Paper 114-27

Predicting Customer Churn in the Telecommunications Industry — An Application of Survival Analysis Modeling Using SAS®

Junxiang Lu, Ph.D.
Sprint Communications Company
Overland Park, Kansas

ABSTRACT

Conventional statistical methods (e.g. logistics regression, decision tree, and etc.) are very successful in predicting customer churn. However, these methods could hardly predict when customers will churn, or how long the customers will stay with. The goal of this study is to apply survival analysis techniques to predict customer churn by using data from a telecommunications company. This study will help telecommunications companies understand customer churn risk and customer churn hazard in a timing manner by predicting which customer will churn and when they will churn. The findings from this study are helpful for telecommunications companies to optimize their customer retention and/or treatment resources in their churn reduction efforts.

INTRODUCTION

In the telecommunication 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 less 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 high 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.

In order to support telecommunications companies manage churn reduction, not only do we need to predict which customers are at high risk of churn, but also we need to know how soon these high-risk customers will churn. Therefore the telecommunications companies can optimize their marketing intervention resources to prevent as many customers as possible from churning. In other words, if the telecommunications companies know which customers are at high risk of churn and when they will churn, they are able to design customized customer communication and treatment programs in a timely efficient manner.

Conventional statistical methods (e.g. logistics regression, decision tree, and etc.) are very successful in predicting customer 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 churn.

OBJECTIVES

The objectives of this study are in two folds. The first objective is to estimate customer survival function and customer hazard function to gain knowledge of customer churn over the time of customer tenure. The second objective is to demonstrate how survival analysis techniques are used to identify the customers who are at high risk of churn and when they will churn.

DEFINITIONS AND EXCLUSIONS

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

Churn – In the telecommunications industry, the broad definition of churn is the action that a customer's telecommunications service is canceled. This includes both service-provider initiated churn and customer initiated churn. 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 the reasons behind vary. In this study, only customer initiated churn is considered and it is defined by a series of cancel reason codes. Examples of reason codes are: unacceptable call quality, more favorable competitor's pricing plan, misinformation given by sales, customer expectation not met, billing problem, moving, change in business, and so on.

High-Value Customers – Only customers who have received at least three monthly bills are considered in the study. High-value customers are these with monthly average revenue of $X or more for the last three months. If a customer's first invoice covers less than 30 days of

service, then the customer monthly revenue is prorated to a full month's revenue.

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

Exclusions – This study does not distinguish international customers from domestic customers. However it is desirable to investigate international customer churn separately from domestic customer churn in the future. Also, this study does not include employee accounts, since churn for employee accounts is not of a problem or an interest for the company.

SURVIVAL ANALYSIS AND CUSTOMER 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). Typically, 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 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 has 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

On August 16, 2000, a sample of 41,374 active high-value customers was randomly selected from the entire customer base from a telecommunications company. All these customer were followed for the next 15 months. Therefore August 16, 2000 is the *origin of time* and November 15, 2001 is the *observation termination time*. During this 15-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* (August 16, 2000). A second variable of 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 15 (months) for the variable DUR.

DATA SOURCES

There are four major data sources for this study: block level marketing and financial information, customer level demographic data provided through a third party vendor, customer internal data, and customer contact records. A brief description of some of the data sources follows.

Demographic Data – Demographic dada is from a third party vendor. In this study, the following are examples of customer level demographic information:

- Primary household member's age
- Gender and marital status
- Number of adults
- Primary household member's occupation
- Household estimated income and wealth ranking
- Number of children and children's age
- Number of vehicles and vehicle value
- Credit card
- Frequent traveler
- Responder to mail orders
- Dwelling and length of residence

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, bill agency, customer segmentation code, ownership of the company's other products, dispute, late fee charge, discount, promotion/save promotion, additional lines, toll free services, 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
- Percentage change of minutes
- Share of domestic/international revenue

Customer Contact Records – The Company's Customer Information System (CIS) stores detailed records of customer contacts. This basically includes customer calls to service centers 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

Model process includes the following four major steps.

Explanatory Data Analysis (EDA) – Explanatory data analysis was conducted to prepare the data for the survival analysis. An 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 a big issue, except for those demographic variables. The demographic variables with more than 20% of missing values were eliminated. 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 212 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 categorical variables to be included in the next modeling step. All the categorical variables with a chi-square value or t statistics of 0.05 or less were kept. This step reduced the number of variables to 115 (&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 29 exploratory 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;
```

Model Estimation – With only 29 exploratory variables, the final data set has reasonable number of variables to perform survival analysis. 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 is to gain knowledge of customer churn hazard characteristics. From the shape of hazard function, customer churn in this study demonstrates a typical hazard function of a Log-Normal model. As previously discussed, since the shape of survival distribution and hazard function was known, PROC LIFEREG produces 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;
```

The final step is to estimate customer 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;
```

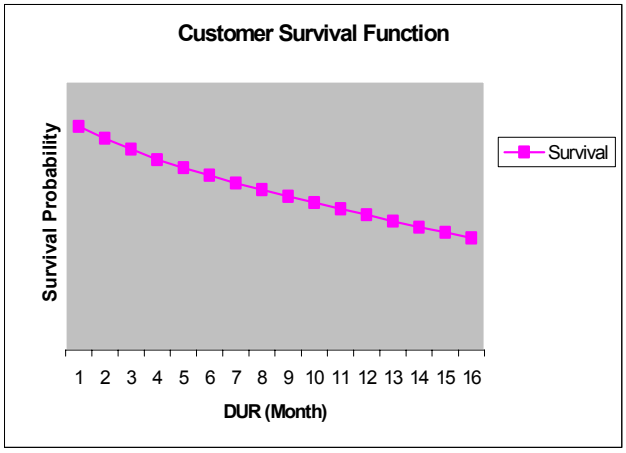

Figure 1. Customer Survival Function.

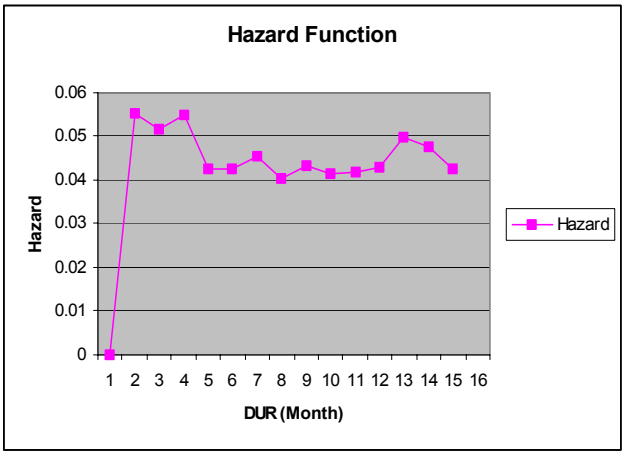

Figure 2, Customer Hazard Function.

By appending the validation data set to the model data set like 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 at 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.

```
%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 a specified time (that is, number of months since the *origin of time* - August 16, 2000), each customer will have P_0 through P_{15} as its predicted survival probabilities. P_0 is the predicted survival probability at the beginning of month 1, which is the *origin of time*. P_1 is the predicted survival probability at end of month 1; P_2 is the predicted survival probability at end of month 2, and so on. By sorting customers in ascending predicted survival probabilities for a specified time, if the model works, customers with the lowest predicted survival probabilities will have the highest likelihood to churn for this specified time period.

There are two ways to validate the successfulness of the model. One way is to rank the predicted survival probabilities for a specified time in ascending order into deciles and then to compare the number of churned customer during 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 churned during month 3. The other way is to put the predicted survival probabilities in the same order and then compare the number of churned customers up to this specified time in each decile. An example is to rank all customers predicted survival probabilities P_3's into deciles and then count the number of customers who have churned up to the end month 3 from the *origin of time*.

Table 1 describes the number of churners for each decile during a specified time period, while Table 2 describes the number of churners up to a specified time since the *origin of time* for each decile. Information about cumulative percentage of churners is also included in the tables. Only selected time periods are reported in the tables.

Table 1, Number of Churners by Decile During a Specified Month

	During Month 1		During Month 3		During Month 6	
Dec.	Num of Churn	Cum. %	Num of Churn	Cum. %	Num Of Churn	Cum. %
1	441	36.9%	329	34.9%	114	15.9%
2	269	59.4%	236	59.9%	190	42.5%
3	179	74.3%	125	73.2%	149	63.3%
4	117	84.1%	90	82.7%	100	77.2%
5	95	92.1%	73	90.5%	85	89.1%
6	55	96.7%	49	95.7%	47	95.7%
7	23	98.6%	26	98.4%	19	98.3%
8	9	99.3%	5	98.9%	7	99.3%
9	5	99.7%	4	99.4%	0	99.3%
10	3	100%	6	100%	5	100%

Table 2, Number of Churners by Decile Up to a Specified Month

	Up to Month 3		Up to Month 6		Up to Month 9	
Dec.	Num of Churn	Cum. %	Num of Churn	Cum. %	Num Of Churn	Cum. %
1	1099	35.2%	1594	30.3%	1889	26.8%
2	732	58.7%	1302	55.1%	1757	51.8%
3	446	73.0%	834	70.9%	1226	69.2%
4	318	83.2%	578	81.9%	811	80.8%
5	253	91.3%	459	90.6%	623	89.6%
6	150	96.1%	273	95.8%	378	95.0%
7	73	98.5%	133	98.4%	200	97.8%
8	23	99.2%	47	99.3%	77	98.9%
9	11	99.6%	14	99.5%	29	99.3%
10	14	100%	25	100%	47	100%

The *lift* measurement is preferred among marketers for evaluating and comparing model performance. At each decile it demonstrates the model's power to beat the random approach or average performance, which is visualized in a *lift gain chart*. Figure 3 is the cumulative lift gain chart for identifying the customers who churned during the third month since the *origin of time*. From the chart, we see that the top two deciles capture about 60% of churners, while the top five deciles captured about 90% of churners. Figure 4 is the cumulative lift gain chart for identifying the customers who have churned in the next 6 months since the *origin of time*. The top two deciles capture about 55% of churners, while the top five deciles capture about 90% of churners.

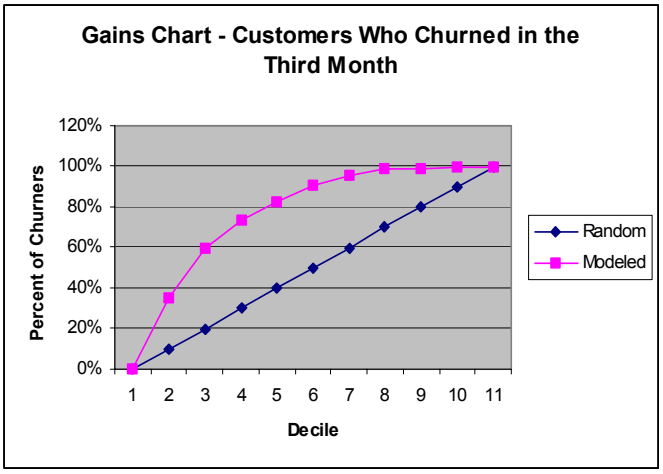

Figure 3, Model Lift Chart for Identifying Churners During the Third Month.

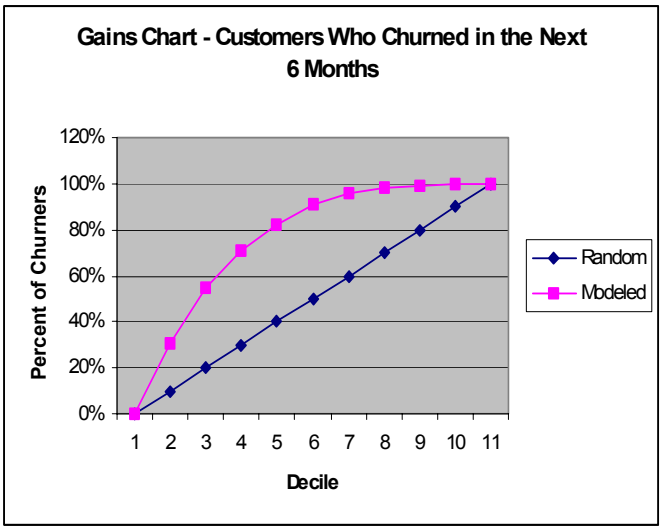

Figure 4, Model Lift Chart for Identifying Churners In the Next 6 Months.

CONCLUSION

This study provides another very powerful statistical tool – survival analysis to predict customer churn. By ranking the customers predicted survival probabilities in ascending order, the top two deciles capture 55-60% of churners and the top five deciles capture almost 90% of churners. The findings of this study will help telecommunications companies understand customer churn risk and customer churn hazard over the time of customer tenure. Overall, this study is helpful in customizing marketing communications and customer treatment programs to optimally time their marketing intervention efforts.

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.

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.

ACKNOWLEDGMENTS

I am grateful for the valuable inputs and thoughtful comments in this study by Robert Klein, group manager of the modeling development group at Sprint's Mass Markets Organization.

I wish to thank Paul Garcia, Sr. director of Database Marketing and CRM Implementation group at Sprint PCS, for his encouragement and full support in my modeling endeavors.

CONTACT INFORMATION

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

Junxiang Lu, Ph.D.
Sprint Communications Company
6130 Sprint Pkwy
Overland Park, KS 66251
Work Phone: (913) 762-5621
Fax: (913) 762-0804
Email: jlu02@sprintspectrum.com