CRISP-DM
Discussion Paper
March, 1999

The CRISP-DM Process Model

Pete Chapman (NCR)

Randy Kerber (NCR)

Julian Clinton (SPSS)

Thomas Khabaza (SPSS)

Thomas Reinartz (DaimlerChrysler)

Rüdiger Wirth (DaimlerChrysler)

This discussion paper describes the enhanced CRISP-DM Process Model, including an introduction to the CRISP-DM methodology, the CRISP-DM Reference Model, the CRISP-DM User Guide, and the CRISP-DM outputs, as well as an appendix with additional useful and related information.

Conventions Used in This Document

The following notational conventions are used in this document:

Trademarks

All trademarks and service marks mentioned in this document are marks of their respective owners and are as such acknowledged by the members of the CRISP-DM consortium.

Control Information

Page 91 is the last page of this document.

Contents

I Introduction

1 The CRISP-DM Methodology

1.1 Hierarchical Breakdown

The CRISP-DM data mining methodology is described in terms of a hierarchical process model, consisting of sets of tasks described at four levels of abstraction (from general to specific): *phase*, *generic task*, *specialised task*, and *process instance* (see figure 1).

At the top level, the data mining process is organized into a number of *phases*; each phase consists of several second-level *generic tasks*. This second level is called generic, because it is intended to be general enough to cover all possible data mining situations. The generic tasks are intended to be as *complete* and *stable* as possible. Complete means covering both the whole process of data mining and all possible data mining applications. Stable means that the model should be valid for yet unforeseen developments like new modelling techniques.

The third level, the *specialised task* level, is the place to describe how actions in the generic tasks should be carried out in certain specific situations. For example, at the second level there might be a generic task called *clean data*. The third level would describe how this task differed in different situations, such as cleaning numeric values versus cleaning categorical values, or whether the problem type is clustering or predictive modeling.

The description of phases and tasks as discrete steps performed in a specific order represents an idealised sequence of events. In practice, many of the tasks can be performed in a different order and it will often be necessary to repeatedly backtrack to previous tasks and repeat certain actions. Our process model does not attempt to capture all of these possible routes through the data mining process because this would require an overly complex process model.

The fourth level, the *process instance*, is a record of the actions, decisions, and results of an actual data mining engagement. A process instance is organized according to the tasks defined at the higher levels, but represents what actually happened in a particular engagement, rather than what happens in general.

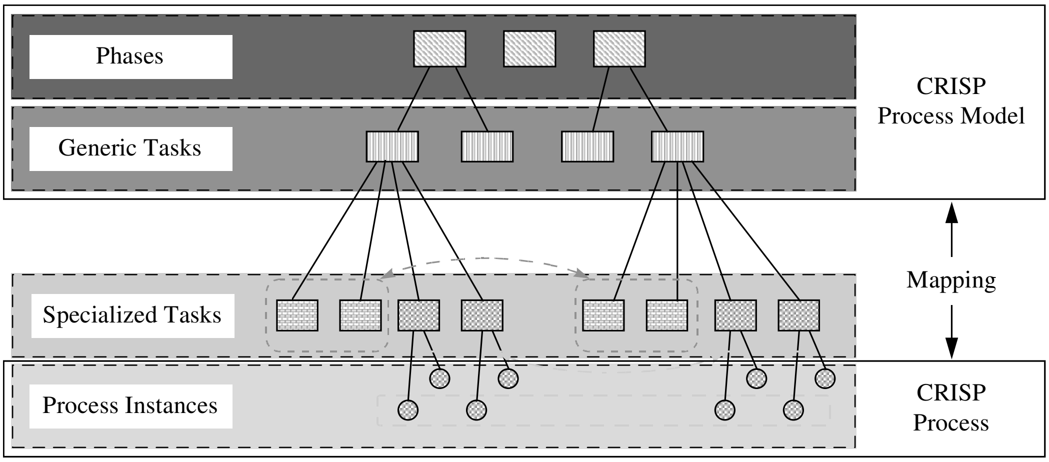


Figure 1: Four Level Breakdown of the CRISP-DM Methodology

1.2 Reference Model and User Guide

Horizontally, the CRISP-DM methodology distinguishes between the *Reference Model* and the *User Guide*. Whereas the Reference Model presents a quick overview of phases, tasks, and their outputs, and describes *what to do* in a data mining project, the User Guide gives more detailed tips and hints for each phase and each task within a phase and depicts *how to do* a data mining project.

This document covers both the Reference Model and the User Guide at the generic level.

2 Mapping Generic Models to Specialized Models

2.1 Data Mining Context

Mapping between the generic and the specialized level in CRISP-DM is driven by the *Data Mining Context*. Currently, we distinguish between four different dimensions of data mining contexts:

1. The *application domain* is the specific area in which the data mining project takes place.
2. The *data mining problem type* describes the specific class(es) of objective(s) which the data mining project deals with (see also appendix V.2).
3. The *technical aspect* covers specific issues in data mining which describe different (technical) challenges that usually occur during data mining.
4. The *tool and technique* specifies which data mining tool(s) and/or techniques are applied during the data mining project.

Table 1 below summarizes these dimensions of data mining contexts and shows specific examples for each dimension.

Table 1: Dimensions of Data Mining Contexts and Examples

	Data Mining Context			
Dimension	*Application Domain*	*Data Mining Problem Type*	*Technical Aspect*	*Tool and Technique*
Examples	Response Modeling	Description and Summarization	Missing Values	Clementine
	Churn Prediction	Segmentation	Outliers	MineSet
	...	Concept Description	...	Decision Tree
		Classification		...
		Prediction		
		Dependency Analysis		

A specific data mining context is a concrete value for one or more of these dimensions. For example, a data mining project dealing with a classification problem in churn prediction constitutes one specific context. The more values for different context dimensions are fixed, the more concrete is the data mining context.

2.2 Mappings with Contexts

We distinguish between two different types of mapping between generic and specialized level in CRISP-DM:

1. "Mapping for the Presence":
   If we only apply the generic process model to perform a single data mining project and attempt to map generic tasks and their descriptions to the specific project as required, we talk about a single mapping for (probably) only one usage.

2. "Mapping for the Future":
   If we systematically specialise the generic process model according to a pre-defined context (or similarly systematically analyse and consolidate experiences of a single project towards a specialised process model for future usage in comparable contexts), we talk about explicitly writing up a specialized process model in terms of CRISP-DM.

Which type of mapping is appropriate for your own purposes depends on your specific data mining context and the needs of your organization.

2.3 How to map?

The basic strategy for mapping the generic process model to the specialized level is the same for both types of mappings:

- Analyse your specific context.
- Remove any details not applicable to your context.
- Add any details specific to your context.
- Specialize (or instantiate) generic contents according to concrete characteristics of your context.
- Possibly rename generic contents to provide more explicit meanings in your context for the sake of clarity.

3 Description of Parts

3.1 Contents

The CRISP-DM Process Model (this document) is organized into five different parts:

- Part I is this introduction into the CRISP-DM methodology and provides some general guidelines for mapping the generic process model to specialised process models.
- Part II describes the CRISP-DM Reference Model, its phases, generic tasks, and outputs.
- Part III presents the CRISP-DM User Guide which goes beyond the pure description of phases, generic tasks, and outputs, and contains more detailed advice how to perform data mining projects including check lists.
- Part IV focuses on concrete specifications of each output and its components, and provides template documents for outputs if appropriate and shows cross references among outputs and tasks.
- Finally, part V is the appendix which covers a glossary of important terminology, as well as a characterization of data mining problem types.

3.2 Purpose

Users and readers of this document should be aware of the following instructions:

- If you start reading the CRISP-DM Process Model for the first time, it is worth to initially begin with part I, the introduction, in order to understand the CRISP-DM methodology and all its concepts and how different concepts, and hence different parts of this document, relate to each other.

In further readings, you might skip the introduction and only get back to it if necessary for clarification.

- If you need fast access to an overview of the CRISP-DM Process Model, you should refer to part II, the CRISP-DM Reference Model, either to begin with a data mining project quickly or to get an introduction to the CRISP-DM User Guide.
- If you need detailed advice in performing your data mining project, part III, the CRISP-DM User Guide, is the most valuable part of this document. Note, if you have not read the introduction or the Reference Model first, it might be helpful to step back and start reading with these two first parts.
- If you are in the state of data mining when you write up your reports and descriptions of deliverables, jump into part IV, the details on outputs and their template documents if appropriate.
  If you prefer to generate your deliverable descriptions during the project, you possibly move back and forth between parts III and IV as desired.
- Finally, the appendix is useful as additional background information on CRISP-DM and data mining. You can use the appendix to lookup various terms if you are not yet an expert in the field.

II The CRISP-DM Reference Model

The current process model for data mining provides an overview of the life cycle of a data mining project. It contains the corresponding phases of a project, their respective tasks, and relationships between these tasks. At this description level, it is not possible to identify all relationships. Essentially, there possibly exist relationships between all data mining tasks depending on the goals, the background and interest of the user, and most importantly on the data.

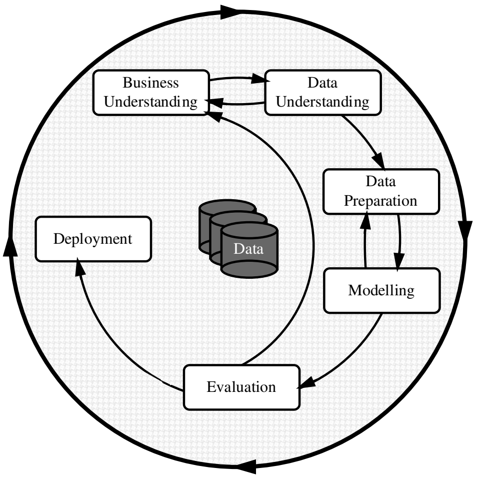


Figure 2: Phases of the CRISP-DM Reference Model

The life cycle of a data mining project consists of six phases. Figure 2 shows the phases of a data mining process. The sequence of the phases is not strict. Moving back and forth between different phases is always required. It depends on the outcome of each phase which phase, or which particular task of a phase, has to be performed next. The arrows indicate the most important and frequent dependencies between phases.

The outer circle in Figure 2 symbolizes the cyclic nature of data mining itself. Data mining is not over once a solution is deployed. The lessons learned during the process and from the deployed solution can trigger new, often more focused business questions. Subsequent data mining processes will benefit from the experiences of previous ones.

In the following, we outline each phase briefly:

- *Business Understanding*

  This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives.

- *Data Understanding*

  The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.

- *Data Preparation*

  The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools.

- *Modeling*

  In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed.

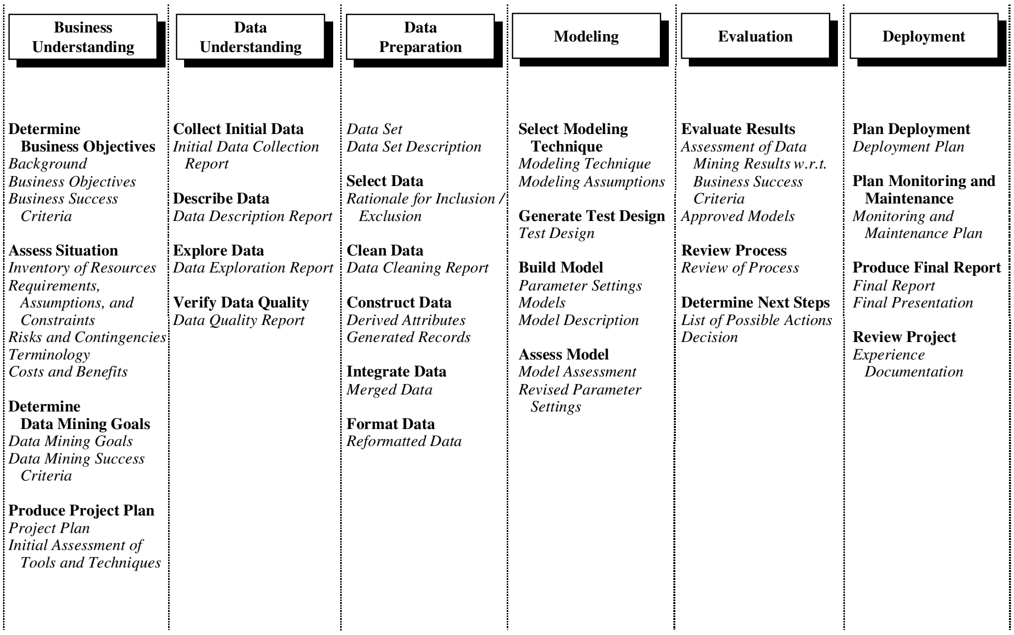
- *Evaluation*

  At this stage in the project you have built a model (or models) that appears to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

- *Deployment*

  Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. In many cases it will be the customer, not the data analyst, who will carry out the deployment steps. However, even if the analyst will not carry out the deployment effort it is important for the customer to understand up front what actions will need to be carried out in order to actually make use of the created models.

Figure 3 presents an outline of phases accompanied by generic tasks (bold) and outputs (italic). In the following sections, we describe each generic task and its outputs in more detail.



Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria*	**Collect Initial Data** *Initial Data Collection Report*	*Data Set* *Data Set Description*	**Select Modeling Technique** *Modeling Technique* *Modeling Assumptions*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models*	**Plan Deployment** *Deployment Plan*
Assess Situation *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits*	**Describe Data** *Data Description Report*	**Select Data** *Rationale for Inclusion / Exclusion*	**Generate Test Design** *Test Design*	**Review Process** *Review of Process*	**Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan*
Determine Data Mining Goals *Data Mining Goals* *Data Mining Success Criteria*	**Explore Data** *Data Exploration Report*	**Clean Data** *Data Cleaning Report*	**Build Model** *Parameter Settings* *Models* *Model Description*	**Determine Next Steps** *List of Possible Actions* *Decision*	**Produce Final Report** *Final Report* *Final Presentation*
Produce Project Plan *Project Plan* *Initial Assessment of Tools and Techniques*	**Verify Data Quality** *Data Quality Report*	**Construct Data** *Derived Attributes* *Generated Records*	**Assess Model** *Model Assessment* *Revised Parameter Settings*		**Review Project** *Experience Documentation*
		Integrate Data *Merged Data*			
		Format Data *Reformatted Data*			

Figure 3: Generic Tasks (bold) and Outputs (italic) of the CRISP-DM Reference Model

We focus our attention on task overviews and summaries of outputs.

1 Business Understanding

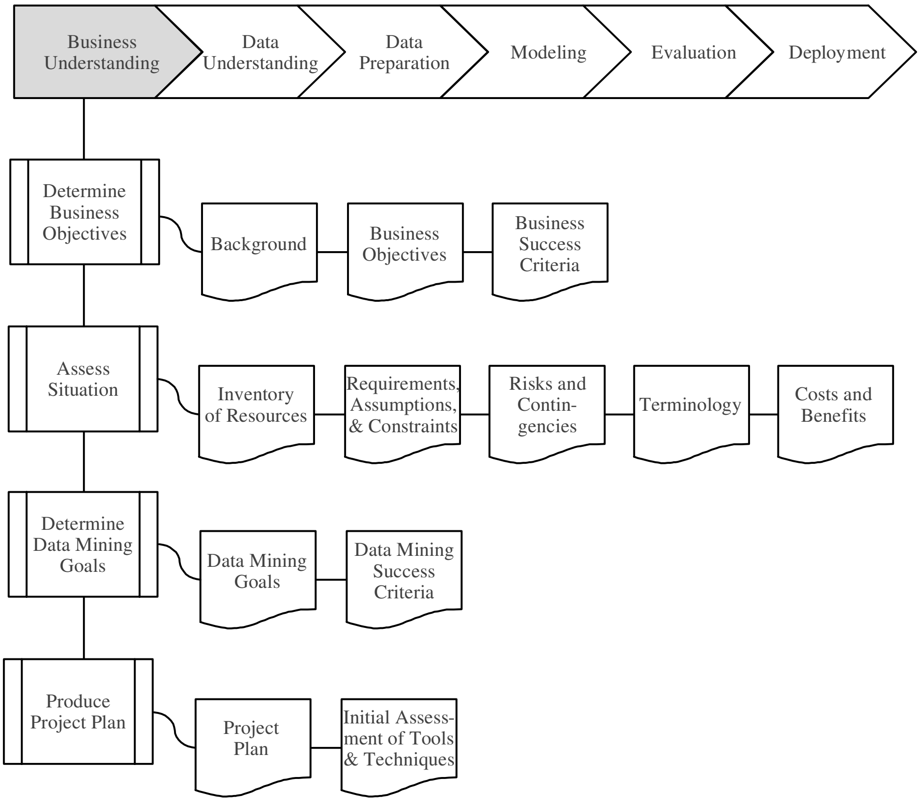


Figure 4: Business Understanding

1.1 Determine Business Objectives

Task **Determine Business Objectives**

The first objective of the data analyst is to thoroughly understand, from a business perspective, what the client really wants to accomplish. Often the client will have many competing objectives and constraints that must be properly balanced. The analyst's goal is to uncover important factors, at the beginning, that can influence the outcome of the project. A possible consequence of neglecting this step is to expend a great deal of effort producing the right answers to the wrong questions.

Outputs **Background**

Details the information that is known about the organization's business situation at the beginning of the project.

Business Objectives

Describe the customer's primary objective, from a business perspective. In addition to the primary business objective, there are typically a large number of related business questions that the customer would like to address. For example, the primary business goal might be to keep current customers by predicting when there are prone to move to a competitor. Examples of related business questions are things like "How does the primary channel (e.g., ATM, visit branch, internet) a bank customer uses affect whether they will stay or go?" or "Will lower ATM fees significantly reduce the number of high-value customers who leave?"

Business Success Criteria

Describe the criteria for a successful or useful outcome to the project from the business point of view. This might be quite specific, such as reduction of customer churn to a certain level, or general and subjective such as "give useful insights into the relationships". In the latter case it should be indicated who will make the subjective judgment.

1.2 Assess Situation

Task

Assess Situation

This task involves more detailed fact-finding about all of the resources, constraints, assumptions, and other factors that should be considered in determining the data analysis goal and project plan. In the previous task, your objective is to quickly get to the crux of the situation. Here, you want to flesh out the details.

Outputs

Inventory of Resources

List the resources available to the project, including: personnel (business experts, data experts, technical support, data mining personnel), data (fixed extracts, access to live warehoused or operational data), computing resources (hardware platforms), software (data mining tools, other relevant software).

Requirements, Assumptions, and Constraints

List all requirements of the project including schedule of completion, comprehensibility and quality of results, and security as well as legal issues. As part of this output, make sure that you are allowed to use the data.

List the assumptions made by the project. These may be assumptions about the data which can be checked during data mining, but may also include non-checkable assumptions about the business upon which the project rests. It is particularly important to list the latter if they form conditions on the validity of the results.

List the constraints on the project. These may be constraints on the availability of resources, but may also include technological constraints such as the size of data which it is practical to use for modeling.

Risks and Contingencies

List the risks, that is events which might occur to delay the project or cause it to fail. List the corresponding contingency plans; what action will be taken if the risks happen.

Terminology

A glossary of terminology relevant to the project. This may include two components:
(1) A glossary of relevant business terminology, which forms part of the business understanding available to the project. Constructing this glossary is a useful "knowledge elicitation" and education exercise.
(2) A glossary of data mining terminology, illustrated with examples relevant to the business problem in question.

Costs and Benefits

A cost-benefit analysis for the project; compare the costs of the project with the potential benefit to the business if it is successful. The comparison should be as specific as possible, for example using monetary measures in a commercial situation.

1.3 Determine Data Mining Goals

Task **Determine Data Mining Goals**

A *business goal* states objectives in business terminology. A *data mining goal* states project objectives in technical terms. For example, the business goal might be "Increase catalog sales to existing customers" while a data mining goal might be "Predict how many widgets a customer will buy, given their purchases over the past three years, demographic information (age, salary, city, etc.), and the price of the item".

Outputs **Data Mining Goals**

Describe the intended outputs of the project which will enable the achievement of the business objectives.

Data Mining Success Criteria

Define the criteria for a successful outcome to the project in technical terms, for example a certain level of predictive accuracy, or a propensity to purchase profile with a given degree of "lift". As with business success criteria, it may be necessary to describe these in subjective terms, in which case the person or persons making the subjective judgment should be identified.

1.4 Produce Project Plan

Task **Produce Project Plan**

Describe the intended plan for achieving the data mining goals, and thereby achieving the business goals. The plan should specify the anticipated set of steps to be performed during the rest of the project including an initial selection of tools and techniques.

Outputs **Project Plan**

List the stages to be executed in the project, together with duration, resources required, inputs, outputs and dependencies. Where possible make explicit the large-scale iterations in the data mining process, for example repetitions of the modeling and evaluation phases.

As part of the project plan, it is also important to analyse dependencies between time schedule and risks. Mark results of these analyses explicitly in the project plan, ideally with actions and recommendations if the risks appear.

Note, the project plan contains detailed plans for each phase. For example, decide at this point which evaluation strategy will be used in the evaluation phase.

The project plan is a dynamic document in the sense that at the end of each phase a review of progress and achievements is necessary and an update of the project plan accordingly is recommended. Specific review points for these reviews are part of the project plan, too.

Initial Assessment of Tools and Techniques

At the end of the first phase, the project also performs an initial assessment of tools and techniques. Here, you select a data mining tool which supports various methods for different stages of the process, for example. It is important to assess tools and techniques early in the process since the selection of tools and techniques possibly influences the entire project.

2 Data Understanding

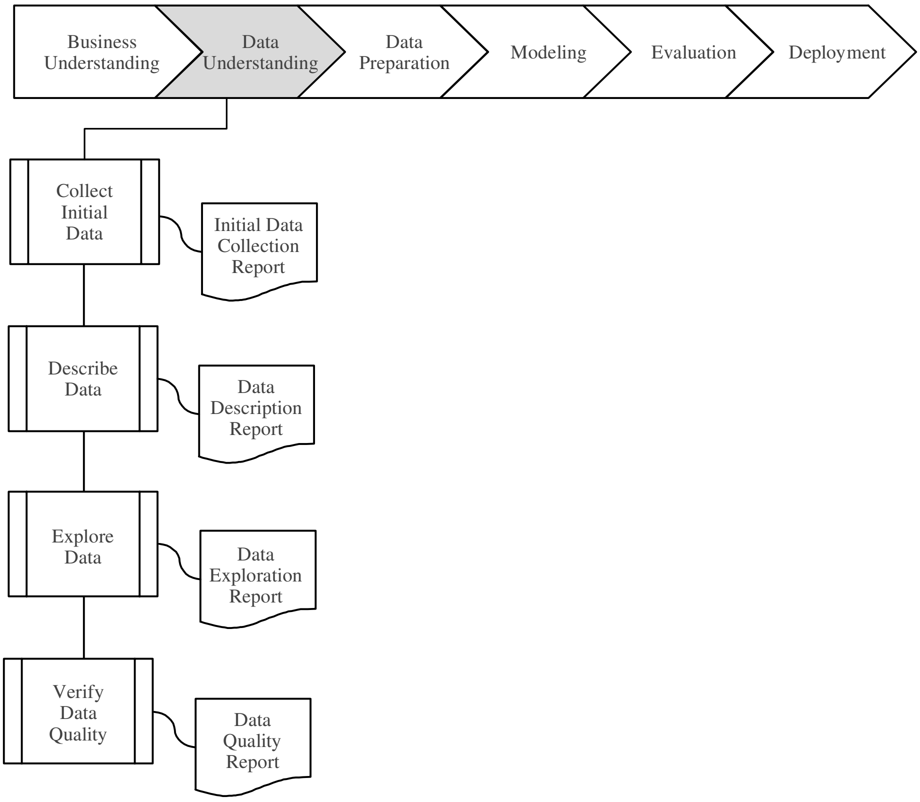


Figure 5: Data Understanding

2.1 Collect Initial Data

Task **Collect Initial Data**

Acquire within the project the data (or access to the data) listed in the project resources. This initial collection includes data loading if necessary for data understanding. For example, if you apply a specific tool for data understanding, it makes perfect sense to load your data into this tool. This effort possibly leads to initial data preparation steps. Note, if you acquire multiple data sources, integration is an additional issue, either here or in data preparation later.

Output **Initial Data Collection Report**

List the data set (or data sets) acquired, together with their locations within the project, the methods used to acquire them and any problems encountered. (Problems encountered, and any solutions achieved, are listed to aid with future replication of this project or with the execution of similar future projects.)

2.2 Describe Data

Task **Describe Data**

Examine the "gross" or "surface" properties of the acquired data and report on the results.

Output **Data Description Report**

Describe the data which has been acquired, including: the format of the data, the quantity of data, for example number of records and fields in each table, the identities of the fields, and any other surface features of the data which have been discovered. Does the data acquired satisfy the relevant requirements?

2.3 Explore Data

Task **Explore Data**

This task tackles the data mining questions which can be addressed using querying, visualisation and reporting. These include: Distribution of key attributes, for example the target attribute of a prediction task; Relations between pairs or small numbers of attributes; Results of simple aggregations; Properties of significant sub-populations; Simple statistical analyses. These analyses may address directly the data mining goals; they may also contribute to or refine the data description and quality reports, and feed into the transformation and other data preparation needed for further analysis.

Output **Data Exploration Report**

Describes results of this task including first findings or initial hypothesis and their impact on the remainder of the project. The report possibly also covers graphs and plots which indicate data characteristics or lead to interesting data subsets for further examination.

2.4 Verify Data Quality

Task **Verify Data Quality**

Examine the quality of the data, addressing questions such as: Is the data complete (does it cover all the cases required)? Is it correct or does it contain errors, and if there are errors how common are they? Are there missing values in the data? If so how are they represented, where do they occur and how common are they?

Output **Data Quality Report**

List the results of the data quality verification; if quality problems exist, list possible solutions. Solutions to data quality problems will generally depend heavily of both data and business knowledge.

3 Data Preparation

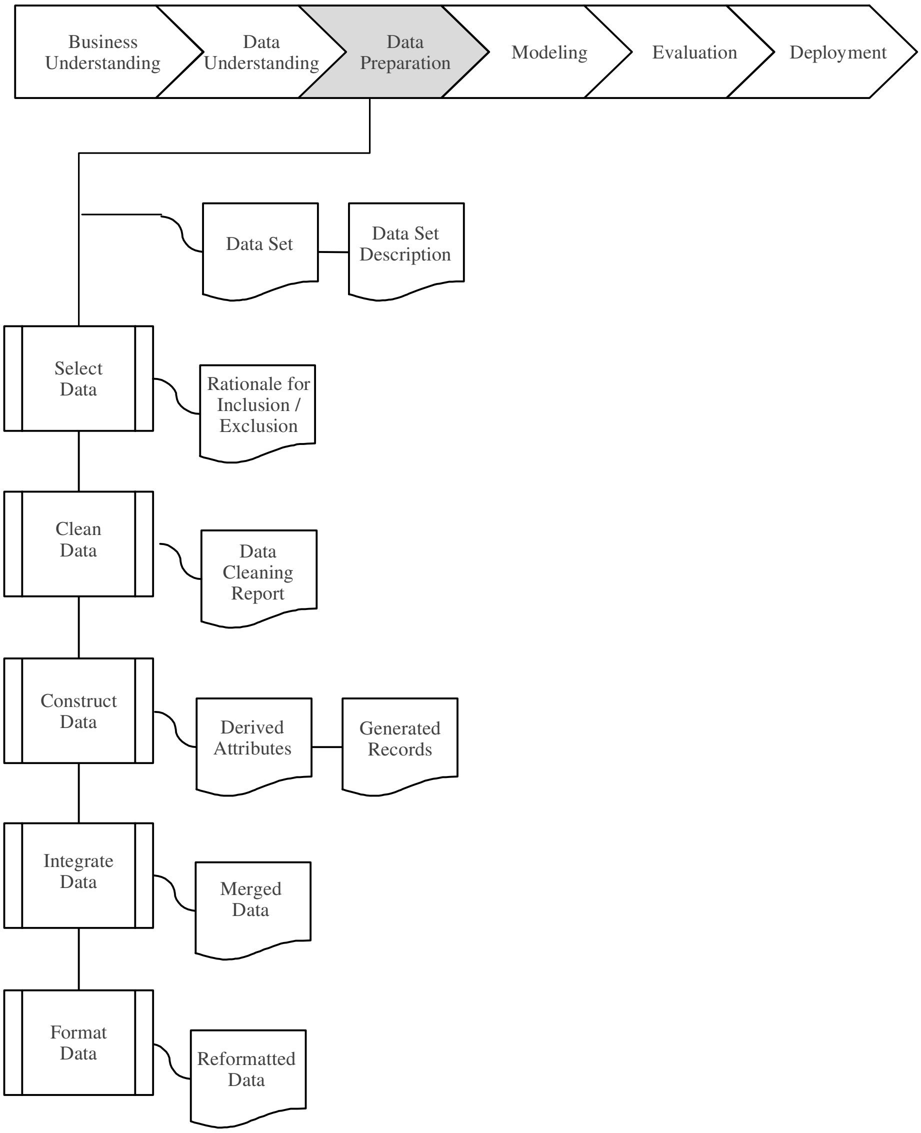


Figure 6: Data Preparation

Outputs

Data Set

This is the data set (or data sets) produced by the data preparation phase, which will be used for modeling or the major analysis work of the project.

Data Set Description

Describe the dataset (or datasets) which will be used for the modeling or the major analysis work of the project.

3.1 Select Data

Task **Select Data**

Decide on the data to be used for analysis. Criteria include relevance to the data mining goals, quality, and technical constraints such as limits on data volume or data types. Note that data selection covers selection of attributes (columns) as well as selection of records (rows) in a table.

Output **Rationale for Inclusion / Exclusion**

List the data to be included / excluded and the reasons for these decisions.

3.2 Clean Data

Task **Clean Data**

Raise the data quality to the level required by the selected analysis techniques. This may involve selection of clean subsets of the data, the insertion of suitable defaults or more ambitious techniques such as the estimation of missing data by modeling.

Output **Data Cleaning Report**

This set of reports describes what decisions and actions were taken to address the data quality problems reported during the *verify data quality* task of the *data understanding* phase. Transformations of the data for cleaning purposes and the possible impact on the analysis results should be considered.

3.3 Construct Data

Task **Construct Data**

This task includes constructive data preparation operations such as the production of derived attributes, entire new records, or transformed values for existing attributes.

Outputs **Derived Attributes**

Derived attributes are new attributes that are constructed from one or more existing attributes in the same record. Examples: *area* = *length* * *width.*

Generated Records

Describe the creation of completely new records. Examples: Create records for customers who made no purchase during the past year. There was no reason to have such records in the raw data, but for modeling purposes it might make sense to explicitly represent the fact that certain customers made zero purchases.

3.4 Integrate Data

Task **Integrate Data**

These are methods whereby information is combined from multiple tables or records to create new records or values.

Output **Merged Data**

Merging tables refers to joining together two or more tables that have different information about the same objects. Example: A retail chain has one table with information about each store's general characteristics (e.g., floor space, type of mall), another table with summarized sales data (e.g., profit, percent change in sales from previous year), and another with information about the demographics of the surrounding area. Each of these tables contains one record for each store. These tables can be merged together into a new table with one record for each store, combining fields from the source tables.

Merged data also covers aggregations. Aggregation refers to operations where new values are computed by summarizing together information from multiple records and/or tables. For example, converting a table of customer purchases where there is one record for each purchase into a new table where there is one record for each customer, with fields such as *number of purchases, average purchase amount, percent of orders charged to credit card, percent of items under promotion, etc.*

3.5 Format Data

Task **Format Data**

Formatting transformations refer to primarily *syntactic* modifications made to the data that do not change its meaning, but might be required by the modeling tool.

Output **Reformatted Data**

Some tools have requirements on the order of the attributes, such as the first field being a unique identifier for each record or the last field being the outcome field the model is to predict.

It might be important to change the order of the records in the dataset. Perhaps the modeling tool requires that the records be sorted according to the value of the outcome attribute. Most common situation is that the records of the dataset are initially ordered in some way but the modeling algorithm needs them to be in a fairly random order. For example, when using neural networks it is generally best for the records to be presented in a random order although some tools will hand this automatically without explicit user intervention.

Additionally, there are purely syntactic changes made to satisfy the requirements of the specific modeling tool. Examples: removing commas from within text fields in comma-delimited data files, trimming all values to a maximum of 32 characters.

4 Modeling

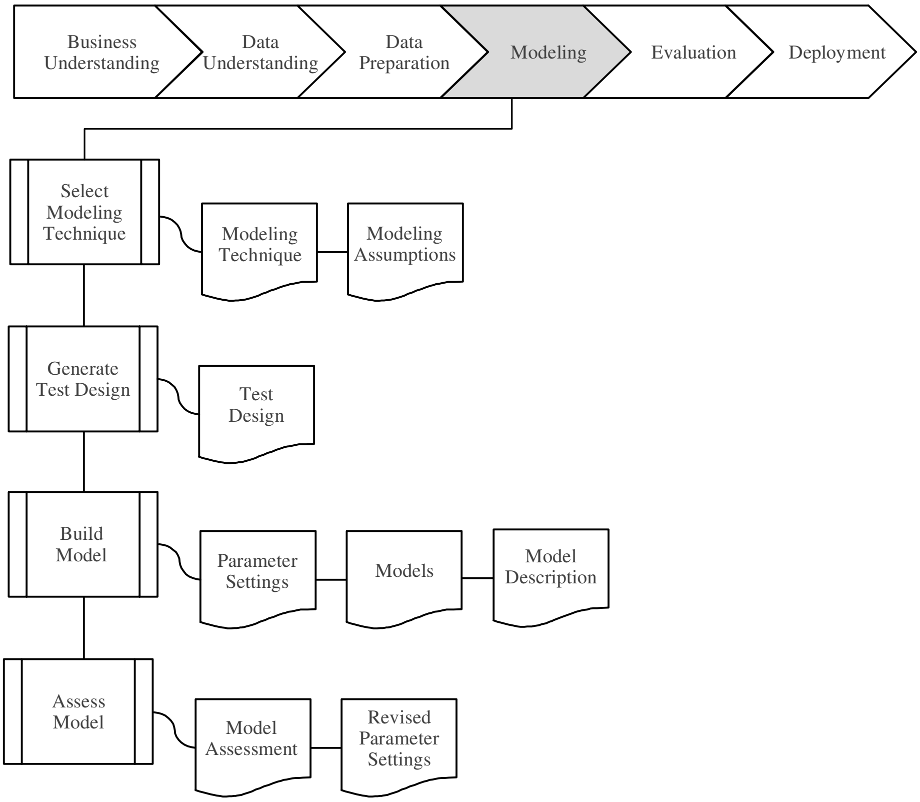


Figure 7: Modeling

4.1 Select Modeling Technique

Task **Select Modeling Technique**

As the first step in modeling, select the actual modeling technique which is used now. Whereas you possibly already selected a tool in business understanding, this task refers to the specific modeling technique, e.g., decision tree building with C4.5 or neural network generation with back propagation. If multiple techniques are applied, perform this task for each technique separately.

Outputs **Modeling Technique**

This output refers to the actual modeling technique which is used.

Modeling Assumptions

Many modeling techniques make specific assumptions on the data, e.g., all attributes have uniform distributions, no missing values allowed, class attribute must be symbolic etc.

4.2 Generate Test Design

Task **Generate Test Design**

Before we actually build a model, we need to generate a procedure or mechanism how 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 separated the data set into train and test set, build the model on the train set, and estimate its quality on the separate test set.

Output **Test Design**

This deliverable describes the intended plan for training, testing, and evaluating the models. A primary component of the plan is to decide how to divide the available data set into training data, test data, and validation data sets.

4.3 Build Model

Task **Build Model**

Run the modeling tool on the prepared data set 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. This report lists the parameters and their chosen value, along with rationale for the choice of parameter settings.

Models

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

Model Description

Describe the resultant model. Report on the interpretation of the models and any difficulties encountered with their meanings.

4.4 Assess Model

Task **Assess Model**

The data mining engineer interprets the models according to his domain knowledge, data mining success criteria, and the desired test design. This task interferes with the subsequent evaluation phase. Whereas the data mining engineer judges the success of the application of modeling and discovery techniques more technically, he contacts business analysts and domain experts later in order to discuss the data mining results in the business context. Moreover, this task only considers models whereas the evaluation phase also takes into account all other results which were produced in the course of the project.

The data mining engineer tries to rank the results. He assesses the models according to the evaluation criteria. As far 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 different alternative techniques. In this task, he also compares all results according to the evaluation criteria.

Outputs **Model Assessment**

Summarizes results of this task, lists qualities of generated models (e.g., in terms of accuracy), and ranks 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 task build model. Iterate model building and assessment until you strongly believe that you found the *best* model(s).

5 Evaluation

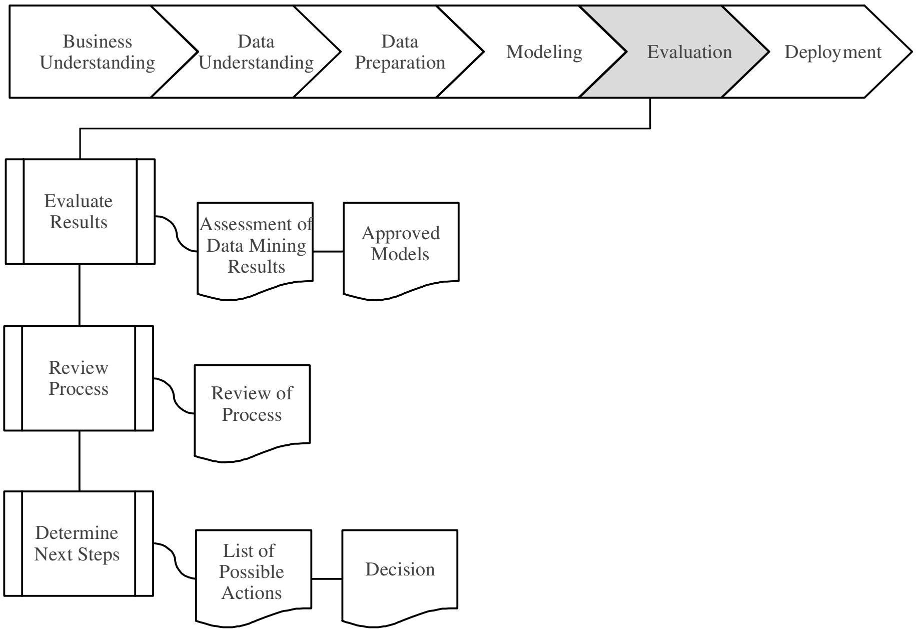


Figure 8: Evaluation

5.1 Evaluate Results

Task **Evaluate Results**

Previous evaluation steps dealt with factors such as the accuracy and generality of the model. This step assesses the degree to which the model meets the business objectives and seeks to determine if there is some business reason why this model is deficient. Another option of evaluation is to test the model(s) on test applications in the real application if time and budget constraints permit.

Moreover, evaluation also assesses other data mining results generated. Data mining results cover models which are necessarily related to the original business objectives, and all other findings which are not necessarily related to the original business objectives but might also unveil additional challenges, information, or hints for future directions.

Outputs **Assessment of Data Mining Results w.r.t. Business Success Criteria**

Summarizes assessment results in terms of business success criteria including a final statement whether the project already meets the initial business objectives.

Approved Models

After model assessment w.r.t. business success criteria, you eventually get approved models if the generated models meet the selected criteria.

5.2 Review Process

Task **Review Process**

At this point the resultant model hopefully appears to be satisfactory and to satisfy business needs. It is now appropriate to do a more thorough review of the data mining engagement in order to determine if there is any important factor or task that has somehow been overlooked. This review also covers quality assurance issues, e.g., did we correctly build the model? Did we only use attributes that we are allowed to use and that are available for future analyses?

Output **Review of Process**

Summarizes the process review and gives hints for activities that have been missed and/or should be repeated.

5.3 Determine Next Steps

Task **Determine Next Steps**

According to the assessment results and the process review, the project decides how to proceed at this stage. The project needs to decide whether to finish this project and move on to deployment if appropriate, or whether to initiate further iterations or set up new data mining projects. This task include analyses of remaining resources and budget which influences the decisions.

Outputs **List of Possible Actions**

A list of possible further actions along with the reasons for and against each option.

Decision

Describes the decision as to how to proceed along with the rationale.

6 Deployment

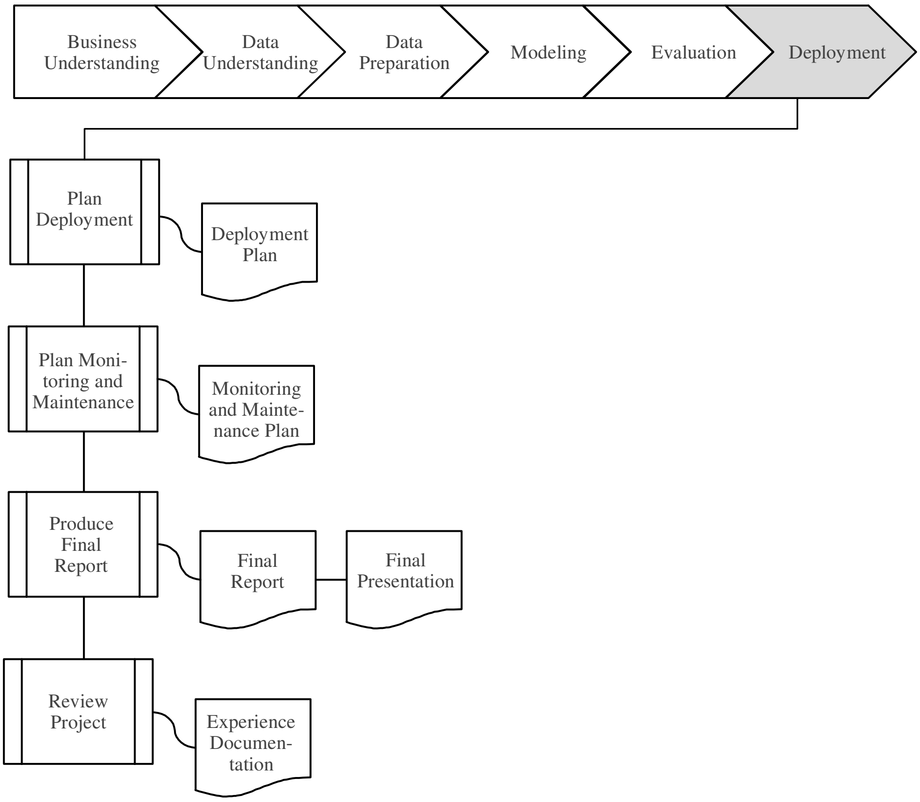


Figure 9: Deployment

6.1 Plan Deployment

Task **Plan Deployment**

In order to deploy the data mining result(s) into the business, this task takes the evaluation results and concludes a strategy for deployment. If there has been identified a general procedure to create the relevant model(s), this procedure is documented here for deployment later.

Output **Deployment Plan**

Summarizes deployment strategy including necessary steps and how to perform them

6.2 Plan Monitoring and Maintenance

Task **Plan Monitoring and Maintenance**

Monitoring and maintenance are important issues if the data mining result becomes part of the day-to-day business and its environment. A careful preparation of a maintenance strategy helps to avoid unnecessarily long periods of incorrect usage of data mining results. In order to monitor the deployment of the data mining result(s), the project needs a detailed plan on the monitoring process. This plan takes into account the specific type of deployment.

Output **Monitoring and Maintenance Plan**

Summarizes monitoring and maintenance strategy including necessary steps and how to perform them.

6.3 Produce Final Report

Task **Produce Final Report**

At the end of the project, the project leader and his team write up a final report. It depends on the deployment plan, if this report is only a summary of the project and its experiences (if they have not already been documented as an ongoing activity), or if this report is a final presentation of the data mining result(s).

Outputs **Final Report**

The final written report of the data mining engagement. Includes all of the previous deliverables, plus summarizing and organizing the results.

Final Presentation

There will also often be a meeting at the conclusion of the project where the results are verbally presented to the customer.

6.4 Review Project

Task **Review Project**

Assess what went right and what went wrong, what was done well and what needs to be improved.

Output **Experience Documentation**

Summarizes important experiences made during the project. For example, pitfalls, misleading approaches or hints for selecting the best suited data mining techniques in similar situations could be part of this documentation. In ideal projects, experience documentation covers also any reports that have been written by individual project members during the project phases and their tasks.

III The CRISP-DM User Guide

1 Business Understanding

1.1 Determine Business Objectives

Task **Determine Business Objectives**

The first objective of the analyst is to thoroughly understand, from a *business* perspective, what the client really wants to accomplish. Often the customer will have many competing objectives and constraints that must be properly balanced. The analyst's goal is to uncover important factors at the beginning of the project that can influence the final outcome. A likely consequence of neglecting this step would be to expend a great deal of effort producing the correct answers to the wrong questions.

Output **Background**

This gives details about the information that is known about the organization's business situation at the start of the project. These details will not only serve to more closely identify the business goals that will be solved, but will also serve to identify resources, both human and material, that may be used or needed during the course of the project.

Activities

Organization

- Develop organization chart identifying divisions, departments, and project groups. The chart should also identify managers names and responsibilities.
- Identify key persons in the business and their roles
- Identify an internal sponsor (financial sponsor and primary user/domain expert)
- Is there a steering committee and who are the members?

- Identify the business units which will be impacted by the data mining project (e.g., Marketing, Sales, Finance)

Problem Area

- Identify the problem area (e.g., Marketing, Customer Care, Business Development, …)
- Describe the problem in general terms
- Check the current status of the project (e.g., Check if it is already clear within the business unit that we are performing a data mining project, or do we need to advertise data mining as a key technology in the business?)
- Clarify prerequisites of the project (e.g., What is the motivation of the project? Does the business already use data mining?)
- If necessary, prepare presentations and present data mining to the business
- Identify target groups for the project result (e.g., Do we expect a written report for top management, or do we expect a running system that is used by naïve end users?)
- Identify the users' needs and expectations

Current Solution

- Describe the solution currently in use for the problem
- Describe the advantages and disadvantages of the current solution, and the level to which it is accepted by the users

Output **Business Objectives**

Describe the customer's primary objective, from a business perspective, in the data mining project. In addition to the primary business objective, there are typically a large number of related business questions that the customer would like to address. For example, the primary business goal might be to keep current customers by predicting when they are prone to move to a competitor, whilst secondary business objectives might be to determine whether lower fees will affect only one particular segment of customers.

Activities

- Informally describe the problem which is supposed to be solved with data mining
- Specify all business questions as precisely as possible
- Specify any other business requirements (e.g., The business does not want to lose any customers)
- Specify expected benefits in business terms

Beware! Beware of setting unattainable goals – make them as realistic as possible

Output **Business Success Criteria**

Describe the criteria for a successful or useful outcome to the project from the business point of view. This might be quite specific, such as reduction of customer churn to a certain level, or general and subjective such as "give useful insights into the relationships". In the latter case it should be indicated who would make the subjective judgment.

Activities

- Specify business success criteria (e.g., improve response rate in a mailing campaign by 10%, and sign-up rate increased by 20%)
- Identify who will assess the success criteria

Remember! Each of the success criteria should relate to *at least one* of the specified Business Objectives

Good Idea! Before starting with Situation Assessment, you might consider previous experiences of this problem – either internally using CRISP-DM or externally using pre-packaged solutions.

1.2 Assess Situation

Task **Assess Situation**

This task involves more detailed fact-finding about all of the resources, constraints, assumptions, and other factors that should be considered in determining the data analysis goal and project plan.

Output **Inventory of Resources**

List the resources available to the project, including: Personnel (business and data experts, technical support, data mining personnel), Data (fixed extracts, access to live warehoused or operational data), Computing resources (hardware platforms), Software (data mining tools, other relevant software).

Activities

<u>Hardware Resources</u>

- Identify the base hardware
- Establish the availability of the base hardware for the data mining project
- Check if the hardware maintenance schedule conflicts with the availability of the hardware for the data mining project
- Identify the hardware available for the data mining tool to be used (if the tool is known at this stage)

<u>Sources of Data & Knowledge</u>

- Identify data sources
- Identify type of data sources (on-line sources, experts, written documentation, ...)
- Identify knowledge sources
- Identify type of knowledge sources (on-line sources, experts, written documentation, ...)
- Check available tools and techniques
- Describe the relevant background knowledge (informally or formally)

<u>Personnel Sources</u>

- Identify project sponsor (if different from internal sponsor as in Section 1.1.1)
- Identify system administrator, database administrator and technical support staff for further questions
- Identify market analysts, data mining experts and statisticians and check their availability
- Check availability of domain experts for later phases

Remember! Remember that the project may need technical staff at odd times throughout the project. E.g. during Data Transformation

Output **Requirements, Assumptions and Constraints**

List all requirements of the project including schedule of completion, comprehensibility and quality of results, and security as well as legal issues. As part of this output, make sure that you are allowed to use the data.

List the assumptions made by the project. These may be assumptions about the data, which can be checked during data mining, but may also include non-checkable assumptions about the business upon which the project rests. It is particularly important to list the latter if they form conditions on the validity of the results.

List the constraints made on the project. These constraints might involve lack of resources to carry out some of the tasks in the project within the timescale required or there may be legal or ethical constraints on the use of the data or the solution needed to carry out the data mining task.

Activities

<u>Requirements</u>

- Specify target group profile
- Capture all requirements on scheduling
- Capture requirements on comprehensibility, accuracy, deployability, maintainability and repeatability of the Data Mining project and the resulting model(s)
- Capture requirements on security, legal restrictions, privacy, reporting and project schedule

<u>Assumptions</u>

- Clarify all assumptions (including implicit ones) and make them explicit (e.g., To address the business question, a minimum number of customers with age above 50 is necessary)
- List assumptions on data quality (e.g., accuracy, availability)
- List assumptions on external factors (e.g., economic issues, competitive products, technical advances)
- Clarify assumptions that lead to any of the estimates (e.g., the price of a specific tool is assumed to be lower than $1000)
- List all assumptions on whether it is necessary to *understand* and describe or explain the model. (e.g., How should the model and results be presented to senior management/sponsor.)

Constraints

- Check general constraints (e.g., legal issues, budget, timescales and resources)
- Check access rights to data sources (e.g., access restrictions, password required)
- Check technical accessibility of data (operating systems, data management system, file or database format)
- Check whether relevant knowledge is accessible
- Check budget constraints (Fixed costs, implementation costs, etc.)

Remember! The list of assumptions also includes assumptions at the beginning of the project, i.e., what has been the starting point of the project

Output **Risks and Contingencies**

List the risks, that is, events that might occur, impacting schedule, cost or result. List the corresponding contingency plans; what action will be taken to avoid or minimize the impact or recover from the occurrence of the foreseen risks.

Activities

Identify Risks

- Identify business risks (e.g., competitor comes up with better results first)
- Identify organisational risks (e.g., department requesting project not having funding for project)
- Identify financial risks (e.g., further funding depends on initial data mining results)
- Identify technical risks
- Identify risks that depend on data and data sources (e.g. poor quality and coverage)

Develop Contingency Plans

- Determine conditions when each risk may occur
- Develop contingency plans

Output **Terminology**

Make a glossary of terminology relevant to the project. This should include at least two components:
(1) A glossary of relevant business terminology, which forms part of the business understanding available to the project.

(2) A glossary of data mining terminology, illustrated with examples relevant to the business problem in question.

Activities

- Check prior availability of glossaries, otherwise begin to draft glossaries
- Talk to domain experts to understand their terminology
- Become familiar with the business terminology

Output **Costs and Benefits**

Prepare a cost-benefit analysis for the project, comparing the costs of the project with the potential benefit to the business if it is successful.

Good Idea! The comparison should be as specific as possible, as this enables a better business case to be made.

Activities

- Estimate costs for data collection
- Estimate costs of developing and implementing a solution
- Identify benefits when a solution is deployed (e.g. improved customer satisfaction, ROI and increase in revenue)
- Estimate operating costs

Beware! Remember to identify hidden costs such as repeated data extraction and preparation, changes in work flows, and training time during learning

1.3 Determine Data Mining Goals

Task **Determine Data Mining Goals**

A *business goal* states objectives in business terminology, a *data mining goal* states project objectives in technical terms. For example, the business goal might be "Increase catalogue sales to existing customers" while a data mining goal might be "Predict how many widgets a customer will buy, given their purchases over the past three years, relevant demographic information, and the price of the item".

Output **Data Mining Goals**

Describe the intended outputs of the project that will enable the achievement of the business objectives. Note that these will normally be *technical* outputs.

Activities

- Translate the business questions to data mining goals (e.g., a marketing campaign requires segmentation of customers in order to decide whom to approach in this campaign; it should be specified the level/size of the segments)
- Specify data mining problem type (e.g., classification, description, prediction and clustering) For more details about Data Mining Problem Types, see Appendix V.2, where they are described in more detail.

Good Idea! It may be wise to re-define the problem. E.g. Modeling product retention rather than customer retention since targeting customer retention may be too late to affect the outcome!

Output **Data Mining Success Criteria**

Define the criteria for a successful outcome to the project in technical terms, for example a certain level of predictive accuracy, or a propensity to purchase profile with a given degree of "lift". As with business success criteria, it may be necessary to describe these in subjective terms, in which case the person or persons making the subjective judgment should be identified.

Activities

- Specify criteria for model assessment (e.g., model accuracy, performance and complexity)
- Define benchmarks for evaluation criteria

- Specify criteria which address subjective assessment criteria (e.g. model explainability and data and marketing insight provided by the model)

Beware! Remember that the Data Mining Success Criteria will be different to the Business Success Criteria defined earlier.

Remember that it is wise to plan for deployment already at the start of the project.

1.4 Produce Project Plan

Task **Produce Project Plan**

Describe the intended plan for achieving the data mining goals and thereby achieving the business goals.

Output **Project Plan**

List the stages to be executed in the project, together with duration, resources required, inputs, outputs and dependencies. Wherever possible make explicit the large-scale iterations in the data mining process, for example, repetitions of the modeling and evaluation phases. As part of the project plan, it is also important to analyse dependencies between time schedule and risks. Mark results of these analyses explicitly in the project plan, ideally with actions and recommendations for actions if the risks appear.

Remember that, although this is the only task in which the Project Plan is directly named, nevertheless it should be continually consulted and reviewed throughout the project. This should be a document which should at least be consulted whenever a new Task is started or a further iteration of a task or activity is begun.

Activities

- Define the initial process plan & discuss the feasibility with all involved personnel
- Put all identified goals and selected techniques together into a coherent procedure that solves the business questions and meets the business success criteria
- Estimate effort and resources needed to achieve and deploy the solution (It is useful to consider other peoples experience when estimating timescales for data mining projects. For example, it is often postulated that 50%-70% of the time and effort in a data mining project is used in the Data Preparation Phase, and 20%-30% in the Data Understanding Phase, whilst only 10%-20% is spent in each of the Modeling, Evaluation and Business Understanding Phases and 5%-10% in the Deployment Phase.)
- Identify critical steps
- Mark decision points
- Mark review points
- Identify major iterations

Output **Initial Assessment of Tools and Techniques**

At the end of the first phase, the project also performs an initial assessment of tools and techniques. Here, you select a data mining tool which supports various methods for different stages of the process, for example. It is important to assess tools and techniques early in the process since the selection of tools and techniques possibly influences the entire project.

Activities

- Create a list of selection criteria for tools and techniques (or use an existing one if available)
- Choose potential tools and techniques
- Evaluate appropriateness of techniques
- Review and prioritise applicable techniques according to the evaluation of alternative solutions

2 Data Understanding

2.1 Collect Initial Data

Task **Collect Initial Data**

Acquire within the project the data (or access to the data) listed in the project resources. This initial collection includes data loading if necessary for data understanding. For example, if you intend to use a specific tool for data understanding, it is logical to load your data into this tool.

Output **Initial Data Collection Report**

List all the various data that will be used within the project, together with any selection requirements for more detailed data. The Data Collection Report should also define whether some attributes are relatively more important than others.

Remember that any assessment of Data Quality should be made not just of the individual data sources but also of any data that comes from merging data sources. Merged data may present problems that do not exist in the individual data sources because of inconsistencies between the sources.

Activities

Data Requirements Planning

- Plan which information is needed (e.g. only given attributes, additional information)
- Check if all the information needed (to solve the data mining goals) is actually available

Selection Criteria

- Specify selection criteria (e.g., Which attributes are necessary for the specified data mining goals? Which attributes have been identified as being irrelevant? How many attributes can we handle with the chosen techniques?)
- Select tables / files of interest
- Select data within a table / file
- Think about how long history one should use even if available (e.g. even if 18 months data is available, maybe only 12 months is needed for the exercise)

Beware! Be aware that data collected from different sources may give rise to quality problems when merged (e.g. address files merged with own customer base may show up inconsistencies of format, invalidity of data, etc.)

Insertion of Data

- If the data contains free text entries, do we need to encode them for modelling, or do we want to group specific entries?
- How can missing attributes be acquired?
- Describe how to extract the data

Good Idea! Remember that some knowledge about the data may be on non-electronic sources (e.g., People, Printed text, etc.)

Remember that it may be necessary to pre-process the data (time series data, weighted averages, etc.)

2.2 Describe Data

Task **Describe Data**

Examine the "gross" properties of the acquired data and report on the results.

Output **Data Description Report**

Describe the data which has been acquired including: the format of the data, the quantity of the data (e.g. the number of records and fields within each table), the identities of the fields and any other surface features of the data which have been discovered.

Activities

Volumetric Analysis of Data

- Identify data and method of capture
- Access data sources
- Utilise statistical analyses if appropriate
- Report tables and their relations
- Check data volume, number of tuples, complextity
- Does the data contain free text entries?

Attribute Types & Values

- Check accessibility and availability of attributes
- Check attribute types (numeric, symbolic, taxonomy etc.)
- Check attribute value ranges
- Analyse attribute correlations
- Understand the meaning of each attribute and attribute value in business terms
- For each attribute, compute basic statistics (e.g., compute distributions, averages, max, min, standard deviations, variances, modi, skewness etc.)
- Analyse basic statistics and relate the results to their meaning in business terms
- Is the attribute relevant for the specific data mining goal?
- Is the attribute meaning used consistently?
- Interview domain expert on his opinion of attribute relevance
- Is it necessary to balance the data? (Depending on the modelling technique used)

Keys

- Analyse key relations
- Check amount of overlaps of key attribute values across tables

Review Assumptions/Goals

- Update list of assumptions if necessary

2.3 Explore Data

Task **Explore Data**

This task tackles the data mining questions, which can be addressed using querying, visualisation and reporting. These analyses may address directly the data mining goals. However, they may also contribute to or refine the data description and quality reports, and feed into the transformation and other data preparation needed for further analysis.

Output **Data Exploration Report**

Describes results of this task including first findings or initial hypotheses and their impact on the remainder of the project. The report possibly also covers graphs and plots which indicate data characteristics or lead to interesting data subsets for further examination.

Activities

Data Exploration

- Analyse properties of interesting attributes in detail (e.g. basic statistics interesting sub-populations)
- Identify characteristics of sub-populations

Form Suppositions for Future Analysis

- Consider and evaluate information and findings in the Data Descriptions Report
- Form hypothesis and identify actions
- Transform hypothesis into a data mining goal if possible
- Clarify data mining goals or make them more precise. Blind search is not necessarily useless but a more directed search towards business objectives is preferable
- Perform basic analysis to verify the hypothesis

2.4 Verify Data Quality

Task **Verify Data Quality**

Examine the quality of the data, addressing questions such as: Is the data complete (does it cover all the cases required)? Is it correct or does it contain errors, and if there are errors how common are they? Are there missing values in the data? If so, how are they represented, where do they occur and how common are they?

Output **Data Quality Report**

List the results of the data quality verification; if there are quality problems, list possible solutions.

Activities

- Identify special values and catalogue their meaning

<u>Review Keys, Attributes</u>

- Check coverage (e.g. are all possibly values represented)
- Check keys
- Do the meanings of attributes and contained values fit together?
- Identify missing attributes and blank fields
- Check for attributes with different values that have similar meanings (e.g., low fat, diet)
- Check spelling of values (e.g., same value but sometime beginning with a lower case letter, sometimes with an upper case letter)
- Check for deviations, decide whether a deviation is noise or may indicate an interesting phenomenon
- Check for plausibility of values, e.g. all fields have the same or nearly the same values.

Good Idea! Make a review of any attributes that may give answers which conflict with common sense (e.g. teenagers with high income)

Use visualisation plots, histograms, etc. to show up inconsistencies in the data

<u>Data Quality in Flat Files</u>

- If the data is stored in flat files, check which delimiter is used and if it is used consistently within all attributes
- If the data is stored in flat files, check number of fields in each record. Do they coincide?

Noise and inconsistencies between sources

- Check consistencies and redundancies between different sources
- Plan how to deal with noise
- Detect type of noise and which attributes are affected

Good Idea!

Remember that it may be necessary to exclude some data since they do not exhibit either positive or negative behaviour (e.g. to check on customers loan behaviour, exclude all those who have never loaned, do not finance a home mortgage, those whose mortgage is nearing maturity, etc.

Review all assumptions whether they are valid or not given the current information on data and knowledge

3 Data Preparation

Output **Data Set**

This is the data set (or data sets) produced by the data preparation phase, which will be used for modeling or the major analysis work of the project.

Output **Data Set Description**

This is the description of the data set(s) which will be used for the modeling or the major analysis work of the project.

3.1 Select Data

Task **Select Data**

Decide on the data to be used for analysis. Criteria include relevance to the data mining goals, quality, and technical constraints such as limits on data volume or data types.

Output **Rationale for Inclusion / Exclusion**

List the data to be used/excluded and the reasons for these decisions.

Activities

- Collect appropriate additional data (from different sources -- in-house as well as externally);
- Perform significance and correlation test to decide if fields should be included
- Reconsider Data Selection Criteria (See Task 2.1) in light of experiences of data quality, data exploration, (i.e. may wish include/exclude other sets of data)
- Reconsider Data Selection Criteria (See Task 2.1) in light of experience of modelling (i.e. model assessment may show that other data sets are needed)
- Select different data subsets (e.g., Different attributes, only data which meet certain conditions)
- Consider use of sampling techniques (e.g. a quick solution may involve the reduction of the size of the test data set or the tool may not be able to handle the full data set, splitting test and training data sets). It may also be useful to have weighted samples to give different importance to different attributes or different values of the same attribute.
- Document the rationale for inclusion/exclusion
- Check available techniques for sampling data

Good Idea! Based on Data Selection Criteria, decide if one or more attributes are more important than others and weight the attributes accordingly. Decide, based on the context (i.e. application, tool, etc.) how to handle the weighting.

3.2 Clean Data

Task **Clean Data**

Raise the data quality to the level required by the selected analysis techniques. This may involve selection of clean subsets of the data, the insertion of suitable defaults or more ambitious techniques such as the estimation of missing data by modeling.

Output **Data Cleaning Report**

This report describes the decisions and actions that were taken to address the data quality problems reported during the Verify Data Quality Task.

Activities

- Reconsider how to deal with observed type of noise
- Correct, remove or ignore noise
- Decide how to deal with special values and their meaning. The area of special values can give rise to many strange results and should be carefully examined. Examples of special values could arise through taking results of a survey where some questions were not asked or nor answered. This might result in a value of ‘99’ for unknown data e.g. 99 for marital status or political affiliation. Special values could also arise when data is truncated – e.g. ‘00’ for 100 year old people, or all cars with 100,000 km on the clock.
- Reconsider Data Selection Criteria (See Task 2.1) in light of experiences of data cleaning (i.e. one may wish include/exclude other sets of data)

Good Idea! Remember that some fields may be irrelevant to the data mining goals and therefore noise in those fields has no significance. However, if noise is ignored for these reasons, it should be fully documented as the circumstances may change later!

3.3 Construct Data

Task **Construct Data**

This task includes constructive data preparation operations such as the production of derived attributes, complete new records, or transformed values for existing attributes.

Activities

- Check available construction mechanisms with the list of tools suggested for the project
- Decide whether it is best to perform the construction inside the tool or outside (i.e. which is more efficient, exact, repeatable)
- Reconsider Data Selection Criteria (See Task 2.1) in light of experiences of data construction (i.e. may wish include/exclude other sets of data)

Output **Derived Attributes**

Derived Attributes are new attributes that are constructed from one or more existing attributes in the same record. An example might be area = length * width.

Why should we need to construct derived attributes during the course of a data mining investigation? It should not be thought that only data from databases or other sources is the only type of data that should be used in constructing a model. Derived attributes might be constructed because:

Background knowledge convinces us that some fact is important and ought to be represented although we have no attribute currently to represent it.

The modelling algorithm in use handles only certain types of data, e.g. we are using linear regression and we suspect that there are certain non-linearities that will be not be included in the model.

The outcome of the modelling phase may suggest that certain facts are not being covered.

Activities

Derived Attributes

- Decide if any attribute should be normalized (e.g. when using a clustering algorithm with age and income in lire, the income will dominate)

- Consider adding new information on the relevant importance of attributes by adding new attributes (e.g. attribute weights, weighted normalization)
- How can missing attributes be constructed? (decide type of construction (e.g., aggregate, average, induction))
- Add new attributes to the accessed data

Good Idea! Before adding Derived Attributes, try to determine if and how they will ease the model process or facilitate the modelling algorithm, Perhaps “income per head” is a better/easier attribute to use that “income per household”. Do not derive attributes simply to reduce the number of input attributes.

Another type of derived attribute is single-attribute transformations, usually performed to fit the needs of the modelling tools.

Activities

Single-Attribute Transformations

- Specify necessary transformation steps in terms of available transformation facilities (e.g. change a discretisation of a numeric attribute)
- Perform transformation steps

Hint! Transformations may be necessary to transform ranges to symbolic fields (e.g. ages to age ranges) or symbolic fields (“definitely yes”, “yes”, “don’t know”, “no”) to numeric values. They are often required by the modelling tools or algorithm.

Output **Generated Records**

Generated Records are completely new records which add new knowledge or represent new data which is not otherwise represented, e.g., having segmented the data, it may be useful to generate a record to represent the prototypical member of each segment for further processing.

Activities

- Check for available techniques if needed (e.g., mechanisms to construct prototypes for each segment of segmented data)

3.4 Integrate Data

Task **Integrate Data**

These are methods whereby information is combined from *multiple* tables or other information sources to create new records or values.

Output **Merged Data**

Merging tables refers to joining together two or more tables that have different information about the same objects. At this stage it may also be advisable to generate new records. It may also be recommended to generate aggregate values.

Aggregation refers to operations where new values are computed by summarizing together information from multiple records and/or tables.

Activities

- Check integration facilities if they are able to integrate the input sources as required
- Integrate sources and store result
- Reconsider Data Selection Criteria (See Task 2.1) in light of experiences of data integration (i.e. may wish include/exclude other sets of data)

Good Idea! Remember that some knowledge may be contained in non-electronic format.

3.5 Format Data

Task **Format Data**

Formatting transformations refer to primarily *syntactic* modifications made to the data that do not change its meaning, but might be required by the modeling tool.

Output **Reformatted Data**

Some tools have requirements on the order of the attributes, such as the first field being a unique identifier for each record or the last field being the outcome field the model is to predict.

Activities

Rearranging Attributes

- Some tools have requirements on the order of the attributes, such as the first field being a unique identifier for each record or the last field being the outcome field the model is to predict.

Reordering Records

- It might be important to change the order of the records in the data set. Perhaps the modeling tool requires that the records be sorted according to the value of the outcome attribute.

Reformatted Within-Value

- These are purely syntactic changes made to satisfy the requirements of the specific modeling tool.
- Reconsider Data Selection Criteria (See Task 2.1) in light of experiences of data cleaning (i.e. may wish include/exclude other sets of data)

4 Modeling

4.1 Select Modeling Technique

Task **Select Modeling Technique**

As the first step in modeling, select the actual modeling technique that is to be used initially. If multiple techniques are applied, perform this task for each technique separately.

It should not be forgotten that not all tools and techniques are available for each and every task. For certain problems, only some techniques are appropriate (See Appendix V.2 where techniques appropriate for certain data mining problem types are discussed in more detail). From amongst these tools and techniques there are "Political Requirements" and other constraints which further limit the choice available to the miner. It may be that only one tool or technique is available to solve the problem in hand – and even then the tool may not be the absolutely technical best for the problem in hand.

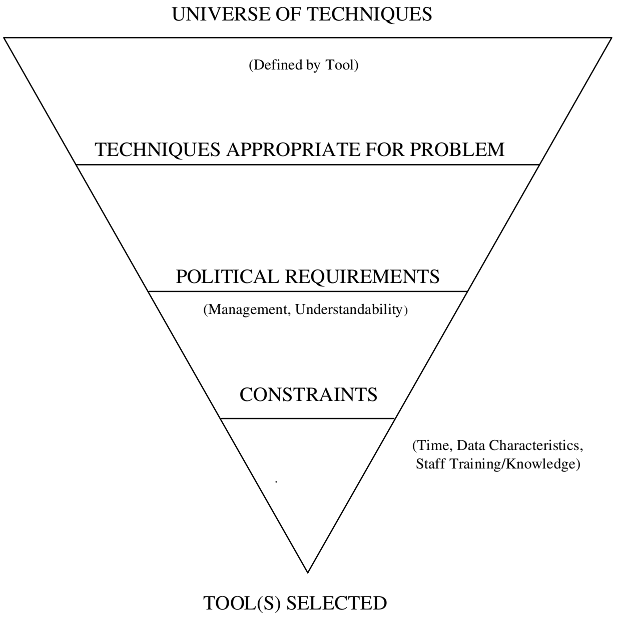

UNIVERSE OF TECHNIQUES
(Defined by Tool)
TECHNIQUES APPROPRIATE FOR PROBLEM
POLITICAL REQUIREMENTS
(Management, Understandability)
CONSTRAINTS
(Time, Data Characteristics,
Staff Training/Knowledge)
TOOL(S) SELECTED

Figure 10: Universe of Techniques

Output **Modeling Technique**

This output refers to the actual modeling technique that is used.

Activities

- Decide on appropriate technique for exercise bearing in mind the tool selected

Output **Modeling Assumptions**

Many modeling techniques make specific assumptions about the data, data quality or the data format.

Activities

- Define any built-in assumptions made by the technique about the data (e.g. quality, format, distribution)
- Compare these assumptions with those in the Data Description Report
- Make sure that these assumptions hold and step back to Data Preparation Phase if necessary

4.2 Generate Test Design

Task **Generate Test Design**

Prior to building a model, a procedure needs to be defined 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 the test design will specify that the data set should be separated into training and test set, the model will be built on the training set, and its quality estimated on the test set.

Output **Test Design**

This deliverable describes the intended plan for training, testing and evaluating the models. A primary component of the plan is to decide how to divide the available data set into training data, test data and validation test sets.

Activities

- Check existing test designs for each data mining goal separately
- Decide on necessary steps (number of iterations, number of folds etc.)
- Prepare data required for test

4.3 Build Model

Task **Build Model**

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

Output **Parameter Settings**

With any modeling tool, there are often a large number of parameters that can be adjusted. This report lists the parameters and their chosen values, along with the rationale for the choice.

Activities

- Set initial parameters
- Document reasons for choosing those values

Output **Models**

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

Activities

- Run the selected technique on the input data set to produce the model
- Post-process data mining results (e.g. Editing rules, Display trees)

Output **Model Description**

Describe the resultant model and assess its expected accuracy, robustness, and possible shortcomings. Report on the interpretation of the models and any difficulties encountered.

Activities

- Describe any characteristics of the current model that may be useful for the future

4.4 Assess Model

Task **Assess Model**

The model should now be assessed to ensure that it meets the data mining success criteria and the passes the desired test criteria. This is a purely technical assessment based on the outcome of the modeling tasks.

Output **Model Assessment**

Summarizes results of this task, lists qualities of generated models (e.g. in terms of accuracy), and ranks their quality in relation to each other.

Activities

- Evaluate result w.r.t. evaluation criteria

Good Idea! “Lift Tables” and “Gain Tables” can be constructed to determine how well the model is predicting.

- Test result according to a test strategy (e.g.: Train & Test, Cross-validation, bootstrapping etc.)
- Compare evaluation results and interpretation
- Create ranking of results w.r.t. success and evaluation criteria
- Select best models
- Interpret results in business terms (as far as possible at this stage)
- Check plausibility of model
- Check impacts for data mining goal
- Check model against given knowledge base to see if the discovered information is novel and useful
- Check reliability of result
- Analyse potentials for deployment of each result
- If there is a verbal description of the generated model (e.g. via rules), assess the rules; are they logical, are they feasible, are there too many or too few, do they offend common sense?
- Assess results

Output **Revised Parameter Settings**

According to the model assessment, revise parameter settings and tune them for the next run in task ‘Build Model’. Iterate model building and assessment until you find the best model.

Activities

- Reset parameters to give better model.

5 Evaluation

Previous evaluation steps dealt with factors such as the accuracy and generality of the model. This step assesses the degree to which the model meets the business objectives and seeks to determine if there is some business reason why this model is deficient. The results need to compared with the evaluation criteria defined at the start of the project.

A good way of defining the total outputs of a Data Mining project is to use the equation:

RESULTS = MODELS + FINDINGS

In this equation we are defining that the total output of the Data Mining project is not just the models (although they are, of course, important) but also findings which we define as anything *(apart from the model)* that is important in meeting objectives of the business (or important in leading to new questions, line of approach or side effects (e.g. data quality problems uncovered by the data mining exercise) Note that although the model is directly connected to the business questions, but the findings need not be related to any questions or objective, only that they are important to the initiator of the project.

5.1 Evaluate Results

Task **Evaluate Results**

Previous evaluation steps dealt with factors such as the accuracy and generality of the model. This step assesses the degree to which the model meets the business objectives and seeks to determine if there is some business reason why this model is deficient. Another option of evaluation is to test the model(s) on test applications in the real application if time and budget constraints permit.

Moreover, evaluation also assesses other data mining results generated. Data mining results cover models which are necessarily related to the original business objectives, and all other findings which are not necessarily related to the original business objectives but might also unveil additional challenges, information, or hints for future directions.

Output **Assessment of Data Mining Results w.r.t. Business Success Criteria**

Summarizes assessment results in terms of business success criteria including a final statement whether the project already meets the initial business objectives.

Activities

- Understand the data mining result
- Interpret the results in terms of the application
- Check impacts for data mining goal
- Check the data mining result against the given knowledge base to see if the discovered information is novel and useful
- Evaluate and assess result w.r.t. business success criteria
- Compare evaluation results and interpretation
- Create ranking of results w.r.t. business success criteria
- Check impacts of result for initial application goal

Output **Approved Models**

After model assessment w.r.t. business success criteria, you eventually get approved models if the generated models meet the selected criteria.

5.2 Review Process

Task **Review Process**

At this point the resultant model appears to be satisfactory and appears to satisfy business needs. It is now appropriate to make a more thorough review of the data mining engagement in order to determine if there is any important factor or task that has somehow been overlooked. At this stage of the Data Mining exercise, the Process Review takes on the form of a Quality Assurance Review.

Output **Review of Process**

Summarizes the process review and gives hints for activities that have been missed and/or should be repeated.

Activities

- Analyse data mining process
- Identify failures
- Identify misleading steps
- Identify possible alternative actions, unexpected paths in the process

5.3 Determine Next Steps

Task **Determine Next Steps**

According to the assessment results and the process review, the project decides how to proceed at this stage. The project needs to decide whether to finish this project and move onto deployment, or whether to initiate further iterations or whether to set up new data mining projects.

Output **List of Possible Actions**

A list of possible further actions along with the reasons for and against each option.

Activities

- Analyse potential for deployment of each result
- Estimate potential for improvement of current process
- Check remaining resources to determine if they allow additional process iterations (or whether additional resources can be made available)
- Recommend alternative continuations
- Refine process plan

Output **Decision**

Describes the decision as to how to proceed along with the rationale.

Activities

- Rank the possible actions
- Select one of the possible actions
- Document reasons for the choice

6 Deployment

6.1 Plan Deployment

Task **Plan Deployment**

This task takes the evaluation results and concludes a strategy for deployment of the data mining result(s) into the business.

Output **Deployment Plan**

Summarizes deployment strategy including necessary steps and how to perform them.

Activities

- Develop and evaluate alternative plans for deployment
- Identify possible problems when deploying the data mining results (pitfalls of the deployment)

6.2 Plan Monitoring and Maintenance

Task **Plan Monitoring and Maintenance**

Monitoring and maintenance are important issues if the data mining result becomes part of the day-to-day business and its environment. A careful preparation of a maintenance strategy helps to avoid unnecessarily long periods of incorrect usage of data mining results. In order to monitor the deployment of the data mining result(s), the project needs a detailed plan on the monitoring process. This plan takes into account the specific type of deployment.

Output **Monitoring and Maintenance Plan**

Summarizes monitoring and maintenance strategy including necessary steps and how to perform them.

Activities

- Check for dynamic aspects (i.e. what things could change in the environment?)
- When should the data mining result or model not be used any more? Identify criteria (validity, threshold of accuracy, new data, change in the application domain, etc.)? What should happen if the model or result can no longer be used? (Update model, set up new data mining project, etc.)
- Will the business objectives of the use of the model change over time? – fully document the initial problem the model was attempting to solve
- Develop Monitoring and Maintenance Plan

6.3 Produce Final Report

Task **Produce Final Report**

At the end of the project, the project leader and his team write up a final report. It depends on the deployment plan, if this report is only a summary of the project and its experiences, or if this report is a final presentation of the data mining result(s).

Output **Final Report**

At the end of the project, there will be (at least one) final report where all the threads will be brought together. As well as identifying the results obtained, the report should also describe the process, show which costs have been incurred, define any deviations from the original plan, describe implementation plans and make any recommendations for future work. The actual detailed content of the report depends very much on the audience for the particular report.

Activities

- Identify what reports are needed (slide presentation, management summary, detailed findings, explanation of models ....)
- Analyse how well initial data mining goals have been met
- Identify target groups for report
- Outline structure and contents of report(s)
- Select findings to be included in the reports
- Write a report

Output **Final Presentation**

As well as a Final Report, it may be necessary to make a Final Presentation to summarize the project – maybe to the management sponsor, for example. The Presentation will normally contain a subset of the information contained in the Final Report, but structured in a different way.

Activities

- Decide on target group for final presentation (will they already have received final report?)
- Select which items from final report should be included in final presentation

6.4 Review Project

Task **Review Project**

Assess what went right and what went wrong, what was done well and what needs to be improved.

Output **Experience Documentation**

Summarizes important experiences made during the project. For example, pitfalls, misleading approaches or hints for selecting the best-suited data mining techniques in similar situations could be part of this documentation. In ideal projects, experience documentation covers also any reports that have been written by individual project members during the project phases and their tasks.

Activities

- Interview all significant people involved in the project and ask them about their experiences during the project
- If end users in the business work with the data mining result(s), interview them - are they satisfied? What could have been done better? Do they need additional support?
- Summarise feedback and write the experience documentation
- Analyse the process (things that worked well, mistakes made, lessons learned,...)
- Document the specific data mining process (How can the results and the experience of applying the model be fed back into the process?)
- Abstract from details to make the experience useful for future projects.

IV The CRISP-DM Outputs

This part will contain document templates for each output in the CRISP-DM Process Model accompanied by brief descriptions of its main purpose and contents. For other outputs than documents, e.g., models and data sets, this part will include additional hints how to generate these types of outputs as well.

Unfortunately, this part is not completely finished yet!

V Appendix

1 Glossary / Terminology

- *CRISP-DM Methodology*

is the general term for all concepts developed and defined in CRISP-DM

- *Process Model*

defines the structure of data mining projects and provides guidance for their execution, consists of Reference Model and User Guide

- *Phase*

high-level term for part of the process model, consists of related tasks

- *Task*

part of a phase, series of activities to produce one or more outputs

- *Generic ~*

a task which holds across all possible data mining projects, as complete, i.e., cover both the whole data mining process and all possible data mining applications, and stable, i.e., valid for yet unforeseen developments like new modeling techniques, as possible

- *Specialized ~*

a task which makes specific assumptions in specific data mining contexts

- *Output*

tangible result of performing a task, decomposed into output components

- *Output Component*

part of an output

- *Activity*

part of a task in user guide, describes actions to perform a task

- *Process Instance*

a specific project described in terms of the process model

- *Reference Model*

decomposition of data mining projects into phases, tasks, and outputs

- *User Guide*

specific advice on how to perform data mining projects

- *Frame*

information structure to record activities and outputs

- *Template*

set of frames or set of templates

- *Data Mining Context*

set of constraints and assumptions such as problem type, techniques or tools, application domain

- *Data Mining Problem Type*

class of typical data mining problems such as data description and summarization, segmentation, concept descriptions, classification, prediction, dependency analysis

- *Model*

ability to apply to a data set to predict a target attribute, executable

2 Data Mining Problem Types

Usually, the data mining project involves a sequence of different problem types which together solve the business problem.

2.1 Data Description and Summarisation

Data Description and Summarisation aims at the concise description of characteristics of the data, typically in elementary and aggregated form. This gives the user an overview of the structure of the data. Sometimes, data description and summarisation alone can be an objective of a data mining project. For instance, a retailer might be interested in the turnover of all outlets broken down by categories. Changes and differences to a previous period could be summarised and highlighted. This kind of problems would be at the lower end of the scale of data mining problems.

However, in almost all data mining projects data description and summarisation is a subgoal in the process, typically in early stages. At the beginning of a data mining process, the user often knows neither the precise goal of the analysis nor the precise nature of the data. Initial exploratory data analysis can help to understand the nature of the data, and to find potential hypotheses for hidden information. Simple descriptive statistical and visualisation techniques provide first insights in the data. For example, the distribution of customer age and their living areas gives hints which parts of a customer group need to be addressed by further marketing strategies.

Data description and summarisation typically occurs in combination with other data mining problem types. For instance, data description may lead to the postulation of interesting segments in the data. Once segments are identified and defined a description and summarisation of these segments is useful. It is advisable to carry out data description and summarisation before any other data mining problem type is addressed. In this document, this is reflected by the fact that data description and summarisation is a task in the data understanding phase.

Summarisation also plays an important role in the presentation of final results.
The outcomes of the other data mining problem types (e.g., concept descriptions or prediction models) may also be considered summarisations of data, but on a higher conceptual level.

Many reporting systems, statistical packages, OLAP and EIS systems can cover data description and summarisation but do usually not provide any methods to perform more advanced modeling. If data description and summarisation is considered a stand alone problem type and no further modeling is required, these tools are also appropriate to carry out data mining engagements.

2.2 Segmentation

The data mining problem type *segmentation* aims at the separation of the data into interesting and meaningful subgroups or classes. All members of a subgroup share common characteristics. For instance, in shopping basket analysis one could define segments of baskets depending on the items they contain.

Segmentation can be performed manually or (semi-)automatically. The analyst can hypothesise certain subgroups as relevant for the business question based on prior knowledge or based on the outcome of data description and summarisation. However, there are also automatic clustering techniques that can detect previously unsuspected and hidden structures in data that allow segmentation.

Segmentation can be a data mining problem type of its own. Then the detection of segments would be the main purpose of data mining. For example, all addresses in zip code areas with higher than average age and income might be selected for mailing advertisements on home nursing insurance.

However, very often segmentation is a step towards solving other problem types. Then the purpose can be to keep the size of the data manageable or to find homogeneous data subsets which are easier to analyse. Typically, in large data sets various influences overlay each other and obscure the interesting patterns. Then, appropriate segmentation makes the task easier. For instance, analysing dependencies between items in millions of shopping baskets is very hard. It is much easier (and more meaningful, typically) to identify dependencies in interesting segments of shopping baskets, for instance high-value baskets, baskets containing convenience goods, or baskets from a particular day or time.

Note: In the literature there is a confusion of terms. Segmentation is sometimes called clustering or classification. The latter term is confusing because some people use it to refer to the creation of classes while others mean the creation of models to predict known classes for previously unseen cases. In this document, we restrict the term classification to the latter meaning (see below) and use the term segmentation for the former meaning, though classification techniques can be used to elicit descriptions of the segments discovered.

Appropriate techniques:

- Clustering techniques
- Neural nets
- Visualisation

Example:

A car company regularly collects information about its customers concerning their socio-economic characteristics like income, age, sex, profession etc. Using cluster analysis, the company can divide its customers into more understandable subgroups and analyse the structure of each subgroup. Specific marketing strategies are deployed for each group separately.

2.3 Concept descriptions

Concept description aims at an *understandable* description of concepts or classes. The purpose is not to develop complete models with a high prediction accuracy but to gain insights. For instance, a company may be interested to learn more about their loyal and disloyal customers. From a concept description of these concepts (loyal and disloyal customers) the company might infer what could be done to keep customers loyal or to transform disloyal customers to loyal customers.

Concept description has a close connection to both segmentation and classification. Segmentation may lead to an enumeration of objects belonging to a concept or class without any understandable description. Typically, there is a segmentation before concept description is performed. Some techniques, e.g., conceptual clustering techniques, perform segmentation and concept description at the same time.

Concept descriptions can also be used for classification purposes. On the other hand, some classification techniques produce understandable classification models which can then be considered as concept descriptions. The important distinction is that classification aims to be complete in some sense. The classification model needs to apply to *all* cases in the selected population. On the other hand, concept descriptions need not be complete. It is sufficient if they describe important parts of the concepts or classes. In the example above, it may be sufficient to get concept descriptions of those customers who are clearly loyal.

<u>Appropriate techniques:</u>

- Rule induction methods
- Conceptual clustering

<u>Example:</u>

Using data about the buyers of new cars and using a rule induction technique, a car company could generate rules which describe its loyal and disloyal customers. Below are examples of the generated rules:

If SEX = male and AGE > 51 then CUSTOMER = loyal
If SEX = female and AGE > 21 then CUSTOMER = loyal
If PROFESSION = manager and AGE < 51 then CUSTOMER = disloyal
If FAMILY STATUS = bachelor and AGE < 51 then CUSTOMER = disloyal

2.4 Classification

Classification assumes that there is a set of objects - characterised by some attributes or features - which belong to different classes. The class label is a discrete (symbolic) value and is known for each object. The objective is to build classification models (sometimes called classifiers) which assign the correct class label to previously unseen and unlabeled objects. Classification models are mostly used for predictive modelling.

The class labels can be given in advance, for instance defined by the user, or derived from a segmentation.

Classification is one of the most important data mining problem types that occurs in a wide range of various applications. Many data mining problems can be transformed to classification problems. For example, credit scoring tries to assess the credit risk of a new customer. This can be transformed to a classification problem by creating two classes, good and bad customers. A classification model can be generated from existing customer data and their credit behaviour. This classification model can then be used to assign a new potential customer to one of the two classes and hence accept or reject him.

Classification has connections to almost all other problem types. Prediction problems can be transformed to classification problems by discretising continuous class labels since discretisation techniques allow to transform continuous ranges into discrete intervals. These discrete intervals are then used as class labels rather than the exact numerical values and hence lead to a classification problem. Some classification techniques produce understandable class or concept descriptions. There is also a connection to dependency analysis because classification models typically exploit and elucidate dependencies between attributes.

Segmentation can either provide the class labels or restrict the data set such that good classification models can be built.

It is useful to analyse deviations before a classification model is built. Deviations and outliers can obscure the patterns which would allow a good classification model. On the other hand, a classification model can also be used to identify deviations and other problems with the data.

Appropriate techniques:

- Discriminant analysis
- Rule induction methods
- Decision tree learning
- Neural nets
- k Nearest Neighbour
- Case-based reasoning
- Genetic algorithms

Example:

Banks generally have information on the payment behaviour of their credit applicants. Combining this financial information with other information about the customers like sex, age, income etc. it is possible to develop a system to classify new customers as good or bad customers, i.e., the credit risk in acceptance of a customer is either low or high, respectively.

2.5 Prediction

Another important problem type that occurs in a wide range of applications is *prediction*. Prediction is very similar to classification. The only difference is that in prediction the target attribute (class) is not a qualitative discrete attribute but a continuous one. It means that the aim of prediction is to find the numerical value of the target attribute for unseen objects. In the literature, this problem type is sometimes called regression. If prediction deals with time series data then it is often called forecasting.

Appropriate techniques:

- Regression analysis
- Regression trees
- Neural nets
- k Nearest Neighbour
- Box-Jenkins methods
- Genetic algorithms

Example:

The annual revenue of an international company is correlated with other attributes like advertisement, exchange rate, inflation rate etc. Having these values (or their reliable estimations for the next year) the company can predict its expected revenue for the next year.

2.6 Dependency Analysis

Dependency analysis consists of finding a model which describes significant dependencies (or associations) between data items or events. Dependencies can be used to predict the value of a data item given information on other data items. Although dependencies can be used for predictive modelling, they are mostly used for understanding. Dependencies can be strict or probabilistic.

Associations are a special case of dependencies which have recently become very popular. Associations describe affinities of data items (i.e., data items or events which frequently occur together). A typical application scenario for associations is the analysis of shopping baskets. There, a rule like "In 30% of all purchases, beer and peanuts have been bought together." is a typical example for an association.

Algorithms for detecting associations are very fast and produce many associations. Selecting the most interesting ones is a challenge.

Dependency analysis has close connections to prediction and classification, where dependencies are implicitly used for the formulation of predictive models. There is also a connection to concept descriptions which often highlight dependencies.

In applications, dependency analysis often co-occurs with segmentation. In large data sets, dependencies are seldom significant because many influences overlay each other. In such cases it is advisable to perform a dependency analysis on more homogeneous segments of the data.

Sequential patterns are a special kind of dependencies where sequences of events are considered. In the shopping basket domain, associations describe dependencies between items at a given time. Sequential patterns describe shopping patterns of one particular customer or a group of customers over time.

Appropriate Techniques:

- Correlation analysis
- Regression analysis
- Association rules
- Bayesian networks
- Inductive Logic Programming
- Visualisation techniques

Example 1:

Using regression analysis, a business analyst has found that there is a significant dependency between the total sales of a product and its price and the amount of the total expenditures for the advertisement. Once the analyst discovered this knowledge, he can reach the desired level of the sales by changing the price and/or the advertisement expenditure accordingly.

Example 2:

Applying association rule algorithms to data about car accessories, a car company has found that if a radio is ordered, an automatic gearbox is ordered as well in 95 % of all cases. Based on this dependency, the car company decides to offer these accessories as a combination which leads to cost reduction.