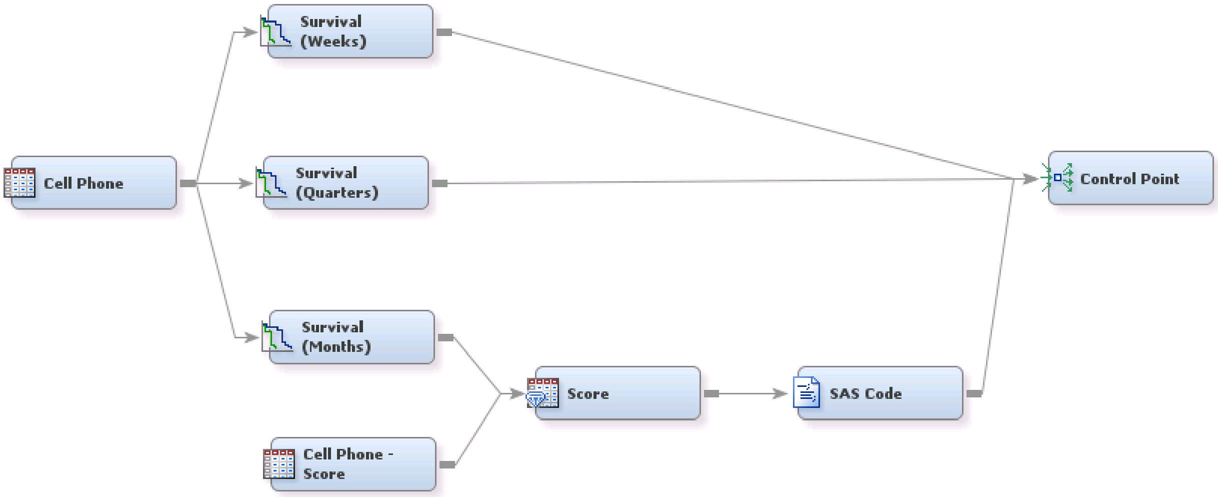
Survival Analysis Using SAS® Enterprise Miner™



Data:

This example uses SAMPSIO.CELLPHONE as the training data and SAMPSIO.CELLPHONE_SCORE as the score data; both data sets contain information about customers of a cell phone service provider. The variable Target in SAMPSIO.CELLPHONE is the target variable; it indicates whether a customer is active, has voluntarily discontinued service, or involuntarily discontinued service. For the Survival node, a value of 0 for the target indicates that the observation is "censored"—that is, no event or failure has occurred. The Time ID variables activation_date and deactivation_date contain the customer's start date and end date, respectively. The end date is missing if the customer is still active (Target=0). An additional variable, _t_, is found in the score data because the Survival node requires it when scoring. The following section discusses how to create _t_.

Data Preparation:

When you score survival data, you need to do some preliminary data preparation. As in any scoring situation, the scoring data set needs to contain all the input variables that the training data set contains. For the case of survival scoring, an extra variable, _t_, also needs to be included. The variable _t_ represents the number of time intervals between a customer's activation date and the current date. In this example, _t_ is in months.

The following example code was used to create the SAMPSIO.CELLPHONE_SCORE data set and has already been run. You do not need to run the SAS code in this example. Typically the score data isn't created from the training data; instead it is new data in which only the start date and inputs are known.

```
data sampsio.cellphone_score;
   set sampsio.cellphone;
   format _currentdate MMDDYY10.0;
   _currentdate=input("31DEC2000", anydtdte10.);
   _t_=intck("MONTH",activation_date, _currentdate);
   drop Target _currentdate;
run;
```

Goal:
The goal is to estimate when customers will cancel their subscription (churn). Estimating when customers will cancel subscriptions (as opposed to predicting which customers will cancel or whether they will cancel) enables you to view trends in time about churn and make decisions accordingly. Survival analysis can also be used to model other types of events or failures (for example, when objects will break or become unusable).

This process flow diagram examines the use of the Survival node without the use of time-varying covariates.

Flow:
The Cell Phone node (an Input Data node) in this example is connected to three different Survival nodes. In practice you would only need one Survival node, but three are shown here to illustrate differences.

The Survival (Weeks) node is the Survival node in which the **Time Interval** property has been set to **Week**. All other properties of this node have been left with their default values. Some of the other properties include user specification of the regression spline model, survival validation, and score. Finally, the **Report** properties enable you to output High Risk tables, which indicate the customers who are at a higher risk of cancellation.

The Survival (Months) node uses the same settings as the Survival (Weeks) node, except that the **Time Interval** property is set to **Month**. Likewise, the Survival (Quarters) node uses the same settings except for the use of **Quarter** as the **Time Interval**.

You are encouraged to use the time interval that makes the most sense to you. When there are multiple possible time intervals, you might want to view the results for each Survival model, because changing the time interval can drastically change your ability to visually understand the data.

After running each of the Survival nodes, you can view the results to see multiple graphs. Two graphs of especially high importance are the Empirical Sub-Hazard Function for Training Data and the Empirical Survival Function for Training Data.

The Empirical Sub-Hazard Function for Training Data plots the various subhazards, which are the different possible churn events. In this case, Subhazard Function 1 refers to voluntary churn, whereas Subhazard Function 2 refers to involuntary churn (because of overdue payments, for example). The Survival (Months) node shows small spikes in the Subhazard Function 1 at month 14 and month 20. A possible explanation is that contracts last for 12 and 18 months and customers who want to cancel usually do so about 2 months after the contract expires (this is merely supposition).

The Empirical Sub-Hazard Function for Training Data in the Survival (Quarters) node appears relatively flat, whereas the same graph in the Survival (Weeks) node has many spikes. This difference illustrates the important point that the time interval can make a large difference in your understanding of the data. A time interval of Quarters makes it hard to see a trend because the time interval is too large, but a time interval of Weeks is too short, making it hard to draw inferences.

The Empirical Survival Function for Training Data graph plots the survival probability and hazard probability as a function of time. The Survival Function is always decreasing, but any sudden drops indicate a point of interest. In this example, the Survival Function decreases at a mostly steady rate. The Hazard Function is an aggregate of all of the subhazards and represents the overall hazard probability for churn. Again, the time interval that you use can make a difference in how you visualize the data.

If the **High Risk Account Tables** option was changed from **No**, then you can click **View >> Model**, and choose the High Risk table that you want to inspect.

The Model Validation plot and Model Validation Statistics table can help you see how well the survival model is fitting your data.

After you choose a survival model that you like, you must make sure that the scoring data has the _t_ variable set such that the time interval used to create the _t_ variable matches the **Time Interval** property in the Survival node that you are using to score. For this example, the _t_ variable in SAMPSIO.CELLPHONE_SCORE was created with months, and the Survival node that is used to score these data also has the **Time Interval** property set to **Month**.

To continue the flow, both the SAMPSIO.CELLPHONE_SCORE scoring data and the Survival (Months) node are connected to a Score node. The Score node uses the previously modeled hazard and survival functions to score the scoring data, enabling you to inspect which new customers might have a high risk of churn during the time that spans the desired number of forecast intervals. You can specify the number of forecast intervals under the property **Number of Forecast Intervals** if you choose **No** for the property **Default Forecast Intervals**. The default of three months is used here.

In this example, the SAS Code node (which enables you to incorporate custom SAS code into SAS Enterprise Miner diagrams) is used to further inspect which customers have the lowest survival probability.   This diagram uses SAS code to print, in the results of the SAS Code node, the 100 customers who have the lowest survival probability three months in the future (from the current date that is used to generate _t_). This code can be edited to display whatever statistic you want; for example, you could use the variable EM_SURVEVENT instead to show which customers are most likely to have an event during the time interval from the current date to the forecast date.