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 meaningful patterns and rules."
- Berry and Linoff (1997), p.5

"Data Mining [is] the process of efficient discovery of nonobvious valuable 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 recognition 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*.
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

Model Management & Monitoring

Discovery

Model Development
Cross-Industry Data Mining Applications

**Customer Analytics**

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

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

- 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 their interest payments
Clustering/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
Clustering/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
- Input Data
- File Import
- Sample
- Data Partition
- Merge
- Filter
- Append
- Time Series

Explore
- Association
- Cluster
- Variable Selection
- Market Basket
- StatExplore
- Variable Clustering
- MultiPlot
- Path Analysis

Modify
- DMDB
- SOM/Kohonen
- Graph Explore
- Transform Variables
- Impute
- Replacement
- Interactive Binning
- Rules Builder
- Drop
- Principal Components

Model
- Decision Tree
- AutoNeural
- Dmine Regression
- DMNeural
- Ensemble
- Gradient Boosting
- LARS
- MBR

Assess
- Neural Network
- SVM
- Partial Least Squares
- Regression
- Rule Induction
- TwoStage
- Model Import
- Model Comparison
- Score
- Segment Profile
- Decisions
- Cutoff
Utility
- Metadata
- SAS Code
- Start Groups
- End Groups
- Control Point
- Reporter
- Score Code Export
- Ext Demo

Applications
- Survival
- Ratemaking

Time Series
- TS Similarity
- TS Exponential Smoothing
- TS Data Preparation
SAS® Enterprise Miner™ 7.1
SEMMA in Action – Repeatable Process
Sample and Explore

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

- **Data evaluation**
  - Statistical measures
  - Visualization
  - Identifying outliers
  - Analytical segmentation
  - Variable creation & selection
Modify and Model

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