MLDM

The CRISP-DM methodology

Claudio Sartori

DISI
Department of Computer Science and Engineering – University of Bologna, Italy
claudio.sartori@unibo.it

Standard Process Model

- Can Data Mining be a push–button technology?

Standard Process Model

- Can Data Mining be a push–button technology? No

Standard Process Model

- Can Data Mining be a push–button technology? No
- Data Mining is a process

Standard Process Model

- Can Data Mining be a push–button technology? No
- Data Mining is a process
- The process has steps and complex choices

Standard Process Model

- Can Data Mining be a push–button technology? No
- Data Mining is a process
- The process has steps and complex choices
- The standard defines the steps in a precise way

Benefits of a Standard Process Model I

DM requires

- a mix of good tools and skilled analysts
- a sound methodology
- project management
- a process model to manage interactions along the process

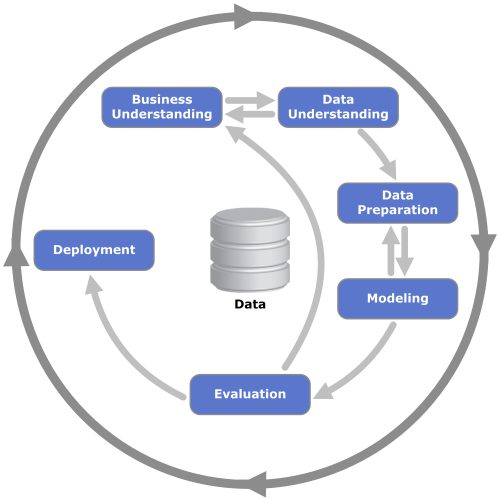
Benefits of a Standard Process Model II

Standardisation provides

- a common reference point for discussions
- a common understanding between the designers and the customers
- a basis for good engineering practice
- checklists
- clarity for expectations

The CRISP-DM methodology

From the problem to the application - https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining



Business understanding

- reformulate the problem in many ways, as necessary
- think about the scenario
- iterative refinement of problem formulation and scenario

Business understanding – Tasks I

- Determine
  - Business Objectives
  - Background Business Objectives
  - Business Success Criteria
- Assess Situation
  - Inventory of Resources
  - Requirements, Assumptions, and Constraints
  - Risks and Contingencies Terminology
  - Costs and Benefits

Business understanding – Tasks II

- Determine Goals
  - Data Mining Goals
  - Data Mining Success Criteria
- Produce Plan
  - Project Plan
  - Initial Assessment of Tools and Techniques

Data understanding

- which raw data are available?
  - they match rarely the problem needs
  - they are usually collected for different purposes (or for no purpose at all)
    - a customer database, a transaction database, and a marketing response database contain different information, may cover different intersecting populations, and may have varying degrees of reliability
- at which cost?
  - internal data are for free, external data may be not
  - interesting information may need to be collected with ad–hoc campaign
- possible forks in the project choices, according to the collected data

Data Understanding – Tasks

- Collect Initial Data
  - Initial Data Collection Report
- Describe Data
  - Data Description Report
- Explore Data
  - Data Exploration Report
- Verify Data Quality
  - Data Quality Report

Data preparation

- some analysis technique may require data transformations
  - converting to tabular format
  - converting between data types
    - e.g. from numeric to symbolic and viceversa
- some transformation can improve the quality of the results
  - normalization, scaling, guessing missing data, cleaning wrong data
  - ...
- *data leaks*
  - it is the case for supervised cases: the information necessary for the decision is not available at the decision time
- this task is usually very expensive and time consuming

Data Preparation – Tasks

- Data Set
  - Data Set Description
- Select Data
  - Rationale for Inclusion / Exclusion
- Clean Data
  - Data Cleaning Report
- Construct Data
  - Derived Attributes
  - Generated Records
- Integrate Data
  - Merged Data
- Format Data
  - Reformatted Data

Modeling

Capture patterns hidden in data

Modeling – Tasks

- Select Modeling Technique
  - Modeling Technique
  - Modeling Assumptions
- Generate Test Design
  - Test Design
- Build Model
  - Parameter Settings
  - Models
  - Model Description
- Assess Model
  - Model Assessment
  - Revised Parameter Settings

Evaluation

- rigorous assessment of the results of the data mining process
- compare different choices on a *qualitative* and *quantitative* basis
- evaluate the confidence of the derived models
- estimate the expected impact on the business
  - e.g. how many wrong decisions can we expect? which will be the cost of wrong decisions?

Evaluation – Tasks

- Evaluate Results
  - Assessment of Data Mining results w.r.t Business Success Criteria
  - Approved models
- Review Process
  - Review of Process
- Determine next steps
  - List of possible actions
  - Decisions

Deployment

The results of the DM process (i.e. the models) are used in software systems to obtain some return of investments

- e.g. in *churn* analysis the model for predicting likelihood of churn can be integrated with a package for churn management, for instance sending special offers to selected customers considered *high–risk of churn*

Deployment – Tasks

- Plan Deployment
  - Deployment Plan
- Plan Monitoring and Maintenance
  - Monitoring and Maintenance Plan
- Produce Final Report
  - Final Report Final Presentation
- Review Project
  - Experience Documentation

Bibliography

- Shearer, C. (2000).
  The CRISP-DM model: The new blueprint for data mining.
  Journal of Data Warehousing, 5:13–22.
- Wirth, R. and Hipp, J. (2000).
  CRISP-DM: Towards a standard process model for data mining.
  Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining.