Data Mining for Model Creation

Presentation by Paul Below, EDS
2500 NE Plunkett Lane
Poulsbo, WA USA 98370
paul.below@eds.com
Agenda

• Data Mining and Estimating Model Creation Challenges
• Types of Data Mining Models and Examples of Each
• Data Mining Issues
What is Data Mining?

Every book has a different definition, but the common themes are:

- Use of very large databases
- Use of tools and a process
- Results have to be useful

The hard thing is not figuring out which algorithm to use, the hard thing is to figure out what to do with the results.
Data Mining Myths

- Find answers to unasked questions
- Continuously monitor your data for interesting patterns
- Eliminate the need to understand your business
- Eliminate the need to collect good data
- Eliminate the need to have good data analysis skills
Model Creation

“All models are wrong, but some are useful.”

“Statisticians, like artists, have the bad habit of falling in love with their models.”

George Box

People love to interpret noise.
Model Creation Challenges

- Databases are already built, and not designed for our purposes
- Databases were designed by committee, and everything anyone thought of is in there
- Data structure is often horrible, keys not appropriate

The model is no better than the data
Model Creation Challenge: Getting Started

- Dozens of input variables usually available, which should I use to build estimating models?
- It is common for our variables to exhibit colinearity
- Which relationships do I explore first?
Data Mining Can Help

- Thin out the forest, so we can examine the important trees
- Data mining tools can identify the variables to look at first
- Success depends more on the way you mine than on the specific tool
Data Mining Can Help

• Data mining can aid in conducting hypothesis testing or getting started with exploratory analysis

• Classification trees can be useful for exploratory analysis:
  – Which variables does the tool split on first?
  – Which variable does the tool think is most important?
  – What variables does it pick for the first 5 or 6?

• Some data mining techniques are supported in basic statistical tools (e.g., SPSS, SPLUS, JMP)

• Many data mining specific tools exist in the marketplace
Types of Data Mining Models

- Classification
  - Usually on discrete variables, predicts a response variable
- Regression
  - Usually on continuous variables, predicts a response variable
- Clustering
  - Grouping the cases by similarity
- Association
  - Grouping the variables by similarity
Types of Models

- Some data mining techniques are black box, others are white box
- Black box is used for prediction (examples are neural networks and k nearest neighbors)
- White box is used for interpretation (classification trees and regression are examples)
- Users generally dislike black box because they cannot see how the model works
Example Output: Classification

Tree-based models are useful for both classification and regression

All Rows
Count 841
Group I .068
Group II .205
Group III .598
Group IV .128

Variable A = 1, 2, 3
Count 269
Group I .167
Group II .483
Group III .335
Group IV .015

Variable A = 4, 5
Count 572
Group I .021
Group II .075
Group III .722
Group IV .182
Example Output: Regression

- Stepwise Regression enters variables one by one and tests them for removal
- Good method when independent variables are correlated
- This example went through 5 steps to build a model
- There were 14 variables excluded from the final model

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
<th>Change Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.248a</td>
<td>.062</td>
<td>.061</td>
<td>25.13423</td>
<td>R Square Change</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F Change</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.062</td>
<td>df1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>df2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sig. F Change</td>
</tr>
<tr>
<td>2</td>
<td>.291b</td>
<td>.084</td>
<td>.083</td>
<td>24.83528</td>
<td>.023</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>41.133</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.023</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1656</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.000</td>
</tr>
<tr>
<td>3</td>
<td>.322c</td>
<td>.104</td>
<td>.102</td>
<td>24.58177</td>
<td>.019</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>35.332</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.019</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1655</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.000</td>
</tr>
<tr>
<td>4</td>
<td>.337d</td>
<td>.113</td>
<td>.111</td>
<td>24.45412</td>
<td>.010</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>18.323</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.010</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1654</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.000</td>
</tr>
<tr>
<td>5</td>
<td>.347e</td>
<td>.121</td>
<td>.118</td>
<td>24.36170</td>
<td>.007</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13.573</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.007</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1653</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.000</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), Log of adjusted function points
b. Predictors: (Constant), Log of adjusted function points, Max Team Size
c. Predictors: (Constant), Log of adjusted function points, Max Team Size, Resource Level
d. Predictors: (Constant), Log of adjusted function points, Max Team Size, Resource Level, Changed count
e. Predictors: (Constant), Log of adjusted function points, Max Team Size, Resource Level, Changed count, Enquiry count
# Example Output: Regression

<table>
<thead>
<tr>
<th></th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
<th>95% Confidence Interval for B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
<td>Lower Bound</td>
</tr>
<tr>
<td>(Constant)</td>
<td>-13.488</td>
<td>3.733</td>
<td></td>
<td>-3.613</td>
<td>.000</td>
</tr>
<tr>
<td>Log of adjusted function</td>
<td>18.465</td>
<td>1.619</td>
<td>.310</td>
<td>11.403</td>
<td>.000</td>
</tr>
<tr>
<td>Max Team Size</td>
<td>-.395</td>
<td>.061</td>
<td>-.152</td>
<td>-6.467</td>
<td>.000</td>
</tr>
<tr>
<td>Resource Level</td>
<td>-3.717</td>
<td>.642</td>
<td>-.135</td>
<td>-5.795</td>
<td>.000</td>
</tr>
<tr>
<td>Changed count</td>
<td>.013</td>
<td>.003</td>
<td>.117</td>
<td>4.842</td>
<td>.000</td>
</tr>
<tr>
<td>Enquiry count</td>
<td>-.028</td>
<td>.008</td>
<td>-.099</td>
<td>-3.684</td>
<td>.000</td>
</tr>
</tbody>
</table>

Dependent Variable: Function Points per Person Month
Example: Correlation

- Independent variables can be correlated with a dependent variable, for example:
  - Nominal: chi-square on crosstabs
  - Ordinal: Kendall’s Tau-B correlation
  - Ratio: Pearson correlation

- Thin out list of variables, examine those that show significant correlation

- Remember that correlation might be non-linear
Example Output: Clustering

- Detects groupings in data
- K-Means iteratively moves from initial to final cluster centers, used with large number of cases
- Another type, hierarchical, finds the closest pair of objects then continues iteratively until all objects are in one cluster, results in stages of clusters which can be examined

### Final Cluster Centers

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Function Points per Person Month</th>
<th>Function Points per Calendar Month</th>
<th>Average Team Size</th>
<th>Max Team Size</th>
<th>Resource Level</th>
<th>Normalised Work Effort</th>
<th>Project Inactive Time</th>
<th>Effort Specify</th>
<th>Effort Plan</th>
<th>Effort Build</th>
<th>Effort Test</th>
<th>Effort Implement</th>
<th>Effort unphased</th>
<th>Minor defects</th>
<th>Major defects</th>
<th>Total Defects Delivered</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.43</td>
<td>173.56</td>
<td>36.98</td>
<td>38.37</td>
<td>2.20</td>
<td>42370.17</td>
<td>.10</td>
<td>7107.88</td>
<td>4423.53</td>
<td>15565.42</td>
<td>7324.58</td>
<td>19405.94</td>
<td>19045.94</td>
<td>12.40</td>
<td>23.33</td>
<td>29.88</td>
</tr>
<tr>
<td>2</td>
<td>23.39</td>
<td>78.84</td>
<td>6.75</td>
<td>7.06</td>
<td>1.39</td>
<td>4038.12</td>
<td>.13</td>
<td>725.77</td>
<td>350.19</td>
<td>1695.08</td>
<td>746.55</td>
<td>1282.80</td>
<td>743.22</td>
<td>20.70</td>
<td>16.48</td>
<td>33.66</td>
</tr>
</tbody>
</table>

### Number of Cases in each Cluster

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Number of Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>75,000</td>
</tr>
<tr>
<td>2</td>
<td>1,584,000</td>
</tr>
<tr>
<td>Valid</td>
<td>1,659,000</td>
</tr>
<tr>
<td>Missing</td>
<td>.000</td>
</tr>
</tbody>
</table>
Example Output: Association

- Two types are factor analysis or principal components
- Study correlations between large number of interrelated quantitative variables by grouping the variables into factors
- Interpret each factor according to the meaning of the variables
- Summarize many variables by a few factors

Scree Plot

Component Number

Eigenvalue
Example Output: Association

- In this example, Functional Size is correlated with Factor 1, not with the other 3 factors.
- The table is useful for naming the factors.

**Rotated Component Matrix**

<table>
<thead>
<tr>
<th></th>
<th>Component</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Functional Size</td>
<td>.949</td>
</tr>
<tr>
<td>Adjusted Function Points</td>
<td>.950</td>
</tr>
<tr>
<td>Value Adjustment Factor</td>
<td>.091</td>
</tr>
<tr>
<td>Normalised Work Effort</td>
<td>.541</td>
</tr>
<tr>
<td>Function Points per Person Month</td>
<td>.259</td>
</tr>
<tr>
<td>Function Points per Calendar Month</td>
<td>.766</td>
</tr>
<tr>
<td>Total Defects Delivered</td>
<td>.171</td>
</tr>
<tr>
<td>Resource Level</td>
<td>.094</td>
</tr>
<tr>
<td>Max Team Size</td>
<td>.132</td>
</tr>
<tr>
<td>Average Team Size</td>
<td>.086</td>
</tr>
</tbody>
</table>

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

*a. Rotation converged in 5 iterations.*
Causation: Post Hoc

Will the estimation model continue to work?

Retrospective studies (in absence of DOE) must meet these criteria to make a good case for causality:

- Association
- Temporal Priority
- Non-spuriousness
- Theoretical Adequacy

There are two clocks that keep perfect time. When “a” points to the hour, “b” strikes. Did “a” cause “b” to strike?
Causation: Confounding Factors

• Apparent causation could be due to:
  – A third factor, correlated to the supposed cause
  – Interaction between two or more factors (higher order effects)
• Therefore, potential confounding factors must be investigated
Data Mining Issue: Overfitting

Five Data Points

- Observed
- Cubic

Sequence

Observed and Cubic curves for five data points.
Data Mining Issue: the Data

- Check for normality
- Check for outliers
- Check for missing values
- Consider transformation
- Know how the tool is dealing with missing data
Data Mining Issue: Exploration

• Do not explore a database
• Identify the business question first
• Otherwise you will go mining and never come back
Summary

Topics covered:
- Data mining and estimating model creation challenges
- Types of data mining models and examples of each
- Data mining issues

Consider the use of data mining to aid in filtering many variables down to a vital few to improve model based estimates.
Final Data Mining Issue: The Laugh Test

Software cannot discriminate between an important strong association and something that is obvious and trivial.

Your conclusions will have to pass the “laugh test” with the project team.

Twyman’s Law: If it looks interesting, it is probably wrong.
Resources

- [www.twocrows.com](http://www.twocrows.com) (free 36 page introductory booklet in Adobe format)
- [www.kdnuggets.com](http://www.kdnuggets.com) (extensive data mining industry website including links to free evaluation software)