Paper 132-2012

It's About Time: Discrete Time Survival Analysis Using SAS® Enterprise Miner™

Sascha Schubert, SAS Institute Inc., Heidelberg, Germany

Susan Haller and Taiyeong Lee, SAS Institute Inc., Cary, NC

ABSTRACT

The new survival analysis algorithm in SAS Enterprise Miner 7.1 provides analysts with an alternate approach to modeling the probability of customer behavior events. Traditional binary classification approaches provide a snapshot view of event propensities, while survival analysis can generate a time based function of event probabilities. This time-based view can help organizations to optimize their customer strategies by gaining a more complete picture of customer event likelihoods.

This paper briefly explains the theory of survival analysis and provides an introduction to its implementation in SAS Enterprise Miner. An example illustrates the usage of this analytical algorithm using a customer churn data set.

SHORT INTRODUCTION TO SURVIVAL DATA MINING

Survival data mining is the application of survival analysis to a data mining problem. The outcome that is modeled is no longer whether an event will occur in a certain time interval, but when the next event will occur. This modeling is useful when there are time-dependent outcomes, such as the time to event for a customer activity.

In traditional data mining, customer churn is modeled with classification algorithms such as logistic regression or decision trees, which predict a probability that an event will occur within a predefined time window. Based on the results of this model, organizations can select customer groups that are determined by cutoff thresholds for critical churn likelihoods. For example, all customers with a probability higher than 80% of canceling their services within the next three months should be included in a retention campaign.

Survival data mining, on the other hand, enables you to extend the model to look at a time-dependent outcome and to model the event likelihood over time. For example, now you can look at predicting the likelihood of churn at each month over the next 12 months. Characteristics such as longer-term trends of the churn probability over time can be exposed, which enables organizations to develop better long-term customer service strategies.

The most common approach in survival data mining is to take discrete time steps (daily, weekly, monthly, and so on) and calculate the event probability at each of these time steps. Discrete event times are represented by nonnegative integer values and represent the duration from the inception (start) time until the censoring date.

In survival analysis, censoring always needs to be taken into account. The time to event is the main characteristic that is analyzed. However, you will always have entities for which the analyzed event has not yet occurred at the time of the analysis. This phenomenon is called censoring. For example, in the case of churn, when you analyze the data at a certain point in time, called the censoring time, all customers that are still active have not churned yet, so their event has not yet occurred at the censoring time.

Two major functions are calculated based on the time-dependent outcome (event): the hazard function $h(x)$ and the survival function $S(x)$. The hazard function $h(x)_t$ at time point t is the probability that an event that has not occurred at time $t - 1$ will occur at time t. For example:

- 1% of all light bulbs will have stopped working after 1,000 hours
- 20% of all customers will have churned after one year

The hazard function *h(x),* also known as the failure rate, hazard rate, or force of mortality, is the ratio of the probability density function *P(x)* to the survival function *S(x)*, given by

$$h(x) = \frac{P(x)}{S(x)}$$

$$h(x)_t = \frac{\sum Customer\ exhibiting\ event\ at\ time\ t-1}{\sum Customers\ at\ time\ t}$$

The survival function is the probability that an event that has not occurred at time *t – 1* will also not occur at time *t*. For example:

- 99% of all light bulbs will still work after 1,000 hours
- 80% of all customers will still be active after one year

The survival function captures the probability that the system will survive beyond a specified time. It is the remainder of the proportion of entities that have not exhibited the event at time *t* and will survive to time *t = 1*; thus it can be expressed as:

$$s(x)_t = 1 - h(x)_t$$

Survival data mining can also take so-called external factors into account. For example, it is possible to analyze whether the number of products owned by a customer has an impact on the likelihood to churn. Rather than modeling an overall function of the event likelihood over time, you can group customers into segments based on their characteristics and calculate subhazards for the data segments. This approach works best with categorical or numeric variables with a few values, such as number of products in a customer's possession, type of products, or type of tariff. Numeric variables with more possible values, such as customer age, tenure, annual income, minutes on air, monthly revenue, and so on, can be easily binned for segmentation. Figure 1 shows an example of subhazards that are based on type of product (Gordon and Linoff 2009). This graphic shows that the expected lifetime of a customer who has the regular product is lower than the expected lifetime of a customer who uses the high-end product.

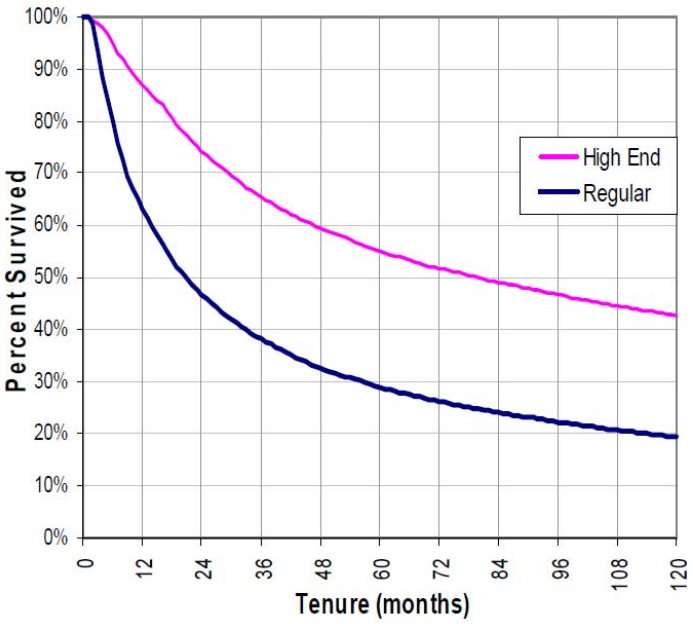


Figure 1. Impact of External Factors on Survival Function

The concept of competing risks is another concept from traditional survival analysis that can be applied to this type of data mining analysis. Sometimes the same outcome can be caused by different events. To distinguish between these different types of events (competing risks), the event needs to be categorized in the data.

For example, in the case of phone churn, a customer might leave voluntarily or involuntarily. The outcome is the same: the customer has left. However, the drivers for the occurrence of this event are very different. Each customer can experience only one of the events: that is, the events are mutually exclusive. So, if the organization records the type of event, the survival analysis can segment the analysis for each event type and provide insights into the different subhazard functions and the drivers for each event type.

APPLICATIONS OF SURVIVAL ANALYSIS

Survival data mining can be applied to all data that contain a time-dependent target. Survival analysis has become popular in customer behavior modeling, especially in customer retention, which is probably the most dramatic event in the lifetime of a customer. However, you can imagine applying this technique to other customer events in the customer lifecycle, such as product purchases, upgrades, or downgrades.

Generally, all customer events have a time dimension that has been included in traditional classification models only on the proxy level, through including time-based input data and by predefining the time window of the prediction. So in all areas where organizations use classification models today to predict the occurrence of events, survival data mining can be applied and can provide a different perspective. Whereas classification models produce the probability of the event at a single point in time, survival mining calculates a function of the event probability over time. Survival mining enables organizations to create strategies that are more optimized in relation to the expected customer behavior over time. Some examples of time-dependent outcomes follow:

- customer behavior:
  - customer churn
  - cancellation of all products and services
  - customer lifetime projections (Potts 2005)
  - unprofitable behavior
  - service downgrade or extreme inactivity
  - new product purchase (especially in relation to existing product portfolio)
  - existing product upgrade
  - missing bill payments and collection optimization

- operations:
  - product maintenance
  - equipment failure

THE SURVIVAL MINING NODE IN SAS ENTERPRISE MINER

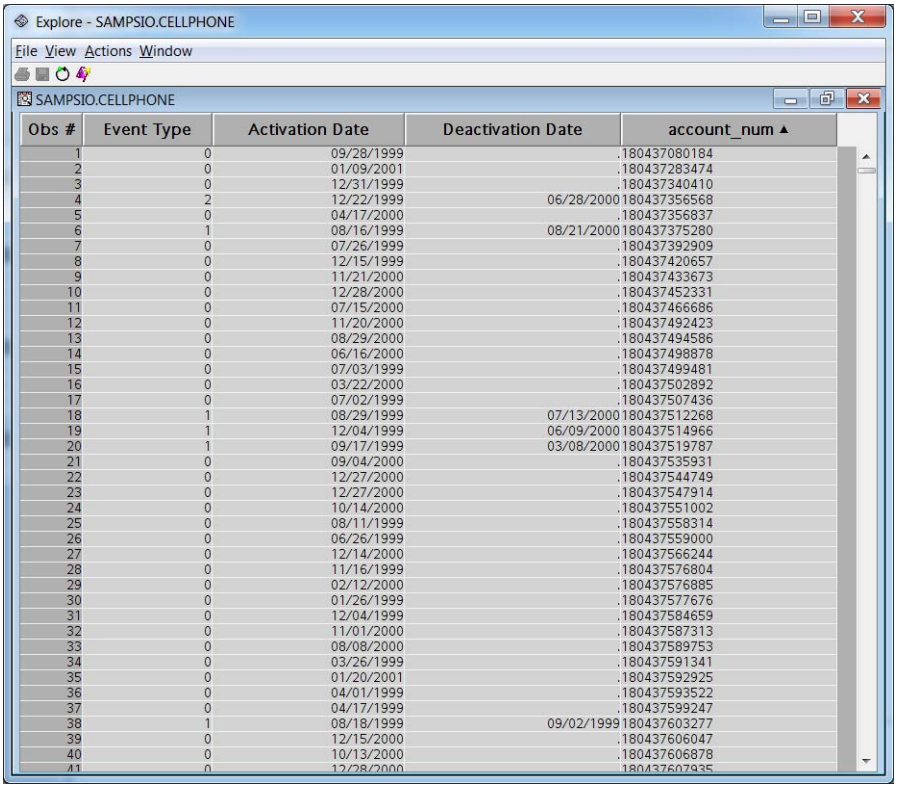
As part of its data mining functions, SAS Enterprise Miner 7.1 includes the new Survival node. The Survival node covers discrete time-to-event modeling, so you can model the customer lifetime in integer-valued time steps (such as 10 weeks, 12 months, and so on).

Functions of the discrete event time (in the form of cubic spline basis functions) serve as predictors in a multinomial regression model. The probabilistic hazard and subhazard functions that are generated by the Survival node are based on the multinomial logistic regression. The hazard function is a conditional probability of an event at time *t*, and often this discrete-event time function is nonlinear in nature. Transforming the event time function with cubic spline basis functions enables the hazard and subhazard functions to be more flexible, resulting in a greater ability to detect and model customer behavior patterns.

DATA PREPARATION IN SAS ENTERPRISE MINER

The data to be mined for survival modeling must be configured for survival analysis. The raw data need to be in the following format, which is illustrated in Figure 2:

- An entity ID (for example, a customer ID or a product ID) needs to be available for the analysis.
- Additional variables that contain the start date and the end date of the observation period must be included and mapped to the variable role of TIMEID in SAS Enterprise Miner. For a customer event survival analysis, the start date is usually the date when the relationship for the customer or the product under analysis was initiated, and the end date is the date when the relationship for the entity was terminated. In Figure 2, the Activation Date variable represents the start date, and the Deactivation Date represents the end date.
- The organization needs to carefully define the characteristics of the event. For example, was it voluntary or involuntary churn, and how is churn defined? The example data set in Figure 2 identifies two types of churn that present competing risks in this analysis: voluntary churn (flagged with Event Type = 1) and involuntary churn (flagged with Event Type = 2). A value of 0 for Event Type represents a right-censored event, which means the event has not occurred at the end date. A value of 0 corresponds to an empty cell in the end-date variable. For example, in Figure 2, the first observation has a value of 0 for Event Type and an empty cell in the Deactivation Date column.



Explore - SAMPSIO.CELLPHONE

File View Actions Window

SAMPSIO.CELLPHONE

Obs #	Event Type	Activation Date	Deactivation Date	account_num ▲
1	0	09/28/1999		.180437080184
2	0	01/09/2001		.180437283474
3	0	12/31/1999		.180437340410
4	2	12/22/1999	06/28/2000	180437356568
5	0	04/17/2000		.180437356837
6	1	08/16/1999	08/21/2000	180437375280
7	0	07/26/1999		.180437392909
8	0	12/15/1999		.180437420657
9	0	11/21/2000		.180437433673
10	0	12/28/2000		.180437452331
11	0	07/15/2000		.180437466686
12	0	11/20/2000		.180437492423
13	0	08/29/2000		.180437494586
14	0	06/16/2000		.180437498878
15	0	07/03/1999		.180437499481
16	0	03/22/2000		.180437502892
17	0	07/02/1999		.180437507436
18	1	08/29/1999	07/13/2000	180437512268
19	1	12/04/1999	06/09/2000	180437514966
20	1	09/17/1999	03/08/2000	180437519787
21	0	09/04/2000		.180437535931
22	0	12/27/2000		.180437544749
23	0	12/27/2000		.180437547914
24	0	10/14/2000		.180437551002
25	0	08/11/1999		.180437558314
26	0	06/26/1999		.180437559000
27	0	12/14/2000		.180437566244
28	0	11/16/1999		.180437576804
29	0	02/12/2000		.180437576885
30	0	01/26/1999		.180437577676
31	0	12/04/1999		.180437584659
32	0	11/01/2000		.180437587313
33	0	08/08/2000		.180437589753
34	0	03/26/1999		.180437591341
35	0	01/20/2001		.180437592925
36	0	04/01/1999		.180437593522
37	0	04/17/1999		.180437599247
38	1	08/18/1999	09/02/1999	180437603277
39	0	12/15/2000		.180437606047
40	0	10/13/2000		.180437606878
41	0	12/28/2000		.180437607935

Figure 2. Sample Data for Phone Churn Survival Analysis

The Survival node includes functional modules that prepare data for the analysis by expanding data from one record per ID to one record per time unit and that perform sampling to reduce the size of the expanded data with minimized information loss. The Survival node also performs survival model training, validation, scoring, and reporting.

The data are transformed automatically based on the properties defined in the Survival Mining node. Also, the start and end dates for the analysis are identified automatically from the data. You use the variable metadata to specify the start and end time variables. The Survival node then calculates the start date as the minimum date and the end date as the maximum date from these variables. As described earlier, the end date defines the censoring date. If the time interval in the data and the time interval that you choose for the analysis do not match, the Survival node automatically adjusts the dates.

For example, when the date is collected on a daily time interval and the analysis is carried out on a daily time interval, the start and end dates can be directly extracted from the data as they are. When the data are stored on a daily time interval and you choose to run the analysis on a monthly time interval, the end of the month previous to the maximum date becomes the censoring date to avoid incomplete intervals. For example, if the data end at January 15 and the selected time interval is monthly, the censoring date is December 31 and all observations after that date are treated as censored.

The Time interval property enables you to specify a time interval as Day, Week, Month, Quarter, Semi-year, and Year for censoring and reporting.

Figure 3 illustrates the transformation from data that are formatted with one record per ID to data that are formatted with one record per time unit for a monthly time interval. The left panel shows different activation and deactivation dates for different customer IDs. Customer 1 started on Jan 01, Customer 2 started on Apr 01, and Customer 3 started on Feb 01. Customer 1 deactivated the relationship on Oct 01 and Customer 3 left on Sep 01. Customer 2 is still active, so this is a censored observation. In the transformed “tenure” view, the data are expanded to one record per time unit, which is monthly. For each customer, the time units are counted since the start time (tenure = 0). Customer 1 has stayed 10 time units (the event occurred at time step 10), Customer 3 has stayed eight time units (the event occurred at time step 6), and Customer 2 is still active after nine time units.

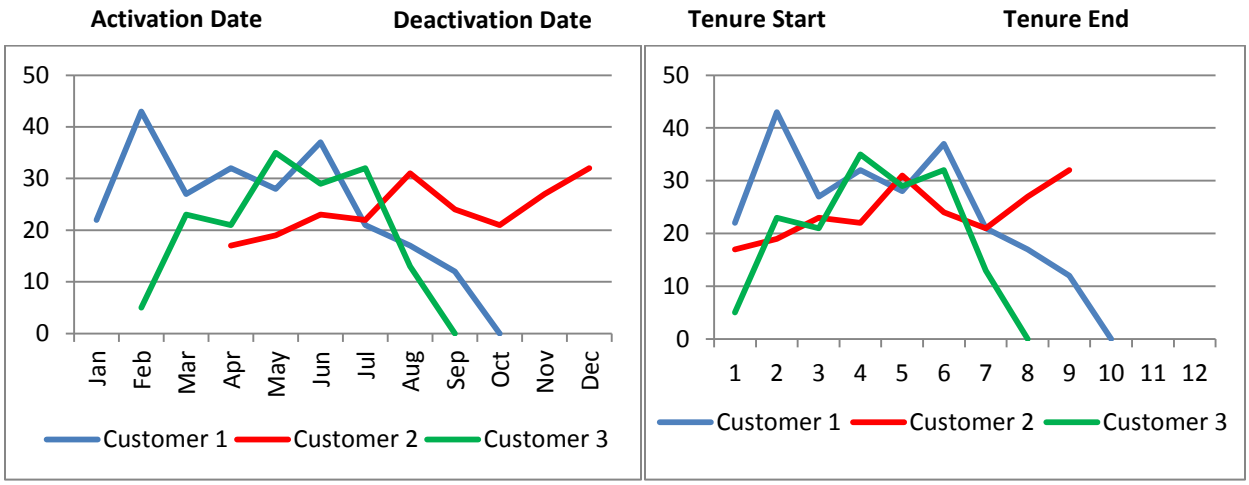


Figure 3. Transformation from Time Stamp View to Tenure View

Sampling and Validation

The Survival Mining node provides built-in sampling and model validation for optimal model fitting. Sampling usually becomes necessary with survival analysis because expansion of the data for the analysis from one-record per entity ID to one record per time unit increases sample size exponentially. The sampling method enables you to apply choice-based sampling to increase the event rate in the training data by eliminating from the expanded data observations that do not have an event. Training the model with a biased sample enables you to estimate the predicted subhazard functions more precisely than you could if you used a random sample. To correct the bias caused by sampling, you can automatically adjust the subhazard functions after you build the model.

Model validation is very important for identifying whether the model accuracy is acceptable for basing business decisions on the model results. For survival mining, the time dimension is the most important influence factor. For this reason, the data must be partitioned into training data and validation data with respect to time. For both the model training and the model validation, the Survival node automatically creates an unbroken time period by creating a hold-out sample that uses the last quarter of the time period in the data.

For example, if the data span eight years in time on a monthly scale, the model is trained on 75% of the data. That is, the first six years and the last two years are held out for model validation. You can define customized settings for the start time and the length of the validation data.

Survival Mining Algorithms

Based on the expanded data, the Survival Mining node calculates the hazard and subhazard functions as described in the section “Short Introduction to Survival Data Mining.”

The hazard and subhazard functions represent the chance that an event that has not occurred for a given span of time is going to occur before the next unit of time. The subhazard function simply represents the conditional probability of an event occurrence of type *x* at time *t*, given that no event has occurred before time *t*. The Survival node uses observations that have a target value of 0 (observations with no observed event, known as censored observations) as the reference level.

Approximation of the Hazard Function Shape

The hazard function is often a nonlinear function of time. To approximate this nonlinear relationship, the Survival node in SAS Enterprise Miner uses cubic splines with a predefined number of knots that you can specify; the default is five. The cubic spline approximations then become predictors in the multinomial regression in addition to the optional covariates that you choose to include.

The cubic spline basis functions are of the following form:

$$csb(t, k_j) = \begin{cases} -t^3 + 3k_j t^2 - 3k_j^2 t & if\ t \leq k_j \\ -k_j^3 & if\ t > k_j \end{cases}$$

where *j* is the number of knots and *k* is the value of the knot.

A cubic spline is a segmented function that consists of third-degree (cubic) polynomial functions that are joined together so that the whole curve and its first and second derivatives are continuous. The join points between the segments are called knots. The knots are points where the function makes a transformation. For example, a knot is the point at which one of the cubic spline basis functions changes from a cubic function to a constant function. Figure 4 shows examples of cubic spline functions with different numbers of knots.

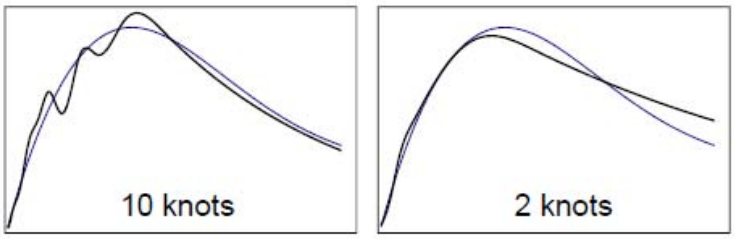


Figure 4. Examples of Cubic Spline Functions with Different Numbers of Knots (Potts 2005)

A discrete-time multinomial logistic regression fits a model to the shape of the hazard function. The regression model uses discrete time, cubic spline basis functions, and covariates for model fitting. You can apply stepwise selection to identify the significant predictors for the function fitting. Then you can use the resulting regression model to score each customer beyond the current time.

Scoring

The Survival node includes scoring options. Scoring in the context of survival mining is different from scoring with the predictive classification models that are usually applied to customer behavior modeling. Although the traditional models predict the probability of the occurrence of an event at a predefined future time window (within the next month, the next three months, and so on), the survival analysis calculates a function of the event occurrence probability over time. The Survival node also produces the Mean Residual Lifetime for each entity by two extrapolation methods of hazard beyond the observed time period. The Mean Residual Lifetime is the expected time that remains until the observed event, and it is calculated for each customer. You specify the number of time intervals to forecast into the future. The following defaults are defined for different time intervals:

- Day: 30 days into the future
- Week: four weeks into the future
- Month: three months into the future
- Quarter: four quarters into the future
- Semi-Year: two half-years into the future
- Year: one year into the future

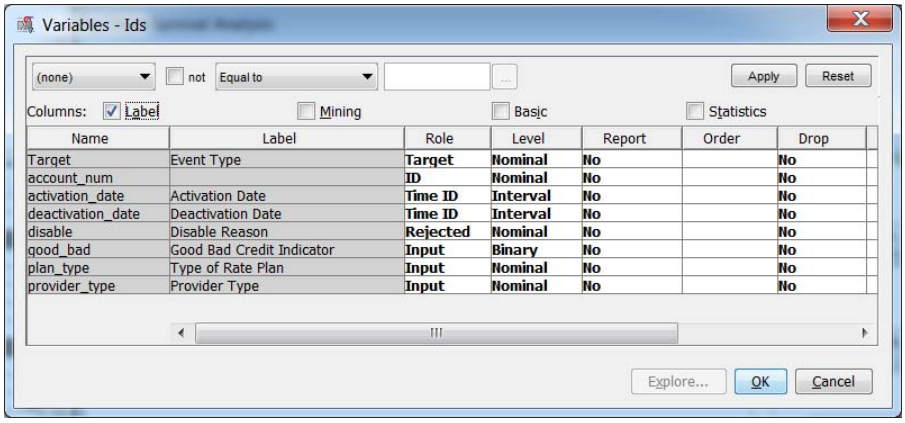
You can use the Score node in the process flow to generate scoring code that can be deployed in production. The generated score code is Base SAS® code, and it can be implemented on a different SAS server without a SAS Enterprise Miner license. With the deployed score code, the statistics on probabilities for customer events can be calculated in regular time intervals (scheduled batch runs) and fed into customer service strategy systems for selection of the optimal customer segments. For example, you can select the top 10% of customers with the highest churn event probability within the next six months for customer retention campaigns, and you can select the top 10% of customers with the highest survival probability over the next six months for loyalty campaigns.

Example

Data

This example uses data from a cellphone provider. These data are also included in the sample data library of SAS Enterprise Miner 7.1. As explained earlier, for a survival mining analysis you need the date of service activation and service termination (which can be censored) for each entity. Here, the entity is the customer. As shown in Figure 5, the data recorded the activation date and the deactivation date (for churn) for the sample of customers. Churn can be voluntary and involuntary, which is indicated in the data by different values for the target variable. A value of 0 represents a customer who is still active (censored), a value of 1 represents a customer who churned voluntarily, and a value of 2 represents a customer who was forced to churn by company rules. In addition to these required input variables, the sample data also contain some covariates that you can use to analyze the impact of customer characteristics on the churn event probability functions.

Certain settings of the input variables for survival mining are required. Figure 5 shows the required ID variable, which identifies the entity—in this case a unique customer account. The activation and termination dates are specified as Time ID, and the event variable is specified as Target. The covariates are included as optional Input variables.



Name	Label	Role	Level	Report	Order	Drop
Target	Event Type	Target	Nominal	No		No
account_num		ID	Nominal	No		No
activation_date	Activation Date	Time ID	Interval	No		No
deactivation_date	Deactivation Date	Time ID	Interval	No		No
disable	Disable Reason	Rejected	Nominal	No		No
good_bad	Good Bad Credit Indicator	Input	Binary	No		No
plan_type	Type of Rate Plan	Input	Nominal	No		No
provider_type	Provider Type	Input	Nominal	No		No

Figure 5. Variable Settings for Survival Analysis

Figure 6 displays the distribution of the target variable Target. You can see that about 81% of the customer accounts are still active at the censoring time, while about 15% have churned voluntarily and 4% were forced to churn. The observation time window in these data is January 1999 to January 2001, and the time interval is Month. The Survival node automatically uses the maximum termination date in the context of the time interval as a censoring date. In this example, the censoring date is set to January 2001. All customers who have not left before this date (as represented by a missing value in the termination date variable) are still active.

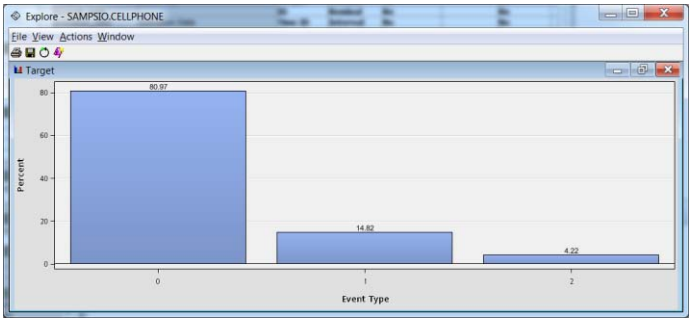


Figure 6. Distribution of Target Event Type

Process Flow

Two different settings of the Survival node are used for demonstration purposes:

- Multinomial regression model without covariates
- Stepwise regression model selection with covariates

Figure 7 illustrates the case study process flow in SAS Enterprise Miner 7.1. This paper assumes that you are familiar with the process of creating a project, a diagram, and a data source in SAS Enterprise Miner; see SAS Enterprise Miner online Help for more details. Since the Survival node combines several analysis steps (such as data preparation, function calculation, model fitting and model validation, and reporting) in one single tool, the process flow is rather simple. Because the Survival node also includes implicit model validation, data partition is not required. The Survival node automatically splits the data into training and validation, based on the time sequence. By default, the last 25% of the time period is used for validation. For this time period, the model that is fit to the training data is scored and validated. However, you can optionally partition the data for additional model validation. In this case the defined data are used for model training and model validation.

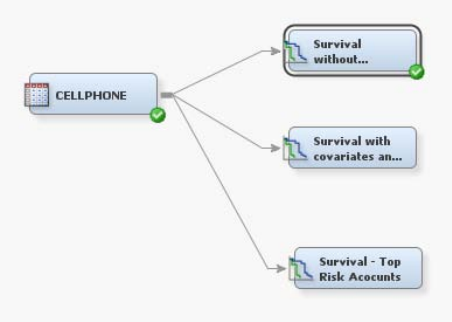


Figure 7. SAS Enterprise Miner Process Flow for Survival Mining

Multinomial Regression Model without Covariates

This analysis excludes the covariates from the model by setting their Use variable to No in the variables property window of the Survival node. (See Figure 8.) Then the Survival node is run with the default settings.

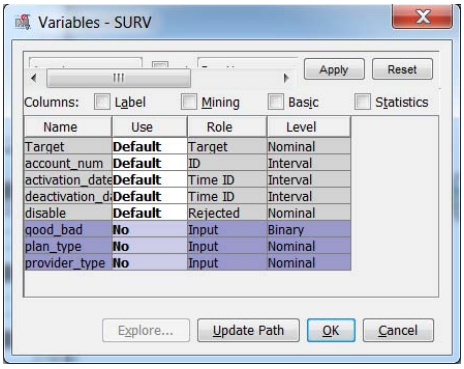


Figure 8. Reject Covariates from Analysis

The main information gained from a survival analysis are the hazard and subhazard plots. As explained earlier, the hazard function describes the risk of an event occurring at time t+1 if it has not happened yet at time t as a function of time t. Where the hazard function displays peaks in the shape of the function provides insight into high-risk periods in a customer lifecycle (as in this example of a churn event analysis). Examining where peaks occur enables businesses to determine the best time to intervene with customers they want to retain in the customer base.

The subhazard functions help businesses understand the impact of competing events. The drivers for the event of voluntary and involuntary churn are different, and thus you would expect different hazard functions for the competing risks. The segmentation enables you to drill deeper and to define the correct actions for each of the event categories.

The empirical survival plot (blue line) calculated from the training data shows an almost linear behavior. (See Figure 9.) From this you can conclude that the monthly churn rate in the customer base is constant, which means that a similar proportion of customers are leaving every month (between about 1.2% and 2.7%). From the survival function, you can also see that after a tenure of 24 months, about 65% of the customers are still active, which means that about 35% of the customer base has left during that time period.

The hazard function (red line) is also quite telling. You can see that there are peaks in the hazard function around months 12 and 20. It seems logical in the cellphone industry that these peaks might correspond to the ending of promotional periods in which users discontinue service after the initial discount expires.

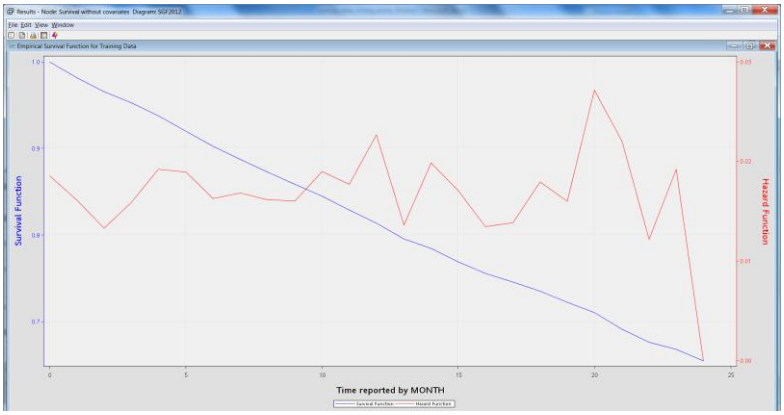


Figure 9. Hazard and Survival Functions of Case Study Data

Figure 10 shows the hazard and survival functions for an analysis on weekly and quarterly time scales. As expected, the hazard function shows a much higher volatility on a smaller time scale (such as weekly) as compared to a lower time resolution (such as quarterly). The peak in the hazard rate at the end of the observation period is more pronounced on a weekly time scale, suggesting that churn is driven by the end of the contract period and that there was a period when a larger proportion of the customer base reached the end of the contract within the same week. When you look at the quarterly resolution, you see that these short time effects are smoothed out.

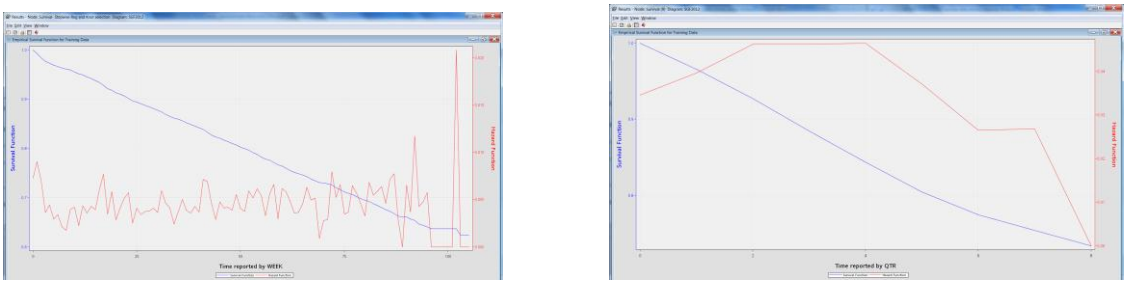


Figure 10. Empirical Hazard and Survival Functions on a Weekly and a Quarterly Time Scale

You can run this complex analysis on different time scales just by changing one parameter in the property sheet of the Survival node, and you can easily find the time scale that creates the best model and also best fits the business requirements. Although different time scales are used in this example to show the capabilities of the survival application, a resolution of a monthly time scale for churn usually fits the business requirements best, because telephone companies usually have a monthly billing cycle. With a monthly time scale, the organization can run the

model regularly on a monthly cycle and target the highest-risk accounts based on event occurrence probabilities. The rest of the example focuses the analysis on a monthly time scale.

The empirical subhazard functions calculated from the training data display the time impact on the competing risks of churn. (See Figure 11.) Subhazard function1 (blue line, upper line) relates to the risk of a customer leaving voluntarily, and subhazard function 2 (red line, lower line) relates to the risk of a customer being forced to leave because of payment failure. Fortunately, for this organization, the risk of having customers forced to leave is lower than the risk that they are churning voluntarily; this means that at least the vast majority of customers pay their bills. There is an area of concern at the end of the observation period (23 months), where you see a spike in the proportion of customers who are forced to leave. This spike might point to a problem that the business needs to pay more attention to in its business process.

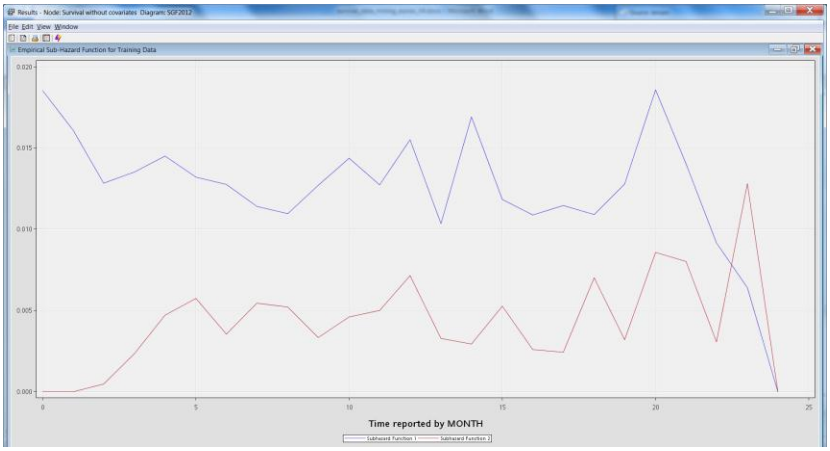


Figure 11. Empirical Subhazard Functions for Competing Risks

This example uses only the discrete time points as input to the multinomial regression model. You can see that a lift of 1.037 was achieved, which is not really satisfactory. Another indicator of the rather weak model performance is the flat concentration curve. (See Figure 12.) A large benefit value indicates the depth at which the model is doing the best job of predicting the outcome. The benefit curve can be used to establish an appropriate cutoff value for the decision about whether a customer is going to churn based on the probability that the predicted event occurs.

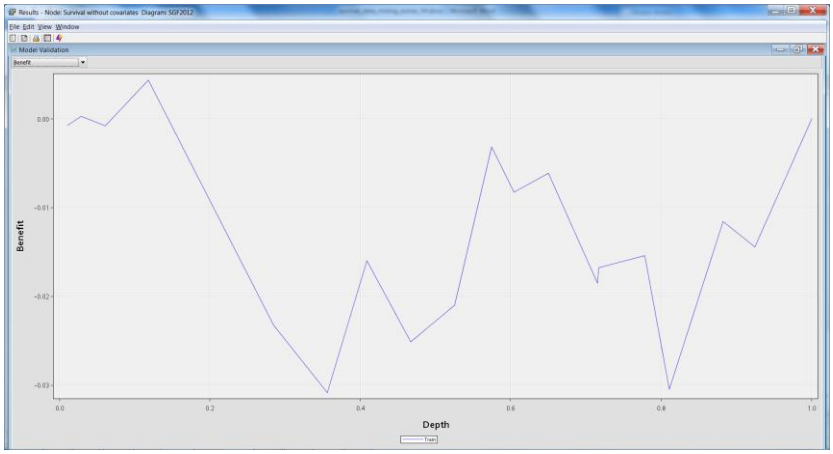


Figure 12. Benefit Curve of Churn Event Analysis

Stepwise Regression Model Selection with Covariates

This analysis includes the covariates into the regression model by leaving the default settings in the variables property window of the Survival node and selecting **Stepwise Regression** in the property sheet of the Survival node. (See Figure 13.)

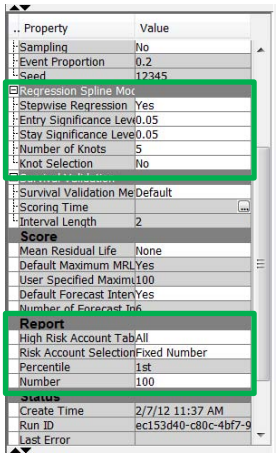


Figure 13. Property Settings for the Survival Node for Stepwise Regression Including Covariates

The results are shown in Figure 14. Although the empirical functions calculated from the training data are exactly the same (as expected because the same training data are used), the model fitting has improved significantly, as shown by the increased lift value of 1.50 in the "Model Fit Statistics" table. Also, the benefit curve in the model validation graph shows significant improvement. Instead of a flat line (which indicates weak model performance), the line is curved, which shows an area with improved model predictions.

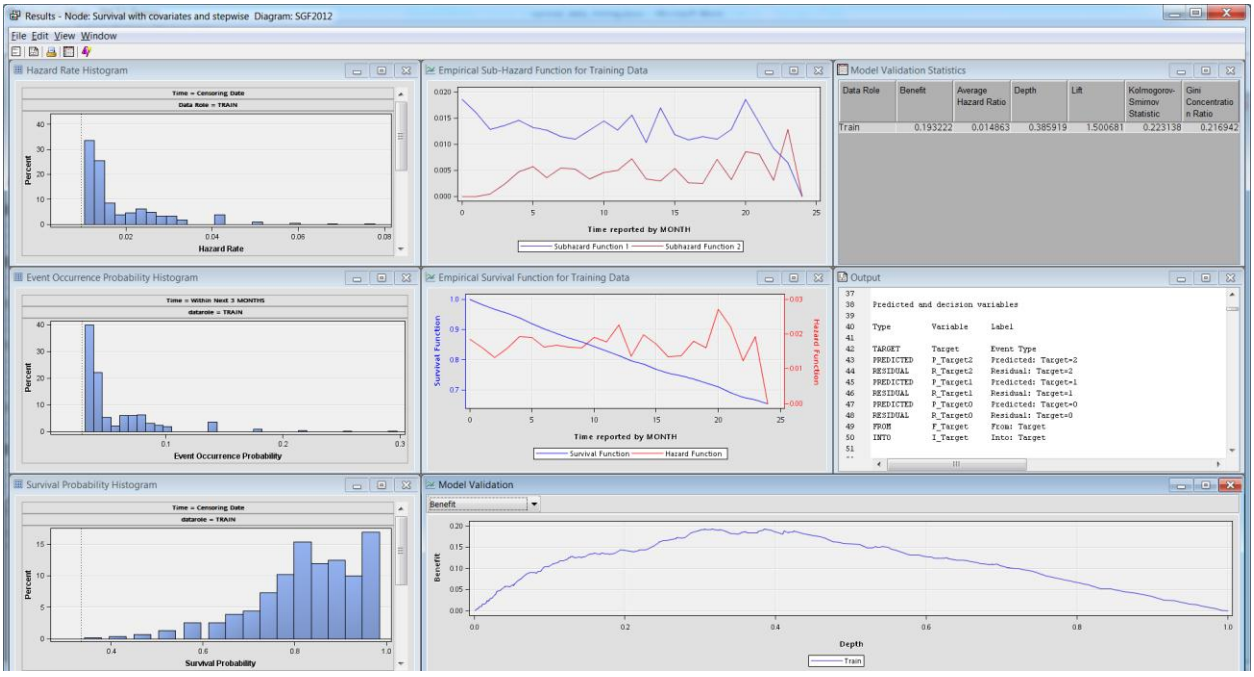
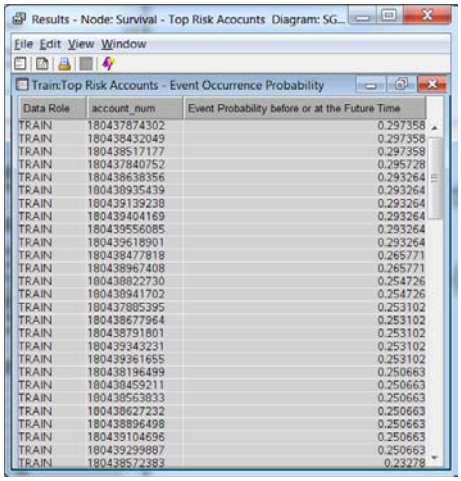


Figure 14. Result Browser of the Survival Node with Benefit Curve

However, the Survival Mining node provides even more insight into event occurrence risk on an entity level, such as the customer. A very important and easy-to-use metric is the mean residual lifetime, which is the time expected to remain until the occurrence of an event, such as an existing customer terminating a business relationship. For each entity, this metric is calculated and a histogram of the distribution across the training and the validation data is generated by default in the result browser.

You can also get a report of the highest-risk accounts based on one or more statistics such as event probability, survival probability, and the hazard rate before or at the selected future time. This report ranks the customer records based on the selected statistics and generates a table with the selected proportion of the customer base for further business actions, such as communication activities or retention campaigns. (To create a high-risk account table for all probabilities—event, survival, and hazard probability—select the property **All** for the item **High-Risk Account Table** in the Report section of the property sheet for the Survival node, as shown in Figure 13.)

In the Results browser of the Survival node, you can view the high-risk account table for the different probabilities. (See Figure 15.) This table enables the business to immediately target different customer segments for different business actions.



Data Role	account_num	Event Probability before or at the Future Time
TRAIN	180437874302	0.297358
TRAIN	180438432049	0.297358
TRAIN	180438517177	0.297358
TRAIN	180437840752	0.295728
TRAIN	180438638356	0.293264
TRAIN	180438935439	0.293264
TRAIN	180439139238	0.293264
TRAIN	180439404169	0.293264
TRAIN	180439556085	0.293264
TRAIN	180439618901	0.293264
TRAIN	180438477818	0.265771
TRAIN	180438967408	0.265771
TRAIN	180438822730	0.254726
TRAIN	180438941702	0.254726
TRAIN	180437885395	0.253102
TRAIN	180438677964	0.253102
TRAIN	180438791801	0.253102
TRAIN	180439343231	0.253102
TRAIN	180439361655	0.253102
TRAIN	180438196499	0.250663
TRAIN	180438459211	0.250663
TRAIN	180438563833	0.250663
TRAIN	180438627232	0.250663
TRAIN	180438896498	0.250663
TRAIN	180439104696	0.250663
TRAIN	180439299887	0.250663
TRAIN	180438572383	0.23278

Figure 15. The 100 Highest-Risk Accounts Based on Event Occurrence Probability

Creating and Deploying the Score Code

As data miners are used to doing in SAS Enterprise Miner, you can easily create the score code for model deployment by adding a Score node to the Survival node. The Base SAS code that is generated includes all the logic required to immediately deploy the score code in a production SAS environment. The scoring data set also needs to contain a tenure variable _T_, which contains the number of time units since activation for each record to score. This is different from most scoring processes in SAS Enterprise Miner, where the scoring code that is created can be applied to new data directly without any required additions or modifications. The SAS Enterprise Miner online Help contains an example of how to create this variable for the scoring data.

One caveat to keep in mind when you use covariates in the survival model is that in the current version the model supports only time-independent variables. This example uses variables that do not change over time; thus, they can be included in the scoring code directly from the training data.

Since survival models differ from traditional event classification models, the output variables that are created are also different. The survival score code generates these output variables:

- survival probability at censoring time
- survival probability at future time
- event probability before or at the future time
- hazard function at censoring time
- hazard function at future time
- subhazard function at censoring time (if competing risks are involved)
- subhazard function at future time (if competing risks are involved)

With these metrics, organizations can select customer segments from different perspectives for business actions. For example, they not only can select the highest-risk accounts at a specific point in time (censoring or future time) but also can look at the changes in probabilities and select accounts with the fastest-growing probability of the event occurrence.

CONCLUSION

The new Survival node in SAS Enterprise Miner provides analysts with a new perspective on predicting the likelihood of events occurring in relation to a discrete time interval. With traditional data mining classification approaches, the data mining model was applied to a predefined time window in the future, such as the next one, three, or six months. The model then predicted the likelihood of the event occurring for this single point in time. With survival analysis, organizations now are able to create a longitudinal view of the event probability rather than a snapshot in time. The longitudinal view provides the organization with additional information about significant events that need to be embedded into business process decisions. For example, with the application of survival analysis to customer churn in telecommunications, organizations can define the best time for intervention with greater confidence because they have a much more complete picture of the churn risk for each customer.

REFERENCES

1. Gordon, M., and Linoff, G. 2009. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*, 3rd ed. New York: John Wiley and Sons Inc.
2. Potts, W. 2005. “Predicting Customer Value.” *Proceedings of the SAS Global Forum 2005 Conference*. Cary, NC: SAS Institute Inc. Available at: http://www2.sas.com/proceedings/sugi30/073-30.pdf

CONTACT INFORMATION

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

Sascha Schubert
SAS Institute Inc.
In der Neckarhelle 162
Heidelberg, Germany 69120
E-mail: Sascha.Schubert@sas.com
Web: www.sas.com

Susan Haller
SAS Institute Inc.
100 SAS Campus Drive
Cary, NC 27513
E-mail: Susan.Haller@sas.com
Web: www.sas.com

Taiyeong Lee
SAS Institute Inc.
100 SAS Campus Drive
Cary, NC 27513
E-mail: Taiyeong.Lee@sas.com
Web: www.sas.com