UNIVERSITÀ DEGLI STUDI DI BERGAMO
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE
E METODI MATEMATICI°

QUADERNI DEL DIPARTIMENTO

Department of Information Technology and Mathematical Methods

Working Paper

Series "*Mathematics and Statistics*"

n. 12/MS – 2012

Survival Analysis for Customer Lifetime Value Estimate

by

Maurizio Toccu

° Viale Marconi. 5, I – 24044 Dalmine (BG), ITALY, Tel. +39-035-2052339; Fax. +39-035-562779

[§] L'accesso alle *Series* è approvato dal Comitato di Redazione. I *Working Papers* della Collana dei Quaderni del Dipartimento di Ingegneria dell'Informazione e Metodi Matematici costituiscono un servizio atto a fornire la tempestiva divulgazione dei risultati dell'attività di ricerca, siano essi in forma provvisoria o definitiva.

Survival Analysis for Customer Lifetime Value Estimate

Maurizio Toccu

Keywords: customer lifetime value, survival analysis, churn prevention, customer retention, sas

Abstract

Traditional marketing metrics such as brand awareness, sales and share are not enough to show a return on marketing investment. Customers are important intangible assets of a company but they are not equally remunerative. Therefore, estimating customer lifetime value (CLV) is becoming increasingly important to identify prospective profitable customers. In particular, thanks to Gupta et al. (2006), using survival analysis approach it is possible to estimate the customer retention rate in term of probability.
This article examines the recent literature of a number of implementable customer lifetime value and survival models and proposes a case study (telecommunication sector) about modeling customer lifetime value using survival analysis. With the increasing development of information technology, the gap between products and services becomes less and less. The wealth of customer information and increasingly sophisticated information technology and statistical modeling have led to a revolution in areas such as customer relationship management or CRM (Winer, 2001). This paper represents a useful tool that allows to help management strategy to estimate CLV and the risk of customer churn.

1 Introduction

The transition from traditional product-centric to a customer-centric approach has sanctioned that the customer is the more important asset of a company. Over the past decade, the customer management has rapidly emerged as an important area of research in marketing. Beginning with work that started to consider the customer as a critical asset of the firm (Blattberg and Deighton 1996; Reinartz and Kumar 2000; Srivastava, Shervani, and Fahey 1998), research on customer management has grown significantly.
If firms know the needs of customers they will be able to survival over time, therefore CLV is the best metric to measure and identify prospective and profitable customers (Chen et al., 2009). Many models of customer lifetime value have emerged (see Jain and Singh (2002) for a review), several models examining the antecedents of CE have been developed (e.g., Blattberg, Getz, and Thomas 2001, Rust, Lemon, and Zeithaml 2004; Rust, Zeithaml, and Lemon 2000; Venkatesan and Kumar 2004), and a strong research tradition has emerged providing insight into the effects of marketing actions on customer retention and customer lifetime value (Bolton 1998; Bolton, Lemon, and Verhoef 2004; Reinartz and Kumar 2003; Venkatesan and Kumar 2004; see also the August 2002 issue of Journal of Service Research).
Recently, researchers have started exploring the link between CLV and financial performance (Gupta and Lehmann 2003; Gupta, Lehmann, and Stuart 2004). The marketing literature has developed and discussed the concept of customer lifetime value, which is the present value of all future profits generated from a customer (Gupta & Lehmann, 2003). Customer profitability models have evolved into important strategic tools for marketers. Traditional customer profitability models implicitly assume that customers can be valued in isolation from one another and that social interactions can be ignored. But these conventional models may be inappropriate for markets involving new products or services because they fail to account for the social effects (e.g., word of mouth and imitation) that can influence future customer acquisitions. Proposers of CLV say that it could be used for customer acquisition (CLV is the upper bound of what one should be to spend acquiring a customer; Farris et al. 2006; Berger and Nasr 1998), customer selection (focusing on customer with high CLV; Venkatesan and Kumar 2004), and resource allocation (marketing resources should be allocated so as to maximize CLV?; Reinartz et al. 2005; Venkatesan and Kumar 2004). There is an increasing interesting in firms to make quantitative marketing, indeed recent studies have showed that not all customers are alike valuable. Therefore, it may be advantageous to select some customers or assign different resources to other group of customers. Such controls are not possible from aggregate financial measures but CLV is a disaggregate metric that is useful to identify valuable customers and allocate resources accordingly. A new CLV evolution uses survival models to estimate customer churn probability.

Customer churn (in particular profitable customers) has become a significant problem for firms in several sectors of economy (advertising, financial services, insurance services, electric utilities, banking services, internet services, telephone services, etc.). In this paper we consider the telecommunications sector and we study churn rate by customer survival probability.

In the introduction of this paper, we discuss some implementable CLV and survival models and propose a

case study about modeling customer lifetime value using survival analysis. The rest of the article is organized as follows. The second section shows the statistical methods that we have used to analyze the data of telephone company. The third section presents a literature review of CLV metrics and future developments. The fourth section presents a case study based on the dataset of a telephone company and relevant pieces of SAS code. The fifth section presents the SAS procedures used to analyze the data of the case study. The sixth section contains the conclusions.

2 Statistical methods

The aim of this study is to calculate customer lifetime value through estimating customer survival time. Survival analysis techniques that, at the beginning, were designed to handle medical data and therefore they are powerful tools to forecast customer survival/churn. Common statistical methods, like logistics regression, decision tree, etc., are very efficacious in predicting customer survival/churn but they don't manage censored data (i.e., when the customer does not die).

Customer lifetime value

CLV is usually defined as the net present value of the cash flows attributed to the relationship with a customer over time. CLV seems like the discounted cash flow approach used in finance but there are two differences. First of all, CLV is usually defined and estimated for each customer. This is useful to choose customers who are more profitable. Second, unlike finance, CLV explicitly incorporates the possibility that a customer may defect to competitors in the future (Gupta et al., 2006). CLV for a customer is (Gupta, Lehmann, and Stuart 2004; Reinartz and Kumar 2003):

$$CLV = \sum_{t=1}^{T} \frac{(p_t - c_t) r_t}{(1+i)^t} - AC \tag{1}$$

p_t = price paid by a consumer at time t
c_t = direct cost of servicing the customer at time t
i = discount rate or cost of capital for the firm
r_t = probability of customer repeat buying or being alive at time t
AC = acquisition cost
T = time horizon for estimating CLV

When this expression contains the acquisition cost (AC) then we are taking into account the CLV of potential customer, else we are computing the expected residual lifetime value of an existing customer.

Furthermore, if a consumer purchases multiple products from a firm, the margin used in Equation 1 is the sum of margins obtained from all products purchased (Gupta et al., 2006).

Survival analysis

Common statistical methods, like logistics regression, decision tree, etc., are very efficacious in predicting customer survival/churn but they could hardly forecast when customers will churn. In particular, one possibility is to perform a logistic regression with a dichotomous dependet variable: event or not event. But this analysis ignores information on the timing of the events (Allison, 2004). Morever, survival analysis allows to manage censoring observations.

Survival function and hazard function are used to describe the status of customer survival during the observation time.

The survival function calculate the probability to survive beyond a definite time point t.

$$S(t) = P(T > t) \tag{2}$$

Conversely, the hazard function describes the risk of event (in this case, customer churn) in an interval time after time t, assuming that the individual has survived to the beginning of the interval.

$$h(t) = \lim_{\Delta t \to 0} \frac{P[t \leq T < t + \Delta t | T \geq t]}{\Delta t} \tag{3}$$

where

$P[t \leq T < t+\Delta t | T \geq t]$: probability that an individual fail in the time interval $(t, t + \Delta t)$ given the individual has survived to t.

Then, the hazard function attempts to quantify the instantaneous risk that customer will churn at t given that the customer already has survived until t.

For survival analysis we begin observing a set of customers starting from the origin of time and then follow them for some period of time, recording the times at which customers churn.

Survival analysis, that originally has been designed for longitudinal data on the occurrence of events and it has been applied to manage medical data, is a family of statistical methods that are useful to study the time of events. Paul Allison (2004) defines an event as a qualitative change that can be situated in time. Qualitative change means a transition from one discrete state to another. A marriage, for example, is a transition from the state of being unmarried to the state of being married. Some researchers also define events as occurring when a quantitative variable crosses a threshold (Paul Allison, 2004). Survival time can be definited as the time to the occurrence of a given event (Lee E., Wang J., 2003). For survival analysis, the best observation plan is prospective. You begin observing a set of individuals at some well-defined point in time, and you follow them for some substantial period of time, recording the times at which the events of interest occur. It's not

necessary that every individual experiences the event (Allison, 2004).
Survival data have two features that are difficult to manage with conventional statistical methods: censoring and time-dependent covariates. Survival analysis have three function: density, survival and hazard.

There are three methods to analyze survival data: nonparametric, semi-parametric and parametric. In this work we have used the first two.

Nonparametric methods are useful to implement exploratory data analysis. In particular, these methods provide function survival estimates, survival and risk graphs. However, this methods are less efficient than parametric methods when survival times follow a theoretical distribution and they are more efficient when no suitable theoretical distributions are known. Nonparametric methods include two kind of estimates of the survival function: product-limit and life-table. They are essentially the same. The only difference is that the product-limit estimate is based on individual survival times, whereas in the life-table method survival times are grouped into intervals. The product-limit estimate can be considered as a special case of the life-table estimate where each interval contains only one observation.
If some subjects in the study have not experienced the event of interest at the end of the study then they are called censored observations and they can also occur when people are lost to follow-up after a period of study. For instance, some customers may still be alive (or churn-free) at the end of the study period. The exact survival times of these customers are unknown.

There are three types of right censoring.

Right censoring - Type I

In customers studies, survival times, that has recorded for the customers that churn during the study period, are the times from the start of the experiment to their churn. These are called uncensored observations. The survival times of the customer are not known exactly but are recorded as at least the length of the study period. These are called censored observations. In this type of censoring, if there are no accidental losses, all censored observations equal the length of the study period (Lee, Wang, 2003).

Right censoring - Type II

Another option in customers studies is to wait until a fixed percent of the customers have churned. In type II censoring, if there are no accidental losses, the censored observations equal the largest uncensored observation (Lee E., Wang J., 2003).

Right censoring - Type III (random censoring)

In most marketing studies the observation period is fixed and customers enter in the study at distinct times during the period. Some customers may churn before the end of the study and then their exact survival times are known. Others may withdraw before the end of the study and are lost to follow-up. Still others may be alive at the end of the study. For "lost" customers, survival times are at least from their entrance to the last contact (censored observation). For customers still alive, survival times are at least from entry to the end of the study (censored observation). Since the entry times are not simultaneous, the censored times are also different.

Left censoring

It occurs when it is known that the event of interest occurred prior to a certain time t, but the exact time of occurrence is unknown. For example, an epidemiologist wishes to know the age at diagnosis in a follow-up study of diabetic retinopathy. At the time of the examination, a 50-year-old participant was found to have already developed retinopathy, but there is no record of the exact time at which initial evidence was found. Thus the age at examination (i.e., 50) is a left-censored observation. It means that the age of diagnosis for this patient is at most 50 years (Lee, Wang, 2003).

Interval censoring

It occurs when the event of interest is known to have occurred between times a and b. For example, if medical records indicate that at age 45, the patient in the example above did not have retinopathy, his age at diagnosis is between 45 and 50 years (Lee, Wang, 2003).

Now we describe nonparametric and semiparametric methods.

Nonparametric methods

Nonparametric methods are very easy to apply. These methods are less efficient than parametric methods when survival times follow a know distribution. Then, we suggest using nonparametric methods to analyze survival data before attempting to fit a know distribution and to implement a univariate analysis.

Kaplan-Meier estimate is applicable to small and large sample but, if the data have already been grouped into intervals, or the sample size is very large (say in the thousands) or the interest is in a large population, it may be more convenient to perform a life-table analysis.

Many authors use the term life-table estimates for the product-limit (PL) estimates. The only difference is that the PL estimate is based on individual survival times, whereas in the life-table method, survival times are grouped into intervals (Lee E., Wang J., 2003).

$$\hat{S} = \prod_{t_r \leq t} \frac{n-r}{n-r+1} \tag{4}$$

where

- n, total number of individuals.

- r, are runs through those positive integer for which $t_r \leq t$ and t_r is uncensored. The values of r are consecutive integers 1,2, ..., n if there are no censored observation, else they are not.
- $t_1 \leq t_2 \leq ... \leq t_n$, are survival time in ascending order of magnitude.
- $(n-r)/(n-r+1)$, for every uncensored observation t_i, give the proportion of individual surviving up to and then through t_i.

Semiparametric methods

In statistics a semiparametric model has parametric and nonparametric components. Usually, the exact form of the underlying survival distribution is unknown. Therefore, the use of parametric methods is somewhat limited and then is better to use Cox model that not require knowledge of the underlying distribution. In this model, the hazard function can assume any form but the hazard functions of different individuals are assumed to be proportional and time-independent. The likelihood function is replaced by the partial likelihood function but the statistical inference based on the partial likelihood function is similar to that based on the likelihood function.
The Cox model possesses the property that different individuals have hazard functions that are proportional, i.e. $h(t|x_1)/h(t|x_2)$, the ratio of the hazard functions for two individuals with prognostic factors or covariates $x_1 = (x_{11}, x_{21}, ..., x_{p1})$ and $x_1 = (x_{12}, x_{22}, ..., x_{p2})$ is a constant. This means that the ratio of the risk of dying of two individuals is the same no matter how long they survive (Lee E., Wang J., 2003).
The hazard function is:

$$h(t|x) = h_0(t)exp\left(\sum_{j=1}^{p} b_j x_j\right) = h_0(t)exp(b^{'}\mathbf{x}) \quad (5)$$

and this formula is equivalent to:

$$S(t|x) = [S_0(t)]^{exp(b^{'}\mathbf{x})} \quad (6)$$

The regression model is:

$$\log \frac{h_i(t)}{h_0(t)} = b_1 x_{1i} + b_2 x_{2i} + \ldots + b_p x_{pi} \quad (7)$$

If you want to deepening these methods you can consult the book "Statistical Methods for Survival Data Analysis" (Lee E., Wang J., 2003).

3 Recent literature and future developments

Since 2000, customer management (CM) research has evolved and has had a significant impact on the marketing discipline. In an increasingly networked society where customers can interact easily with other customers and firms through social networks and other new media, the customer engagement is an important new development in CM. Customer engagement is considered as a behavior manifestation toward the brand or firm that goes beyond transactions (Verhoef et al., 2010). The conceptual shift from a product-centric to a customer-centric organization has been a topic for discussion for more than a decade (Webster, 1992).
Companies are looking for strategies to manage non-transactional behaviors. The customer engagement consists of multiple behaviors such as word of mouth, customer feedback, etc.

Kumar, Lemon and Parasuraman (2006) identified three kind of challenges that represent strategic aspects of firm decision making. First, the advent of the Web (and the proliferation of distribution channels) has added to the complexity of managing customers across multiple channels. Second, it is often said that it is important for firms to be "customer-centric." However, the process by which firms move toward a customer- centric approach is somewhat ill defined. Thus, the second challenge was to address the issue of how best (and under what conditions) can firms become customer-centric. Third, firms now recognize the need to manage their customers as a critical asset while at the same time managing their brands as critical assets as well. Thus, the third challenge was to investigate the relationship between brand equity and customer engagement.
Aeron H. et al (2008), said that a credit card issuer firm has to take several different decisions regarding a customer like his or her stay with the firm. CLV estimation can help a firm in making some of these crucial decisions. They have presented a conceptual model for revenue from a credit card customer and have further presented a metric for CLV. This metric has been designed specifically for credit card customers. They have simulated different states of a customer to demonstrate how the proposed metric works. In conclusion, the biggest advantage of the proposed metric is that it is represented as a function of components that can be estimated from historical data and analysed both individually and collectively to get an idea of cardholder behaviour. For instance, the transaction amount, the probability of existence in a state at time (t) and the percentage of amount revolved are values that can be estimated from the available data. With the help of this metric, observations regarding customer behavior can be made which can help a bank in making various decisions regarding acquisition, retention and personalization of cards.
Jen L. et al (2009), in their paper writed that the analysis of customer value in direct marketing typically combines customer timing and quantity data into a single statistic that is used to compute lifetime values, to order customers for differential actions and to identify prospects for cross-selling. However, current models assume that purchase timing and quantity decisions are independently realized (i.e., uncorrelated) over time given individual-level parameters. The authors show

that customer value calculations can be severely biased in these models when timing and quantity are dependently related. The authors propose alternative models that lead to substantial gains in profitability in two direct-marketing data sets. The results indicate that the commonly held assumption of independence leads to an overvaluation of customer value.

Zhao et al (2009), in their article shown that all customers are not equally profitable. In the credit card business, all customers are not equally risky. When a customer misses one payment on a credit card bill, a signal is sent to the credit card company. It is important for the card issuer to interpret the signal and to identify whether the customer is a low-risk one, who will eventually pay back the debt and contribute to the card issuers profits by paying interest on the overdue balance, or a high-risk one, who will not pay back the debt. The issuer can then customize its policies to deal with these different consumer types. Zhao et al (2009) in their article develop a dynamic model for debt repayment behavior of new customers in the credit card market that makes it possible to differentiate between low-risk, delinquent customers and high-risk customers. The authors apply the model to a data set of new consumersmonthly spending and repayment records.

Kumar V. et al (2010), said that customers can interact and create value for firms in a variety of ways. Their article proposes that assessing the value of customers based solely upon their transactions with a firm may not be sufficient, and valuing this engagement correctly is crucial in avoiding undervaluation and overvaluation of customers. They proposed four components of a customers engagement value (CEV) with a firm. The first component is customer lifetime value (customers purchase behavior), the second is customer referral value (as it relates to incentivized referral of new customers), the third is customer influencer value (which includes the customers behavior to influence other customers, that is increasing acquisition, retention, and share of wallet through word of mouth of existing customers as well as prospects), and the fourth is customer knowledge value (the value added to the firm by feedback from the customer). CEV provides a comprehensive framework that can ultimately lead to more efficient marketing strategies that enable higher long-term contribution from the customer. Metrics to measure CEV, future research propositions regarding relationships between the four components of CEV are proposed and marketing strategies that can leverage these relationships suggested.

Gupta et. al (2006), based on the state of modeling tools as reflected in the current academic literature and the needs of the leading edge industry practitioners, identified eleven issues that represent opportunities for future research. They are: Moving beyond the limits of transaction data; Moving from a customer to a portfolio of customers; Reconciling top-down versus bottom-up Measurements, Cost allocations, Developing incentive schemes that encourage globally optimal behavior, Understanding the limits of CLV and CE, Understanding the scope of application, appreciating the limits of our theory-based models, Understanding how to model rare events, Recognizing the dangers of endogeneity, Accounting for network effects.

4 Case Study

In telecommunication firms, customer retention has become even more important than customer acquisition. Many telecommunications companies apply retention strategies along with programs and processes to keep profitable customers providing them tailored products and services. With retention strategies many firms start to include churn reduction as one of their business goals. In particular, telecommunications companies distinguish between which customers stay longer and which ones stay shorter, and then they also distinguish between which customers are highly profitable and which ones are not very profitable.

The aim of this section is to calculate customer lifetime value for each customer using Gupta formula (1) that contains survival analysis methods. This methods allow to estimate survival probability for each client. Then, we have aggregated CLV by italian geographical areas (north, centre, south). The results of this study will be useful for telecommunications firms and their decision makers to develop customer loyalty and to maximize customer lifetime value. Finally, in this section we have proposed a statistical analysis of CLV. Now we present the dataset.

A dataset of 30,379 (81.75% censored) active business italian customers was selected from a telecommunications company. The study period is equal to six months (July 2010-December 2010). The variable DUR represents the duration of the customer within the study period. This duration can be effective (the customer leaves the phone company) or censored (the customer does not leave the phone company). The variable EVENT is used to distinguish observed cases (EVENT=1) from censored cases (EVENT=0).

The dataset includes two type of variables:

- time-independent, as customer information variables (customer code, geographical areas and so on);
- time-dependent, as customer behavior variables (duration of calls, number of calls and value of calls).

Exploratory Data Analysis

We have created a variable to classify the Italian regions in geographical zone (57% north, 27% center, 16% south) and to facilitate the representation by Kaplan-Meier graph.

geo_zone	percent	cum percent
Centre-Italy	27.31	27.31
North-Italy	56.39	83.70
South-Italy	16.30	100.00

Table 1: Customers divided by geographical areas

We have conducted exploratory data analysis to prepare the dataset for the survival analysis using Proc Univariate to pinpoint survival time distribution, missing values and outliers. We have filtred outliers to make more stable the parameters estimates.

We have used Proc Freq and Proc Gchart to verify the customers distribution respect to geographical areas. These procedures have provided following results:

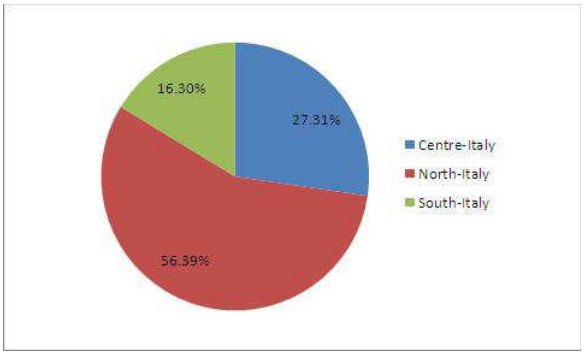

Figure 1: Customers divided by geographical areas

In survival analysis is highly recommended to use Proc Lifetest (discussed in section 5) for all category predictors. Below, we show SAS code.

```
proc lifetest data=tel;
 method=life plot=(s,h) width=1 graphics;
 time dur*event(0);
 strata geo_zone;
run;
```

Using this procedure with Strata statement, we can also produce survival and hazard graphs that are helpful to evaluate the behavior of survival and hazard of different groups. Below, we show these plots stratified by geographical areas.
The survival plot displayed in Figure 2 shows that the customers behavior is the same into several geographical areas. Probably this is due to the short observation period that does not allow to highlight possible changes in customers behavior of different geographical areas.

The hazard plot displayed in Figure 3 shows that the customers risk increases in second month, then suddenly decreases and the risk remains stable until the fifth month. Finally, the risk increases.

Proportionality Assumption Assess

Cox model can be easily extended to allow nonproportional hazards. Time dependent covariates are interactions of the predictors with the time. Whenever you

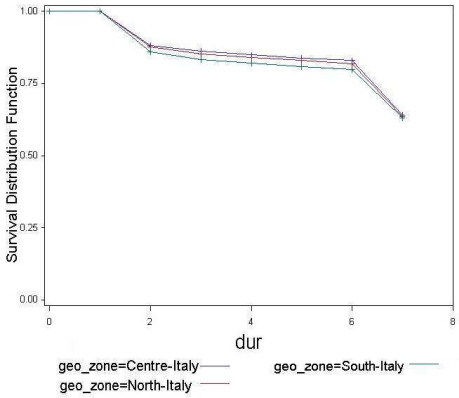

Figure 2: Survival Plot

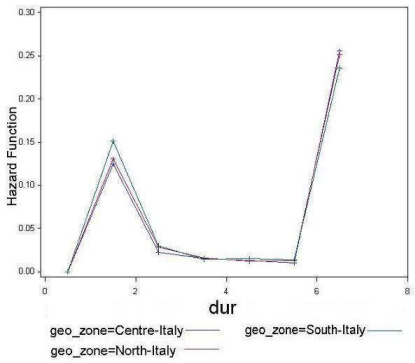

Figure 3: Hazard Plot

introduce time-dependent covariates into a Cox regression model, it's no longer accurate to call it a proportional hazards (PH) model. Why? Because the time-dependent covariates will change at different rates for different individuals, so the ratios of their hazards cannot remain constant (Allison, 2004).
There are numerical and graphical methods to assess the proportionality assumption. In the graphical methods the curves must be approximately parallel but looking the Kaplan-Meier curves is not enough to be certain of proportionality.
Recall that the main purpose of this study is to determine, for each customer, survival probability and customer lifetime value respect to geographical areas. Below, we show SAS code.

```
proc lifetest data=tel;
 method=life plot=(lls) width=1 graphics;
 time dur*event(0);
 strata geo_zone;
run;
```

The plot displayed in Figure 4 shows the log-log survivor functions for each of three geographical areas. The curves are approximately parallel, then we can say

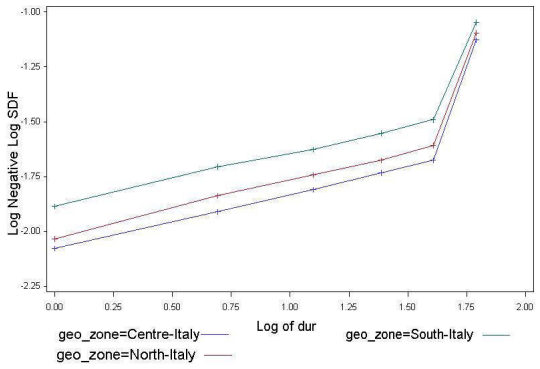

Figure 4: Log-Log Survival Plot

that the hazards respect to geographical areas are proportional.

To confirm this result we can estimate Cox model creating (by Proc Phreg) an interaction variable between geographic zone covariate and time. Then we include this variable in the model. The coefficient for interaction variable is not statistically significant (p-value=0.9583), therefore there is not evidence that the effect of geographical zone changes over time. We can conclude that the PH assumption is not violated for geo_zone variable. PH assumption is respected for the other covariates too.

Cox Model Estimation

We remember that the dataset contains two type of variables: time indipendent variables (customer code, geographical areas and so on) and time dependent variables (duration of calls, number of calls and value of calls). The time dependent variables are collected at regular intervals of time (month).
Typical survival analysis data often take the form of one record per subject. Each of these records is a vector of the sort (T, E) where T has the value of time since the origin and is either the time of an event of the kind being studied, in the case when the indicator variable E takes the value 1, say, or otherwise is a censoring time, in which case E will have the value 0 (Carpenter, Ake, 2003).
On the other hand, data in counting process format may contain more than one record per subject; for any subject with multiple records in the dataset, each such record represents one interval for that subject. Each such record is of the form (T1, T2, E), where T1 represents the time at which the interval started, T2 the time at which the interval ended, and E, as before, is an indicator variable showing the status of the interval. The indicator E could take one value to represent at event occurring at time T2, another to indicate censoring at T2. The actual time interval represented by this record can be represented as (T1,T2], i.e., open on the left, closed on the right, so that the time instant T2 itself is included in the time interval but T1 is not. Thus an event or change in status occurring at T2 would belong to this interval but one occurring at T1 would not; it would belong to the preceding time interval.

Counting process format can easily accommodate a number of special features in one's data, including multiple events of the same type and time-dependent covariates (Carpenter, Ake, 2003).

The coefficients estimates can be obtained using Proc Phreg. Below, we show SAS code.

```
proc phreg data=tel;
  model (entry, exit)*event(0)= &varlist
   /ties=efron selection=s;
run;
```

Entry and Exit are two support variables that represent (T1,T2] and allow to Proc Phreg to pinpont the risk period for each customer.
The following table shows significant covariates and hazard ratios that we have identified by Proc Phreg.

parameter	$\hat{\beta}$	std err	hr
$flag\_preact$	-0.25905	0.03074	0.772
$flag\_bundle$	0.21588	0.08299	1.241
num_calls_out	-0.0301	0.0026	0.970
val_calls_out	1.3417	0.0238	0.261
$firm\_size$	-0.0408	0.0171	0.960

Table 2: Significant Covariates and Hazard Ratios

Now, we describe the significant variables shown in Table 2: flag_preact, indicates if the customer has made the preactivation (i.e. the customer declares its intention to terminate the contract that has with another tlc company and to establish a contractual relationship with new tlc company together with the pre-activation component of the speech in automatic mode selection); flag_bundle, indicates if the customer has the full package (i.e., this package includes several combinations of fixed telephone, mobile and data); num_calls_out, represents the number of total calls from fixed and mobile; val_calls_out, represents the value of total calls from fixed and mobile; firm_size, represents the size of the company compared to the number of seats and number of products.
The hazard ratio represent the risk increase of customer churn per unit increase in the covariates. For example, the estimated risk ratio for the (dummy) variable flag_preact is 0.772. This means that the hazard of chrurn for those who stipulate a contract with preactivation is only 77% of the hazard for those who stipulated a contract without pre-activaction. For (quantitative) covariate val_calls_out the risk ratio is 0.261, which yealds $100(0.261 - 1) = -73.9\%$. Therefore, for each one-euro increase in the consumption of telephone traffic the customer churn goes down by an estimated 73.9%.

Selection of variables have excluded some covariates like: geographical zone, international calls, internet calls from mobile, calls to national mobile, number of sms, call to number of other companies and so on.

Estimation of Customer Lifetime Value

Here we show a case related to particular data (on semester basis). The real aim, however, is not having a definitive result on the specific case but to illustrate the method of calculation and analysis of the CLV. In this section we discuss the CLV in terms of market distribution and show some potentialities of CLV in market strategy analysis. In doing this, we just start a simple exemplifying analysis and discuss preliminary results. A deeper analysis in terms of multivariate analysis and clustering is outside the scope of this paper and will be the subject of a subsequent work.
Assuming that the discount rate is 5%, we can estimate customer lifetime value (i.e., the net present value of the cash flows attributed to the relationship with a customer over time) for each customer applying formula (1). For example: customer1 CLV= €1,269.70, customer2 CLV= €763.34, and so on.

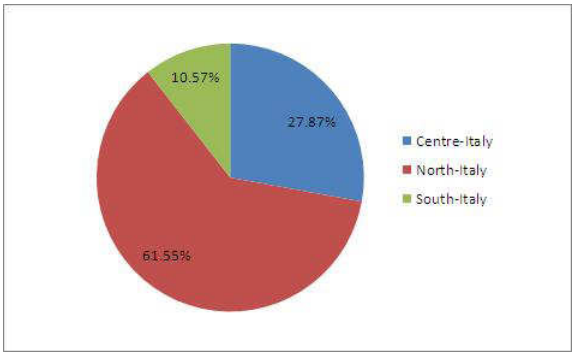

Figure 5: CLV divided by geographical areas

The plot displayed in Figure 5 shows that that most of the CLV come from customers of North-Italy, so the company needs to implement geographical policies to avoid their churn.
Respect to percentiles of CLV we have shared it in three bands:

- band 1, CLV<€1,250;
- band 2, €1,250≤CLV<€3,700;
- band 3, CLV>€3,700.

Thanks to Table 3 it is easy to note that about 95% of customers belong to band 1.
To confirm this, by Gini Index and Lorenz Curve we can see the concentration of CLV. In particular, Gini Index=0.94. This demonstrates that the CLV is very concentrated in few groups of customers. Table 3 shows that this groups represent the customers less profitable for the telephone company.

quantile	estimate (€)
100% Max	81,375
99%	3,670
95%	1,244
90%	750
75% Q3	323
50% Median	111
25% Q1	20
10% Q1	1.24
5% Q1	0.13

Table 3: Percentiles of CLV

Moreover, in Table 3 we can see that the clientele profitability is very much spread between a large audience of custemer with a small six month CLV, the median being €1,113 and a small fraction of customers characterized by high profitability, in particular, we have 1% of customers giving a CLV between €3,670 and €81,375.
In Figure 6 we report Lorenz Curve that confirms that 95% of customers produces 50% of CLV and then the second 50% of CLV is created by 5% of customers.

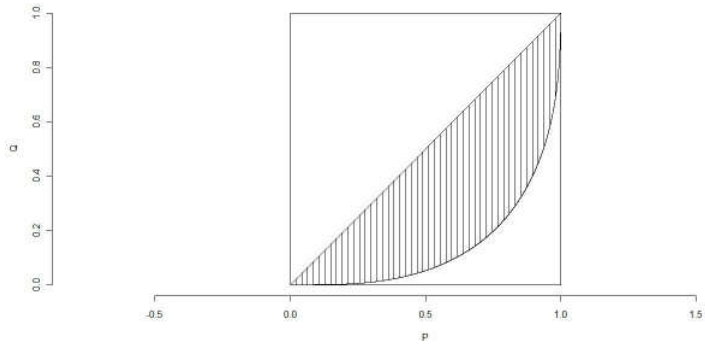

Figure 6: CLV concentration

Table 4 suggests that the strategy of telephone company is oriented, or more effective, towards small businesses (mono-seats) which may be a good way to penetrate the market and then expand to larger companies (pluri-seats).

	CLV		
Firm Size	band1	band2	band3
mono-seats	78.60%	1.99%	0.29%
pluri-seats	16.30%	2.08%	0.73%

Table 4: CLV and Firm Size

By Proc Reg we have assessed how the variable of telephone company dataset influence CLV. In particular we are interested to check the contribution to CLV brought from traffic value generated by fixed and mobile.
Table 5 shows the results of regression. The data are adjusted so well by the model ($R^2 = 0.92$). From

variable	$\hat{\beta}$	std err
intercept	345.06	1.82
val_calls_out_from_mobile	1060.21	1.82
val_calls_out_from_fixed	527.33	2.45
num_minutes_from_fixed	75.12	1.96

Table 5: Covariates that influence CLV

point of view of marketing is known that the CLV of phone company customers depends significantly by two strategic variables: val_calls_out_from_mobile, val_calls_out_from_fixed. But, the very interesting result is the contribute apported to CLV from mobile value: it is about twice respect to the value of calls made from fixed phones. Then, because the covariates are standardized, for each sigma unit of value of mobile or fixed the CLV increases respectively about €1,000 or €500. Therefore, the telephone company seems to aim in particular on mobile as own core business. The competitors analysis could lead to the phone company to be more competitive respect to calls from fixed. So, the managers of company must be careful when they select new customers and when they manage the big customers. For this type of customers the company should implement policy of incentives so that the customer remains with the company.

Finally, using the nonparametric test of Kruskal-Wallis (because the CLV distribution is not normal and variances of groups are not equal) we can evaluate whether there are, on average, differences of CLV in firm size. We remember that firm size has been stratified in four levels: mono-seats and mono-product (level 1), mono-seats and pluri-product (level 2), pluri-seats and mono-product (level 3), pluri-seats and pluri-product (level 4). Amalgamating the customers with mono-seat (level 1 and level 2) and the ones with pluri-seat (level 3 and level 4) it is possible to note that the second (big firms) have activated fewer products Bundle and Mobile than the first.

firm size	N	Mean
mono-seat	24,636	245.60
pluri-seat	5,743	812.95

Table 6: CLV differences in firm size

Table 6 shows the ANOVA results. There is, on average, a statistically significant difference (p-value $<$ .0001) between the CLV of small size companies (mono-seats) and big size companies (pluri-seats). Therefore, the telephone company, through the study of own competitors, could to invest more on customer that have pluri-seats for example designing a dedicated tariff plan.

5 Software

To analyze data and manage dataset we have used SAS. In this section we focus on three procedures that are useful for survival analysis:

Proc LIFETEST

LIFETEST procedure can be used to calculate nonparametric estimates of the survival function either by the product-limit method (also called the Kaplan-Meier method) or by the life-table method. It is designed for univariate analysis and it provides estimates of survivor functions using two methods: Kaplan-Meier method (most suitable for smaller data sets with precisely measured event times) and life-table or actuarial method (most suitable for large data sets or when the measurements of event times are imprecise).
LIFETEST procedure provides survival and hazard graphics (option Plots).
LIFETEST procedure provides two methods (Log-Rank and Wilcoxon) to test the null hypothesis that the survival functions are identical for two or more groups (option Strata).
PROC LIFETEST procedure also tests for associations between event times and time-constant covariates but it does not produce estimates of parameters.

Proc PHREG

PHREG procedure uses Cox's partial likelihood method to estimate regression models with censored data. This method does not require that you choose some particular probability distribution to represent survival times. Cox regression allows to incorporate time-dependent covariates, to adjust for periods of time in which an individual is not at risk of an event and accommodate both discrete and continuous measurement of event times. PROC PHREG incorporates two exact algorithms for tied data.

Proc LIFEREG

LIFEREG procedure produces estimates of parametric regression models with censored survival data (under different alternative distributional assumptions) using the maximum likelihood method. In recent years, PROC LIFEREG has been obscured by the PHREG procedure. PROC LIFEREG allows for several varieties of censoring, but it does not allow for time-dependent covariates. PROC LIFEREG tests hypotheses about hazard function shape but it does not manage time-dependent covariates.

In this case study (with time-dependent covariates) we has used Proc Lifetest to implement univariate analysis and Proc Phreg to estimate the hazard model because Proc Lifereg does not handle time-dependent covariates.

6 Conclusion

This paper presents an application that allows to model customer lifetime value using survival analysis methods. This techniques can help firms and decision makers to develop customer loyalty strategies and to maximize customer lifetime value.
Thanks to Gupta's formula (1) we have calculated CLV for each customer taking into account their survival probabilities. It was noticed that the customers of Northern Italy carry most of the income of the telephone company than those of other geographical areas. So, the telephone company must monitoring these customers in order to avoid their churn.
In particular, the contribute apported to CLV from mobile value is about twice respect to the value of calls made from fixed phones. Therefore, the telephone company seems to aim in particular on mobile as own core business. This result is very important to determinate potential decision of marketing. For example, the phone company could allocate part of the gains from mobile phones to become more competitive in the tariffes of fixed telephony.

In conclusion, a decision maker can use the method proposed in this paper:

- to estimate survival probability for each customer;
- to calculate CLV for each customer;
- to identify the variables that significantly influence CLV.

Acknowledgements

We thank Nunatac Srl for supporting the research and for providing the data of telecommunication company and SAS software.

References

[1] Aeron H., Bhaskar T., Sundararajan R., Kumar A., Moorthy J. (2008), "A metric for customer lifetime value of credit card customers", Journal of Database Marketing & Customer Strategy Management,15, 153-168, DOI: 10.1057/dbm.2008.13.

[2] Allison P. (2004), "*Survival Analysis Using SAS. A Pratical Guide*", SAS Publising.

[3] Ata N., Sozer M. (2007), "Cox Regression Models with Nonproportional Hazards applied to Lung Cancer Survival Data", Hacettepe Journal of Mathematics and Statistics, 36(2), 157-167.

[4] Bijmolt T., Leeflang P., Block F., Eisenbeiss M., Hardie B., Lemmens A. and Saffert P. (2010), "Analytics for Customer Engagement", Journal of Service Research, 13 (3), 341-356, DOI: 10.1177/1094670510375603.

[5] Carpenter A., Ake C. (2003), "Exending the Use of Proc Phreg in Survival Analysis", Proceedings of the 11th Annual Western Users of SAS Software, Inc. Users Group Conference, Cary, NC: SAS Institute Inc.

[6] Chen Y., Zhang H., Zhu P. (2009), "Study of Customer Lifetime Value Model Based on Survival-Analysis Methods", Computer Science and Information Engineering, 2009 WRI World Congress on, 20, 266 - 270, DOI: 10.1109/CSIE.2009.313.

[7] Firebaugh G. (1999). "Empirics of World Income Inequality", American Journal of Sociology, 104 (6), 1597–1630, DOI:10.1086/210218.

[8] Gupta S., Lehmann D. (2003), "Customer as Assets", Journal of Interactive Marketing, 17(1), 9-24, DOI: 10.1002/dir.10045.

[9] Gupta S., Hanssens D., Hardie B., Kahn W., Kumar V., Lin N., Sriram N. (2006), "Modeling Customer Lifetime Value", Journal of Service Research, 9(2), 139-155, DOI: 10.1177/1094670506293810.

[10] Jen L., Chou C., Allenby G. (2009), "The Importance of Modeling Temporal Dependence of Timing and Quantity in Direct Marketing", Journal of Marketing Research, 46(4), 482-493, ISSN: 1547-7193.

[11] Kleinbaum D., Klein M. (2005), "*Survival Analysis - A Self-Learning Text*", Springer.

[12] Kumar V., Katherine N. Lemon, A. Parasuraman (2006), "Managing Customers for Value: An Overview and Research Agenda", Journal of Service Research, 9, 87-94, DOI: 10.1177/1094670506293558.

[13] Kumar V., Aksoy L., Donkers B., Venkatesan R., Wiesel T. and Tillmanns S. (2010), "Undervalued or Overvalued Customers: Capturing Total Customer Engagement Value", Journal of Service Research, 13 (3), 297-310, DOI: 10.1177/1094670510375602.

[14] Lee E., Wang J. (2003), "*Statistical Methods for Survival Data Analysis*", Wiley.

[15] Winer R. (2001), "A Framework for Customer Relationship Management", California Management Review,43(4), 89-105.

[16] Zhao Yi, Zhao Yang, Song I. (2009), "Predicting New Customers' Risk Type in the Credit Card Market", Journal of Marketing Research, 46(4), 506-517, DOI: 10.1509/jmkr.46.4.506.