Data Mining

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

- tools and skills
- methodology
- management
- process model

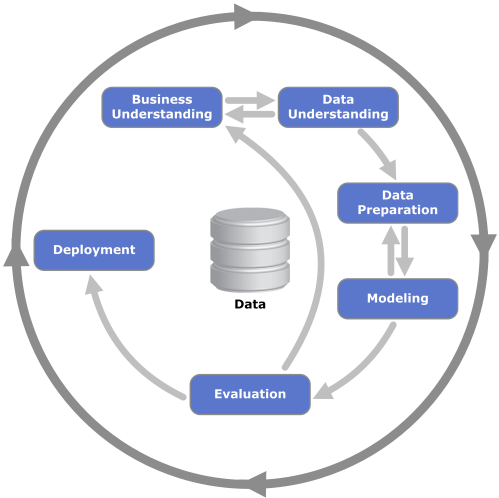
Benefits of a Standard Process Model II

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

General attitude

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

Business understanding – to be determined - I

- Business Objectives
  - specific, action–oriented
  - example: *Increase customer retention by identifying at-risk customers through churn prediction models*
- Background Business Objectives
  - broader strategy, long–term aims
  - example: *Become the market leader in customer satisfaction in the telecom industry*

Business understanding – to be determined - II

Business Success Criteria – Examples

1. Sales increased by 10% after implementing a recommendation engine
2. Customer support costs reduced by 15% through chatbot implementation
3. Churn rate decreased from 20% to 15% over six months
4. Achieve a CSAT score above 90% after improving service delivery
5. Reduce time-to-decision for credit approvals from 3 days to 1 hour
6. Production line efficiency increased by 20% after predictive maintenance
7. Increase market share by 5% in a target region
8. Improve Net Promoter Score (NPS) by 8 points
9. Average session time on a personalized app increased by 25%

Business understanding – Assess Situation

- Inventory of Resources
- Requirements, Assumptions, and Constraints
- Risks and Contingencies Terminology
- Costs and Benefits

Focus: Inventory of resources

Category	Examples
Data Resources	Available datasets, databases, data warehouses, data formats, data quality
Human Resources	Data scientists, domain experts, business analysts, IT staff
Computing Resources	Hardware (servers, GPUs), cloud services, storage, network capacity
Software Tools	Data mining tools (e.g., Python, R, SAS, RapidMiner), database tools, BI tools
Time & Budget	Project timelines, milestones, allocated budget
Other Resources	Access to APIs, third-party data sources, documentation, previous models

Focus: Requirements – examples

- The model must predict customer churn with at least 85% accuracy
- The system must integrate with an existing CRM platform
- Reports must be generated weekly for stakeholders

Focus: Assumptions – examples

- The available data covers all customer segments
- Historical data is representative of future trends
- Data privacy compliance (e.g., GDPR) will be maintained

Focus: Constraints – examples

- Limited budget or time (e.g., project must be completed in 4 weeks)
- Data cannot be transferred outside a specific region due to regulations
- Only open-source tools can be used

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
- Describe Data
- Explore Data
- Verify Data Quality

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 Description
- Select Data
  - Rationale for Inclusion / Exclusion
- Clean Data
- Construct Data
- Integrate Data
- Format Data

All the preparation activities must be traced and documented

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

- Assessment of Data Mining results w.r.t Business Success Criteria
- Review 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

CRISP-DM: Phases vs Actors

Phase	Stakeholders	Business Analysts	Domain Experts	Data Engineers	Data Scientists	DevOps/Developers
Business Understanding	✓	✓	✓			
Data Understanding		✓	✓	✓		
Data Preparation				✓	✓	
Modeling					✓	
Evaluation	✓	✓			✓	
Deployment						✓

Legend: ✓ = Actor involved in phase

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.