SURVIVAL DATA MINING FOR BIG DATA: PRACTITIONER'S GUIDE

ISF 2013 – 24 JUNE 2013

Tae Yoon Lee
Principal Analytical Consultant
TaeYoon.Lee@sas.com

CONTENTS

This presentation is to explain about the Methodology of Survival Data Mining.

- What is Survival Data Mining?
- When do we use Survival Data Mining?
- How can we implement Survival Data Mining?
  - Models
  - Estimation
- How to use SAS EM Survival Data Mining?

WHY?

- When the next event will occur
- Time dependent outcomes
- Discrete event time
- Time-dependent covariates
- Time-varying covariates
- Competing risks

PROGRESSING IN SURVIVAL ANALYSIS AND DATA MINING

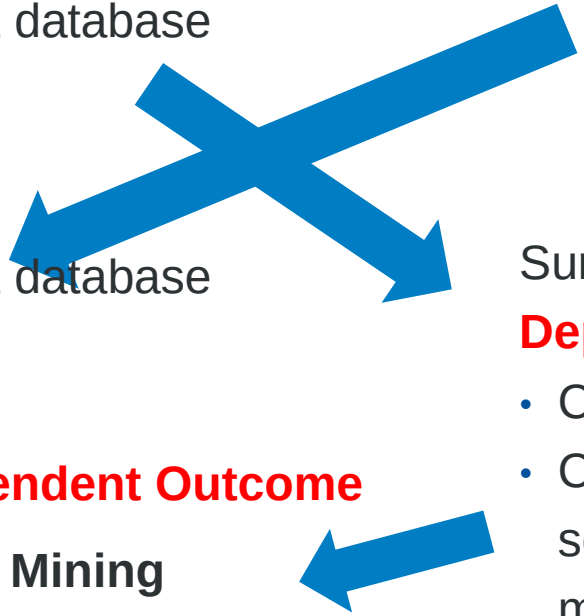
Survival Analysis
- Medical Patient database
- Death event

Survival Analysis
- Medical Patient database
- Death event
- NN, DT, LR
- **Not Time-Dependent Outcome**

Survival Data Mining
- Medical Patient database
- Death event

Data Mining for predictive models
- Commercial Customer database
- Credit Scoring

Survival Data Mining (**Time-Dependent Outcome**)
- Commercial Customer database
- Customer retention, cross selling, other database marketing endeavors



Survival Analysis

- Censored duration data
- Predict a probability that an event will occur within a predefined time window.

Survival Data Mining

- Ultimate purpose of data mining is Prediction, so Predictive Scoring!
- Predict a time-dependent outcome, i.e. model the event likelihood over time.
- 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 (for better long-term strategies.)

WHAT IS SURVIVAL DATA MINING?

- Survival Data Mining = Survival Analysis + Data Mining
  - Predicts when the next event will occur – survival data mining
  - Not whether an event will occur in a certain time interval – survival analysis
  - Key is to forecast survival patterns into the future by extrapolating survival probabilities beyond the time window of available data – predictive scoring in data mining.

- Survival Analysis
  - Parametric models based on PROC LIFEREG – pick up best model / no time dependent covariates / Y=survival time (t or log(t)) / Y= f(time independent covariates).
  - Discrete-time logistic-hazard model using PROC LOGISTIC. Log (ht/ (1- ht)) = f(time, time dependent & time-varying covariates)

- Apply Data mining method to discrete-time logistic-hazard model (DTLHM)
  - Because this model is well suited to the challenging features of survival data mining problems

DISCRETE EVENT TIME (DET)

Discrete Event Times are represented by

- Nonnegative integer values and nonlinear
- Cubic Spline Basis Functions (CSBF)
- used as predictors in the multinomial Logistic Regression

Transforming DET function with CSBF allows the Hazard and Subhazard functions to be more flexible. This results in a greater ability to detect & model Survival Pattern.

SAS ENTERPRISE MINER SURVIVAL NODE USES:

Discrete-Time Logistic-Hazard Model with Cubic Spline Base Functions

$$ln\left[\frac{h(t,m|x)}{1-h(t|x)}\right]=\alpha_{0m}+\left[\alpha_{00}+\alpha_0 t+\sum_{j=1}^{\#knots}\alpha_j\, csb\left(t,k_j\right)\right]+\beta_{1m}x_1+\cdots+\beta_{pm}x_p$$

$$where\quad csb\left(t,k_j\right)=\begin{cases}-t^3+3k_jt^2-3{k_j}^2t & if\ t\le k_j \qquad cubic\\ -{k_j}^3 & if\ t>k_j \qquad constant\end{cases}$$

- No Time Dependent covariates
- No time-varying effects

SAS ENTERPRISE MINER SURVIVAL NODE USES:

For Time Dependent Covariates and Time-Varying Effects

$$ln\left[\frac{h(t,m|x)}{1-h(t|x)}\right]=$$

$$\alpha_{0m}+\left[\alpha_{00}+\alpha_0 t+\sum_{j=1}^{\#knots}\alpha_j\ csb\left(t,k_j\right)\right]+\beta_{1m}(t)x_1(t)+\cdots+\beta_{pm}(t)x_p(t)$$

OVERVIEW OF DISCRTE EVENT TIME LOGISTIC HAZARD MODELING IN SAS ENTERPRISE MINER SURVIVAL NODE

- Account ID
- Time ID
- Time Interval
- Covariates
- Target

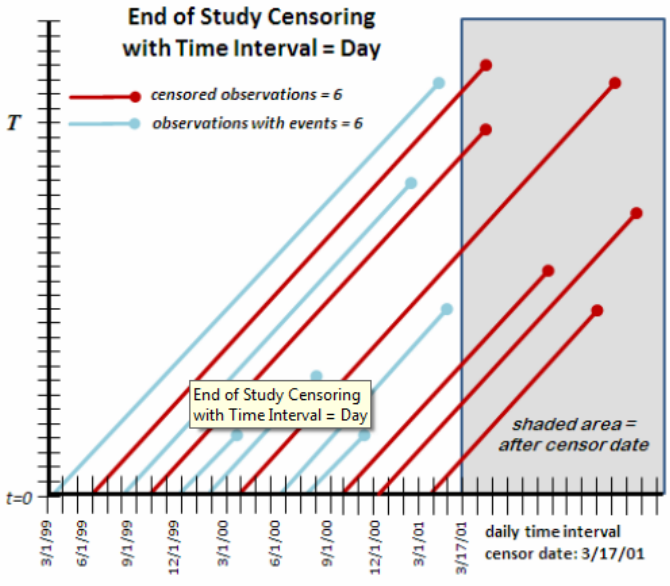
- One observation for each unique patient at each DET

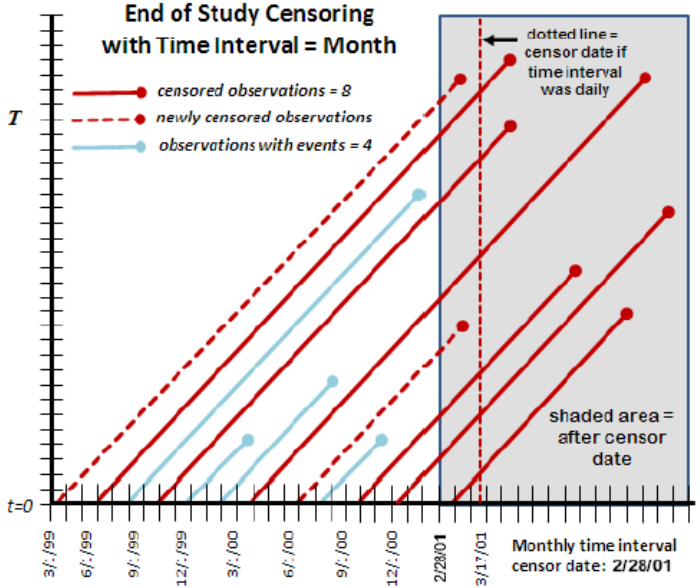
- Case-control sampling

- MRL
- Hazard
- Subhazrd
- Event occurrence probabilty

1. Data Preparation

- Account ID
- Time ID
- Time Interva
- Covariates
- Target





2. Expand Data for Survival Analysis

- One observation for each unique patient at each DET

	Account Number	Event Time	Event Type	Good Bad Credit Indicator	Disable Reason	Type of Rate Plan	Activation Date	Deactivation Date	Provider Type	ctivation Date	Provider Type
1	180437080184	16	0	1		3	09/28/1999	.	PROV1	.	PROV1
2	180437283474	0	0	1		1	01/09/2001	.	PROV1	.	PROV1
3	180437340410	13	0	0		1	12/31/1999	.	PROV1	.	PROV1
4	180437356568	6	2	0	DUE	1	12/22/1999	06/28/2000	PROV2	06/28/2000	PROV2
5	180437356837	9	0	1		1	04/17/2000	.	PROV3	.	PROV3
6	180437375280	12	1	1	TRANSFER	2	08/16/1999	08/21/2000	PROV1	08/21/2000	PROV1
7	180437392909	18	0	1		1	07/26/1999	.	PROV3	.	PROV3
8	180437420657	13	0	0		1	12/15/1999	.	PROV2	.	PROV2
9	180437433673	2	0	0		3	11/21/2000	.	PROV1	.	PROV1
10	180437452331	1	0	0		2	12/28/2000	.	PROV3	.	PROV3

	Account Number	Event Time	Discrete Event Time	Event Type	Good Bad Credit Indicator	Disable Reason	Type of Rate Plan	Activation Date	Deactivation Date	Provider Type
1	180437080184	16	0	0	1		3	09/28/1999	.	PROV1
2	180437080184	16	1	0	1		3	09/28/1999	.	PROV1
3	180437080184	16	2	0	1		3	09/28/1999	.	PROV1
4	180437080184	16	3	0	1		3	09/28/1999	.	PROV1
5	180437080184	16	4	0	1		3	09/28/1999	.	PROV1
6	180437080184	16	5	0	1		3	09/28/1999	.	PROV1
7	180437080184	16	6	0	1		3	09/28/1999	.	PROV1
8	180437080184	16	7	0	1		3	09/28/1999	.	PROV1
9	180437080184	16	8	0	1		3	09/28/1999	.	PROV1
10	180437080184	16	9	0	1		3	09/28/1999	.	PROV1
11	180437080184	16	10	0	1		3	09/28/1999	.	PROV1
12	180437080184	16	11	0	1		3	09/28/1999	.	PROV1
13	180437080184	16	12	0	1		3	09/28/1999	.	PROV1
14	180437080184	16	13	0	1		3	09/28/1999	.	PROV1
15	180437080184	16	14	0	1		3	09/28/1999	.	PROV1
16	180437080184	16	15	0	1		3	09/28/1999	.	PROV1
17	180437080184	16	16	0	1		3	09/28/1999	.	PROV1

3. Sample the Expand Data

- Case-control (or Choice-based) sampling is used, which is a well-known method that is used to handle rare categorical outcomes.
- Choose all events and sample from non-events
- To correct the sampling bias, the subhazard functions are adjusted after the model built.

4. Configure and Run Survival Model

Discrete-Time Logistic-Hazard Model with Cubic Spline Base Functions

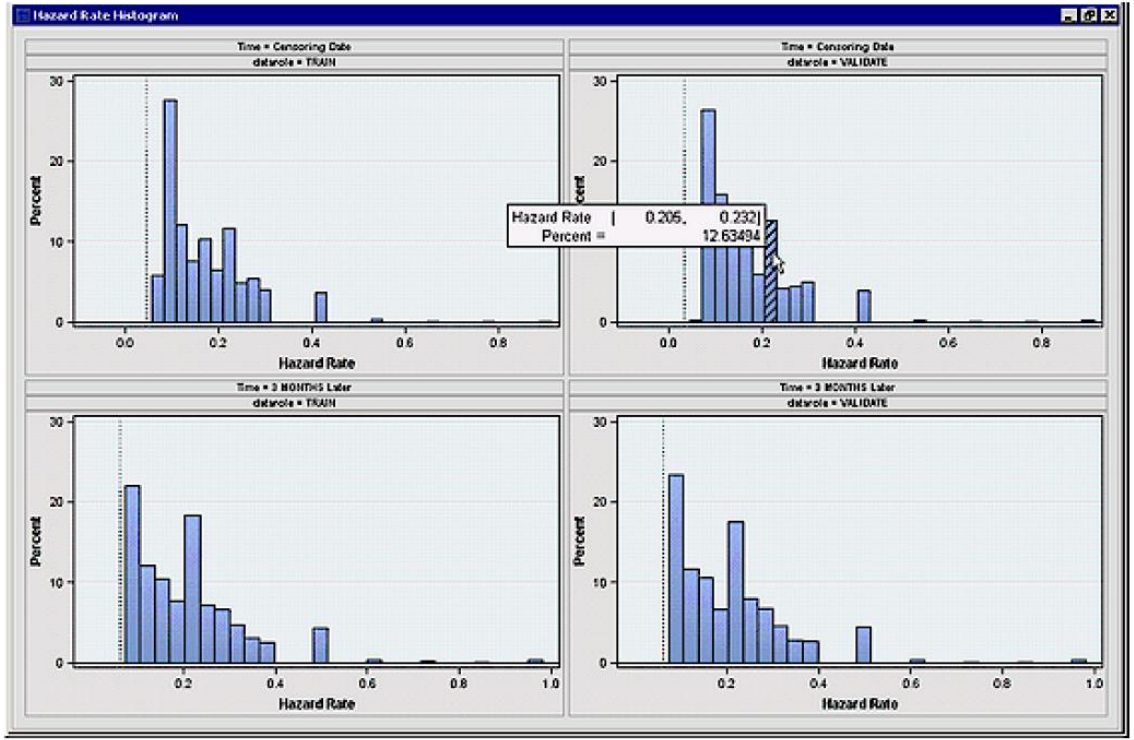
$$ln\left[\frac{h(t,m|x)}{1-h(t|x)}\right]=\alpha_{0m}+\left[\alpha_{00}+\alpha_0 t+\sum_{j=1}^{\#knots}\alpha_j\, csb\left(t,k_j\right)\right]+\beta_{1m}x_1+\cdots+\beta_{pm}x_p$$

Examine and Interpret the output results:

- Mean Residual Life
- Hazard Rate Histogram
- Event Occurrence Probability Histogram
- Survival Probability Histogram

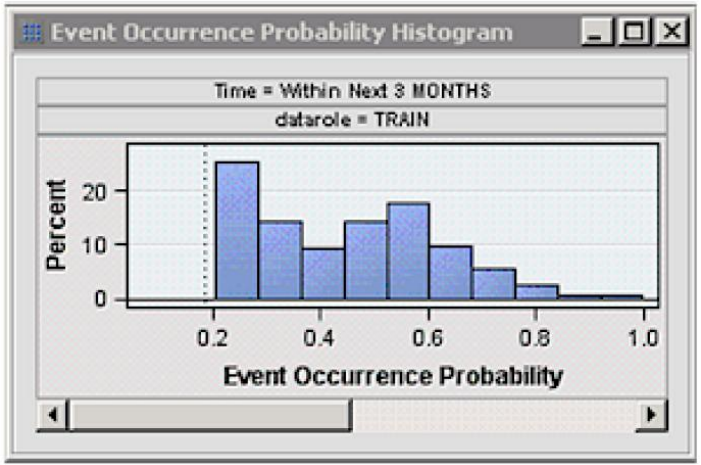
Examine and Interpret the output results:

Hazard Rate Histogram (Time=Censoring Date and Time=3 Time Units Later):



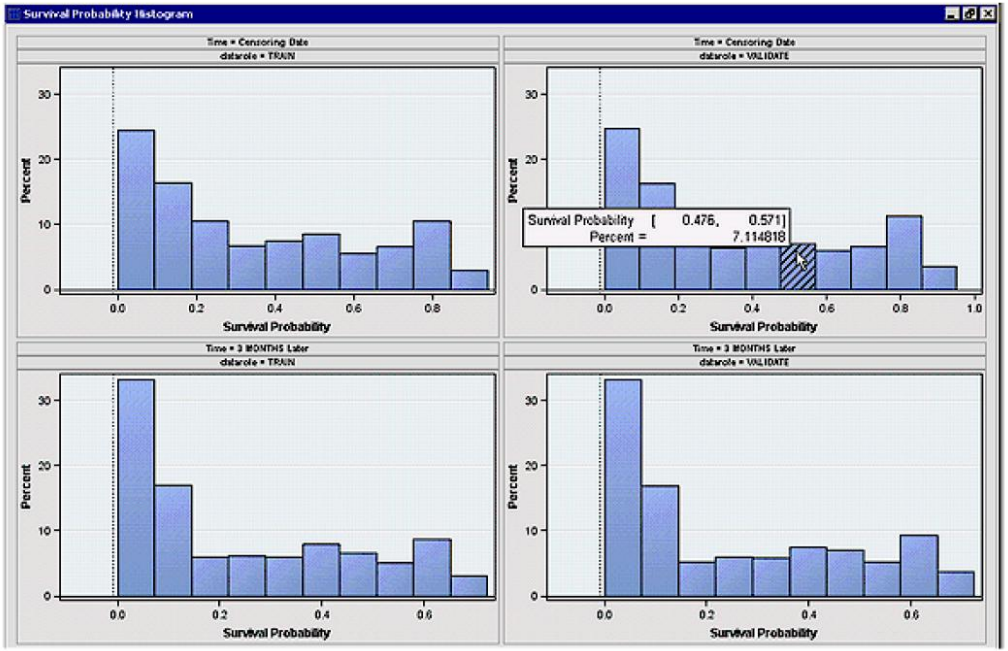
Examine and Interpret the output results:

Event Occurrence Probability Histogram (within next 3 time units):



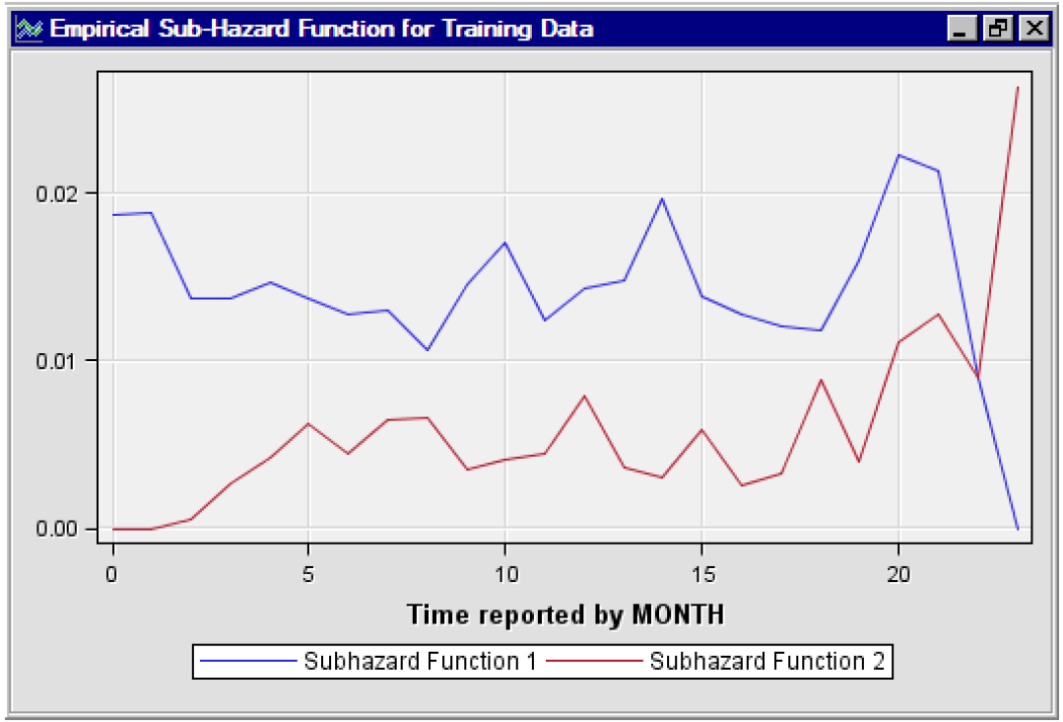
Examine and Interpret the output results:

Survival Probability Histogram (Time=Censoring Date and Time=3 Time Units Later):



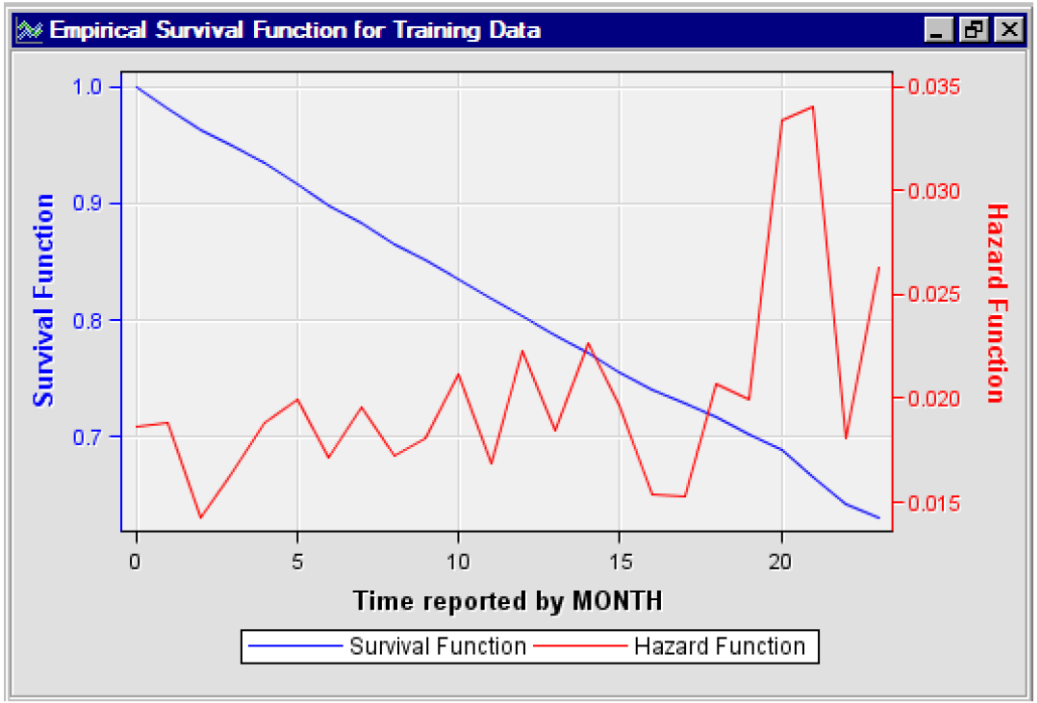
Examine and Interpret the output results:

Empirical Subhazard Function for Training Data:



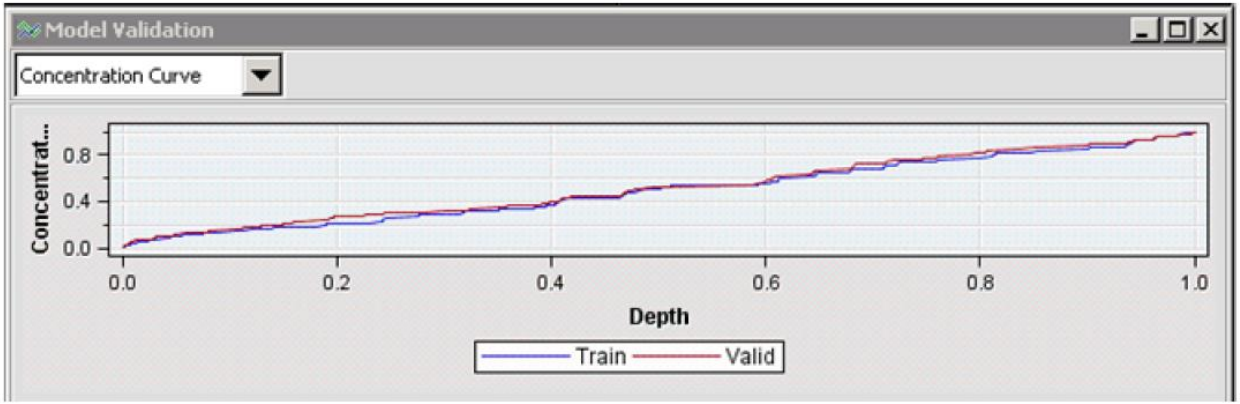
Examine and Interpret the output results:

Empirical Survival Function for Training Data:



Examine and Interpret the output results:

Model Validation Plots



- Concentration Curve
- Lift
- Benefit

5. Predictive Scoring

```
/*-------------------------------------------------*/
/*Survival Score Code*/
/*-------------------------------------------------*/

length _warn_ $ 4;
label _warn_ = "Warnings";
label EM_SURVIVAL = "Survival Probability at Censoring Time";
label EM_SURVFCST = "Survival Probability at Future Time";
label EM_SURVEVENT = "Event Probability before or at the Future Time";
label EM_HAZARD = "Hazard Function at Censoring Time";
label EM_HZRDFCST = "Hazard Function at Future Time";
length _uname $32;
DROP _uname;

if _T_ ne . then do;
T_FCST=_T_+5 ;


/*----------Generate Cubic Spline Basis Functions-------------*/
if _T_ > 8  then
_CSB2 =(_T_-8)**3 - _T_**3 + 24*_T_**2 - 192*_T_;
else
_CSB2=-_T_**3 + 24*_T_**2 - 192*_T_;
```

*Currently only available for Teradata and EMC Greenplum.

** Only available via SAS Enterprise Miner Node

Experimental procedure in HPAS 12.1

5. Predictive Scoring

```
 EM_SUBHZRD2_SURV=
exp(log(EM_SUBHZRD2_SURV/EM_SUBHZRD0_SURV)+log(0.0749796251))/_denom;
 end;
 else do;
  EM_SUBHZRD2_SURV=
exp(log(EM_SUBHZRD2_SURV/0.0001)+log(0.0749796251))/_denom;
 end;
end;
else do;
  EM_SUBHZRD2_SURV= exp(0.0749796251)/_denom;
end;
EM_HZRDFCST=0 + EM_SUBHZRD1_SURV + EM_SUBHZRD2_SURV;
EM_SUBHZRD0_SURV = 1-(0 + EM_SUBHZRD1_SURV + EM_SUBHZRD2_SURV);

EM_SURVFCST=EM_SURVFCST*(1-EM_HZRDFCST);
if _T_=_T0_ then EM_SURVIVAL=EM_SURVFCST;
if EM_SURVIVAL >0 then do;
if _T_=t0_fcst then EM_SURVEVENT=(EM_SURVIVAL-
EM_SURVFCST)/EM_SURVIVAL;
end;
else do;
if _T_=t0_fcst then EM_SURVEVENT=(EM_SURVIVAL-EM_SURVFCST)/0.00001;
end;
```

*Currently only available for Teradata and EMC Greenplum.

** Only available via SAS Enterprise Miner Node

Experimental procedure in HPAS 12.1

Any Questions?

THANK YOU!

ISF 2013 – 24 JUNE 2013

Tae Yoon Lee
Principal Analytical Consultant
TaeYoon.Lee@sas.com

SURVIVAL DATA MINING

SURVIVAL ANALYSIS & DATA MINING CLASSIFICATION