Sept 22 2005

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

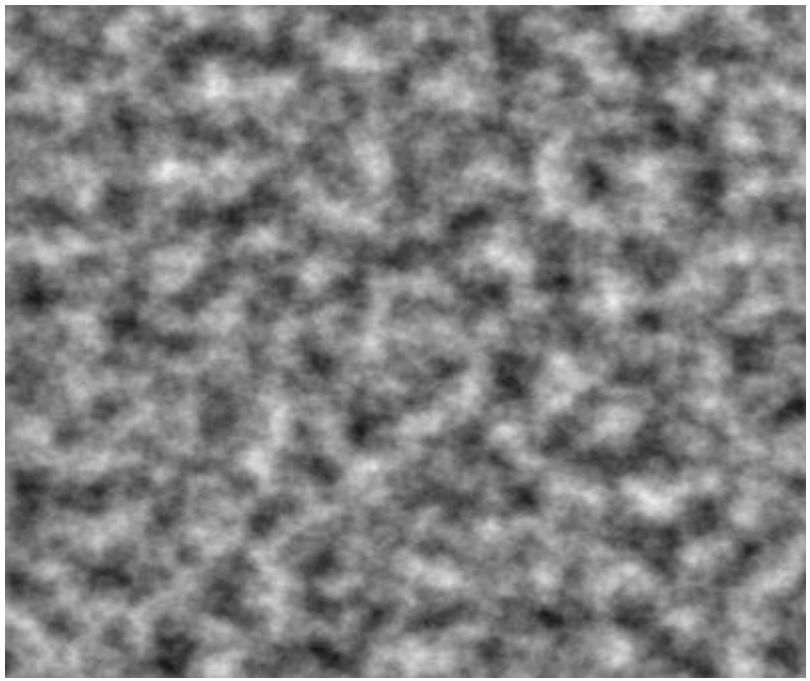
The hard thing is not figuring out which algorithm to use,

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



People love to interpret noise.

"All models are wrong, but some are useful."

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

George Box

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

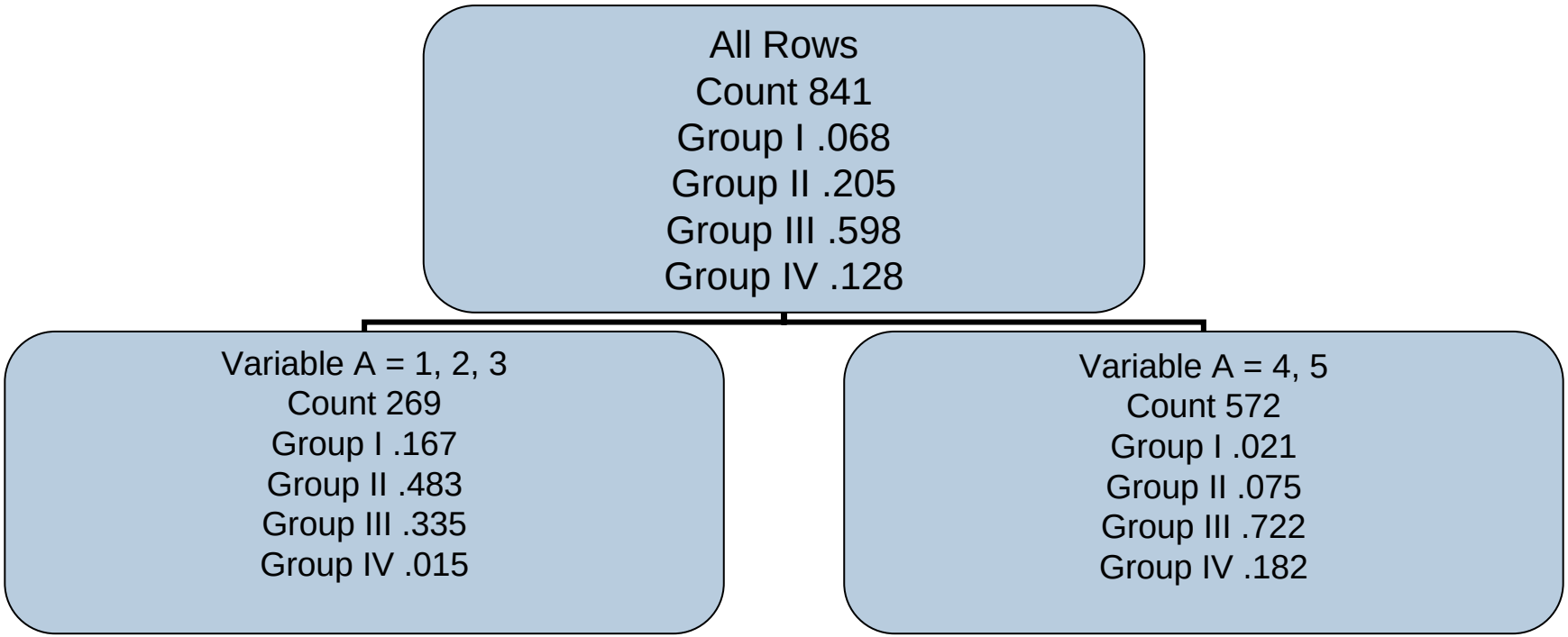
- Classification
  - Usually on discrete variables, predicts a response variable
- Regression
  - Usually on continuous variables, predicts a response variable
- Clustering
  - Grouping the cases by similarity
- Association
  - Grouping the variables by similarity

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



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

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.248[a]	.062	.061	25.13423	.062	108.931	1	1657	.000
2	.291[b]	.084	.083	24.83528	.023	41.133	1	1656	.000
3	.322[c]	.104	.102	24.58177	.019	35.332	1	1655	.000
4	.337[d]	.113	.111	24.45412	.010	18.323	1	1654	.000
5	.347[e]	.121	.118	24.36170	.007	13.573	1	1653	.000

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

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
(Constant)	-13.488	3.733		-3.613	.000	-20.811	-6.166
Log of adjusted function points	18.465	1.619	.310	11.403	.000	15.289	21.642
Max Team Size	-.395	.061	-.152	-6.467	.000	-.515	-.275
Resource Level	-3.717	.642	-.135	-5.795	.000	-4.976	-2.459
Changed count	.013	.003	.117	4.842	.000	.008	.018
Enquiry count	-.028	.008	-.099	-3.684	.000	-.043	-.013

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

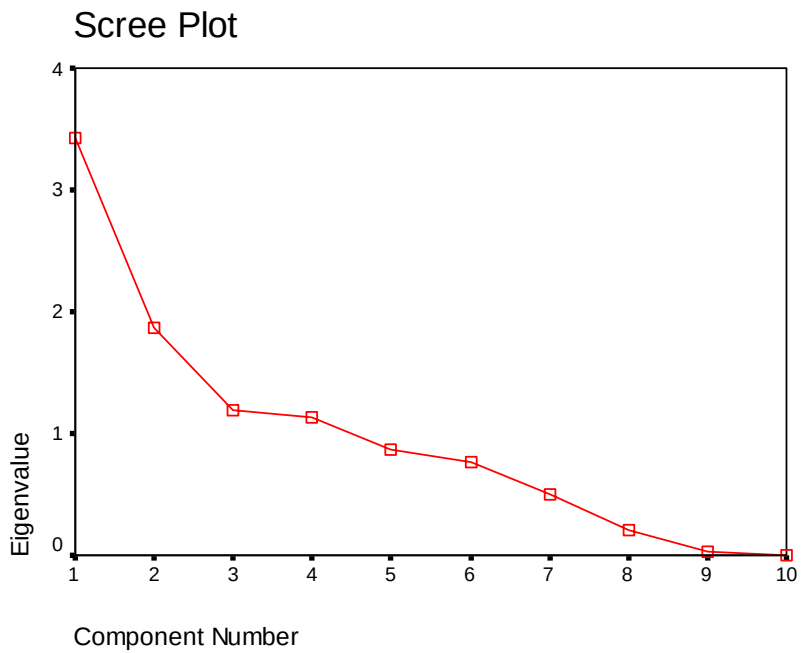
	Cluster	
	1	2
Function Points per Person Month	6.43	23.39
Function Points per Calendar Month	173.56	78.84
Average Team Size	36.98	6.75
Max Team Size	38.37	7.06
Resource Level	2.20	1.39
Normalised Work Effort	42370.17	4038.12
Project Inactive Time	.10	.13
Effort Specify	7107.88	725.77
Effort Plan	4423.53	350.19
Effort Build	15565.42	1695.08
Effort Test	7324.58	746.55
Effort Implement	8571.83	743.22
Effort unphased	19405.94	1282.80
Minor defects	12.40	20.70
Major defects	23.33	16.48
Total Defects Delivered	29.88	33.66

Number of Cases in each Cluster

Cluster	1	75.000
	2	1584.000
Valid		1659.000
Missing		.000

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



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[a]

	Component			
	1	2	3	4
Functional Size	.949	.109	.102	-.009
Adjusted Function Points	.950	.106	.106	.018
Value Adjustment Factor	.091	.033	.085	.702
Normalised Work Effort	.541	.570	.454	-.085
Function Points per Person Month	.259	-.215	-.668	.210
Function Points per Calendar Month	.766	.093	-.233	.032
Total Defects Delivered	.171	-.067	.751	.209
Resource Level	.094	.050	.066	-.778
Max Team Size	.132	.972	.027	.004
Average Team Size	.086	.964	.042	-.006

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?

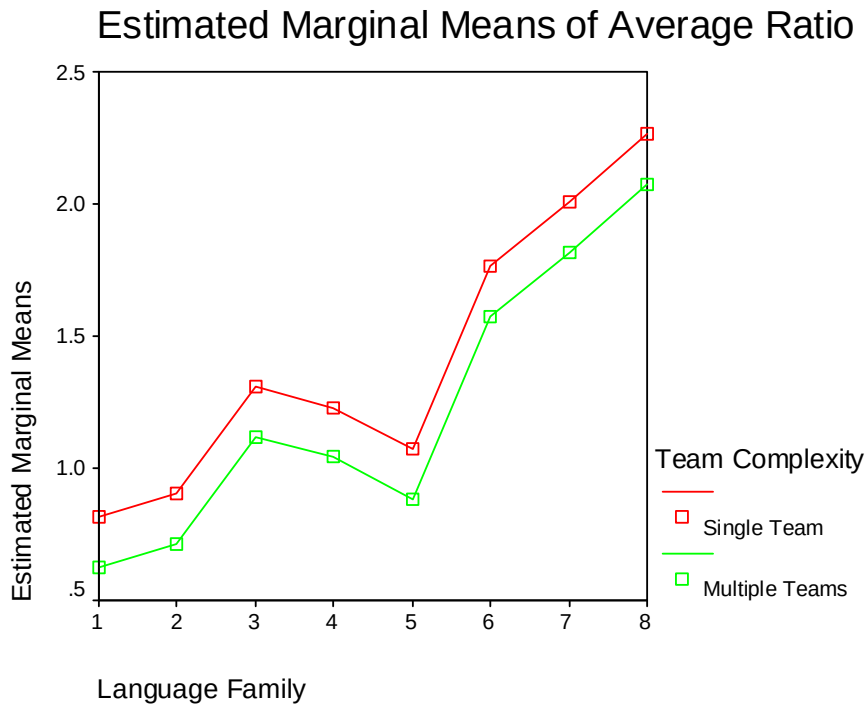
Retrospective studies (in absence of DOE) must meet these criteria to make a good case for causality:

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

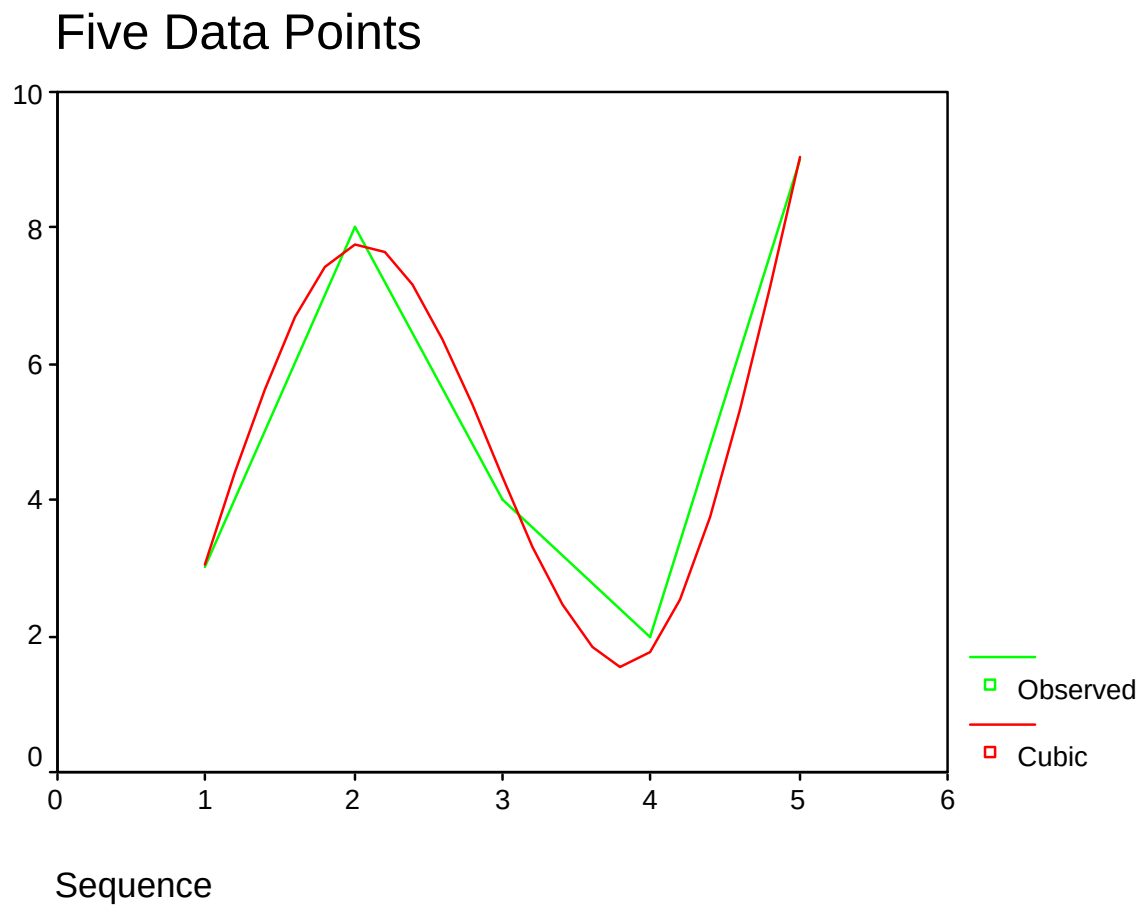
There are two clocks that keep perfect time.
When “a” points to the hour, “b” strikes.
Did “a” cause “b” to strike?

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

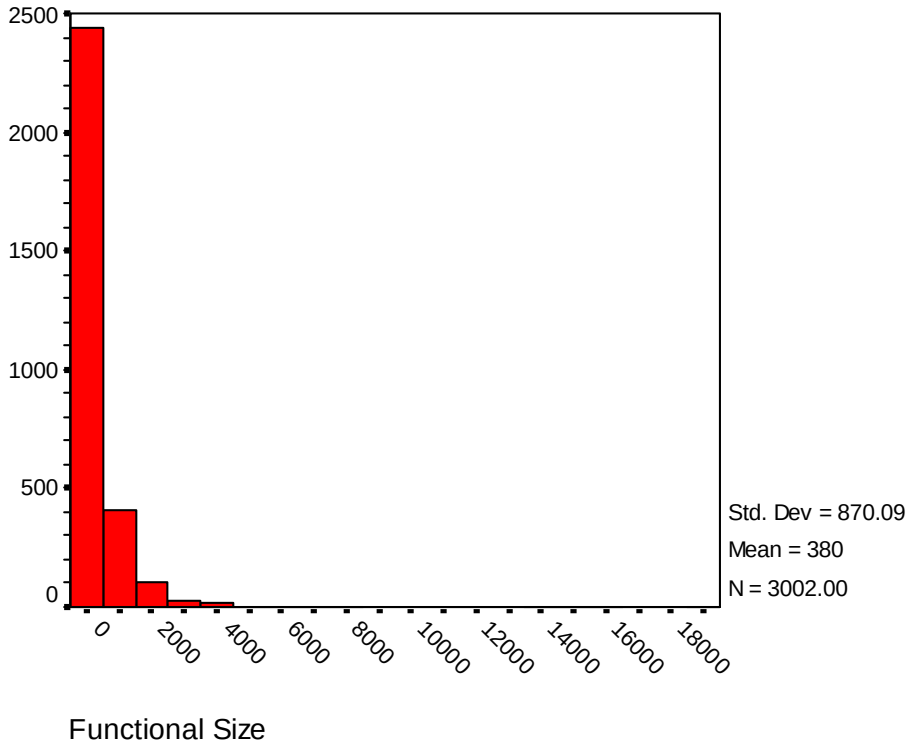


Data Mining Issue: Overfitting



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 (free 36 page introductory booklet in Adobe format)
- www.kdnuggets.com (extensive data mining industry website including links to free evaluation software)
- *Principles of Data Mining*, by David Hand, Keikki Mannila and Padhraic Smyth, MIT Press, 2001.
- *Data Mining – Concepts, Models, Methods and Algorithms*, by Mehmed Kantardzic, IEEE Press, 2003.
- *The Software Metrics Compendium,* by International Software Benchmarking Standards Group, 2002.