understanding the cost and profitability of the old customer who activated his service. No doubt, he is different from the new acquired customer. Thirdly, the nature of covariates that lead to termination of business relations, e.g. type of payment (credit card, cash, bank account), contract (exist or not, is there any penalty for those that leave?), end of promotion, and emergence of new service provider must be considered. It is intuitive to say that each of these covariates leads to a sudden loss of a considerable number of customers on the same date. Again this will result in a different way to see the survival curve as non-smooth with a sharp drop at specific dates representing factors like the end of a contract, end of promotion, emergence of new service, and type of payment. On the other hand a non-smooth spiky hazard curve that represents the probability of churn will be seen.

4 Model construction

For the purpose of measuring the firm's performance using a customer approach and taking into consideration the telecommunication industry, we have constructed the following model. The model is illustrated in Fig. 1

The covariates that can affect customer survival time can be divided into the following groups:

1. Internal factors: customer side
2. Internal factors: company side
3. External factors

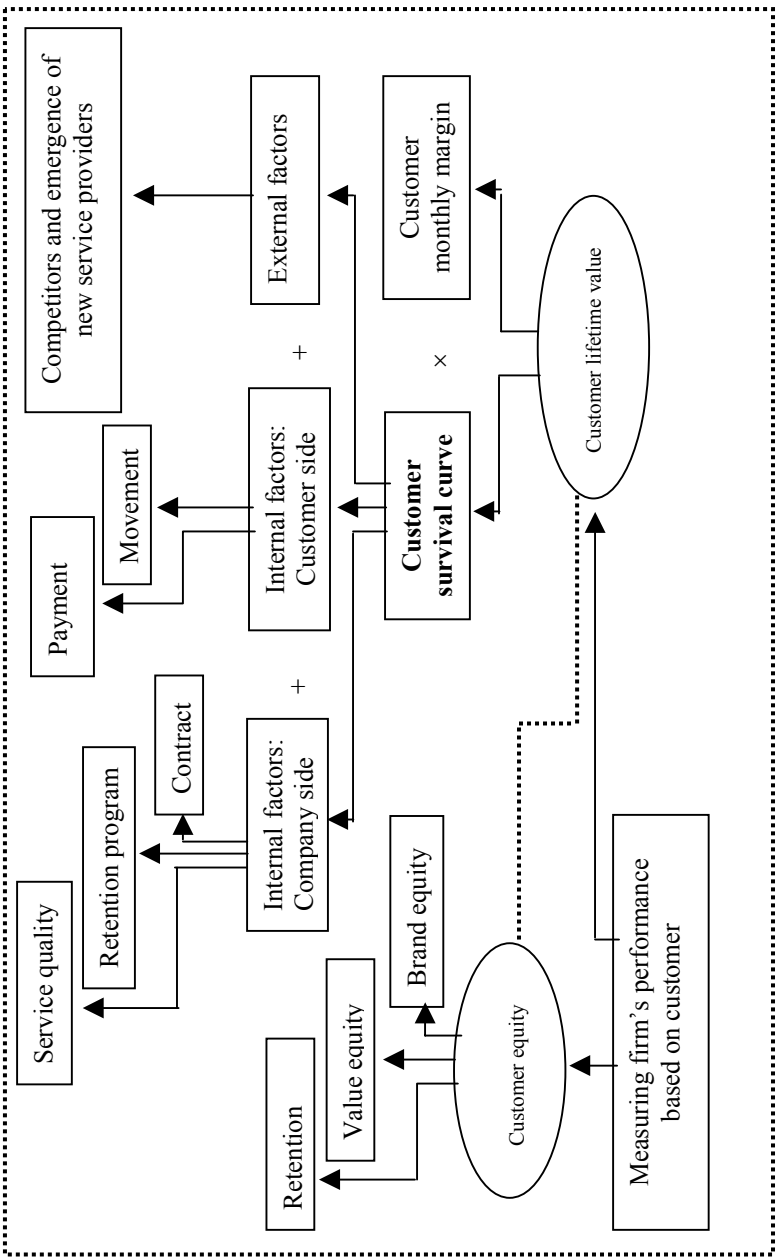


Figure 1: Customer survival time and measuring firm's performance.

The internal factors (covariates) from company side include: service quality, retention programs, and existence of a contract. Firm concerns with customer expectation and needs, and the appropriate incentives and marketing strategies are expected to prolong customer survival time. The existence of a contract, obligation, and penalty make churn and switching difficult within the contract period. Internal covariates from the customer side involve type of payment and movement of the customer. External covariates are those that cannot be managed by the firm or the customer. These include the emergence of a new service provider and all existing competitors' actions. Although the firm is not able to control other firms' strategies and actions, it is able to react positively and appropriately.

The details of how to apply this model are beyond the scope of this study. It is one of our goals to do this in future. However, we sketch here a general idea of how to deal with customer survival time in the telecommunication industry using this model. We recognized the great importance of free thinking when we talk about time-to-event (time from subscription to churn) that can characterize customer survival time. Because not all factors are similar in nature of causing churn, different ways to deal with each factor is required. This gives a chance of better representation of the nature of each factor by a suitable probability distribution or any probabilistic model (e.g. stochastic process). Intuitively we should not consider the customer survival time distribution due to movement in the same way as customer survival time at the end of a contract, as the cause of death cannot be seen as the same event as customers not satisfied with the service. The cause of death is a composition of cause of death of all factors. Factors can be treated as dependent or independent, but strong assumptions of independence of all factors may make life easy.

Suppose that in a business environment covariates are coded as follows: service quality: f_1, retention program: f_2, contract: f_3, payment: f_4, movement: f_5, and emergence of new service provider: f_6. A simple representation, assuming the independence assumption, is given in eqn (5) below

$$1 - S(t) = \Pr ob(\bigcup_{i=1}^{6} f_i) = \sum_{i=1}^{6} \Pr ob(f_i)\,. \tag{5}$$

Where $\Pr ob(f_i)$ is the probability that the churn caused at time t is due to factor f_i and $S(t)$ is the probability that the customer will survive (keep doing business with the company) up to time t. The probability of churn is given by $1 - S(t)$. The hazard rate, which is the probability of churn exactly at t is going to be the sum of all hazard rates of each one of the six factors.

5 Discussion and conclusion

The firm of today has to evaluate its performance by taking the customers into consideration. The proposed model shows how the survival time problem lies in the heart of the process of evaluating a firm's performance using a customer

approach. While the main role of survival analysis in a medical application is to analyse influential factors that can affect the life of patients, it has two important functions in the business field. Firstly, we need to use survival analysis techniques to study factors that can prolong the customer's relation with the firm. Secondly, customer survival time is used to predict the expected future revenue from the customer, i.e. customer lifetime value. This is helpful to identify and target the most profitable customers and to evaluate the firm's value by taking into consideration a customer approach, but it requires more precision, otherwise we are going to lose reality along with a series of estimations which involve approximations and assumptions. Data mining will have an important role to extract such information. The nature of factors that affect customer survival time in the telecommunication industry and modelling objectives (estimating future revenue as well as analysing influential factors) require more precision and careful methodological bases to use. We believe that the proposed model will meet these requirements. But again, each factor needs to be treated differently with respect to methodological bases, i.e. the probability distribution of the survival time of a customer. Further study on how to put the proposed model into practice is of great importance and part of future research.

References

[1] Rust, R. et al. Driving customer equity, The Free Press, 2000.
[2] Blattberg, R. & Deighton, J., Manage marketing by the customer equity test, Harvard Business Review, 74(4), pp. 136-144, 1996.
[3] Bell D., et al. Seven barriers to customer equity management, Journal of service research, 2002.
[4] Oakes, D., Biometrika Centenary: Survival analysis, Biometrika, 88(1), pp. 99-142, 2001.
[5] Cox, D.R., Regression models and life tables (with discussion), J. R. Statist. Soc. B 34, pp. 187-220, 1972.
[6] Potts, W., Survival data mining (white paper), http://www.data-miners.com/resources/Will%20Survival.pdf.
[7] Lu, J., Modelling customer lifetime value using survival analysis, SUGI (28), pp 120-128, 2003.
[8] Linoff, G., Survival data mining for customer insight, Intelligent Enterprise (7), 2004.

