The Data Mining Process

Data mining is a craft. It involves the application of a substantial amount of science and technology, but the proper application still involves art as well. But as with many mature crafts, there is a well-understood process that places a structure on the problem, allowing reasonable consistency, repeatability, and objectiveness.

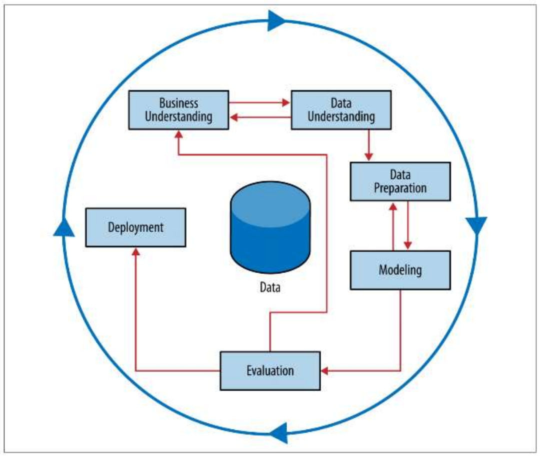

Figure 2-2. The CRISP data mining process.

A useful codification of the data mining process is given by the Cross Industry Standard Process for Data Mining (CRISP-DM; Shearer, 2000), illustrated in Figure 2-2. This process diagram makes explicit the fact that iteration is the rule rather than the exception.

Going through the process once without having solved the problem is, generally speaking, not a failure. Often the entire process is an exploration of the data, and after the first iteration the data science team knows much more. The next iteration can be much more well-informed.

1. **Business Understanding**
   a. The initial formulation may not be complete or optimal so multiple iterations may be necessary for an acceptable solution formulation to appear
   b. Ask questions such as: "What exactly do we want to do?", "How exactly would we do it?", "What parts of this use scenario constitute possible data mining models?"
   c. We will loop back and realize that often the use scenario must be adjusted to better reflect the actual business need
2. **Data Understanding**
   a. If solving the business problem is the goal, the data comprise the available raw material from which the solution will be built.
   b. A critical part of the data understanding phase is estimating the costs and benefits of each data source and deciding whether further investment is merited
   c. Supervised technique: credit card transactions have reliable labels (fraud and legitimate)
   d. Unsupervised technique: Medicare fraud; The perpetrators of fraud—medical providers who submit false claims, and sometimes their patients—are also legitimate service providers and users of the billing system. Such a problem usually requires unsupervised approaches such as profiling, clustering, anomaly detection, and co-occurrence grouping
3. **Data Preparation**
   a. Data are manipulated and converted into forms that yield better results
   b. Data is converted to tabular format, removing, or inferring missing values, and converting data to different types.
   c. Often the quality of the data mining solution rests on how well the analysts structure the problems and craft the variables
   d. Be aware of a 'leak', which is a situation where a variable collected in historical data that gives information on the target variable—information that appears in historical data but is not actually available when the decision has to be made.

4. **Modeling**
   a. The output of modeling is some sort of model or pattern capturing regularities in the data.
   b. The modeling stage is the primary place where data mining techniques are applied to the data
5. **Evaluation**
   a. To assess the data mining results rigorously and to gain confidence that they are valid and reliable before moving on.
   b. The evaluation stage also serves to help ensure that the model satisfies the original business goals.
   c. Evaluating the results of data mining includes both quantitative and qualitative assessments.
   d. To facilitate such qualitative assessment, the data scientist must think about the comprehensibility of the model to stakeholders
6. **Deployment**
   a. The clearest cases of deployment involve implementing a predictive model in some information system or business process.
   b. The results of data mining—and increasingly the data mining techniques themselves—are put into real use in order to realize some return on investment.
   c. Two main reasons for deploying the data mining system itself rather than the models produced by a data mining system are (i) the world may change faster than the data science team can adapt, as with fraud and intrusion detection, and (ii) a business has too many modeling tasks for their data science team to manually curate each model individually.
   d. Regardless of whether deployment is successful, the process often returns to the Business Understanding phase