The CRISP-DM Methodology

Introduction

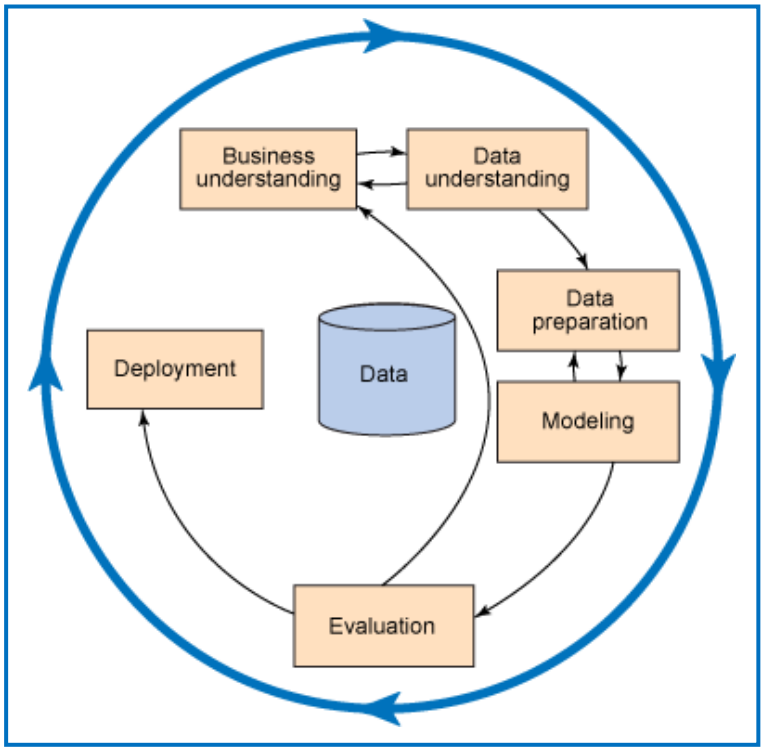
The CRoss-Industry Standard Process for Data Mining (CRISP-DM) was conceived in 1996 by Daimler-Chrysler, SPSS and NCR to be a structured and robust methodology for planning and carrying out data mining projects. A core part of CRISP-DM is ensuring that the data are in the right form to meet the project's business and modelling objectives. CRISP-DM is not linked to any tool or application. PAM Analytics uses CRISP-DM.

The analytics component of a project is the means to the project's ends, not the ends themselves. The ends are the results, conclusions and business advantage gained by using the analytics to leverage the data. Additionally, a frequent result of using analytics is identifying areas for further investigation.

Methodology

Figure 1 shows the six stages of CRISP-DM and how they are related.

Figure 1



As the name suggests and as shown in Figure 1, CRISP-DM is a *process*, not an *event*. The iterative nature of CRISP-DM is clear from Figure 1. The process is iterative because the results of some stages may require the project cycle to go back to earlier stages. For example, a result of the modelling stage may be that more data preparation is required – new data may be needed or the existing data may have to be prepared in a different way. Each stage of the process is described below.

Business Understanding

This is the first stage of the process. Its aims are to:

- specify and agree with the client the business context and objectives of the project
- become familiar with the client's current procedures
- define the modelling objectives and agree the project's priorities and success criteria
- plan the next phases of the project.

Data Understanding

The aims of the Data Understanding stage are to:

- identify the data required to achieve the project's objectives
- decide which data to use. The selection criteria can be based on a time period or on selected units or on both.
- establish the data's availability and accessibility. If some data are difficult to access, for example because they are stored in legacy systems, the use of alternative data should be investigated.
- gather the data and become familiar with them
- review the quality of the data and improve it if necessary (as is invariably the case). Data quality issues must be addressed during this stage of the project to ensure that the data used during later stages of the project are fit for purpose. Failure to do so will have adverse and possible serious consequences on the later stages.
- carry out exploratory data analysis and high-level modelling. This will reveal initial insights, some of which may be unexpected and so require earlier stages of the project plan to be revised, as shown in Figure 1.
- plan the next phases of the project.

Data Preparation

The aims of the Data Preparation stage are to:

- prepare the data for the Modelling stage. Data preparation has many aspects, including:
  - searching for and correcting inconsistencies in the data
  - searching for and removing records with spurious data
  - searching for and removing duplicate records
  - searching for and redefining or removing outliers
  - imputing missing values if possible

- redefining variables as necessary
- calculating new variables.

♦ construct the data repository

♦ plan the next phases of the project.

Modelling

The aims of the Modelling stage are to:

♦ develop a number of models, possibly using a variety of techniques

♦ assess the models' results and conclusions

♦ plan the next phases of the project.

Evaluation

The aims of the Evaluation stage are to:

♦ assess and compare the results of the models with the client

♦ decide which models to use and how they should be used

♦ identify areas for further investigation and modelling

♦ plan the Deployment phase of the project.

Deployment

The Model Deployment varies in complexity from relatively simple report generation to implementing a suite of complex models.

After the models have been used for some time, they should be refreshed. This work should also be carried out using CRISP-DM but it may not require the first three stages (Business Understanding, Data Understanding and Data Preparation) to the same extents as for the first model.

