Segmentation using Decision Trees

Gerhard Held
Product Manager Analytical Applications
SAS Institute Europe
1. The SAS Data Mining Solution
2. Introduction to Decision Trees
3. Assessment Measures
4. Overview of DataSplits Tool within SAS Data Mining Solution
The SAS Data Mining Solution

- Data Mining is a Process
- Data Mining involves close Co-operation of IT, Business, and Data Miners
The SAS Data Mining Solution

SAS Institute

Data Mining Process

Business Problem

DBMS
Data Warehouse

Data Mining
Process

Business Reporting
& Graphics

More Informed
Business Decisions

© copyright SAS Institute 1997
The SAS Data Mining Solution

SAS Institute

Data Mining Process

Business Problem

Sample
- Sampling

Explore
- Visual exploration
- Correlation/factor correspondence

Manipulate
- Variable grouping/subsetting
- Adding/subsetting records

Model
- Neural networks
- Statistical techniques
- Tree-based models
- Time series analysis

Assess
- Data update?
- New questions?

DBMS Data Warehouse

Data Mining Process

Business Reporting & Graphics

More Informed Business Decisions
The SAS Data Mining Solution is unique:

- **IT: DW Access, Scalability**
- **Business Users:** Intuitive Interface and Business Orientation
- **Data Miners:** Analytical Depth and Flexibility
Introduction to Decision Trees

• Searching for Segments
• What is a Decision Tree?
• Common Applications
• Tree Methodologies
Searching for Segments

All Observations
4,000 Resp - 29%
10,000 NonResp - 71%

Split on AGE

TEENS <20
3,000 Resp - 60%
2,000 NonResp - 40%

ADULTS >=20
1,000 Resp - 11%
8,000 NonResp - 89%
Searching for Segments

Total Response %

<table>
<thead>
<tr>
<th>Percentage</th>
<th>No Reply</th>
<th>Bought</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>80%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Search for predictor with large difference in response rate over its categories.
Searching for Segments

Total Response %

- No Reply
- Bought

Young: No or Yes?

- No Reply No
- Bought No
- No Reply Yes
- Bought Yes
Searching for Segments

**Young: No or Yes?**

- No Reply: 0%
- Bought: 20%
- No: 40%
- Yes: 60%
- Not shown: 80%
- Bought: 100%

**UK: No or Yes?**

- No Reply: 0%
- Bought: 20%
- No: 40%
- Yes: 60%
- Not shown: 80%
- Bought: 100%
Searching for Segments

All Observations

4,000 Resp - 29%
10,000 NonResp - 71%

Split on AGE
Searching for Segments

All Observations
4,000 Resp - 29%
10,000 NonResp - 71%
Split on AGE

TEENS <20
3,000 Resp - 60%
2,000 NonResp - 40%

ADULTS >=20
1,000 Resp - 11%
8,000 NonResp - 89%
Searching for Segments

All Observations
- 4,000 Resp - 29%
- 10,000 NonResp - 71%

Split on AGE

TEENS <20
- 3,000 Resp - 60%
- 2,000 NonResp - 40%
- 1 -

ADULTS >=20
- 1,000 Resp - 11%
- 8,000 NonResp - 89%

Split on Country

UK only
- 0 Resp - 0%
- 1,000 NonResp - 100%
- 2 -

“Continent”
- 1,000 Resp - 12.5%
- 7,000 NonResp - 87.5%
- 3 -
Rules Define the Segments

- If the person is a teenager…
- If the person is an adult in the UK …
- Otherwise …

```
IF AGE LT 20 THEN SEGMENT=1;
ELSE IF AGE GE 20 AND CTRY EQ 'UK'
    THEN SEGMENT=2;
ELSE SEGMENT=3;
```
What is a Decision Tree?

- A Set of Rules
- For defining Data Segments
- with similar Values of a Target
Why Trees are Popular?

Easy to Understand
- Segmentation defined without Formulae
- Express Rules using a single Variable
- User can create and modify Rules

Easy to Use
- Accepts ordered and unordered Variables
- No transformation of Predictors
- Uses missing Data
Decision Tree Applications

- Segmentation Using a Target Variable
- Profiling Clusters
- Modeling for Classification
- Variable Selection and Creation

A few Examples...
Profiling Clusters - Why...

- Group Customers with common Preferences
  - No Need to define Target Variable
  - without a specific Product in Mind
- Select Products or Promotions separately
Profiling Clusters - Steps...

- **Cluster:** Form Segments from Data
- **Profile:** Describe Segments intuitively
Clusters & Profiles

Segment 1
• ages: 18-25
• income: 40K

“Startups”
• Foreign Economical Cars
• Junk Food

Segment 2
• ages: 26-35
• income: 100K

“DINKS”
• Sports Cars
• Fashionable Restaurants
Profiling Using Trees

- Cluster on demographic Data
  - Such as: Age, Income, City
  - Mostly Available, generic
- Assign an ID to each Segment
- Train a Decision Tree
  - Segment ID as Target
  - Profiling Data as Input Variables
New Applications: Variable Selection

Include
- AGE
- COUNTRY

Exclude
- FINANCIAL ASSETS
- INCOME
Variable Creation

Include
• AGE
• STATE

Create
• TEENS
• ADULTS
• ADULTS not UK
• ADULTS in UK
Tree Methodologies - The AID Family

AID (1964), CHAID (1978), FIRM, TREEDISC, KnowledgeSeeker

- Automatic Interaction Detection
- Detects complex statistical Relationships
- Number of Branches varies from 2 to number of Categories of Predictor
- Statistical Significance Tests used to determine size of Tree
- Nominal Responses
- Nominal or Ordinal Predictors
CART, Salford Systems, S-Plus (all 1980s)

- Classification And Regression Trees
- Statistical Prediction
- Exactly 2 Branches from each nonterminal Node
- Cross-validation and pruning used to determine Size of Tree
- Quantitative or nominal Response Variable
- Nominal, ordinal and continuous Predictors

```
is x(1)<5.0?
  yes
  is x(3)<3.2?
    yes
    class 2
    no
  no
  is x(1)<7.6?
    yes
    class 1
    no
    class 3
```
Model Assessment

- Classification
- Sensitivity
- Lift
- Profit and Loss Equations
Model Assessment
Alternatives

• Proportion correctly predicted (classified)
• Proportion responding to Promotion
• Expected Profit
### Classification

**PREDICTED**

<table>
<thead>
<tr>
<th></th>
<th>Default</th>
<th>Paid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Paid</td>
<td>0</td>
<td>40</td>
</tr>
</tbody>
</table>

\[
\frac{(10 + 40)}{60} = \frac{50}{60}
\]
## Sensitivity

<table>
<thead>
<tr>
<th>Observed</th>
<th>Predicted</th>
<th>Default</th>
<th>Paid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>10</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Paid</td>
<td>0</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>50</td>
<td>60</td>
</tr>
</tbody>
</table>

Sensitivity: 10 / 20
Sensitivity and Classification

<table>
<thead>
<tr>
<th></th>
<th>Default</th>
<th>Paid</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>20</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>Paid</td>
<td>20</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>Observed</td>
<td>40</td>
<td>20</td>
<td>60</td>
</tr>
</tbody>
</table>

Sensitivity: \( \frac{20}{20} \)

\[
\frac{(20 + 20)}{60} = \frac{40}{60}
\]
## Specificity

### PREDICTED

<table>
<thead>
<tr>
<th></th>
<th>Default</th>
<th>Paid</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Default</strong></td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td><strong>Paid</strong></td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>40</td>
<td>20</td>
</tr>
</tbody>
</table>

Specificity $= \frac{20}{40}$

The Alternative of Sensitivity
### Lift

**PREDICTED**

<table>
<thead>
<tr>
<th>Observed</th>
<th>Respond</th>
<th>No</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Respond</td>
<td>20</td>
<td>0</td>
<td>20/40</td>
</tr>
<tr>
<td>No</td>
<td>20</td>
<td>20</td>
<td>20/40</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>20</td>
<td>60/60</td>
</tr>
</tbody>
</table>

\[
\text{Lift} = \frac{20/40}{20/60}
\]
Profit & Loss Equation

• Given: A scored Data Set A Cost Function
• Then the Assessment Tool produces for each Level of the Target a P&L Equation, such as:

  Purchase..................100
  Cost of Deliver..........- 10
  Net Sales...............90
  Cost of Goods.........50
  Profit..................40
Alternative Objectives

• Classification
  – Proportion Correct
  – Diagonal

• Sensitivity
  – Proportion correct of an observed Event
  – Row

• Lift
  – Proportion correct of a predicted Event
  – Column
Trade Offs

- Classification
  - none
- Sensitivity
  - Specificity
- Lift
  - Size of Subset (Node)
Need for a Stopping Method

- Trees can grow indefinitely
- Probability of an Event depends on Depth of Tree
- More Depth creates Problems to generalise (overfitting)
Current Stopping Methods

- Leaf Size
- Statistical Significance of Split
- Prune to best Generalization Score using Classification Rate
- Problems:
  - Dominant Class dominates Classification
  - Rare Events are a Lift Tradeoff
- SAS Institute researches Stopping Methodology
Overview of DataSplits Tool

• Tool in the SAS Enterprise Miner
  – Icon
  – PROC DATASPLIT
  – Methods Calls (non-display AF Object Class)
Overview of DataSplits Tool

- Seamless Integration of CHAID and CART Methodology
- Nominal and Interval Targets and Inputs
- Multiway Splits on all Input Types
- Up to 32000 Values for a nominal Input
- Allows User defined Split Criteria
- General Profit and Loss Functions
- Suited for Rare Events
Overview of DataSplits Tool - Datasets
Overview of DataSplits Tool - Variables

<table>
<thead>
<tr>
<th>Name</th>
<th>Status</th>
<th>Model Role</th>
<th>Measurement</th>
<th>Type</th>
<th>Format</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>PURCHASE</td>
<td>use</td>
<td>target</td>
<td>binary</td>
<td>num</td>
<td>BEST12.</td>
<td>Purchase</td>
</tr>
<tr>
<td>DOMESTIC</td>
<td>use</td>
<td>input</td>
<td>interval</td>
<td>num</td>
<td>BEST12.</td>
<td>Domestic</td>
</tr>
<tr>
<td>APPAREL</td>
<td>use</td>
<td>input</td>
<td>interval</td>
<td>num</td>
<td>BEST12.</td>
<td>Apparel P</td>
</tr>
<tr>
<td>LEISURE</td>
<td>use</td>
<td>input</td>
<td>interval</td>
<td>num</td>
<td>BEST12.</td>
<td>Leisure Pr</td>
</tr>
<tr>
<td>KITCHEN</td>
<td>use</td>
<td>input</td>
<td>interval</td>
<td>num</td>
<td>BEST12.</td>
<td>Kitchen Pr</td>
</tr>
<tr>
<td>LUXURY</td>
<td>use</td>
<td>input</td>
<td>binary</td>
<td>num</td>
<td>BEST12.</td>
<td>Luxury It</td>
</tr>
<tr>
<td>PROMO7</td>
<td>don't use</td>
<td>input</td>
<td>interval</td>
<td>num</td>
<td>BEST12.</td>
<td>Promos: 1</td>
</tr>
<tr>
<td>PROMO13</td>
<td>don't use</td>
<td>input</td>
<td>interval</td>
<td>num</td>
<td>BEST12.</td>
<td>Promos: 8</td>
</tr>
<tr>
<td>AVHINC</td>
<td>use</td>
<td>input</td>
<td>interval</td>
<td>num</td>
<td>BEST12.</td>
<td>Average H</td>
</tr>
<tr>
<td>AVVAL</td>
<td>use</td>
<td>input</td>
<td>interval</td>
<td>num</td>
<td>BEST12.</td>
<td>Average I</td>
</tr>
</tbody>
</table>
Overview of DataSplits Tool - Options

File Tools

SAS: Data Splits Decision Trees

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Variables</th>
<th>Options</th>
<th>Advanced Options</th>
<th>Decisions</th>
<th>Assessment</th>
<th>Notes</th>
</tr>
</thead>
</table>

Splitting Criterion:
- Chi-Square Test
- Entropy Reduction
- Gini Reduction

Significance Level: 0.20

Related Values:
- Minimum size of a subset: 1
- Subset size required for splitting: 50
- Maximum subsets from a single split: 2
- Maximum depth of tree: 4
Overview of DataSplits Tool - Advanced Options

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Variables</th>
<th>Options</th>
<th>Advanced Options</th>
<th>Decisions</th>
<th>Assessment</th>
<th>Notes</th>
</tr>
</thead>
</table>

**Related Values**

- **Number of splitting rules:** 5
- **Size sufficient for split search:** 1000

**Adjustments to p-values**
- Method
  - KASS
  - DEVILL
  - DEPTH

**Effective Numbers of Inputs:** 34
Overview of DataSplits Tool - Decisions

Intend classification

- YES
- NO

Bonus Variables:

<table>
<thead>
<tr>
<th>Decision Alternatives</th>
<th>Bonus Variable</th>
<th>Profit Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>.</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>.</td>
</tr>
</tbody>
</table>

Maximize Profit
Minimize Misclassification Cost
Overview of DataSplits Tool - Assessment

SAS: Data Splits Decision Trees

<table>
<thead>
<tr>
<th>Target</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Overview of Tree Browser

- Summary Table (upper left)
- Tree-Ring Navigator (upper right)
  - Accessible from here: Tree Diagram + Assessment Statistics
- Assessment Table (lower left)
- Assessment Graph (lower right)
  - blue Training Data, red Validation Data
Overview of Tree Browser

![Image of SAS: Data Splits - Browser interface]

<table>
<thead>
<tr>
<th>Browser</th>
<th>Summary Table</th>
<th>Tree Ring</th>
<th>Assessment Table</th>
<th>Assessment Graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRAIN 0</td>
<td>N 822.12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRAIN 0</td>
<td>Row % 81.4667</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRAIN 1</td>
<td>N 590.12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRAIN 1</td>
<td>Row % 62.3020</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRAIN +</td>
<td>N 1412.12</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Segments</th>
<th>Assessment</th>
<th>Validated</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0.6028</td>
<td>0.6143</td>
</tr>
<tr>
<td>7</td>
<td>0.6028</td>
<td>0.6143</td>
</tr>
<tr>
<td>8</td>
<td>0.6028</td>
<td>0.6143</td>
</tr>
<tr>
<td>9</td>
<td>0.6028</td>
<td>0.6143</td>
</tr>
</tbody>
</table>

Numbers of Segments:

- 0.50 to 0.55
- 0.55 to 0.60
- 0.60 to 0.65

SeuGi 15 - iMadrid
Profit Chart from Assessment Node

Cumulative Profit

Profit

$-2,000  $0  $2,000  $4,000  $6,000

10  20  30  40  50  60  70  80  90  100  DECILE

METHODS

- Neural
- Random
- Tree
Summary: Decision Trees in SAS Enterprise Miner

- Decision Trees are a powerful Tool to segment Customer related Data
- DataSplits is a very comprehensive Implementation of Decision Trees
- Decision Alternatives can be assessed graphically and in Financial Terms associated with each Alternative
- The DataSplits Tool is one of many integrated Tools in the SAS Enterprise Miner.
Segmentation Using Decision Trees

Thank you for your Attention!

Questions?