Data Mining with SAS

Mathias Lanner
mathias.lanner@swe.sas.com

Agenda

- Data mining Introduction
- Data mining applications
- Data mining techniques
- SEMMA mythology
- Survival data mining
- Time series data mining

What is data mining?

”Data mining uses sophisticated statistical analysis and modeling techniques to uncover patterns and relationships hidden in organizational databases-patterns that ordinary methods might miss.”
-Two Crows Corporation (1998),p.1

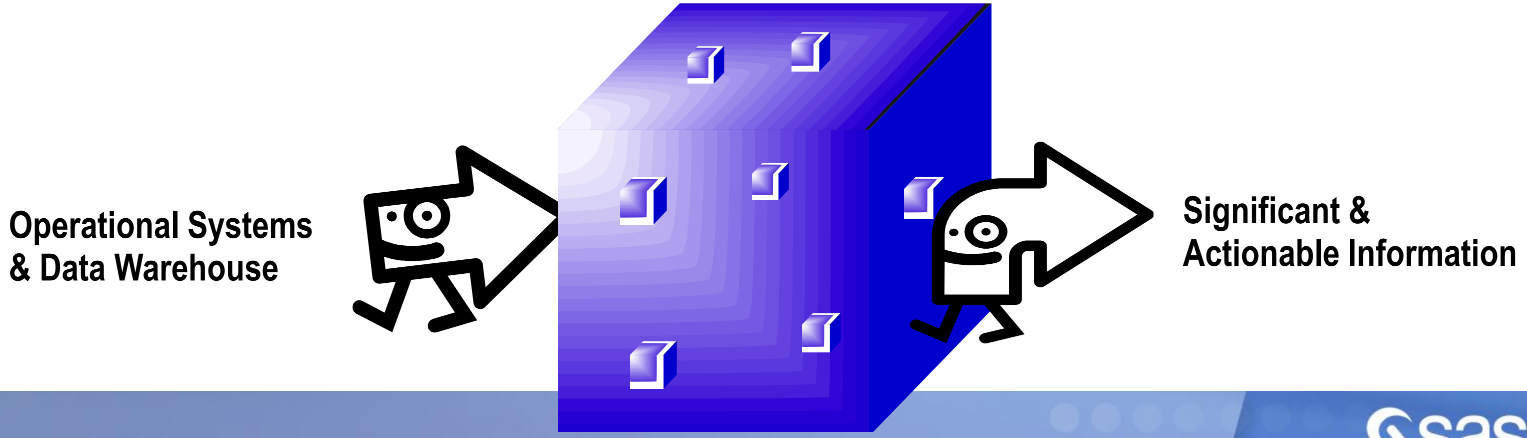
”Data Mining, as we use the term, is the exploration and analysis by automatic or semiautomatic means, of large quantities of data in order to discover meaningsful patterns and rules.”
-Berry and Linoff(1997), p.5

”Data Mining [is] the process of efficient discovery of nonobvious valuble information from large collection of data.”
-Berson and Smith (1997), p.565

”Data Mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognation technologies as well as statistical and mathematical techniques.”
-Erick Brethnoux, Gartner Group

Data Mining Definition :

The *process of selecting, exploring, and modeling* large amounts of data to uncover previously unknown information for a *business advantage*

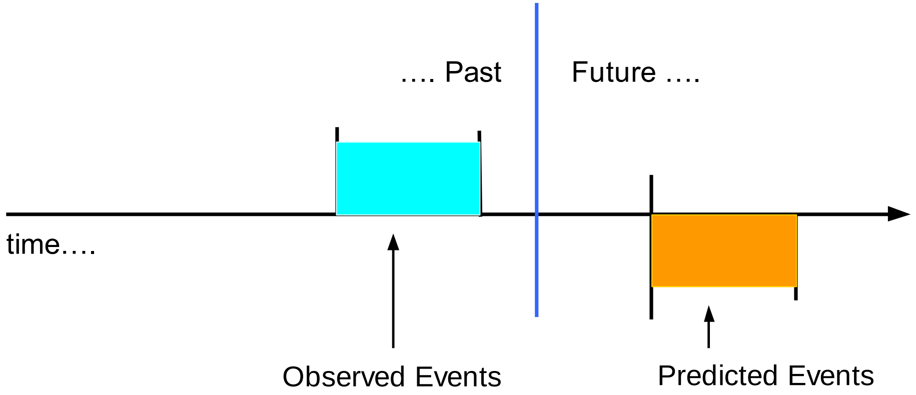


Operational Systems & Data Warehouse

Significant & Actionable Information

Data Mining Is:

- Discovering patterns, trends and relationships represented in data
- Developing models to understand and describe characteristics and activity based on these patterns
- Use insights to help evaluate future options and take fact-based decisions
- Deploy scores and results for timely, appropriate action



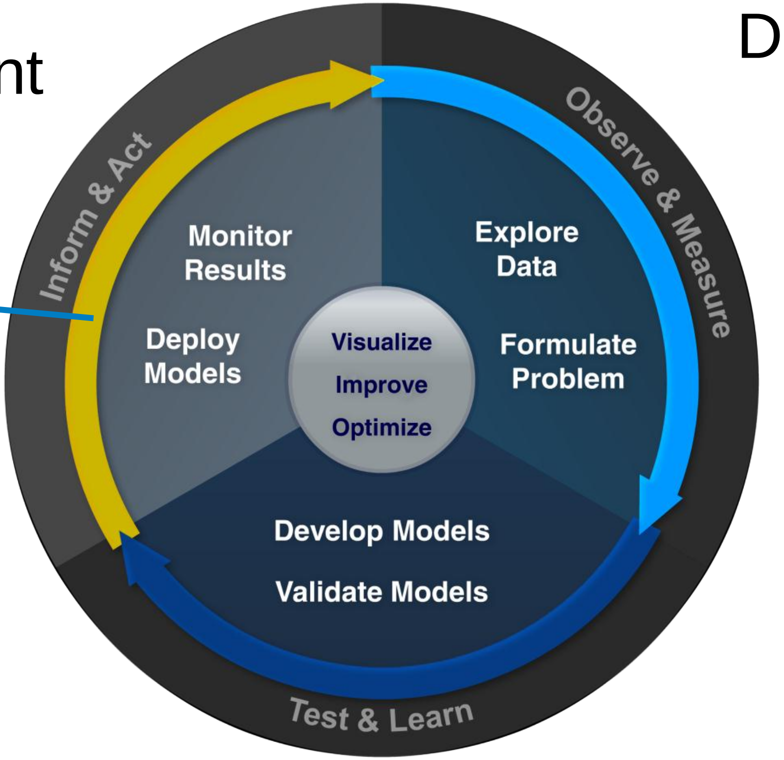
Predictive Analytics and Data Mining

Key Components

Model Deployment

Discovery

Model Management & Monitoring



Model Development

Cross-Industry Data Mining Applications

Customer Analytics

Application	What is Predicted?	Driven Business Decision
Profiling and Segmentation	Customer’s behaviors and needs by segment	How to create better-targeted product/service offers?
Cross-sell and Up-Sell	Identify what will customer’s buy?	Which product/service to recommend?
Acquisition and Retention	Customer’s preferences and purchase patterns	How to grow and maintain valuable customers?
Campaign Management	Evaluate the success of customer communications	How to direct right offer to right person at the right time?
Profitability and Life-time Value	Understand the drivers of future value (margin and retention)	Identify economically valuable channels/demographics and incremental benefits?

Industry Specific Data Mining Applications

Application	What is Predicted?	Driven Business Decision
Credit Scoring *(Banking)*	Measure credit worthiness of new and existing set of customers	How to assess and control risk within existing (or new) consumer portfolios?
Market Basket Analysis *(Retail)*	Which products are likely to purchased together?	How to increase sales with cross-sell/up-sell, loyalty programs, promotions?
Asset Maintenance *(Utilities, Mfg., Oil & Gas)*	Identify real drivers of asset or equipment failure	How to minimize operational disruptions and maintenance costs?
Health & Condition Mgmt. *(Health Insurance)*	Identify patients at risk of a chronic illness & offer treatment program	How can we reduce healthcare costs and satisfy patients?
Fraud Mgmt. (Govt., *Insurance, Banks)*	Detect unknown fraud cases and future risks	How to decrease fraud losses and lower false positives?
Drug Discovery *(Life Science)*	Find compounds that have desirable effects & detect drug behavior during trials	How to bring drugs quickly and effectively to the marketplace?

Data mining techniques - Connect business problems with the right analytical technique

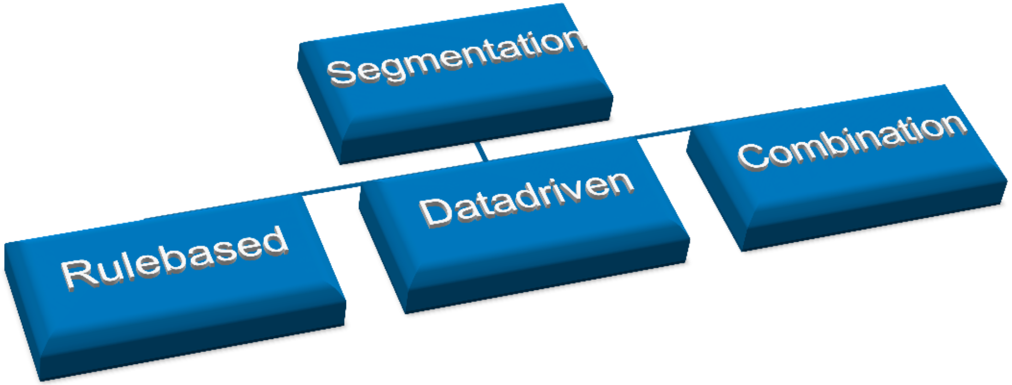
- Basket-/Sequence analysis
  - Find the association between products purchased together or sequentially
- Clustering/Segmentation
  - Divide the customers into different groups for different campaigns
- Predictive modeling
  - Classification – Find potential buyers
  - Prediction – Predict the future value

Predictiv Modeling- how does it works

- Predict how someone / something will behave (prediction / scoring)
  - Will this customer respond to the offer?
  - Will this borrower manage their interest payments?
- By starting from how other individuals / units behaved
  - Database – loyalty card
    - » Information about customers (age, gender, buying amount…)
    - » Information about their previous responds
  - Database – Credit application
    - » Information about customers (age, gender, income, occupation, etc.)
    - » Information about how they have performed theirs interest payments

Clustring/Segmentation

- A strategic map of the customer base
- Communicate it into the business
- The division can be done in different ways
- What information will be used to generate the segments

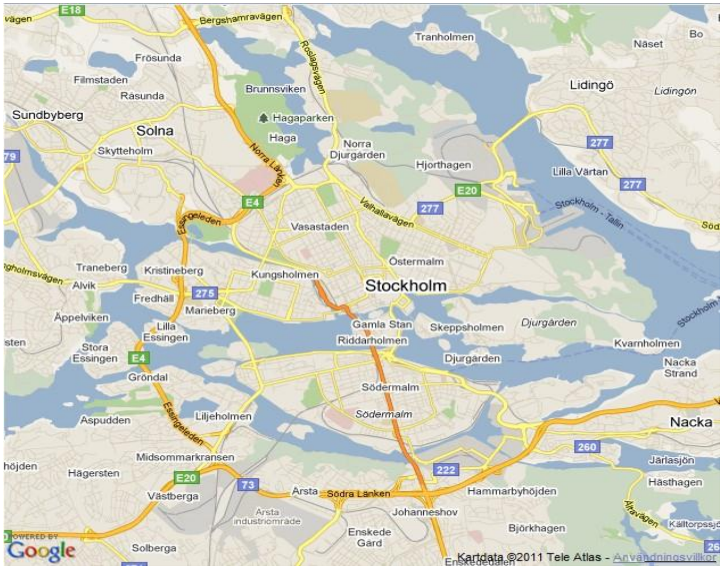


Segmentation

Rulebased

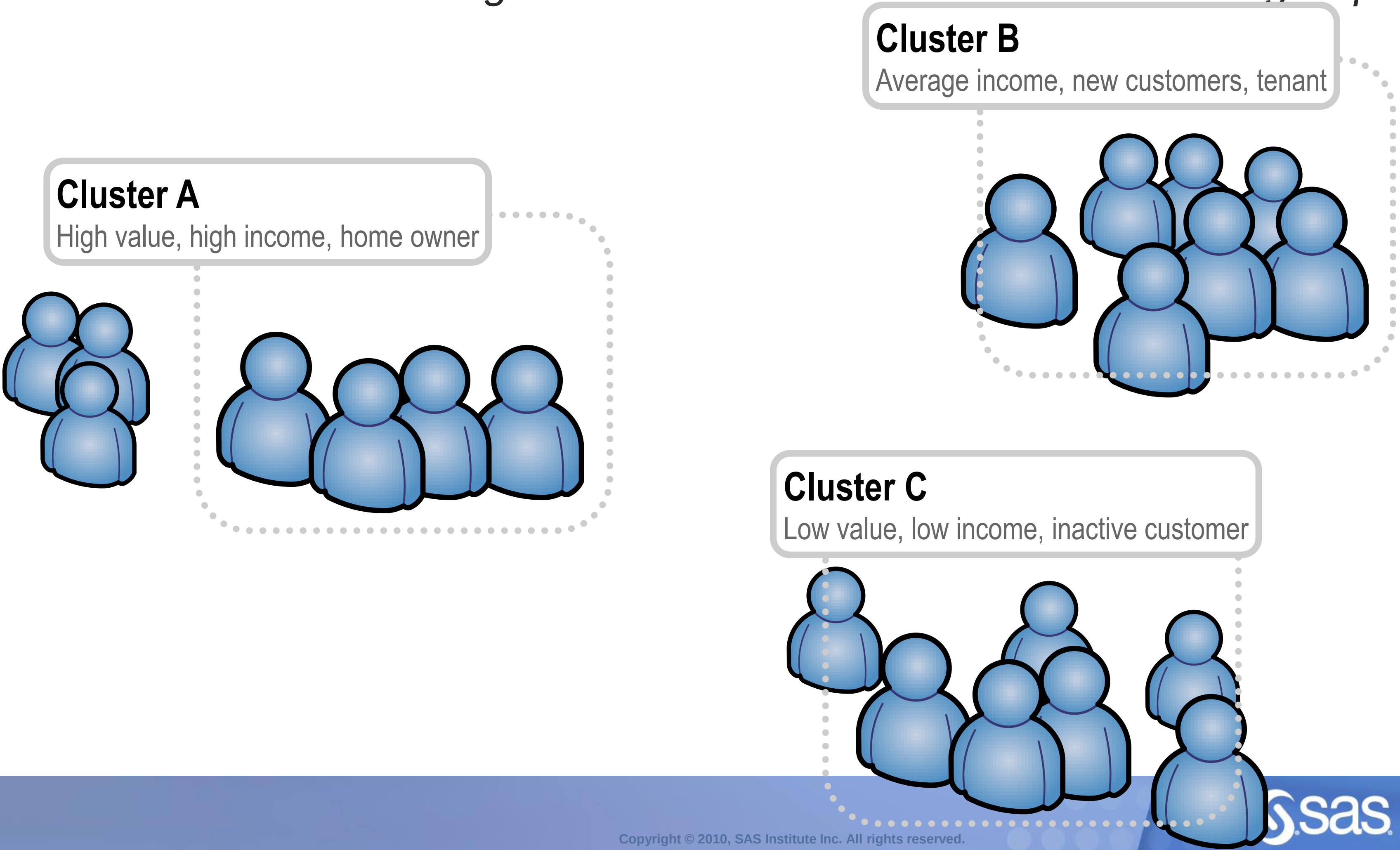
Datadriven

Combination



Clustring/Segmentation

“Create groups that have similar characteristics - also provides a measure of how big the difference is between the different groups



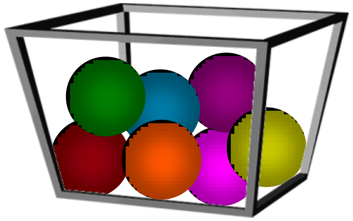
Cluster A
High value, high income, home owner

Cluster B
Average income, new customers, tenant

Cluster C
Low value, low income, inactive customer

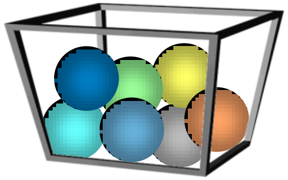
Association / Sequence Analysis

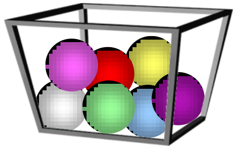
“Identify events that occur in association with each other, possibly in a particular order "



What products customers purchase together or in a certain sequence?

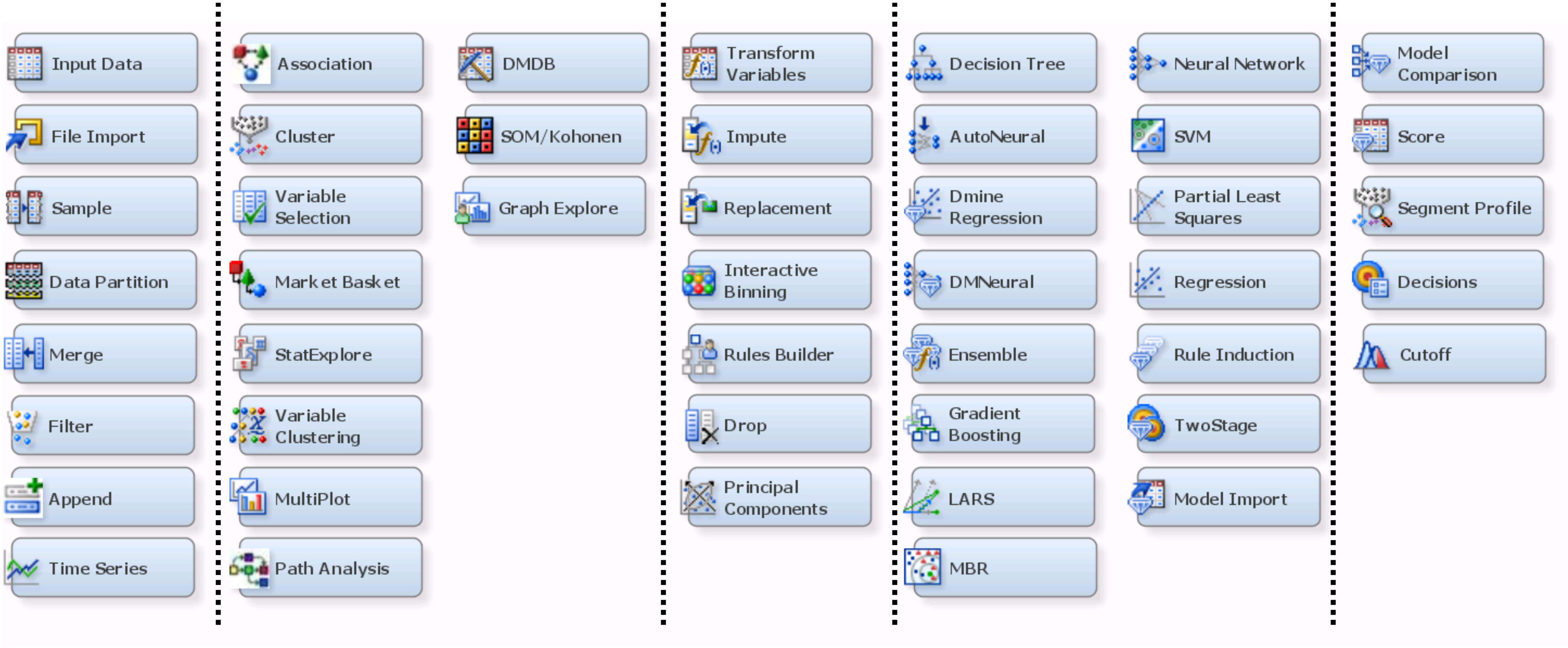
Can we predict what you will have in your shopping cart?





SAS® Enterprise Miner™ 7.1

Model Development Process

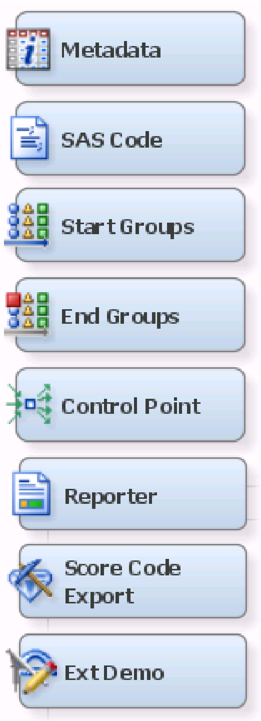


Sample	Explore		Modify	Model		Assess
Input Data	Association	DMDB	Transform Variables	Decision Tree	Neural Network	Model Comparison
File Import	Cluster	SOM/Kohonen	Impute	AutoNeural	SVM	Score
Sample	Variable Selection	Graph Explore	Replacement	Dmine Regression	Partial Least Squares	Segment Profile
Data Partition	Market Basket		Interactive Binning	DMNeural	Regression	Decisions
Merge	StatExplore		Rules Builder	Ensemble	Rule Induction	Cutoff
Filter	Variable Clustering		Drop	Gradient Boosting	TwoStage	
Append	MultiPlot		Principal Components	LARS	Model Import	
Time Series	Path Analysis			MBR		

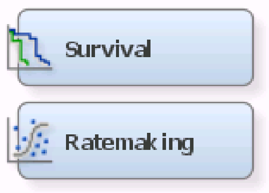
SAS® Enterprise Miner™ 7.1

Model Development Process

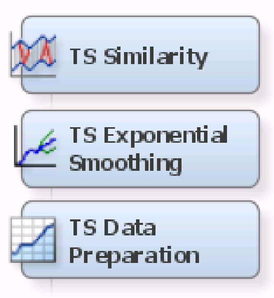
Utility



Applications

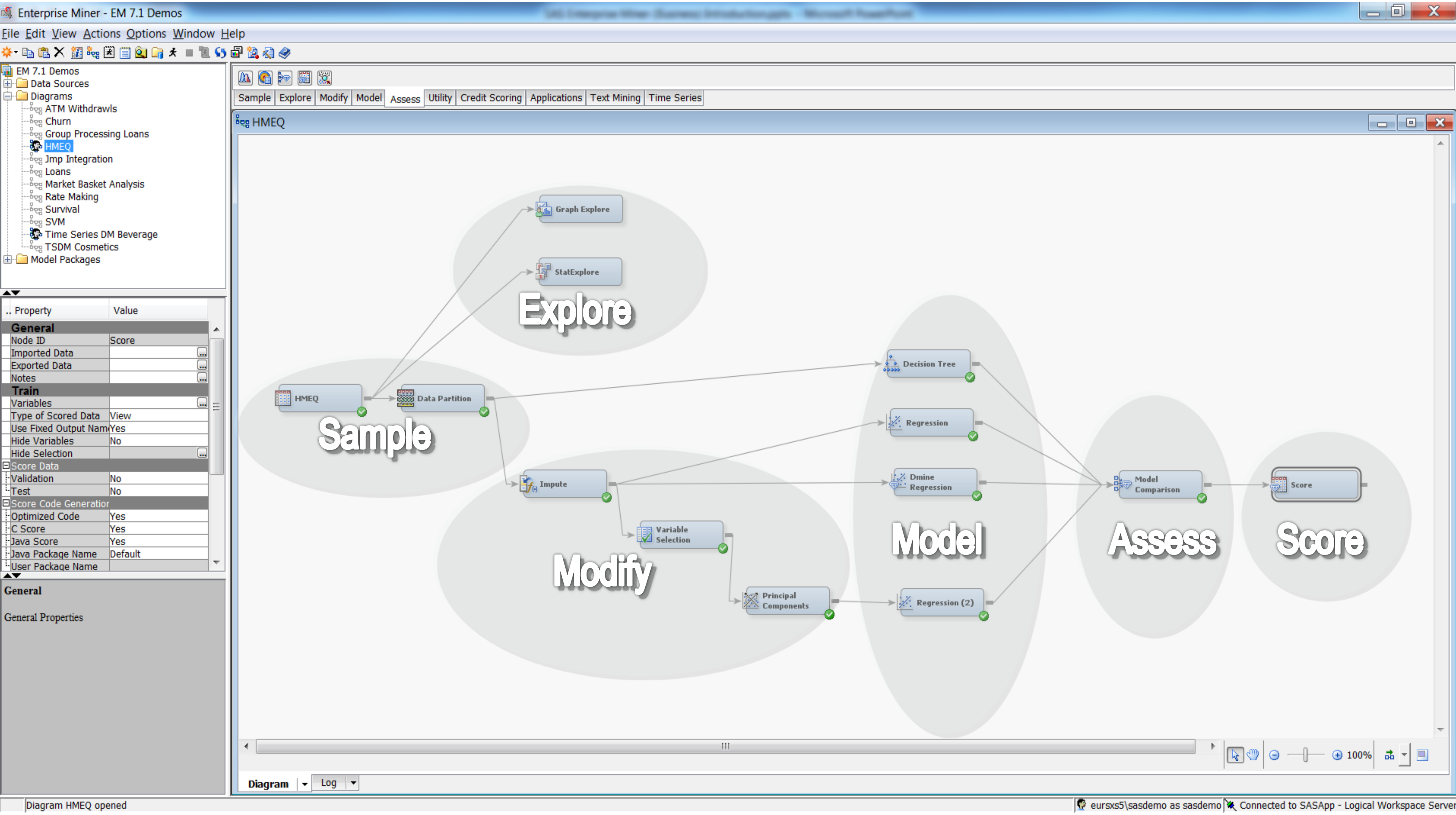


Time Series



SAS® Enterprise Miner™ 7.1

SEMMA in Action – Repeatable Process



- EM 7.1 Demos
  - Data Sources
  - Diagrams
    - ATM Withdrawls
    - Churn
    - Group Processing Loans
    - HMEQ
    - Jmp Integration
    - Loans
    - Market Basket Analysis
    - Rate Making
    - Survival
    - SVM
    - Time Series DM Beverage
    - TSDM Cosmetics
  - Model Packages

.. Property	Value
General	
Node ID	Score
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Type of Scored Data	View
Use Fixed Output Nam	Yes
Hide Variables	No
Hide Selection	...
Score Data	
Validation	No
Test	No
Score Code Generatior	
Optimized Code	Yes
C Score	Yes
Java Score	Yes
Java Package Name	Default
User Package Name	

General

General Properties

Diagram HMEQ opened

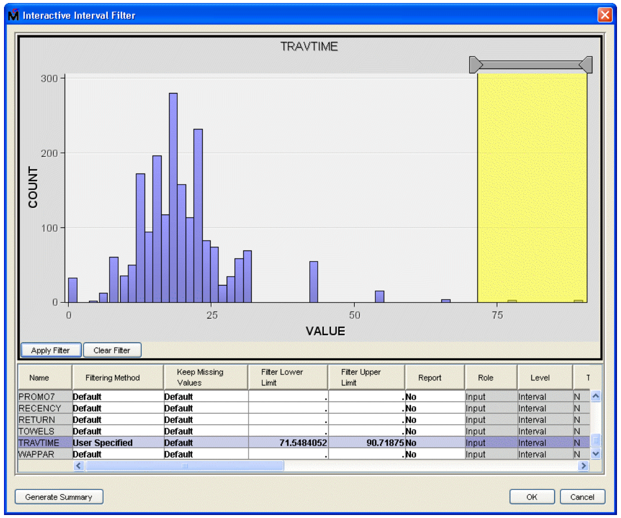
eursxs5\sasdemo as sasdemo

Connected to SASApp - Logical Workspace Server

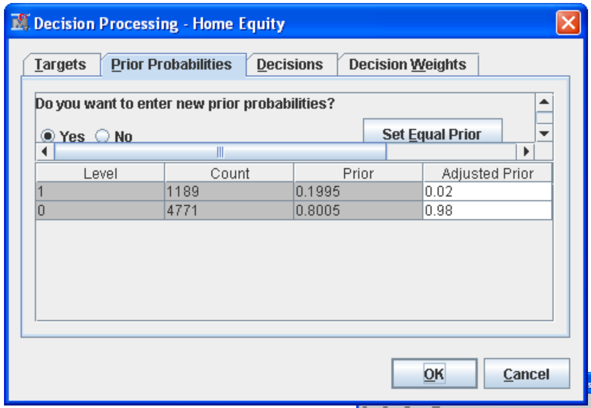
Sample and Explore

- Data selection
  - Required & excluded fields
  - Sample balancing
  - Data partitioning

- Data evaluation
  - Statistical measures
  - Visualization
  - Identifying outliers
  - Analytical segmentation
  - Variable creation & selection



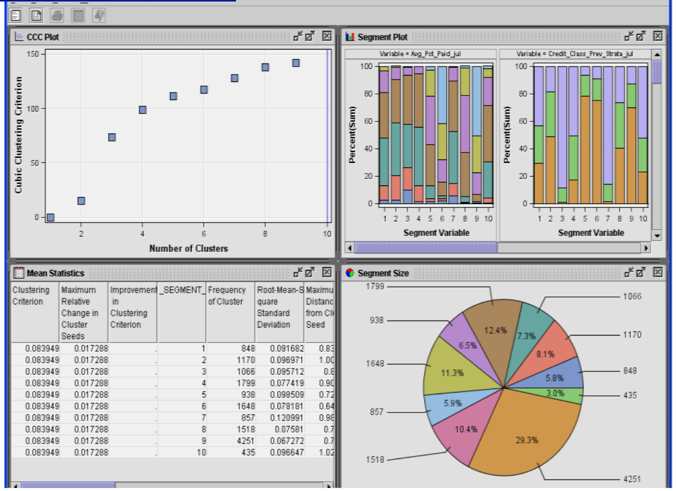
Name	Filtering Method	Keep Missing Values	Filter Lower Limit	Filter Upper Limit	Report	Role	Level	T
PROMO7	Default	Default	.	.	No	Input	Interval	N
RECENCY	Default	Default	.	.	No	Input	Interval	N
RETURN	Default	Default	.	.	No	Input	Interval	N
TOWELS	Default	Default	.	.	No	Input	Interval	N
TRAVTIME	User Specified	Default	71.5484052	90.71875	No	Input	Interval	N
WAPPAR	Default	Default	.	.	No	Input	Interval	N



Do you want to enter new prior probabilities?

◉ Yes ○ No

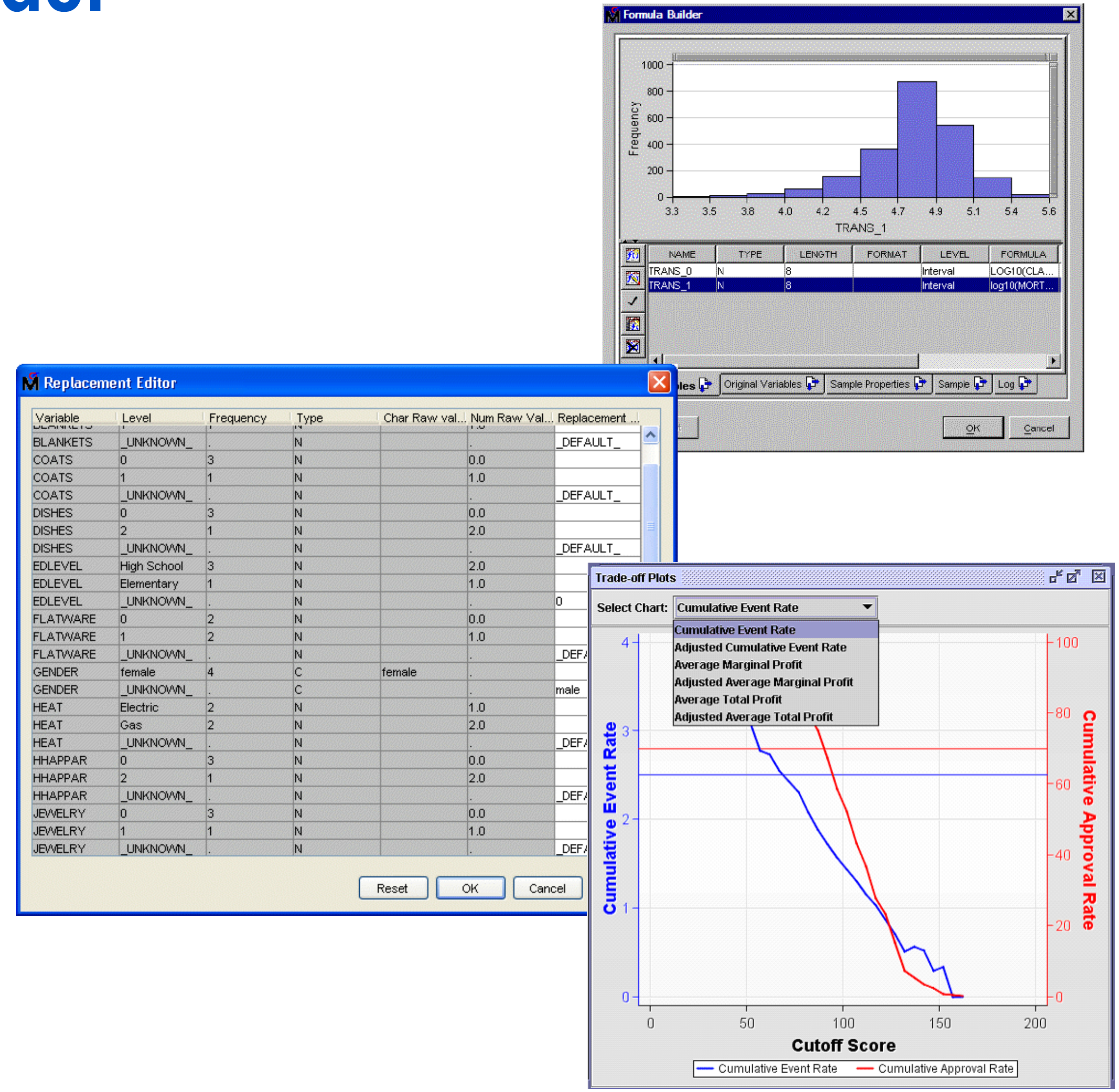
Level	Count	Prior	Adjusted Prior
1	1189	0.1995	0.02
0	4771	0.8005	0.98



Clustering Criterion	Maximum Relative Change in Cluster Seeds	Improvement in Clustering Criterion	_SEGMENT_	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximu Distanc from Cl Seed
0.083949	0.017288	.	1	848	0.091682	0.83
0.083949	0.017288	.	2	1170	0.096971	1.00
0.083949	0.017288	.	3	1066	0.095712	0.8
0.083949	0.017288	.	4	1799	0.077419	0.90
0.083949	0.017288	.	5	938	0.098509	0.72
0.083949	0.017288	.	6	1648	0.078181	0.64
0.083949	0.017288	.	7	857	0.120991	0.96
0.083949	0.017288	.	8	1518	0.07581	0.7
0.083949	0.017288	.	9	4251	0.067272	0.7
0.083949	0.017288	.	10	435	0.096647	1.02

Modify and Model

- Binning
- Scaling
- Imputation
- Replacement/recoding
- Modeling Policies
  - Prediction functions
  - Classification functions
- Modeling Methods



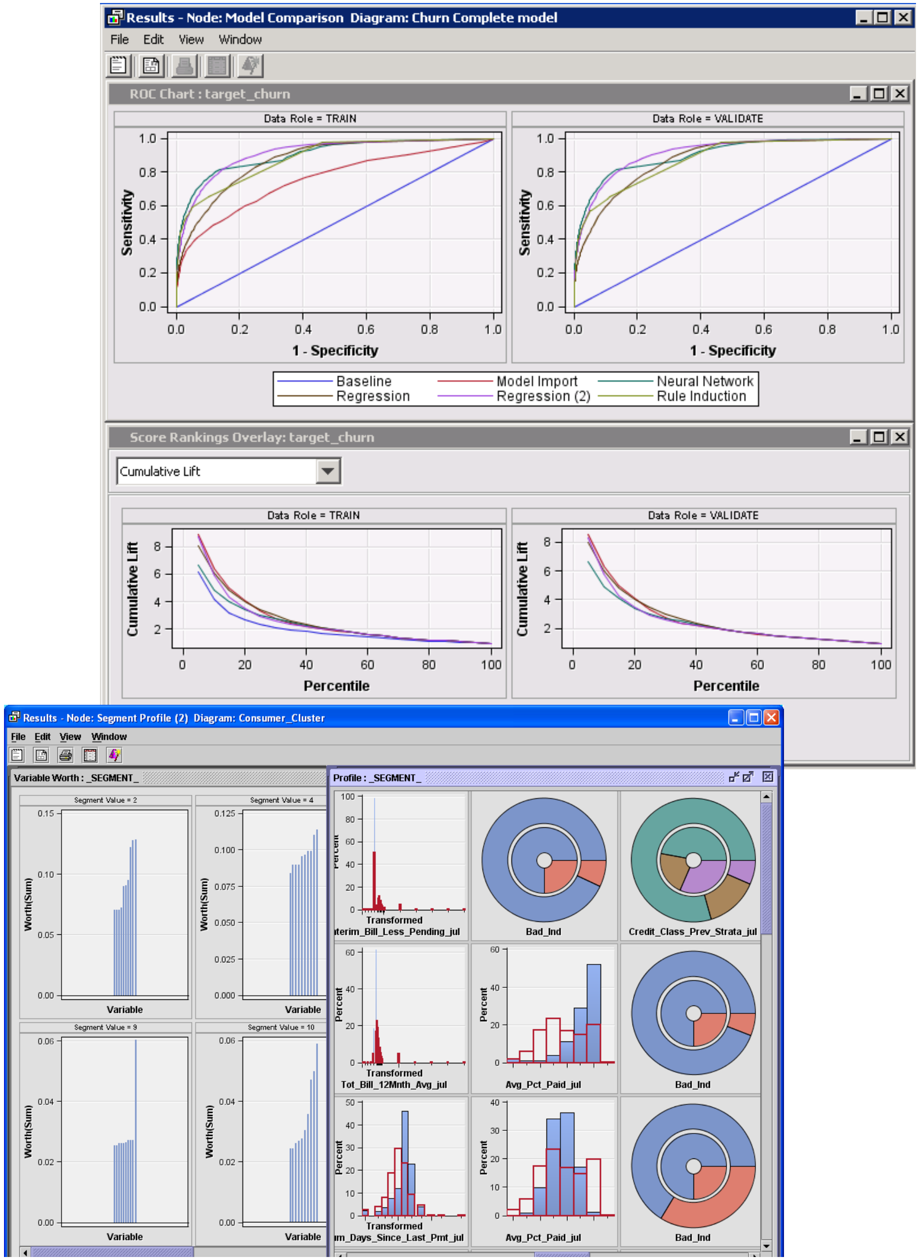
NAME	TYPE	LENGTH	FORMAT	LEVEL	FORMULA
TRANS_0	N	8		Interval	LOG10(CLA...
TRANS_1	N	8		Interval	log10(MORT...

Replacement Editor

Variable	Level	Frequency	Type	Char Raw val...	Num Raw Val...	Replacement ...
BLANKETS	_UNKNOWN_	.	N		.	_DEFAULT_
COATS	0	3	N		0.0	
COATS	1	1	N		1.0	
COATS	_UNKNOWN_	.	N		.	_DEFAULT_
DISHES	0	3	N		0.0	
DISHES	2	1	N		2.0	
DISHES	_UNKNOWN_	.	N		.	_DEFAULT_
EDLEVEL	High School	3	N		2.0	
EDLEVEL	Elementary	1	N		1.0	
EDLEVEL	_UNKNOWN_	.	N		.	0
FLATWARE	0	2	N		0.0	
FLATWARE	1	2	N		1.0	
FLATWARE	_UNKNOWN_	.	N		.	_DEF
GENDER	female	4	C	female	.	
GENDER	_UNKNOWN_	.	C		.	male
HEAT	Electric	2	N		1.0	
HEAT	Gas	2	N		2.0	
HEAT	_UNKNOWN_	.	N		.	_DEF
HHAPPAR	0	3	N		0.0	
HHAPPAR	2	1	N		2.0	
HHAPPAR	_UNKNOWN_	.	N		.	_DEF
JEWELRY	0	3	N		0.0	
JEWELRY	1	1	N		1.0	
JEWELRY	_UNKNOWN_	.	N		.	_DEF

Assess

- Compare training performance
- Champion / Challenger
  - Training and monitoring
- Ensure Generalization
  - Prevent over fitting
- Estimate deployment performance
  - Acquire target measures
- Select final model



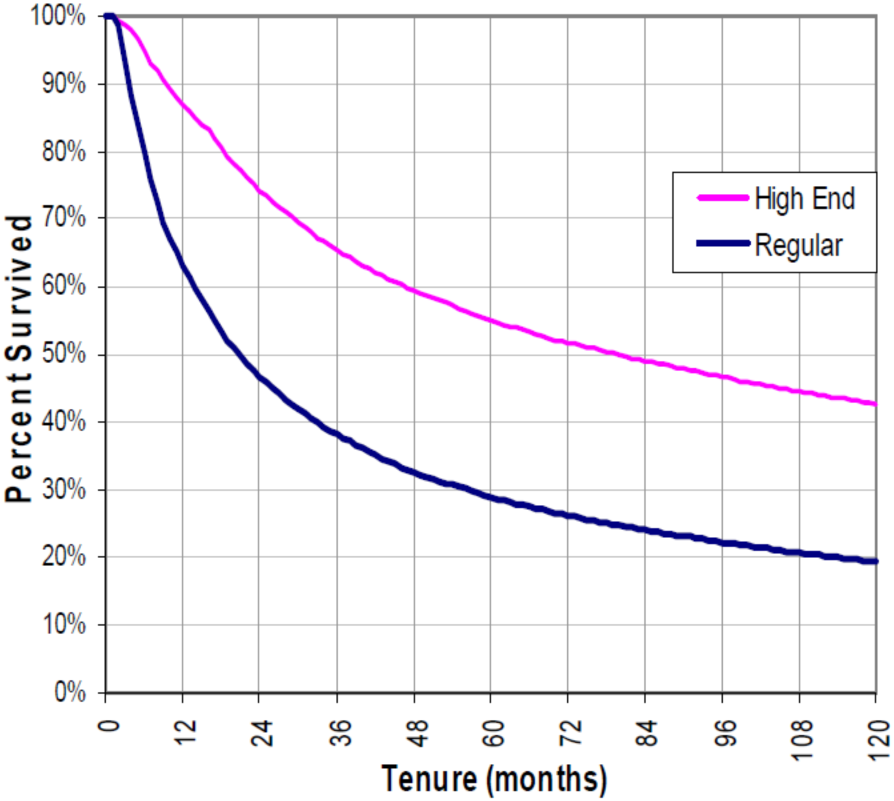
Survival Data Mining Analysis

- Prediction when event will happen, not just if it will happen
- Very popular for customer behavior modeling, such as
  - When will customer churn
  - When will customer upgrade
- Predicts event probability for time intervals for each customer
  - i.e. Customer has 50% chance to cancel next month but 75% chance to cancel the month after
- Can take external factors into account
  - customers with more than 2 products tend to stay longer

Survival Data Mining Analysis

Approach

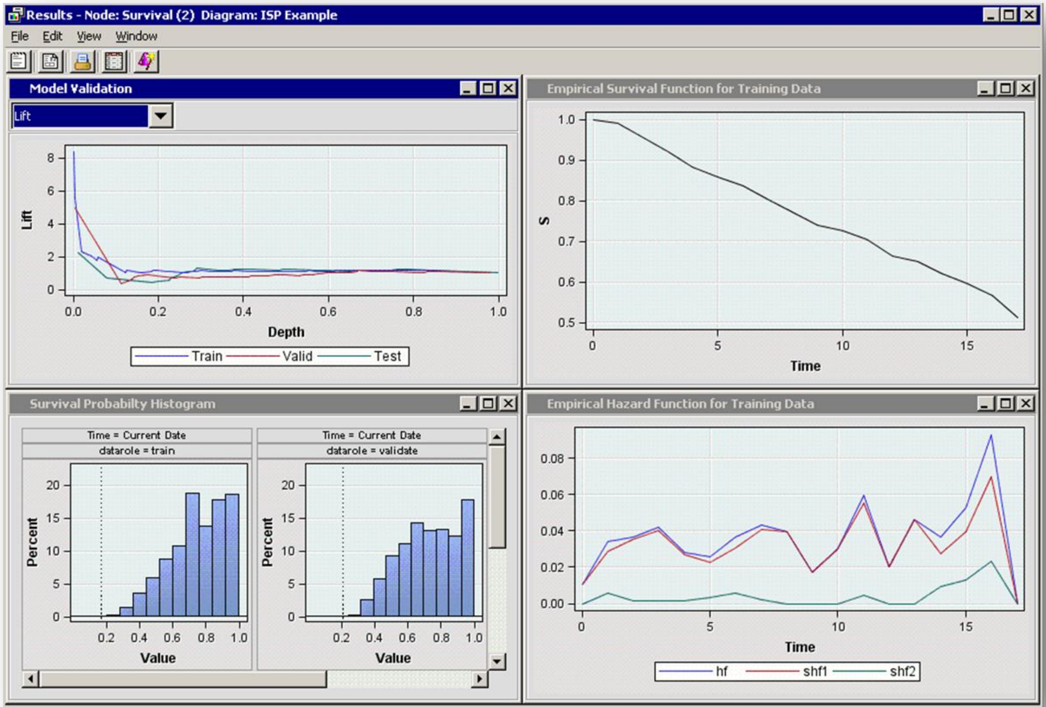
- Look at probability of hazard (Event) at discrete time points
- Time to event is most important feature
- Covariates can be integrated – need to be categorized
  - Customer demographics – age group
  - Product details – bronze, silver, gold tariff
  - Usage history – high, normal, low usage etc.



Survival Data Mining Analysis Node

Reports

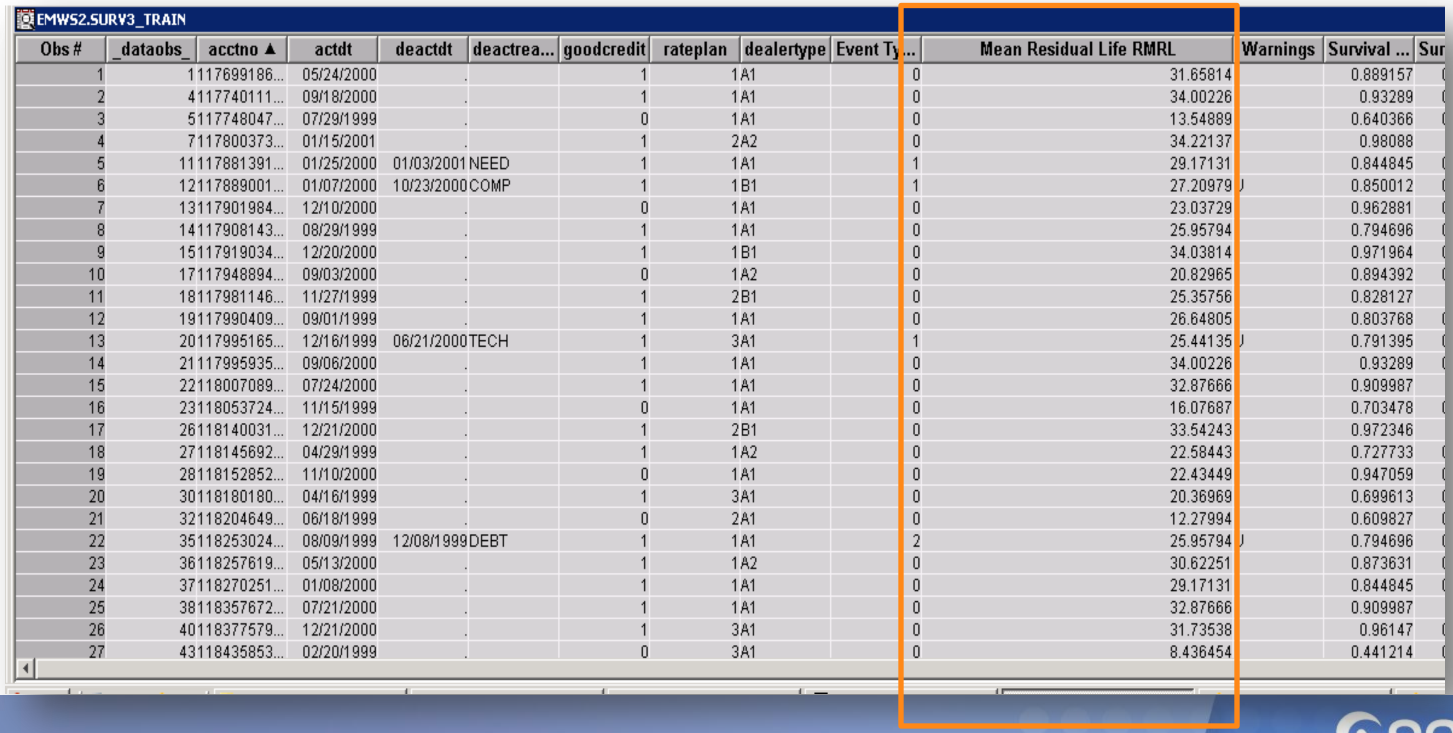
- Discrete time to event regression with additive logistic regression.
- Data preparation to define time interval and time training range (tenure view)
- Time effect is modeled with cubic splines to allow for flexible shapes of the hazard functions.
- Proportional hazard function is fitted with constant covariates.



Survival Data Mining Node

Scoring

- Mean Residual Lifetime: expected time till event occurs based on projecting hazard function into the future
  - Projection based on constant hazard function
  - Projection based on continuing trend of hazard function

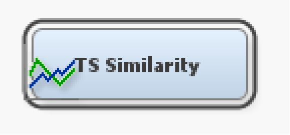


EMWS2.SURV3_TRAIN

Obs #	_dataobs_	acctno ▲	actdt	deactdt	deactrea...	goodcredit	rateplan	dealertype	Event Ty...	Mean Residual Life RMRL	Warnings	Survival ...	Sur
1	1	117699186...	05/24/2000	.		1	1	A1	0	31.65814		0.889157	0
2	4	117740111...	09/18/2000	.		1	1	A1	0	34.00226		0.93289	0
3	5	117748047...	07/29/1999	.		0	1	A1	0	13.54889		0.640366	0
4	7	117800373...	01/15/2001	.		1	2	A2	0	34.22137		0.98088	
5	11	117881391...	01/25/2000	01/03/2001	NEED	1	1	A1	1	29.17131		0.844845	0
6	12	117889001...	01/07/2000	10/23/2000	COMP	1	1	B1	1	27.20979	U	0.850012	0
7	13	117901984...	12/10/2000	.		0	1	A1	0	23.03729		0.962881	0
8	14	117908143...	08/29/1999	.		1	1	A1	0	25.95794		0.794696	0
9	15	117919034...	12/20/2000	.		1	1	B1	0	34.03814		0.971964	0
10	17	117948894...	09/03/2000	.		0	1	A2	0	20.82965		0.894392	0
11	18	117981146...	11/27/1999	.		1	2	B1	0	25.35756		0.828127	
12	19	117990409...	09/01/1999	.		1	1	A1	0	26.64805		0.803768	0
13	20	117995165...	12/16/1999	06/21/2000	TECH	1	3	A1	1	25.44135	U	0.791395	0
14	21	117995935...	09/06/2000	.		1	1	A1	0	34.00226		0.93289	0
15	22	118007089...	07/24/2000	.		1	1	A1	0	32.87666		0.909987	
16	23	118053724...	11/15/1999	.		0	1	A1	0	16.07687		0.703478	0
17	26	118140031...	12/21/2000	.		1	2	B1	0	33.54243		0.972346	
18	27	118145692...	04/29/1999	.		1	1	A2	0	22.58443		0.727733	0
19	28	118152852...	11/10/2000	.		0	1	A1	0	22.43449		0.947059	0
20	30	118180180...	04/16/1999	.		1	3	A1	0	20.36969		0.699613	0
21	32	118204649...	06/18/1999	.		0	2	A1	0	12.27994		0.609827	0
22	35	118253024...	08/09/1999	12/08/1999	DEBT	1	1	A1	2	25.95794	U	0.794696	0
23	36	118257619...	05/13/2000	.		1	1	A2	0	30.62251		0.873631	0
24	37	118270251...	01/08/2000	.		1	1	A1	0	29.17131		0.844845	0
25	38	118357672...	07/21/2000	.		1	1	A1	0	32.87666		0.909987	
26	40	118377579...	12/21/2000	.		1	3	A1	0	31.73538		0.96147	0
27	43	118435853...	02/20/1999	.		0	3	A1	0	8.436454		0.441214	0

Time Series Data Mining Nodes (experimental)

- Integrate time dimension into analysis
- Data is often stored as transactional data with time stamp or in form of time series
- Nodes in SAS Enterprise Miner 7.1
  - Data Preparation
    - » Provides a tool of aggregation, differencing, summarization, etc.
  - Exponential Smoothing
    - » Fits ESM to interval variables
  - Similarity
    - » Computes several similarity measures among time series



sas THE POWER TO KNOW