CRISP-DM stage four – modelling

Select modelling technique

Task

As the first step in modelling, select the actual modelling 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 modelling 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

- **Modelling technique** – document the actual modelling technique that is to be used.
- **Modelling assumptions** – many modelling 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 assumptions made.

Generate test design

Task

Before you build a model you 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, you 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 determining how to divide the available dataset into training, test and validation datasets.

Build model

Task

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

Outputs

- **Parameter settings –** with any modelling 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 modelling 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.

Assess model

Task

Interpret the models according to your domain knowledge, your data mining success criteria and your desired test design. Judge the success of the application of modelling and discovery techniques technically, then contact business analysts and domain experts later in order to discuss the data mining results in the business context. This task only considers models, whereas the evaluation phase also considers all other results that were produced in the course of the project. At this stage you should rank the models and assess them according to the evaluation criteria. As much as possible, take business objectives and business success criteria into account. In most data mining projects, a single technique is applied more than once and data mining results are generated with several different techniques.

Outputs

- **Model assessment –** summarise 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.