

Data Mining Techniques

Chapter 3: Data Mining Methodology & Best Practices

Why have a methodology?	2
Statistical practices	3
Data mining methodology	4
Data mining methodology I	5
Data mining methodology II	6
Data mining methodology III	7

Why have a methodology?

- Learning things that aren't true:
 - patterns may not represent any underlying rule (overfitting);
 - model set may not reflect the relevant population (biased sample);
 - data may be at the wrong level of detail.
- Learning things that are true, but not useful:
 - things that are already known;
 - things that can't be used.

© Iain Pardoe, 2006

2 / 7

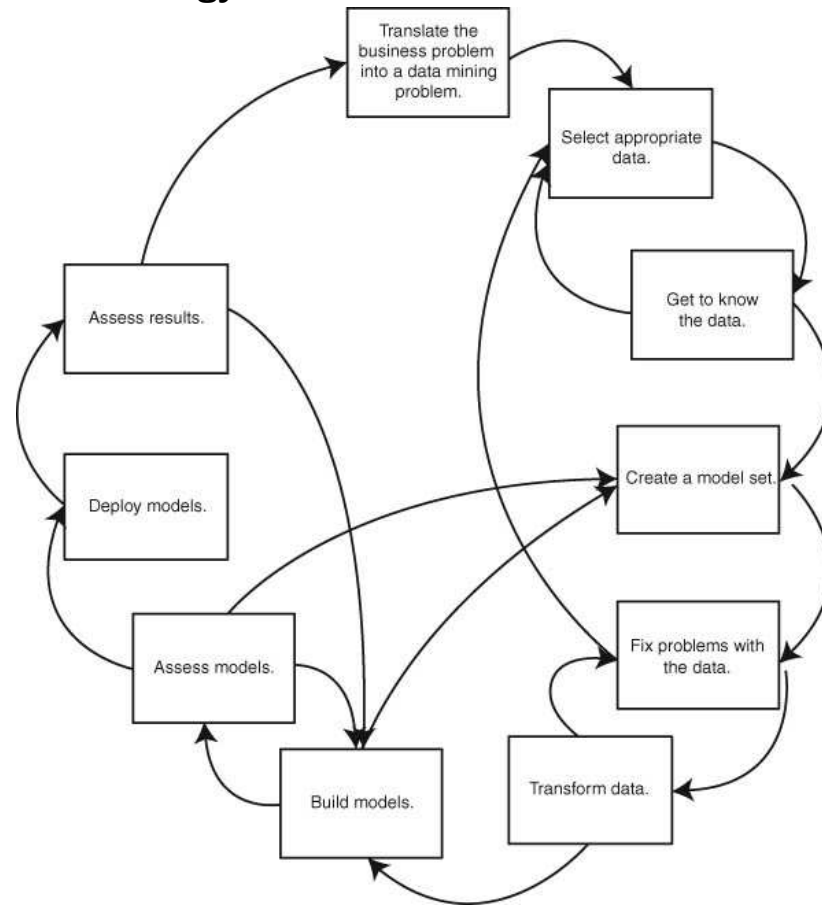
Statistical practices

- Hypothesis testing:
 - generating hypotheses;
 - testing hypotheses.
- Models, profiling, and prediction:
 - models provide a mathematical abstraction of real-life useful for understanding patterns and associations;
 - profiling uses data from the past to describe what happened in the past;
 - prediction uses data from the past to predict what is likely to happen in the future.

© Iain Pardoe, 2006

3 / 7

Data mining methodology



Data mining methodology I

1. Translate business problem into DM problem:
 - what does a DM problem look like?
 - how will results be used and delivered?
 - what is the role of business users and information technology?
2. Select appropriate data:
 - what is available?
 - how much data is enough?
 - how much history is required?
 - how many variables?
 - what must the data contain?
3. Get to know data (see Chapter 17 in the book, and practice dataset on NYtowns on book website):
 - examine distributions;
 - compare values with descriptions;
 - validate assumptions;
 - ask lots of questions.

© Iain Pardoe, 2006

5 / 7

Data mining methodology II

4. Create a model set:
 - assemble customer signatures;
 - create a balanced sample (whether this is a good idea depends on DM technique used);
 - include multiple timeframes (to avoid over-reliance on a particular period);
 - ensure predictions respect logistical time constraints;
 - partition model set: training, validation, test.
5. Fix problems with data:
 - categorical variables with too many categories;
 - skewed distributions and outliers;
 - missing values;
 - definitions that change over time;
 - inconsistent coding.

© Iain Pardoe, 2006

6 / 7

Data mining methodology III

6. Transform data:
 - math functions like logarithms to “normalize” highly skewed data;
 - capture time trends by adding derived ratios of one time period’s value to an earlier time period’s value;
 - use subject-matter knowledge to add other derived ratios and combinations;
 - convert counts and totals to proportions.
7. Build models.
8. Assess models using lift, misclassification errors, root mean square error (RMSE), etc.
9. Deploy models.
10. Assess results.
11. Begin again (go back to 1).

© Iain Pardoe, 2006

7 / 7