**CRISP-DM**

Cross Industry Standard Process for Data Mining

These notes are gathered from a reading of *Journal of Data Warehousing, Volume 5, Number 4, Fall 2000*, a 101communications production. Reference section at end.

**Phases of CRISP-DM**

* Business understanding
* Data understanding
* Data preparation
* Modeling
* Evaluation
* Deployment

**Business Understanding**

1. Determine business objectives
2. Assess the situation
3. Determine the data mining goals
4. Producing the project plan

Understand the project objectives from a business perspective

Converting the knowledge into a data mining problem definition

Then developing a preliminary plan designed to achieve objectives

**Determine the Business Objectives**

* What is the client’s true goal?
* Determine the measure(s) of success and how it (they) relate to the business objective(s)

**Access the situation**

* Outline the resources (personnel, software, etc.) available to accomplish the data mining project
* What data is available to meet the primary goal?
* List assumptions made in the project
  + i.e. a minimum number of customers over age 50 are necessary to address the business question
* List the project risks
* List potential solutions to those risks
* Create a glossary of business and data mining terms
* Construct a cost-benefit analysis for the project (is the project worth it?)
  + <https://www.investopedia.com/terms/c/cost-benefitanalysis.asp>
  + A cost-benefit analysis (CBA) is the process used to measure the benefits of a decision or taking action minus the costs associated with taking that action.

**Determine the Data mining Goals**

* The goals state project objectives in business terms
* Success should also be defined in business terms
* If the business goal cannot be effectively translated into a data mining goal, consider redefining the problem.

**Produce a Project Plan**

* Outline specific steps
* Propose a timeline
  + 50%-70% Data Preparation Phase
  + 20%-30% is Data Understanding Phase
  + 10%-20% in Modeling, Evaluation, Business Understanding Phases
  + 5%-10% spent in Deployment Planning Phase
* Access potential risks
* Access initial tools and techniques needed to support the project

**Data Understanding**

1. Collect the Initial Data
2. Describe the Data
3. Explore the Data
4. Verify Data Quality

**Collect the Initial Data**

* Acquire the data (load it if needed)
* Is there any lag time from any data sources to note of?

**Describe the Data**

* Examine the gross or surface properties of the data and report on results
  + Format? Quantity? Number of rows and columns (records and fields)?
* Does the data acquired satisfy the relevant requirements?

**Explore the Data**

* Query and visualize the data to see what’s going on
* Report on your findings

**Verify Data Quality**

* Is the data complete?
* Missing or blank values?
* Spelling correct?
* Does it make common sense?

**Data Preparation**

1. Select Data
2. Clean Data
3. Construct Data
4. Integrate Data
5. Format Data

**Select Data**

* Select data based on relevance to goals, quality and technical constraints (like limits on data volume)
* Explain why certain data is included or excluded
* Which attribute(s) are the most important?

**Clean Data**

* Choose the right subset of data or
* Estimate missing data if you need to through modeling analysis

**Construct Data**

* Produce derived attributes if needed

**Integrate Data**

* Combine data from multiple tables
* You can aggregate data (min, max, average, first, etc.) in during this step

**Format Data**

* Example: remove illegal characters, trimming length, change how a date is displayed, etc.

**Modeling**

1. Select the Modeling Technique
2. Generate Test Design
3. Build the Model
4. Assess the Model

**Select the Modeling Technique**

* Decision Tree? (list other techniques here)

**Generate Test Design**

* placeholder

**Build the Model**

* placeholder

**Assess the Model**

* Interpret models based on your domain knowledge

**Evaluation**

1. Evaluate Results
2. Review Process
3. Determine Next Steps

**Evaluate Results**

* Is it accurate?
* Does it meet the business objectives?

**Review Process**

* Was the model built correctly? Is there any data category that was overlooked?

**Determine Next Steps**

* Do we move onto deployment now or do we initiate further iterations?

**Deployment**

1. Plan Deployment
2. Plan Monitoring and Maintenance
3. Produce Final Report
4. Review Project

**Plan Deployment**

* Take evaluation results and develop a strategy for deployment

**Plan Monitoring and Maintenance**

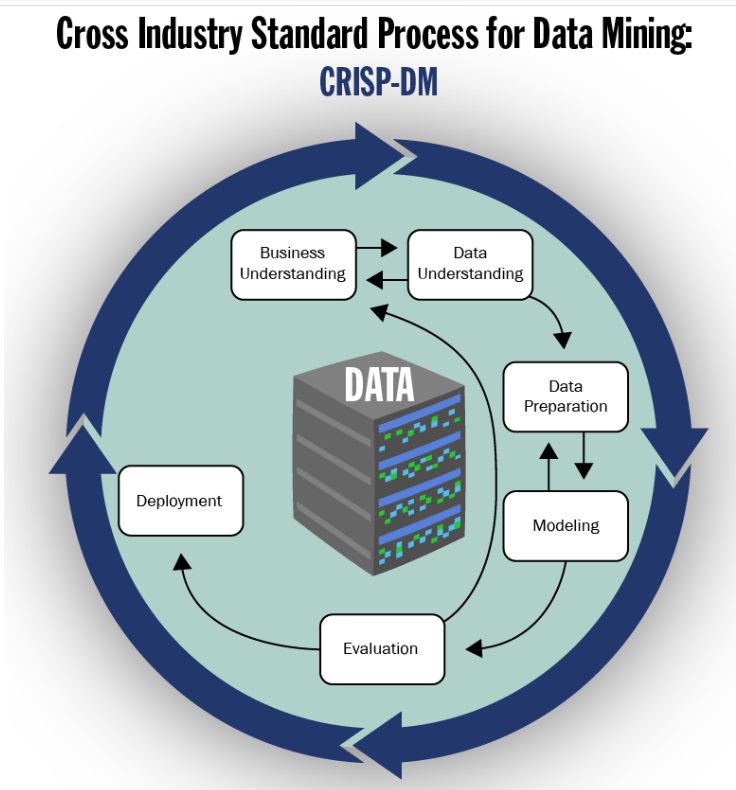
* Enables correct usage of data mining results

**Produce Final Report**

* Could be a summary report or comprehensive review depending on nature of the project

**Review Project**

* Assess failures, successes, and areas of improvement



**Reference**

Shearer, C. (October, 2000). The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing, Volume 5 (Number 4), pages 13-22.