



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 customer really wants to accomplish. Often the customer has 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

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

Activities

Organization

- Develop organizational charts 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)
- Indicate if there is a steering committee and list members
- Identify the business units which are affected by the data mining project (e.g., Marketing, Sales, Finance)

Problem area

- Identify the problem area (e.g., marketing, customer care, business development, etc.)
- 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 a data mining project is to be performed, or whether data mining needs to be promoted 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., Are we expected to deliver a report for top management or an operational system to be used by naive end users?)
- Identify the users' needs and expectations

Current solution

- Describe any solution currently used to address 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 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, while a secondary business objective might be to determine whether lower fees affect only one particular segment of customers.

Activities

- Informally describe the problem to be solved
- 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 and readily measurable, 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, be sure to indicate who would make the subjective judgment.

Activities

- Specify business success criteria (e.g., Improve response rate in a mailing campaign by 10 percent and sign-up rate by 20 percent)
- Identify who assesses the success criteria



Remember! Each of the success criteria should relate to at least one of the specified business objectives.

Good idea! Before starting the situation assessment, you might analyze 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 in developing the project plan.

Output **Inventory of resources**

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

Activities **Hardware resources**

- 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)

Sources of data and knowledge

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

Personnel sources

- 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, for example 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 verified during data mining, but may also include non-verifiable assumptions related to the project. It is particularly important to list the latter if they will affect 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 in the time 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**Requirements**

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

Assumptions

- 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 \$1,000)
- List all assumptions regarding 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 the starting point of the project has been.

Output

Risks and contingencies

List the risks, that is, the 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 organizational risks (e.g., department requesting project doesn't have funding for the 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 under which each risk may occur
- Develop contingency plans

Output

Terminology

Compile 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 benefits to the business if it is successful

Activities

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

Good idea!

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

Beware!

Remember to identify hidden costs, such as repeated data extraction and preparation, changes in workflows, and time required for training.

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, relevant demographic information, and the price of the item.”

Output**Data mining goals**

Describe the intended outputs of the project that enable the achievement of the business objectives. Note that these are normally 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; the level/size of the segments should be specified).
- Specify data mining problem type (e.g., classification, description, prediction, and clustering). For more details about data mining problem types, see Appendix 2.



Good idea!	It may be wise to re-define the problem. For example, modeling product retention rather than customer retention when targeting customer retention delivers results 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	<ul style="list-style-type: none">■ 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 explain ability and data and marketing insight provided by the model)
Beware!	Remember that the data mining success criteria are different than the business success criteria defined earlier. Remember it is wise to plan for deployment from 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 their 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 analyze 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 are manifested. Although this is the only task in which the project plan is directly named, it nevertheless should be consulted continually and reviewed throughout the project. The project plan should be consulted at minimum whenever a new task is started or a further iteration of a task or activity is begun.

Activities

- Define the initial process plan and discuss the feasibility with all involved personnel
- Combine all identified goals and selected techniques in a coherent procedure that solves the business questions and meets the business success criteria
- Estimate the effort and resources needed to achieve and deploy the solution. (It is useful to consider other people's experience when estimating timescales for data mining projects. For example, it is often postulated that 50-70 percent of the time and effort in a data mining project is used in the Data Preparation Phase and 20-30 percent in the Data Understanding Phase, while only 10-20 percent is spent in each of the Modeling, Evaluation, and Business Understanding Phases and 5-10 percent 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 team performs an initial assessment of tools and techniques. Here, it is important to select a data mining tool that supports various methods for different stages of the process, since the selection of tools and techniques may influence 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 prioritize applicable techniques according to the evaluation of alternative solutions

2 Data understanding

2.1 Collect initial data

Task

Collect initial data

Acquire 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

Describe all the various data used for the project, and include 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 results from merging data sources. Because of inconsistencies between the sources, merged data may present problems that do not exist in the individual data sources.

Activities

Data requirements planning

- Plan which information is needed (e.g., only for given attributes, or specific 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 a history one should use (e.g., even if 18 months of data are available, only 12 months may be 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 a customer database may show inconsistencies of format, invalidity of data, etc.).

Insertion of data

- If the data contain free text entries, do we need to encode them for modeling or do we want to group specific entries?
- How can missing attributes be acquired?
- How can we best extract the data?

Good idea!

Remember that some knowledge about the data may be available from non-electronic sources (e.g., from people, printed text, etc.).

Remember that it may be necessary to preprocess 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 that 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 that have been discovered.

Activities**Volumetric analysis of data**

- Identify data and method of capture
- Access data sources
- Use statistical analyses if appropriate
- Report tables and their relations
- Check data volume, number of multiples, complexity
- Note if the data contain free text entries

Attribute types and values

- Check accessibility and availability of attributes
- Check attribute types (numeric, symbolic, taxonomy, etc.)
- Check attribute value ranges
- Analyze attribute correlations
- Understand the meaning of each attribute and attribute value in business terms
- For each attribute, compute basic statistics (e.g., compute distribution, average, max, min, standard deviation, variance, mode, skewness, etc.)
- Analyze basic statistics and relate the results to their meaning in business terms
- Decide if the attribute is relevant for the specific data mining goal



- Determine if the attribute meaning is used consistently
- Interview domain experts to obtain their opinion of attribute relevance
- Decide if it is necessary to balance the data (based on the modeling techniques to be used)

Keys

- Analyze key relationships
- 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 that can be addressed using querying, visualization, and reporting techniques. These analyses may directly address 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 steps needed before further analysis can occur.

Output

Data exploration report

Describe the results of this task, including first findings or initial hypotheses and their impact on the remainder of the project. The report may also include graphs and plots that indicate data characteristics or point to interesting data subsets worthy of further examination.

Activities

Data exploration

- Analyze 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 a hypothesis and identify actions
- Transform the hypothesis into a data mining goal, if possible
- Clarify data mining goals or make them more precise. A “blind” search is not necessarily useless, but a more directed search toward 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? 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 catalog their meaning

Review keys, attributes

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

Good idea!

Review any attributes that give answers that conflict with common sense (e.g., teenagers with high income levels).

Use visualization plots, histograms, etc. to reveal inconsistencies in the data.

Data quality in flat files

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



Noise and inconsistencies between sources

- Check consistencies and redundancies between different sources
- Plan for dealing with noise
- Detect the 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 behavior (e.g., to check on customers' loan behavior, exclude all those who have never borrowed, do not finance a home mortgage, those whose mortgage is nearing maturity, etc.).

Review whether assumptions are valid or not, given the current information on data and business knowledge.

3 Data preparation

Output

Dataset

These are the dataset(s) produced by the data preparation phase, used for modeling or for the major analysis work of the project.

Output

Dataset description

This is the description of the dataset(s) used for the modeling or for 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 tests to decide if fields should be included
- Reconsider Data Selection Criteria (See Task 2.1) in light of experiences of data quality and 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 modeling (i.e., model assessment may show that other datasets are needed)
- Select different data subsets (e.g., different attributes, only data which meet certain conditions)

- Consider the use of sampling techniques (e.g., A quick solution may involve splitting test and training datasets or reducing the size of the test dataset, if the tool cannot handle the full dataset. 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 the 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

Describe the decisions and actions that were taken to address the data quality problems reported during the Verify Data Quality Task. If the data are to be used in the data mining exercise, the report should address outstanding data quality issues and what possible effect this could have on the results.

Activities

- Reconsider how to deal with any 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 not answered. This might result in a value of 99 for unknown data. For example, 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 odometer.
- Reconsider Data Selection Criteria (See Task 2.1) in light of experiences of data cleaning (i.e., you may wish to 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., you 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: $\text{area} = \text{length} * \text{width}$.

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 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 modeling algorithm in use handles only certain types of data—for example 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 modeling phase suggests 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 certain currencies, the income will dominate)
- Consider adding new information on the relevant importance of attributes by adding new attributes (for example, attribute weights, weighted normalization)

- How can missing attributes be constructed or imputed? [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 ease the model process or facilitate the modeling algorithm. Perhaps “income per person” is a better/easier attribute to use than “income per household.” Do not derive attributes simply to reduce the number of input attributes.

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

Activities

Single-attribute transformations

- Specify necessary transformation steps in terms of available transformation facilities (for example, change a binning of a numeric attribute)
- Perform transformation steps

Good idea!

Transformations may be necessary to change ranges to symbolic fields (e.g., ages to age ranges) or symbolic fields (“definitely yes,” “yes,” “don’t know,” “no”) to numeric values. Modeling tools or algorithms often require them.

Output

Generated records

Generated records are completely new records, which add new knowledge or represent new data that 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 for combining information 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 information from multiple records and/or tables.

Activities

- Check if integration facilities are able to integrate the input sources as required
- Integrate sources and store results
- Reconsider Data Selection Criteria (See Task 2.1) in light of experiences of data integration (i.e., you may wish to 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 refers primarily to 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 dataset. 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., you may wish to 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 initial modeling technique. If multiple techniques are to be applied, perform this task separately for each technique.

Remember that not all tools and techniques are applicable to each and every task. For certain problems, only some techniques are appropriate (See Appendix 2, where techniques appropriate for certain data mining problem types are discussed in more detail). “Political requirements” and other constraints further limit the choices available to the data mining engineer. It may be that only one tool or technique is available to solve the problem at hand—and that the tool may not be absolutely the best, from a technical standpoint.

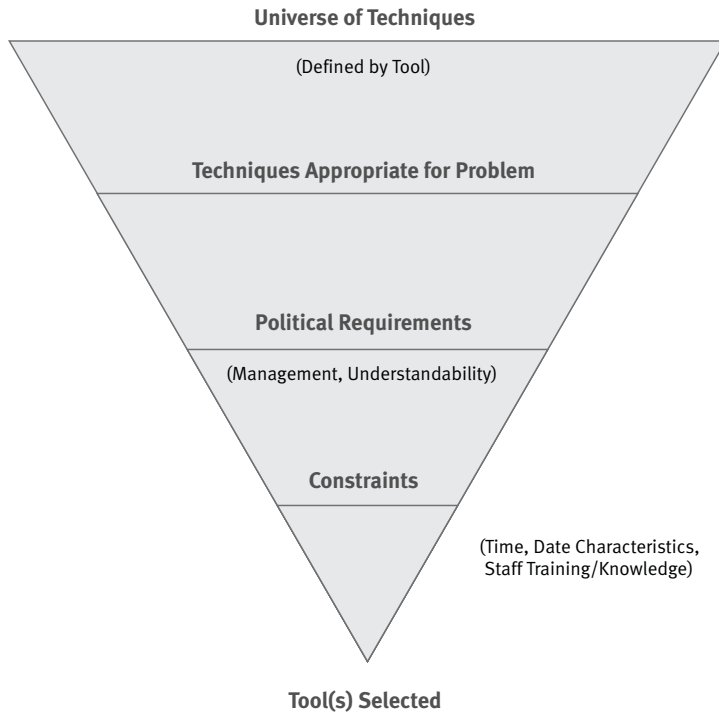


Figure 10:
Universe of Techniques

Output	Modeling technique Record 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.
Activities	<ul style="list-style-type: none">■ 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 go back to the Data Preparation Phase, if necessary

4.2 Generate test design

Task **Generate test design**

Prior to building a model, it is necessary to define a procedure 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 specifies that the dataset should be separated into training and test sets. The model is built on the training set and its quality estimated on the test set.

Output **Test design**

Describe the intended plan for training, testing, and evaluating the models. A primary component of the plan is to decide how to divide the available dataset 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. List 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 dataset to create one or more models.

- Activities**
- Run the selected technique on the input dataset to produce the model
 - Post-process data mining results (e.g., edit rules, display trees)



Output

Model description

Describe the resulting 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
- Record parameter settings used to produce the model
- Give a detailed description of the model and any special features
- For rule-based models, list the rules produced, plus any assessment of per-rule or overall model accuracy and coverage
- For opaque models, list any technical information about the model (such as neural network topology) and any behavioral descriptions produced by the modeling process (such as accuracy or sensitivity)
- Describe the model's behavior and interpretation
- State conclusions regarding patterns in the data (if any); sometimes the model reveals important facts about the data without a separate assessment process (e.g., that the output or conclusion is duplicated in one of the inputs)

4.4 Assess model

Task

Assess model

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

Output

Model assessment

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

Activities

- Evaluate results with respect to evaluation criteria
- Test result according to a test strategy (e.g.: Train and Test, Cross-validation, bootstrapping, etc.)
- Compare evaluation results and interpretation
- Create ranking of results with respect to success and evaluation criteria
- Select best models
- Interpret results in business terms (as far as possible at this stage)
- Get comments on models by domain or data experts
- Check plausibility of model

- Check effect on data mining goal
- Check model against given knowledge base to see if the discovered information is novel and useful
- Check reliability of result
- Analyze potential 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
- Get insights into why a certain modeling technique and certain parameter settings lead to good/bad results

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

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

Activities Adjust parameters to produce better models.

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. It compares results 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:

$$\text{RESULTS} = \text{MODELS} + \text{FINDINGS}$$

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 the findings, which we define as anything (apart from the model) that is important in meeting the objectives of the business or important in leading to new questions, lines of approach, or side effects (e.g., data quality problems uncovered by the data mining exercise). Note: Although the model is directly connected to the business questions, the findings need not be related to any questions or objectives, as long as they are important to the initiator of the project.



5.1 Evaluate results

Task

Evaluate results

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 is to test the model(s) on test applications in the real application, if time and budget constraints permit.

Moreover, evaluation also assesses other generated data mining results. Data mining results cover models that are related to the original business objectives and all other findings. Some are related to the original business objectives while others might unveil additional challenges, information, or hints for future directions.

Output

Assessment of data mining results with respect to business success criteria

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

Activities

- Understand the data mining results
- Interpret the results in terms of the application
- Check effect on 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 results with respect to business success criteria (i.e., has the project achieved the original Business Objectives)
- Compare evaluation results and interpretation
- Rank results with respect to business success criteria
- Check effect of result on initial application goal
- Determine if there are new business objectives to be addressed later in the project, or in new projects
- State recommendations for future data mining projects

Output

Approved models

After accessing models with respect to business success criteria, select and approve the generated models that meet the selected criteria.

5.2 Review process

Task

Review process

At this point, the resulting 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 the form of a Quality Assurance Review.

Output

Review of process

Summarize the process review and list activities that have been missed and/or should be repeated.

Activities

- Provide an overview of the data mining process used
- Analyze the data mining process. For each stage of the process ask:
 - Was it necessary?
 - Was it executed optimally?
 - In what ways could it be improved?
- Identify failures
- Identify misleading steps
- Identify possible alternative actions and/or unexpected paths in the process
- Review data mining results with respect to business success criteria

5.3 Determine next steps

Task

Determine next steps

Based on the assessment results and the process review, the project team decides how to proceed. Decisions to be made include whether to finish this project and move on to deployment, to initiate further iterations, or to set up new data mining projects.

Output

List of possible actions

List possible further actions along with the reasons for and against each option.



- Activities**
- Analyze the 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**
Describe the decisions made, along with the rationale for them.

- 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 starts with the evaluation results and concludes with a strategy for deployment of the data mining result(s) into the business.

Output **Deployment plan**
Summarize the deployment strategy, including necessary steps and how to perform them.

- Activities**
- Summarize deployable results
 - Develop and evaluate alternative plans for deployment
 - Decide for each distinct knowledge or information result
 - Determine how knowledge or information will be propagated to users
 - Decide how the use of the result will be monitored and its benefits measured (where applicable)
 - Decide for each deployable model or software result
 - Establish how the model or software result will be deployed within the organization's systems
 - Determine how its use will be monitored and its benefits measured (where applicable)
 - Identify possible problems during deployment (pitfalls to be avoided)

6.2 Plan monitoring and maintenance

Task **Plan monitoring and maintenance**

Monitoring and maintenance are important issues if the data mining results become 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 for monitoring and maintenance. This plan takes into account the specific type of deployment.

Output **Monitoring and maintenance plan**

Summarize 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?)
 - Decide how accuracy will be monitored
 - Determine when the data mining result or model should not be used any more. Identify criteria (validity, threshold of accuracy, new data, change in the application domain, etc.), and what should happen if the model or result could 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 team writes up a final report. Depending on the deployment plan, this report may be only a summary of the project and its experience, or 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 in which all the threads are 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 intended audience.



- Activities**
- Identify what reports are needed (slide presentation, management summary, detailed findings, explanation of models, etc.)
 - Analyze 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 normally contains a subset of the information contained in the final report, structured in a different way.

- Activities**
- Decide on target group for the final presentation and determine if they will already have received the final report
 - Select which items from the 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**

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

- Activities**
- Interview all significant people involved in the project and ask them about their experience 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?
 - Summarize feedback and write the experience documentation
 - Analyze the process (things that worked well, mistakes made, lessons learned, etc.)
 - Document the specific data mining process (How can the results and the experience of applying the model be fed back into the process?)
 - Generalize from the details to make the experience useful for future projects