info@sv-europe.com — 020 7786 3568

Cart / £0.00 0

What is the CRISP-DM methodology?

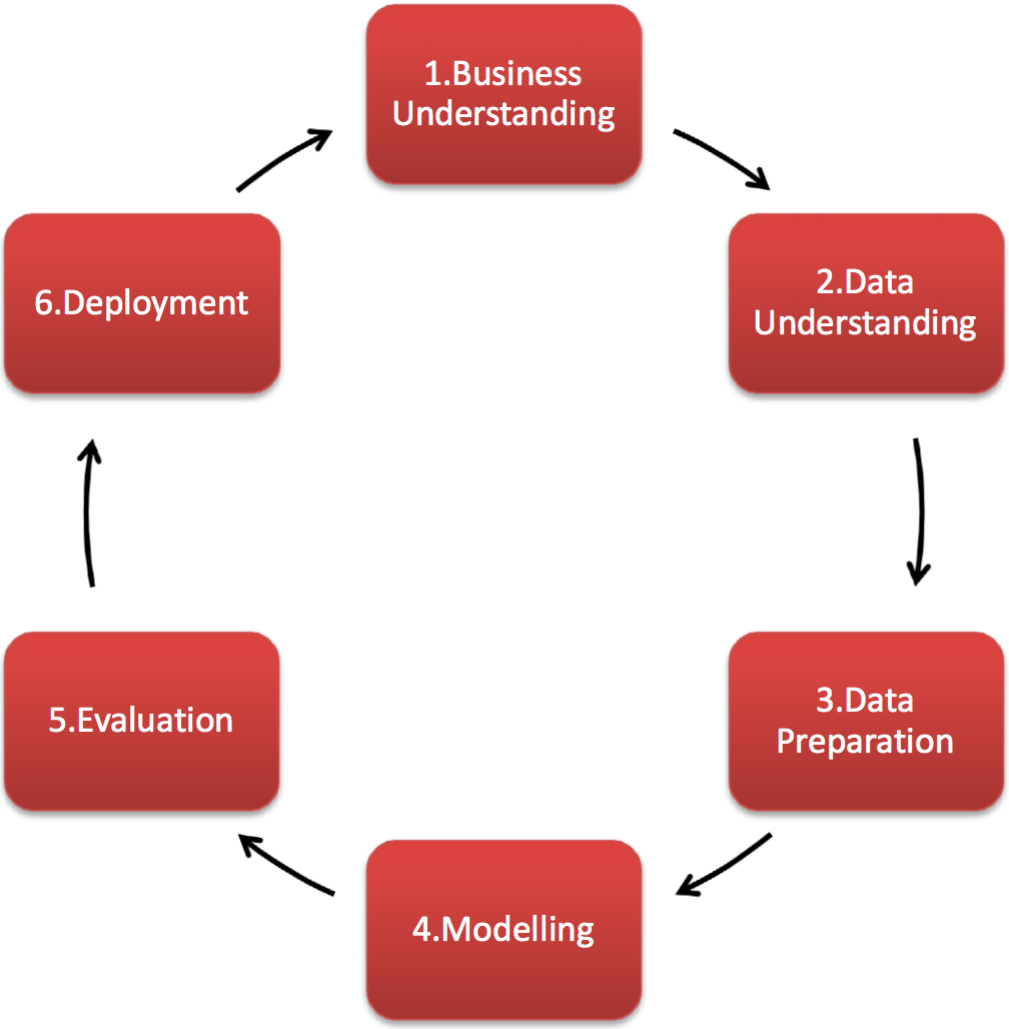
CRISP-DM stands for cross-industry process for data mining. The CRISP-DM methodology provides a structured approach to planning a data mining project. It is a robust and well-proven methodology. We do not claim any ownership over it. We did not invent it. We are however evangelists of its powerful practicality, its flexibility and its usefulness when using analytics to solve thorny business issues. It is the golden thread than runs through almost every client engagement. The CRISP-DM model is shown on the right.

This model is 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 backtrack to previous tasks and repeat certain actions. The model does not try to capture all possible routes through the data mining process.

You can jump to more information about each phase of the process here:

1. Business understanding
2. Data understanding
3. Data preparation
4. Modeling
5. Evaluation
6. Deployment

We have other CRISP DM resource available to help you with your data mining projects. You can download our free guide to using CRISP DM to evaluate data mining tools or you can watch the recording of our introduction to CRISP DM webinar.



STAGE ONE – DETERMINE BUSINESS OBJECTIVES

The first stage of the CRISP-DM process is to understand what you want to accomplish from a business perspective. Your organisation may have competing objectives and constraints that must be properly balanced. The goal of this stage of the process is to uncover important factors that could influence the outcome of the project. Neglecting this step can mean that a great deal of effort is put into producing the right answers to the wrong questions.

What are the desired outputs of the project?

1. **Set objectives** – This means describing your primary objective from a business perspective. There may also be other related questions that you would like to address. For example, your primary goal might be to keep current customers by predicting when they are prone to move to a competitor. Related business questions might be "Does the channel used affect whether customers stay or go?" or "Will lower ATM fees significantly reduce the number of high-value customers who leave?"
2. **Produce project plan** – Here you'll describe the plan for achieving the data mining and business goals. The plan should specify the steps to be performed during the rest of the project, including the initial selection of tools and techniques.

3. **Business success criteria** – Here you'll lay out the criteria that you'll use to determine whether the project has been successful from the business point of view. These should ideally be specific and measurable, for example reduction of customer churn to a certain level, however sometimes it might be necessary to have more subjective criteria such as "give useful insights into the relationships." If this is the case then it needs to be clear who it is that makes the subjective judgment.

Assess the current situation

This involves more detailed fact-finding about all of the resources, constraints, assumptions and other factors that you'll need to consider when determining your data analysis goal and project plan.

1. **Inventory of resources** – List the resources available to the project including:
   - Personnel (business experts, data experts, technical support, data mining experts)
   - Data (fixed extracts, access to live, warehoused, or operational data)
   - Computing resources (hardware platforms)
   - Software (data mining tools, other relevant software)
2. **Requirements, assumptions and constraints** – List all requirements of the project including the schedule of completion, the required comprehensibility and quality of results, and any data security concerns as well as any legal issues. Make sure that you are allowed to use the data. List the assumptions made by the project. These may be assumptions about the data that can be verified during data mining, but may also include non-verifiable assumptions about the business related to the project. It is particularly important to list the latter if they will affect 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 set that it is practical to use for modelling.
3. **Risks and contingencies** – List the risks or events that might delay the project or cause it to fail. List the corresponding contingency plans – what action will you take if these risks or events take place?
4. **Terminology** – Compile a glossary of terminology relevant to the project. This will generally have two components:
   - 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.
   - A glossary of data mining terminology, illustrated with examples relevant to the business problem in question.
5. **Costs and benefits** – Construct a cost-benefit analysis for the project which compares the costs of the project with the potential benefits to the business if it is successful. This comparison should be as specific as possible. For example, you should use financial measures in a commercial situation.

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." 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.”

1. **Business success criteria** – describe the intended outputs of the project that enable the achievement of the business objectives.
2. **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.

Produce project plan

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

1. **Project plan** – List the stages to be executed in the project, together with their duration, resources required, inputs, outputs, and dependencies. Where possible, try and make explicit the large-scale iterations in the data mining process, for example, repetitions of the modelling 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 if the risks are manifested. Decide at this point which evaluation strategy will be used in the evaluation phase. Your project plan will be a dynamic document. At the end of each phase you’ll review progress and achievements and update the project plan accordingly. Specific review points for these updates should be part of the project plan.
2. **Initial assessment of tools and techniques** – At the end of the first phase you should undertake an initial assessment of tools and techniques. Here, for example, you select a data mining tool that supports various methods for different stages of the process. It is important to assess tools and techniques early in the process since the selection of tools and techniques may influence the entire project.

STAGE TWO – DATA UNDERSTANDING

The second stage of the CRISP-DM process requires you to acquire the data listed in the project resources. This initial collection includes data loading, if this is necessary for data understanding. For example, if you use a specific tool for data understanding, it makes perfect sense to load your data into this tool. If you acquire multiple data sources then you need to consider how and when you’re going to integrate these.

- **Initial data collection report** – List the data sources acquired together with their locations, the methods used to acquire them and any problems encountered. Record problems you encountered and any

resolutions achieved. This will help both with future replication of this project and with the execution of similar future projects.

Describe data

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

- **Data description report** – Describe the data that has been acquired including its format, its quantity (for example, the number of records and fields in each table), the identities of the fields and any other surface features which have been discovered. Evaluate whether the data acquired satisfies your requirements.

Explore data

During this stage you’ll address data mining questions using querying, data visualization and reporting techniques. These may include:

- Distribution of key attributes (for example, the target attribute of a prediction task)
- Relationships between pairs or small numbers of attributes
- Results of simple aggregations
- Properties of significant sub-populations
- Simple statistical analyses

These analyses may directly address your 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 steps needed for further analysis.

- **Data exploration report** – Describe results of your data exploration, including first findings or initial hypothesis and their impact on the remainder of the project. If appropriate you could include graphs and plots here to indicate data characteristics that suggest further examination of interesting data subsets.

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?

Data quality report

List the results of the data quality verification. If quality problems exist, suggest possible solutions. Solutions to data quality problems generally depend heavily on both data and business knowledge.

STAGE THREE – DATA PREPARATION

Select your data

This is the stage of the project where you decide on the data that you're going to use for analysis. The criteria you might use to make this decision include the relevance of the data to your data mining goals, the quality of the data, and also 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.

- **Rationale for inclusion/exclusion** – List the data to be included/excluded and the reasons for these decisions.

Clean your data

This task involves raise the data quality to the level required by the analysis techniques that you've selected. This may involve selecting clean subsets of the data, the insertion of suitable defaults, or more ambitious techniques such as the estimation of missing data by modelling.

- **Data cleaning report** – Describe what decisions and actions you took to address data quality problems. Consider any transformations of the data made for cleaning purposes and their possible impact on the analysis results.

Construct required data

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

- **Derived attributes** – These are new attributes that are constructed from one or more existing attributes in the same record, for example you might use the variables of length and width to calculate a new variable of area.
- **Generated records** – Here you describe the creation of any completely new records. For example you might need to 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 modelling purposes it might make sense to explicitly represent the fact that particular customers made zero purchases.

Integrate data

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

- **Merged data** – Merging tables refers to joining together two or more tables that have different information about the same objects. For example a retail chain might have one table with information about each store's general characteristics (e.g., floor space, type of mall), another table with summarised 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.
- **Aggregations** – Aggregations refers to operations in which new values are computed by summarising 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.

STAGE FOUR – MODELLING

Select modeling technique

As the first step in modelling, you'll select the actual modelling technique that you'll be using. Although you may have already selected a tool during the business understanding phase, at this stage you'll be selecting the specific modelling technique e.g. decision-tree building with C5.0, or neural network generation with back propagation. If multiple techniques are applied, perform this task separately for each technique.

- **Modelling technique** – Document the actual modelling technique that is to be used.
- **Modelling assumptions** – Many modelling techniques make specific assumptions about the data, for example that all attributes have uniform distributions, no missing values allowed, class attribute must be symbolic etc. Record any assumptions made.

Generate test design

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

- **Test design** – Describe the intended plan for training, testing, and evaluating the models. A primary component of the plan is determining how to divide the available dataset into training, test and validation datasets.

Build model

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

- **Parameter settings** – With any modelling tool there are often a large number of parameters that can be adjusted. List the parameters and their chosen values, along with the rationale for the choice of parameter settings.
- **Models** – These are the actual models produced by the modelling tool, not a report on the models.
- **Model descriptions** – Describe the resulting models, report on the interpretation of the models and document any difficulties encountered with their meanings.

Assess model

Interpret the models according to your domain knowledge, your data mining success criteria and your desired test design. Judge the success of the application of modelling and discovery techniques technically, then contact business analysts and domain experts later in order to discuss the data mining results in the business context. This task only considers models, whereas the evaluation phase also takes into account all other results that were produced in the course of the project.

At this stage you should rank the models and assess them according to the evaluation criteria. You should take the business objectives and business success criteria into account as far as you can here. In most data mining projects a single technique is applied more than once and data mining results are generated with several different techniques.

- **Model assessment** – Summarise the results of this task, list the qualities of your generated models (e.g.in terms of accuracy) and rank their quality in relation to each other.
- **Revised parameter settings** – According to the model assessment, revise parameter settings and tune them for the next modelling run. Iterate model building and assessment until you strongly believe that you have found the best model(s). Document all such revisions and assessments.

STAGE FIVE – EVALUATION

Evaluate your results

Previous evaluation steps dealt with factors such as the accuracy and generality of the model. During this step you'll assesses the degree to which the model meets your business objectives and seek 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. The evaluation phase also involves assessing any other data mining results you've generated. Data mining results involve models that are necessarily related to the original business objectives and all other findings that are not necessarily related to the original business objectives, but might also unveil additional challenges, information, or hints for future directions.

- **Assessment of data mining results** – Summarise assessment results in terms of business success criteria, including a final statement regarding whether the project already meets the initial business objectives.

- **Approved models** – After assessing models with respect to business success criteria, the generated models that meet the selected criteria become the approved models.

Review process

At this point, the resulting models appear to be satisfactory and to satisfy business needs. It is now appropriate for you 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—for example: did we correctly build the model? Did we use only the attributes that we are allowed to use and that are available for future analyses?

- **Review of process** – Summarise the process review and highlight activities that have been missed and those that should be repeated.

Determine next steps

Depending on the results of the assessment and the process review, you now decide how to proceed.Do you finish this project and move on to deployment, initiate further iterations, or set up new data mining projects? You should also take stock of your remaining resources and budget as this may influence your decisions.

- **List of possible actions** – List the potential further actions, along with the reasons for and against each option.
- **Decision** – Describe the decision as to how to proceed, along with the rationale.

STAGE SIX – DEPLOYMENT

Plan deployment

In the deployment stage you'll take your evaluation results and determine a strategy for their deployment. If a general procedure has been identified to create the relevant model(s), this procedure is documented here for later deployment. It makes sense to consider the ways and means of deployment during the business understanding phase as well, because deployment is absolutely crucial to the success of the project. This is where predictive analytics really helps to improve the operational side of your business.

- **Deployment plan** – Summarise your deployment strategy including the necessary steps and how to perform them.

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. The 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 monitoring process plan. This plan takes into account the specific type of deployment.

- **Monitoring and maintenance plan** – Summarise the monitoring and maintenance strategy, including the necessary steps and how to perform them.

Produce final report

At the end of the project you will write up a final report. Depending on the deployment plan, this report may be only a summary of the project and its experiences (if they have not already been documented as an ongoing activity) or it may be a final and comprehensive presentation of the data mining result(s).

- **Final report** – This is the final written report of the data mining engagement. It includes all of the previous deliverables, summarising and organising the results.
- **Final presentation** – There will also often be a meeting at the conclusion of the project at which the results are presented to the customer.

Review project

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

- **Experience documentation** – Summarise important experience gained during the project. For example, any pitfalls you encountered, 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 also covers any reports that have been written by individual project members during previous phases of the project.

There's more advice on how to manage the deployment phase of the data mining process in this post on our blog.

Terms and conditions
Privacy Policy
Technical support
Frequently asked questions

Connect with us

Contact us

info@sv-europe.com

020 7786 3568

Registered address

Level 17, Dashwood House, 69 Old Broad Street, London, EC2M 1QS

Copyright © 2021 Smart Vision Europe