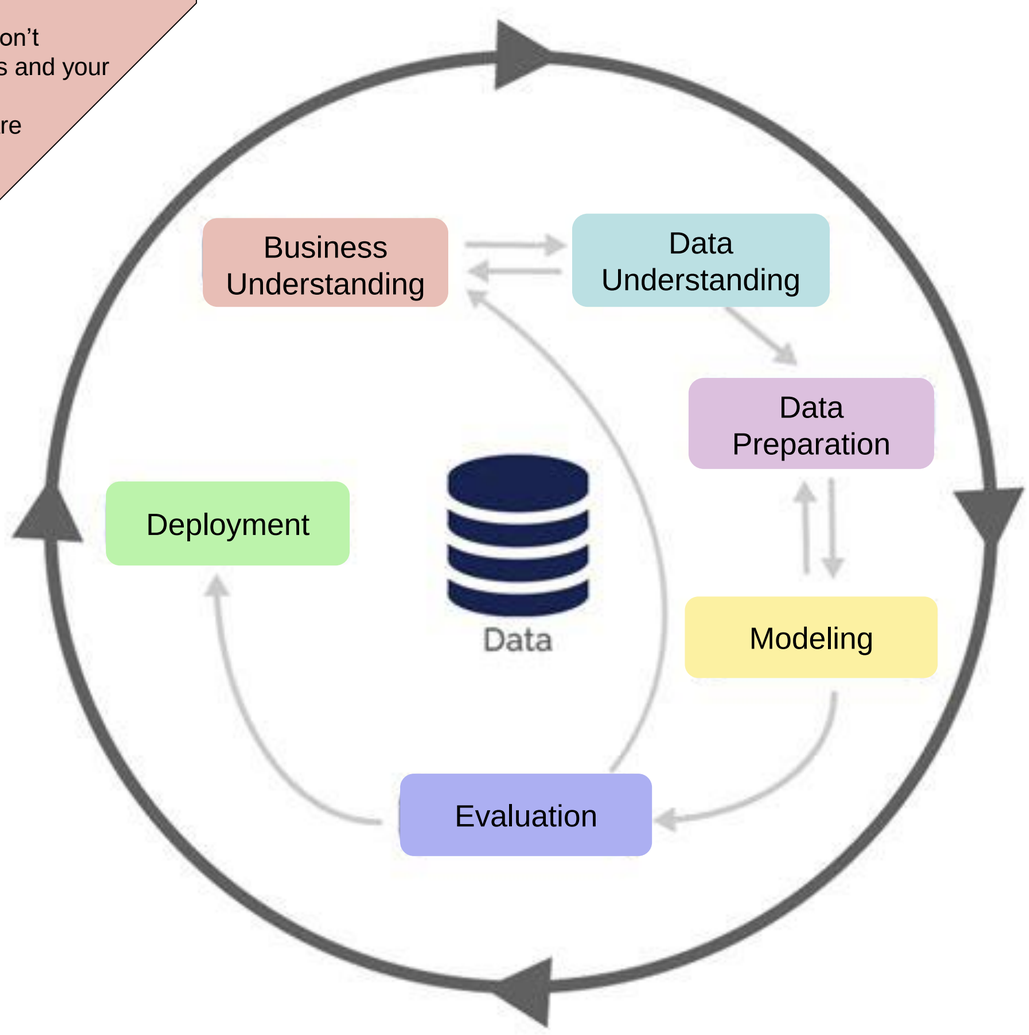
Cross Industry Standard Process for Data Mining

Abstract The terms big data and data mining are often used in the same context. However, it is important to separate the two terms properly.
Big Data deals with particularly large amounts of data, that cannot be processed efficiently and within a reasonable timeframe using conventional methods and tools.
Data mining is often used for large amounts of data, but is not limited to big data.
This is a big challenge for data science. Therefore for companies like Get Your Guide it is important to handle large amounts of data properly, to interpret them correctly and to act accordingly.
CRISP-DM (Cross Industry Standard Process for Data Mining) is a standardized process model that can be used for data mining in order to search databases for patterns, trends and correlations. For this, the standard defines six different phases, which have to be carried out one or more times.
CRISP-DM has established itself worldwide and is one of the most frequently used analytics models in this environment.
The model can be used across industries and is generally available.

Business Understanding:
The first step is trying to get a better idea of what business needs should be extracted from data. The analyst has to understand what the customer really wants from a business perspective. The customer often has several competing goals and restrictions that need to be properly coordinated.
Moreover the business understanding phase is about defining the specific goals and requirements for data mining. The result of this phase is the formulation of the task and the description of the planned rough procedure to achieve both business and data mining goals.
This also includes the initial selection of tools and techniques.
Besides, here you'll lay out the business success criteria that you'll use to determine whether the project has been successful from the business point of view.
Deliverables of this step are some important reports.
The inventory of resources lists all resources available for the project. These may include people (not just data miners, but also those with expert knowledge of the business problem, data managers, technical support, and others), data, hardware, and software.
Requirements, assumptions and constrains are written down. Requirements for example include a schedule for completion, legal and security obligations, and requirements for acceptable finished work.
Besides, you identify causes that could delay completion of the project, and prepare a contingency plan for each of them.
An important part is creating a common understanding of terminology.
You may create a glossary with definitions of business and data-mining terms that are relevant to your project so that everyone involved in it can have a common understanding of those terms.
Moreover, you prepare a cost-benefit analysis. If the benefits don't significantly exceed the costs, stop and reconsider this analysis and your project.
After that you also define data-mining success criteria, which are technical criteria necessary to support the business success criteria. You try to define these in quantitative terms (such as model accuracy or predictive improvement compared to an existing method).

Data Understanding:
As part of the data understanding, an attempt is made to get a first overview of the available data and their quality.
This involves checking whether all the required data (to meet the Data Mining goals) is actually available as well as developing a plan to determine which data is required.
First, you 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. You evaluate whether the data acquired satisfies your requirements.
If some of the data you want is unavailable, you have to decide how to address that issue.
Alternatives could be substituting with an alternative data source, narrowing the scope of the project or gathering new data.
Importing the data into the data-mining platform you'll be using for the project makes it possible to confirm you can do so and that you understand the process. In the course of this trial you may discover software (or hardware) limitations you had not anticipated, such as limits on the number of cases, fields, or on the amount of memory you may use or inability to read the data formats of your sources.
Then you explore the data using querying, data visualization and reporting techniques.
This may include finding out the distribution of key attributes, simple aggregations or simple statistical analyses.
To avoid potential pitfalls the data quality is analysed and evaluated.
Problems with the quality of the existing data in relation to the task defined in the previous phase must be identified.
Typical questions in this phase are:

- 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?



Data Preparation:
In this phase, the data is prepared for the further data mining process. Data preparation is one of the most important and often time-consuming aspects of data mining. In fact, it is estimated that data preparation usually takes 50-70% of a project's time and effort.
Business decisions rely on analytics. But, if the data is inaccurate or incomplete, your analytics inform wrong business decisions. Bad analytics means poor business decisions.

Therefore data from different sources is merged and cleaned up so that there are no duplicate, incorrect or incomplete entries. Changes are made e.g. tracking down sources to make specific data corrections, excluding some cases or individual cells (items of data), or replacing some items of data with default values.
The deliverable for this task is the data-cleaning report, which documents every decision and action used to clean the data. This report should cover and refer to each data quality problem that was identified in the verify data quality task in the data-understanding phase of the process.
The next step is to prepare the content of the data, which means that the data is transferred into usable formats.
You may need to add derived attributes, which are new attributes that are constructed from one or more existing attributes in the same record, aggregate data or add completely new attributes.
Besides selection criteria must be defined.
This includes attributes (columns) as well as selection of records (rows) in a table.
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.

The data preparation is used to create a final data set that forms the basis for the next phase of the modeling.

Deployment:
After data preparation, model building and model verification, the selected model is used in the deployment phase.

Generating a model is generally not the end of the project. Even if the goal was to deepen the knowledge of the data, the knowledge gained must now be processed and presented to the customer so that the customer can use it without any problems.

Depending on the requirements, this development phase can consist of the creation of a simple report or the complex implementation of a repeatable data mining process throughout the company.
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.

Finally, the data-mining team should review its work.
This is where an outline of any work methods that worked particularly well should be done, so that they are documented to use again in the future, and any improvements that might be made to your process. It's also the place to document problems and bad experiences, with your recommendations for avoiding similar problems in the future.

Modeling:
Modeling is the analytical core of the data mining process. This is where the selection and use of modeling techniques take place.
Before actually building a model, you typically separate the dataset into train, test and validation sets.
Then you build the models on the train set.
Many modeling techniques make specific assumptions about the data, for example that all attributes have uniform distributions or no missing values are allowed. Besides, with any modeling tool there are often a large number of parameters that can be adjusted. You have to record any assumptions made and list the parameters and their chosen values.

The found models then run through a technical evaluation phase.
They are reviewed for accuracy and generality.
The found rules are applied to test data records that were not used in the modeling.
Then you list the qualities of your generated models (e.g.in terms of accuracy) and rank their quality in relation to each other.
Iterative steps are used to approach gradually the final model, which only allows marginal improvements from a data analytical perspective.

Evaluation:
The evaluation ensures an exact comparison of the created data models with the task and selects the most suitable model.
The results of the previous steps are evaluated using the business criteria established at the beginning of the project.
So this phase is about checking whether the data mining solution satisfies the business problem and seeking to determine if there is some business reason why a model is deficient.
Another option is to test the model(s) on test applications in the real application, to determine whether it works as well in the workplace as it did in your tests, if time and budget constraints permit.
Now you may take time to review your process. This is an opportunity to spot issues that you might have overlooked and that might draw your attention to flaws in the work that you've done while you still have time to correct the problem before deployment.
Finally, the model may be ready to deploy, or you may judge that it would be better to repeat some steps and try to improve it.

Patricia Sowa

Bachelor-Student (3. Semester)
Ringvorlesung - Practical Data Engineering

E-Mail: patricia.sowa@student.hpi.de

HPI Hasso Plattner Institut