Paper 120-28

# Modeling Customer Lifetime Value Using Survival Analysis – An Application in the Telecommunications Industry

**Junxiang Lu, Ph.D.**
**Overland Park, Kansas**

## ABSTRACT

Increasingly, companies are viewing customers in terms of their lifetime value – the net present value of customers calculated profit over a certain number of months. Customer lifetime value is a powerful and straightforward measure that synthesizes customer profitability and churn (attrition) risk at individual customer level. For existing customers, customer lifetime value can help companies develop customer loyalty and treatment strategies to maximize customer value. For newly acquired customers, customer lifetime value can help companies develop strategies to grow the right customers.

The calculation of customer lifetime value varies across industries. In the telecommunications industry, customer monthly margin and customer survival curve are the two major components to calculate the customer lifetime value. Since customer monthly margin is from accounting models, the key to estimate customer lifetime value is the customer survival curve. In this study, survival analysis is applied to estimate customer survival curve, therefore customer lifetime value is calculated.

## INTRODUCTION

In the telecommunications industry, customers are able to choose among multiple service providers and actively exercise their rights of switching from one service provider to another. In this fiercely competitive market, customers demand tailored products and better services at fewer prices, while service providers constantly focus on acquisitions as their business goals. Given the fact that the telecommunications industry experiences an average of 30-35 percent annual churn rate and it costs 5-10 times more to recruit a new customer than to retain an existing one, customer retention has now become even more important than customer acquisition. For many incumbent operators, retaining highly profitable customers is the number one business pain. Many telecommunications companies deploy retention strategies in synchronizing programs and processes to keep customers longer by providing them with tailored products and services. With retention strategies in place, many companies start to include churn reduction as one of their business goals.

With the telecommunications market getting more and more mature, telecommunications companies do not satisfy themselves with predicting customer churn; they instead start viewing customers in terms of customer lifetime value. Not only do the telecommunications companies distinguish between which customers stay longer and which ones stay shorter, they also distinguish between which customers are highly profitable and which ones are low profitable or not profitable. Customer lifetime value is therefore developed to satisfy telecommunications companies need to evaluate their customer value.

Conventional statistical methods (e.g. logistics regression, decision tree, and etc.) are very successful in predicting customer survival/churn. These methods could hardly predict when customers will churn, or how long the customers will stay with. However, survival analysis was, at the very beginning, designed to handle survival data, and therefore is an efficient and powerful tool to predict customer survival/churn.

The goal of this study is to calculate customer lifetime value through estimating customer survival curve using survival analysis techniques. The results from this study are helpful for telecommunications companies to develop customer loyalty and treatment strategies to maximize customer value. It is also useful for telecommunications companies to develop strategies to grow the right customers.

## OBJECTIVES

The objectives of this study are in two folds. The first objective is to develop the concept of customer lifetime value in the telecommunications industry. The second one is to demonstrate how survival analysis techniques are used to estimate customer lifetime value.

## DEFINITIONS AND EXCLUSIONS

This section clarifies some of the important concepts and exclusions used in this study.

**Survival/Churn** – In the telecommunications industry, the broad definition of churn is the action that a

customer's telecommunications service is cancelled. In this study, both service-provider initiated churn and customer initiated churn are included. An example of service-provider initiated churn is a customer's account being closed because of payment default. Customer initiated churn is more complicated and reasons behind vary from customer to customer. Customer survival is the opposite of customer churn, and both terms are used in the study.

**Active** – Active is a customer status. Customers whose service being involuntarily terminated and are in collection stage are not in "active" status.

**Granularity** – This study examines customer survival/churn at the account level.

**Customer Contract** – This study does not distinguish customers with or without contracts, although separate models may be desirable for each contract status.

**Exclusions** – This study does not include employee accounts.

## CUSTOMER LIFETIME VALUE

The calculation of customer lifetime value (LTV) varies across industries. In telecommunications industry, customer monthly margin and customer survival curve are the two major components of customer lifetime value. The customer lifetime value is the net present value of customers calculated profit over a certain number of months. Here is the formula to calculate customer lifetime value:

$$LTV = MM \times \Sigma^{T}_{i=1} \left( p^{i} / (1 + r/12)^{i-1} \right)$$

Where ***MM*** is the monthly margin for the last three months for existing customers, or the last month's monthly margin for newly acquired customer. ***MM*** is either calculated from accounting models or estimated through a set of regression models. The calculation of monthly margin is not the focus of this study and therefore not covered. ***T*** is the number of months in consideration to calculate customer lifetime value; it could be 24, 36, or some other number that makes the most business sense. ***r*** is the discount rate. $p^{i}$ is the series of customer survival probabilities (customer survival curve) from month 1 through Month ***T***, where $p^{1} = 1$. $p^{i}$ is estimated through customer survival model.

## SURVIVAL ANALYSIS AND CUSTOMER SURVIVAL/CHURN

Survival analysis is a clan of statistical methods for studying the occurrence and timing of events. From the beginning, survival analysis was designed for longitudinal data on the occurrence of events. Keeping track of customer churn is a good example of survival data. Survival data have two common features that are difficult to handle with conventional statistical methods: *censoring* and *time-dependent covariates*.

Generally, survival function and hazard function are used to describe the status of customer survival during the tenure of observation. The survival function gives the probability of surviving beyond a certain time point *t*. However, the hazard function describes the risk of event (in this case, customer churn) in an interval time after time *t*, conditional on the customer already survived to time *t*. Therefore the hazard function is more intuitive to use in survival analysis because it attempts to quantify the instantaneous risk that customer churn will take place at time *t* given that the customer already survived to time *t*.

For survival analysis, the best observation plan is prospective. We begin observing a set of customers at some well-defined point of time (called the *origin of time*) and then follow them for some substantial period of time, recording the times at which customer churns occur. It's not necessary that every customer experience churn (customers who are yet to experience churn are called *censored* cases, while those customers who already churned are called *observed* cases). Not only do we predict the timing of customer churn, we also want to analyze how *time-dependent covariates* (e.g. customers calls to service centers, customers change plan types, customers change billing options, and etc.) impact the occurrence and timing of customer survival/churn.

SAS/STAT® has two procedures for survival analysis: PROC LIFEREG and PROC PHREG. The LIFEREG procedure produces parametric regression models with censored survival data using maximum likelihood estimation. The PHREG procedure is a semi-parametric regression analysis using partial likelihood estimation. PROC PHREG gained popularity over PROC LIFEREG in the last decade since it handles time-dependent covariates. However if the shapes of survival distribution and hazard function are known, PROC LIFEREG produces more efficient estimates (with smaller standard error) than PROC PHREG does.

## SAMPLING STRATEGY

A sample of 64,320 active customers was randomly selected from the entire customer base from a telecommunications company. All these customers were active on January 16, 2001 and their survival/churn behaviors were followed for the next 20 months. Therefore January 16, 2001 is the *origin of time* and September 15, 2002 is the *observation termination time*.

During this 20-month observation period, the timing of customer churn was recorded. For each customer in the sample, a variable of DUR is used to indicate the time that customer churn occurred, or for *censored cases*, the last time at which customers were observed, both measured from the *origin of time* (January 16, 2001). A second variable, STATUS, is used to distinguish the *censored cases* from *observed cases*. It is common to have STATUS = 1 for *observed cases* and STATUS = 0 for *censored cases*. In this study, the survival data are *singly right censored* so that all the *censored cases* have a value of 21 (months) for the variable DUR.

## DATA SOURCES

There are four major data sources for this study: census-block level marketing and financial information, customer level demographic data, customer internal data, and customer contact records. The first two data sources are from third party vendors. A brief description of some of the data follows.

**Customer Internal Data** – Customer internal data is from the company's data warehouse. It consists of two parts. The first part is about customer information like market channel, plan type, payment, bill type, contract status, customer segmentation code, ownership of the company's other products, dispute, late fee charge, discount, promotion/save promotion, additional accounts, rewards redemption, billing dispute, and so on. The second part of customer internal data is customer's telecommunications usage data. Examples of customer usage variables are:

- Weekly average call counts
- Average length of calls
- Call percentage change from one time period to another
- Share of peak/off-peak minutes
- Using of add-on products

**Customer Contact Records** – The company keeps track of detailed records of customer contact. This includes customer calls to service centers, customer calls to the IVR system, customers' web access to their accounts, and the company's mail contacts to customers. The customer contact records are then classified into customer contact categories. Among the customer contact categories are customer general inquiry, customer requests to change service, customer inquiry about cancel, and so on.

## MODELING PROCESS

The modeling process includes the following five major steps.

**Exploratory Data Analysis (EDA)** – Exploratory data analysis was conducted to prepare the data for the survival analysis. The univariate frequency analysis was used to pinpoint value distributions, missing values and outliers. Variable transformation was conducted for some necessary numerical variables to reduce the level of skewness, because transformations are helpful to improve the fit of a model to the data. Outliers are filtered to exclude observations, such as outliers or other extreme values that are suggested not to be included in the data mining analysis. Filtering extreme values from the training data tends to produce better models because the parameter estimates are more stable. Variables with missing values are not severe, except some of the demographic variables. The demographic variables with more than 25% of missing values were replaced with their missing indicators. For observations with missing values, one choice is to use incomplete observations, but that may lead to ignore useful information from the variables that have nonmissing values. It may also bias the sample since observations that have missing values may have other things in common as well. Therefore, in this study, missing values were replaced by appropriate methods. For interval variables, replacement values were calculated based on the random percentiles of the variable's distribution, i.e., values were assigned based on the probability distribution of the nonmissing observations. Missing values for class variables were replaced with the most frequent values (count or mode).

**Variable Reduction –** Started with 328 variables in the original data set, by using PROC FREQ, an initial univariate analysis of all categorical variables crossed with customer churn status (STATUS) was carried out to determine the statistically significant exploratory variables to be included in the next modeling step. All the categorical variables with a chi-square value or t statistics significant at 0.05 or less were kept. This step reduced the number of variables to 225 (&VARLIST1) – including all the numerical variables and the kept categorical variables from the step one.

The next step is to use PROC PHREG to further reduce the number of variables. A stepwise selection method was used to create a final model with statistically significant effects of 42 independent variables on customer churn over time.

```
PROC PHREG DATA = SASOUT2.ALL2 OUTEST =
SASOUT2.BETA;
        MODEL DUR*STATUS(0) = &VARLIST1
        / SELECTION = STEPWISE SLENTRY = 0.0025 SLSTAY
        = 0.0025 DETAILS;
RUN;
```

**Model Estimation –** With only 42 independent variables, the final data set has reasonable number of variables to perform survival analysis regression. Before applying survival analysis procedures to the final data set, the customer survival function and hazard function were estimated using the following code. The purpose of estimating customer survival function and customer hazard function (Figures 1 and 2) is to gain knowledge of

customer survival/churn hazard characteristics. From the shape of hazard function, customer survival/churn in this study demonstrates a typical hazard function of a Log-Normal type model. As previously discussed, since the shapes of survival distribution and hazard function were known beforehand, PROC LIFEREG is expected to produce more efficient estimates (with smaller standard error) than PROC PHREG does.

```
PROC LIFETEST DATA = SASOUT2.ALL3 OUTSURV =
SASOUT2.OUTSURV
        METHOD = LIFE PLOT = (S, H) WIDTH = 1 GRAPHICS;
        TIME DUR*STATUS(0);
RUN;
```

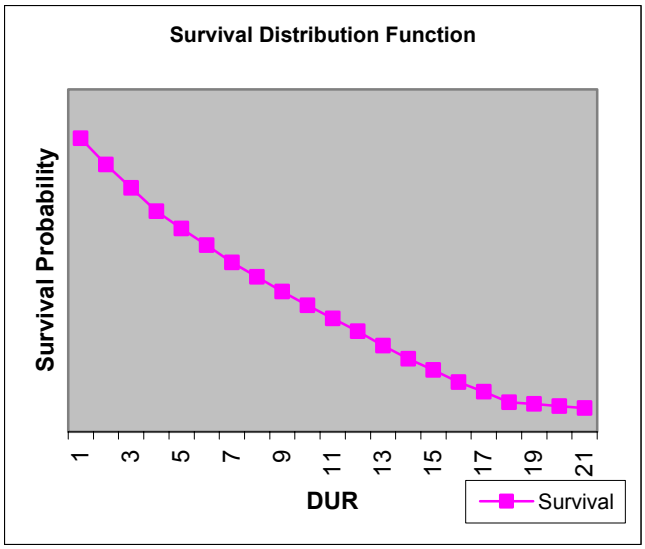


Figure 1. Customer Survival Function

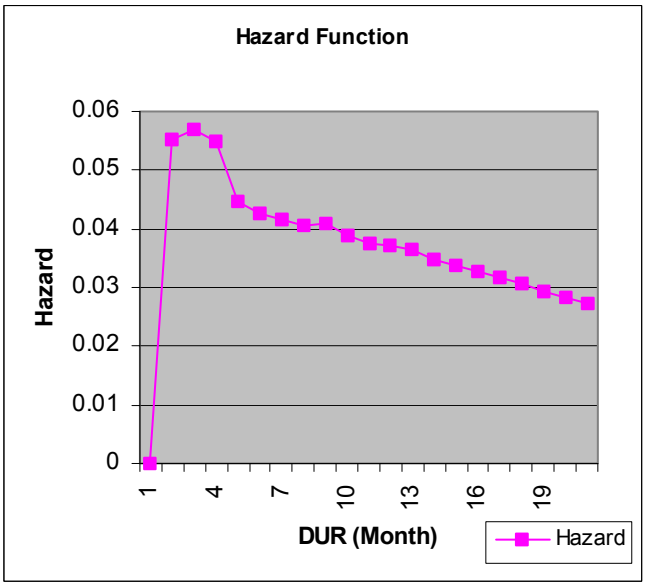


Figure 2. Customer Hazard Function

The Final step is to estimate customer survival/churn. PROC LIFEREG was used to calculate customer survival probability. At this step the final data set was divided 50/50 into two data sets: model data set and validation data set. The model data set is used to fit the model and the validation data set is used to score the survival probability for each customer. A variable of USE is used to distinguish the model data set (set USE = 0) and validation data set (set USE = 1). In the validation data set, set both DUR and STATUS missing so that cases in the validation data set were not to be used in model estimation. The following code is to prepare the model data set and validation data set.

```
IF RANUNI(0) < 0.5 THEN OUTPUT MODEL;
        ELSE OUTPUT VALID;

DATA SASOUT2._MODEL;
        SET MODEL;
        USE = 0;
DATA SASOUT2._VALID;
        SET VALID;
        STATUS = .;
        DUR = .;
        USE = 1;

DATA SASOUT2.APPEND;
        SET SASOUT2._MODEL
            SASOUT2._VALID;
```

By appending the validation data set to the model data set as above, the following SAS code is to score each customer's survival probability in the validation data set. The PREDICT macro produces predicted survival probabilities for each customer for a specified time (for this case, number of months after the *origin of time*), based on the model fitted by PROC LIFEREG. Refer Allison (1995) for details of this macro.

PROC PHREG was also used to calculate the predicted survival probabilities. Comparing with those from PROC LIFERE, PHREG procedure does not provide as good performance as LIFEREG procedure does.

```
%MACRO PREDICT (OUTEST=, OUT=_LAST_,XBETA=,TIME=);
DATA _PRED_;
_P_=1;
SET &OUTEST (KEEP=_DIST_ _SCALE_ _SHAPE1_ ) POINT=_P_;
SET &OUT;
LP=&XBETA;
T=&TIME;
GAMMA=1/_SCALE_;
ALPHA=EXP(-LP*GAMMA);
PROB=0;
IF _DIST_='EXPONENT' OR _DIST_='WEIBULL' THEN
PROB=EXP(-ALPHA*T**GAMMA);
IF _DIST_='LNORMAL' THEN PROB=1-PROBNORM((LOG(T)-
LP)/_SCALE_);
IF _DIST_='LLOGISTC' THEN PROB=1/(1+ALPHA*T**GAMMA);
IF _DIST_='GAMMA' THEN DO;
        D=_SHAPE1_;
        K=1/(D*D);
        U=(T*EXP(-LP))**GAMMA;
        PROB=1-PROBGAM(K*U**D,K);
        IF D LT 0 THEN PROB=1-PROB;
END;
DROP LP GAMMA ALPHA _DIST_ _SCALE_ _SHAPE1_ D K U;
RUN;
```

```
%MEND PREDICT;

PROC LIFEREG DATA=SASOUT2.APPEND OUTEST=A;
        MODEL DUR*STATUS(0) =  &VARLIST1 /
DIST=LNORMAL;
        OUTPUT OUT=B XBETA=LP CONTROL =USE;
RUN;

%PREDICT(OUTEST=A,OUT=B,XBETA=LP,TIME=03)
```

**Model Validation –** With each customer in the validation data set being scored for predicted survival probabilities for specified time periods (that is, number of months since the *origin of time* – January 16, 2001), each customer will have $P_1$ through $P_{21}$ as its predicted survival probabilities. $P_1$ is the predicted survival probability at the beginning of month 1, which is the *origin of time*, is always equal to 1. $P_2$ is the predicted survival probability at end of month 1, $P_3$ is the predicted survival probability at end of month 2, and so on. By sorting customers in descending predicted survival probabilities for a specified time period, if the model works, customers with the highest predicted survival probabilities will have the highest likelihood to survive up to this specified time period.

In order to validate the successfulness of the model, we rank the predicted survival probabilities for a specified time in descending order into deciles and then to compare the number of survived customer up to this specified time period in each decile. For example, we can rank all customers predicted survival probabilities $P_3$'s into deciles and then count the number of customers who survived up to month 3. Figure 3 describes cumulative lift gains for identifying the customers who have survived up to a specified time period. Only selected time periods are reported in the figure. For example, the top three deciles capture about 74% of survivors through month 3, while the top five deciles capture about 92% of survivors through month 3.

**Estimation of Customer Lifetime Value –** With the model successfully estimating customer survival curve (a series of survival probabilities, $P_1$ through $P_{21}$ , for example), customer lifetime value can be estimated for each customer.

Assume an existing customer A's monthly margin (***MM***) for the last three months is $22.50, and the discount rate is 8.5%. From the model, the customer's predicted series of survival probabilities are: 1.000, 0.990, 0.972, 0.958, 0.939, 0.922, 0.906, 0.888, 0.871, 0.858, 0.841, 0.826, 0.812, 0.798, 0.792, 0.787, 0.782, 0.777,0.772, 0.767, 0.762, 0.757, 0.752, 0.747, then this customer's lifetime value is $423.05.

Assume another existing customer B. The customer's monthly margin (***MM***) for the last three months is $45.00, the discount rate is same, and this customer's predicted series of survival probabilities are: 1.000, 0.995, 0.982, 0.012, 0.002, 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0.000. Using the formula described earlier, this customer' lifetime value is $133.65.

Through customer lifetime value, it is clear that customer A has significantly higher value to the company than customer B, even though customer B seems more profitable than customer A.

| | Up to Month 3 | | Up to Month 6 | | Up to Month 9 | |
|---|---|---|---|---|---|---|
| DEC. | # of Sur-vivors | Cum. % | # of Sur-vivors | Cum. % | # of Sur-vivors | Cum. % |
| 1 | -* | 36.3% | - | 30.9% | - | 27.2% |
| 2 | - | 59.5% | - | 55.8% | - | 52.4% |
| 3 | - | 74.2% | - | 72.2% | - | 69.9% |
| 4 | - | 84.7% | - | 83.1% | - | 81.5% |
| 5 | - | 92.2% | - | 91.5% | - | 89.9% |
| 6 | - | 96.3% | - | 96.0% | - | 94.8% |
| 7 | - | 98.5% | - | 98.4% | - | 97.6% |
| 8 | - | 99.2% | - | 99.2% | - | 99.1% |
| 9 | - | 99.6% | - | 99.6% | - | 99.4% |
| 10 | - | 100% | - | 100% | - | 100% |

Figure 3, Model Lift Gains for Identifying Survived Customers for Selected Time Periods (*Number of survived customers are intentionally disguised to protect company proprietary information).

## IMPLEMENTATION

Customer lifetime value can be implemented in multiple ways. It could be used as a customer segmentation tool to segment customer value. It could also be used to segment customers by their churn behavior. Customer lifetime value could be used to measure marketing campaign efficiency as well.

The results from the customer survival curve estimation could be used to devise strategies to extend customers expected life span.

## CONCLUSION

This study presents an application to model customer lifetime value – the net present value of customers calculated profit over a certain number of months, using survival analysis techniques. Customer lifetime value is a powerful and straightforward measure that synthesizes customer profitability and churn (attrition) risk at individual customer level. For existing customers, customer lifetime value can help companies develop customer loyalty and treatment strategies to maximize customer value. For newly acquired customers, customer lifetime value can help companies develop strategies to grow the right customers.

Customer lifetime value in this study is essentially based upon customers "single product" value. It can be extended to incorporate cross-sell probabilities to estimate customer lifetime value in a multiple product scenario.

## REFERENCES

Allison, Paul D. *Survival Analysis Using the SAS® System: A Practical Guide,* Cary, NC: SAS Institute Inc. 1995. 292pp.

Hosmer, JR. DW, and Lemeshow S. *Applied Survival Analysis: Regression Modeling of Time to Event Data*. New York: John Wiley & Sons, 1999.

Lu, Junxiang, *Detecting Churn Triggers for Early Life Customers in the Telecommunications Industry – An Applications of Interactive Tree Training*, Proceedings of the 2nd Data Mining Conference of DiaMondSUG 2001, Chicago, IL, 2001.

Lu, Junxiang, *Predicting Customer Churn in the Telecommunications Industry – An Application of Survival Analysis Techniques*, Proceedings of the 27th Annual SAS Users Group International Conference, Cary, NC: SAS Institute Inc., 2002.

Rud, Olivia Parr, *Data Mining Cookbook*, New York: John Wiley & Sons, 2001.

SAS Institute Inc., *SAS/STAT® Users Guide, Version 6, Forth Edition, Volume 1*, Cary, NC: SAS Institute Inc., 1989. 943pp.

SAS Institute Inc., *SAS/STAT® Users Guide, Version 6, Forth Edition, Volume 2*, Cary, NC: SAS Institute Inc., 1989. 846pp.

SAS Institute Inc., *SAS/STAT® Software: Changes and Enhancements through Release 6.11*, Cary, NC: SAS Institute Inc., 1989. 846pp.

Smith Tyler and Besa Smith, *Survival Analysis and the Application of Cox's Proportional Hazard Modeling Using SAS*, Proceedings of the 26th Annual SAS Users Group International Conference, Cary, NC: SAS Institute Inc., 2001.

## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Junxiang Lu, Ph.D.
8004 W 146 St
Overland Park, KS 66223
Email: lu8004@yahoo.com