An Introduction to CRISP-DM

John McConnell
27th February 2013
Smart Vision Europe Ltd.
Roles required

- **Business**
  - Domain
  - Subject Matter

- **Analytical**
  - Methodologies
  - What to use when

- **Data**
  - Management
  - Structure

- **Technology**
  - Integration
  - Building apps
CRISP-DM characteristics

• Good for projects of different sizes
• Highly iterative
• Clearly defined outputs from each step
  – See the CRISP User Manual for mode detail
• A strong framework for communication between the roles in any Data Mining/Predictive project
The CRISP-DM process

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modelling
5. Evaluation
6. Deployment

www.crisp-dm.org
The CRISP-DM process

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modelling
5. Evaluation
6. Deployment
1. Business understanding

• Get a clear understanding of the business objectives
  – E.g. We need to reduce downtime
  – E.g. We need to get more value from our assets
  – E.g. We need to reduce failures
  – E.g. We need to reduce the cost of maintenance

• Agree success criteria
  – E.g. We need to reduce downtime by 30%

• Translate to analytical objectives (if possible)

• Evaluate the cost/benefit

• Clearly understand how action can be taken based on the likely outcomes
  – How to deploy

• Document relevant resources, constraints, systems
The CRISP-DM process

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modelling
5. Evaluation
6. Deployment
2. Data understanding – High Level

• Identify the **data sources** and **fields** which may have a bearing on the business/analytical objectives
• Review data schemas and any other data documentation
• What looks relevant?
• What are the formats?
  – Databases, text files, excel, etc.
• What are the fieldnames?
  – Metadata
2. Data understanding – Low level

• **Explore** the data
• Typically looking for **patterns** between fields
• Using uni- and bi-variate analyses
  – Examine fields one-by-one or in pairs
  – Often using visualisation tools
• Test hypotheses
  – E.g. Failure is accelerated by average ground temperature
• Validate data
  – Identifies any issues involving anomalies
• Develops understanding and informs modelling
The CRISP-DM process

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modelling
5. Evaluation
6. Deployment
3. Data Preparation

- Data Understanding effectively designs this step
- Together with Data Understanding this can be more time consuming than expected
  - Sometimes 80% of a project
  - Especially for newer projects
- Typically integrates data from different sources
- Create composite measures
  - E.g. band variables
  - Apply formulae e.g. compute annualised figures and other ratios
- Comparable to ETL (Extract Transform Load)
The CRISP-DM process

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modelling
5. Evaluation
6. Deployment
4. Modelling

- Apply a variety of modelling techniques
- Candidate list identified during understanding phase
  - Driven by data types (see later)
  - Constrained by available tools
- 2 broad styles:
  a) **Hypothesis led.** Add the fields/predictors that we believe are driving the outcome
  b) **Data led.** Add more fields at the beginning and incrementally reduce (and/or let the algorithms do that)
- The best performing modelling algorithm is a function of the specific data/problem
The CRISP-DM process

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modelling
5. Evaluation
6. Deployment
5. Evaluation

- Essential that the models are tested against unseen data
- Typically the data is partitioned into 2 (or 3) sets at random e.g. 70%:30%
  1. Training (modelling) set
  2. Test (holdout) set
  3. Evaluation set
- Evaluate against the success criteria agreed in the understanding phase
- Often it is about how well the model performs against a given value criteria e.g. revenue
  - Defined in Data Understanding phase
The CRISP-DM process

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modelling
5. Evaluation
6. Deployment
6. Deployment (1)

• Could be as simple as a list of names and predictions/scores
  – E.g. a mailing list
• Could be as complex as a model encapsulated as a computer program and embedded in an operational system to predict in real time and automate decisions
  – E.g. a model embedded in a system which sends alerts and triggers
• Could be embedded in a What-if? simulator
• Important to distinguish between a model in the modelling and deployment phases
• Typically...
  – In the modelling phase many different models and modelling options are built and evaluated
  – In the deployment phase the winning model(s) are fixed
    • E.g. we deploy a decision tree with a fixed shape
6. Deployment (2) (Monitoring)

- If we did our job properly then the deployed model should correspond to what we saw in evaluation
  - Other factors may intervene
- Ongoing evaluation (“monitoring”) still needs to happen if models are to be used over time
  - Some models have a longer shelf life than others
- More recently there has been some development of models which adapt/correct themselves to changing circumstances
  - Some level of re-modelling to improve accuracy
    - “Self adapting”
  - More commonly this is achieved through the concept of champion/challenger modelling or model refresh approaches
What-if?... scenarios

Part of a broader area of activity under the heading of Decision Management
The CRISP-DM process

The process is highly iterative
CRISP-DM summary points

• A DM project can be more like a Research & Development project
  – Can we build a successful model?
  – Has anyone done this before?
  – What is the risk that we cannot achieve the objectives?

• Hence projects can fail
  – Communication and alignment between the roles are key

• Needs to keep evolving to reflect the changes in practice
  – We’ve proposed “Monitoring” as a separate step