Modeling

4 Modeling

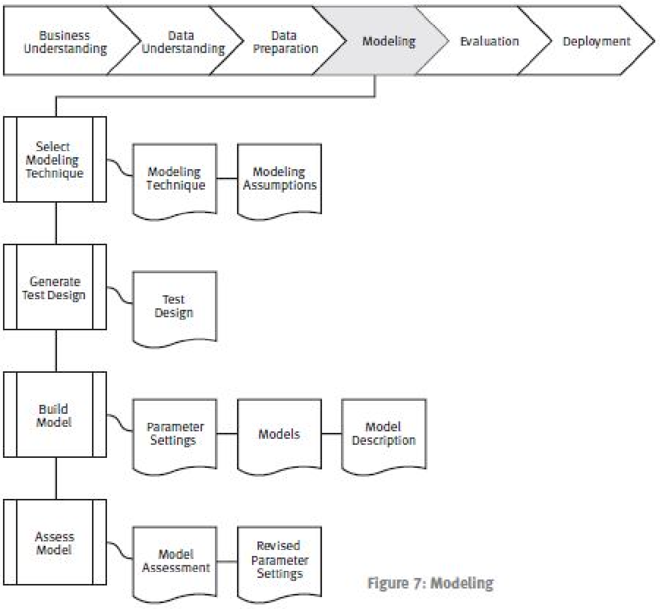


Figure 7: Modeling

Task: Select modeling technique

As the first step in modeling, select the actual modeling technique that is to be used. Although you may have already selected a tool during the Business Understanding phase, this task refers to the specific modeling technique, e.g., decision-tree building with C5.0, or neural network generation with back propagation. If multiple techniques are applied, perform this task separately for each technique.

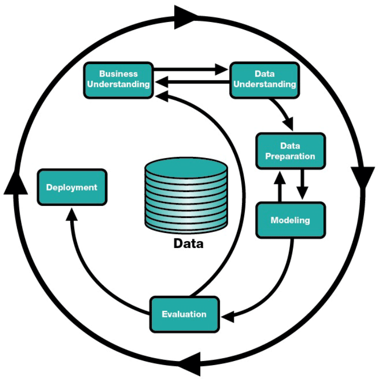
Outputs: Modeling technique
Document the actual modeling technique that is to be used.

Modeling assumptions
Many modeling techniques make specific assumptions about the data—for example, that all attributes have uniform distributions, no missing values allowed, class attribute must be symbolic, etc. Record any such assumptions made.

Task: Generate test design
Before we actually build a model, we need to generate a procedure or mechanism to test the model's quality and validity. For example, in supervised data mining tasks such as classification, it is common to use error rates as quality measures for data mining models. Therefore, we typically separate the dataset into train and test sets, build the model on the train set, and estimate its quality on the separate test set.

Output: Test design
Describe the intended plan for training, testing, and evaluating the models. A primary component of the plan is

Phases of the CRISP-DM reference model



Smart Vision eNewsletter

Click here to sign up

determining how to divide the available dataset into training, test, and validation datasets.

Task: Build model

Run the modeling tool on the prepared dataset to create one or more models.

Outputs: Parameter settings
With any modeling tool, there are often a large number of parameters that can be adjusted. List the parameters and their chosen values, along with the rationale for the choice of parameter settings.

Models
These are the actual models produced by the modeling tool, not a report.

Model descriptions
Describe the resulting models. Report on the interpretation of the models and document any difficulties encountered with their meanings.

Task: Assess model
The data mining engineer (now often referred to as a Data Scientist) interprets the models according to his domain knowledge, the data mining success criteria, and the desired test design. The data mining engineer judges the success of the application of modeling and discovery techniques technically; he contacts business analysts and domain experts later in order to discuss the data mining results in the business context. Please note that this task only considers models, whereas the evaluation phase also takes into account all other results that were produced in the course of the project.

The data mining engineer tries to rank the models. He assesses the models according to the evaluation criteria. As much as possible, he also takes into account business objectives and business success criteria. In most data mining projects, the data mining engineer applies a single technique more than once, or generates data mining results with several different techniques. In this task, he also compares all results according to the evaluation criteria.

Outputs: Model assessment
Summarize results of this task, list qualities of generated models (e.g., in terms of accuracy), and rank their quality in relation to each other.

Revised parameter settings
According to the model assessment, revise parameter settings and tune them for the next run in the Build Model task. Iterate model building and assessment until you strongly believe that you have found the best model(s). Document all such revisions and assessments

Proceed to evaluation