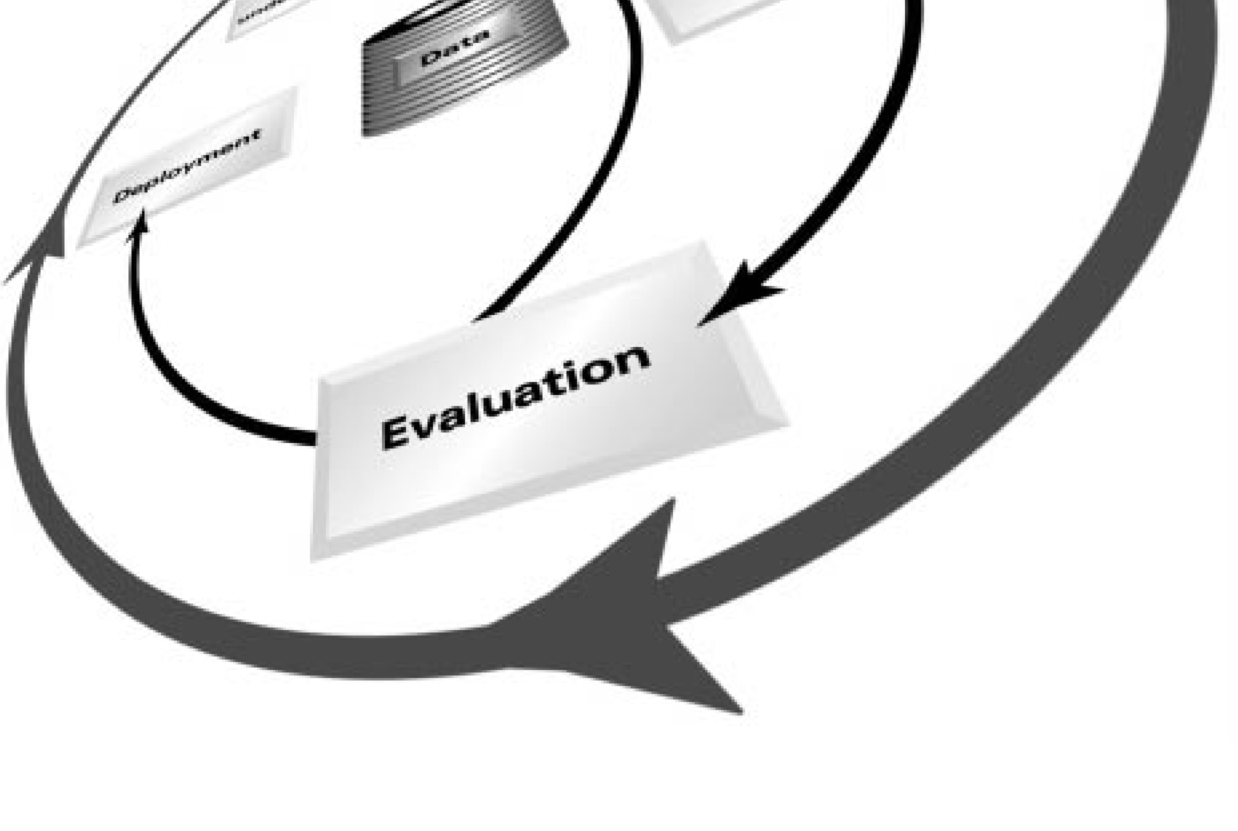
Performing a data mining tool evaluation



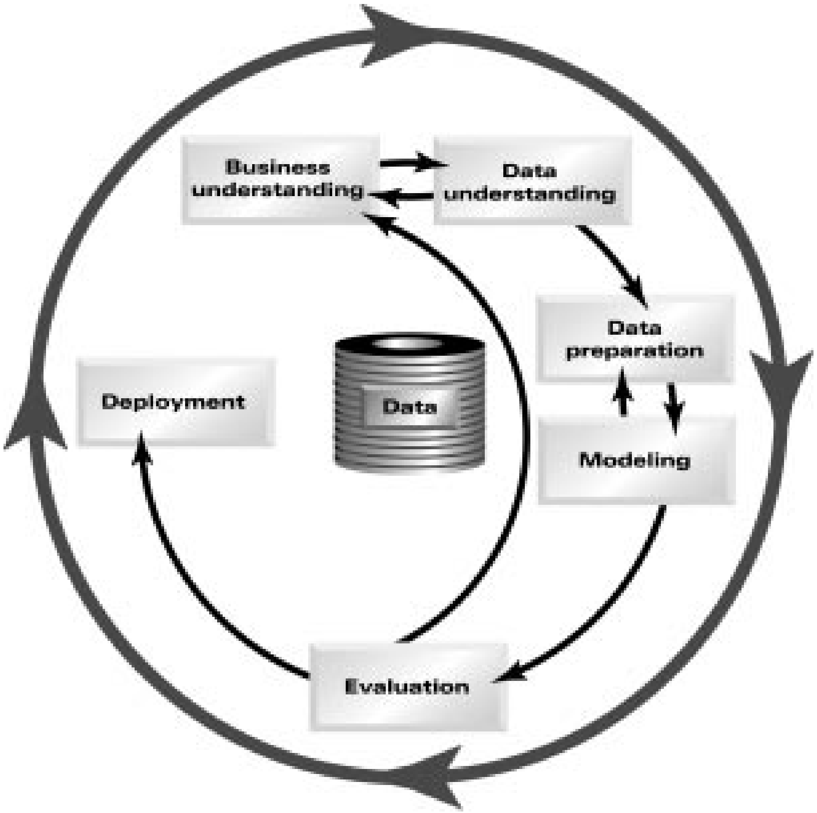
Start with a framework for your evaluation

Data mining helps you make better decisions that lead to significant and concrete results, such as increased revenue and more efficient processes. While the promise of data mining can be dramatic, some of the hype surrounding data mining suggests that incredible results can be attained with minimal effort. Choosing a data mining tool with this expectation can lead to a disappointing return on investment.

To get the results that you expect from your data mining project, assess your business situation and evaluate how tools perform throughout the complete data mining process. To help in your evaluation, the following checklist has been compiled using CRISP-DM, the worldwide cross-industry standard process model for carrying out data mining projects.

CRISP-DM organizes the data mining process into 6 phases, as shown in this diagram. The sequence of the phases is not strict; moving back and forth between different phases is always required. Keep this interactive aspect of data mining in mind in your data mining tool evaluation. Your chosen tool should be flexible enough to make changes to any phase as you work through a project.

The evaluation criteria that follow are organized into these 6 phases, with some additional considerations at the end.



Phases of the CRISP-DM reference model

CRISP-DM phase 1: business understanding

The 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. The checklist items address whether the tool has a track record with a wide range of business problems and how the tool links business understanding to the technical aspects of the data mining process.

- Does the tool have a proven record of solving a wide range of business problems, including the problems that we face?
  - Has the tool shown itself to be useful at solving problems in our industry or with our application? Does the tool have a successful track record with applications that we may want to do?

- How does the tool provide a bridge between business understanding and the technical aspect of data mining?
  - Can the steps for using the tool be clearly mapped to the business needs of data mining? Are data mining concepts presented clearly for the business user? How does the tool integrate with project management or other planning tools? Does the tool require writing applications to bridge this gap?

CRISP-DM phase 2: data understanding

The data understanding phase begins with data collection and proceeds through activities for getting familiar with the data. Open standards for accessing data and visualization techniques are important considerations in this phase.

- How does the tool preserve my existing investment in IT infrastructure?
  - Does the data mining tool work well with your existing data stores? Does the tool work with open data standards or do data need to be migrated to a proprietary data format?

- How does the tool enable interactive exploration and visualization of the data?
  - Does the data mining tool provide visualization techniques to help see patterns in the data? Can visualization be performed interactively by making changes within graphs themselves and by creating new graphs based on different dimensions of the data?

CRISP-DM phase 3: data preparation

The data preparation phase covers all activities to construct the final data set from the initial raw data. Checklist items include the efficiency and ease of data preparation.

- How does the tool prepare data?
  - Is the entire interactive data mining process, including the activities needed to prepare raw data until the data are ready for model building, scaled for efficient data mining? Does the tool present data preparation steps in an easy to follow way?

- Can the tool automatically extract data for preparation?
  - Can the tool extract data automatically or is manual work required to write SQL queries for joins, aggregations, sorts and other data preparation operations?

CRISP-DM phase 4: modeling

In this phase, various modeling techniques are selected and applied and parameters are calibrated to optimal values. Often, analysts jump back to the data preparation phase to meet the requirements of different types of models. Because different models address the same data mining problem type, the key checklist items include the proficiency of the tool at applying and comparing different techniques.

- How does the tool boost analyst productivity?
  - Does the tool enable analysts to develop effective models quickly? How easily can users try different models to come up with the best solution? How easily can data preparation be done to meet specific model needs?

- Does the tool offer a wide range of techniques?
  - Does the tool offer techniques or algorithms for visualization, classification, clustering, association and regression?

- Does the tool enable the combination of techniques?
  - Can different techniques be easily combined to get better results? Can the results of algorithms be incorporated into the data set for post-processing and analysis?

- Does the tool preserve my existing investment in technology such as algorithms and other tools?
  - Can the data mining tool work with your existing algorithms? How well does the data mining tool work with your other data analysis tools?

CRISP-DM phase 5: evaluation

The evaluation phase is a thorough assessment of the model or models before deployment. A key objective is to determine if an important business issue has not been sufficiently considered. The checklist items for the evaluation phase relate to how well input from business users have been incorporated into the model and how well the results speak to the intended audience.

- Does the tool achieve consistently high results?
  - Does your data mining tool create solutions that consistently perform to a high standard or does it only provide good results in certain cases or with some types of data? Do your results accurately reflect all business issues and therefore perform well on test data?

- Does the tool provide results that are easy to understand?
  - Are your results easy for business users to understand? If not, what steps are required to make results persuasive? Has the tool encouraged input from business experts throughout the data mining process?

- Can the full range of visualization be applied to validate the results of a model?
  - Can model predictions, scores and other results be easily analyzed to validate the model's performance?

CRISP-DM phase 6: deployment

In the deployment phase, the data mining process is used, whether the process simply gives insight into a business issue or the process is implemented in an application to provide up to date knowledge to information consumers. Often, deployment requires extensive services, so the key checklist item deals with the ability of the tool to help with this task.

- How can I deploy my data mining solutions (now and in the future)?
  - How can my data mining solutions be integrated into operational applications? Can integration be done cost effectively or will it require a substantial investment in development time and cost? How easily can my solutions be updated? If solutions cannot be easily updated, what work and costs do I incur?

Additional considerations: cost of ownership

In addition to the data mining phases in the CRISP process model, you will want to perform a return on investment analysis.

- What is the cost of ownership?
  - Quantify the costs of ownership over the life of the product and service offering, including required complementary systems. Quantify the return that is anticipated. When is a positive ROI expected and does this result meet the business goals?

- What is the implementation time?
  - How long will it take to implement your data mining tool. Does it require other tools or hardware? How much training, consulting and customization is needed to get results from the project?

- Do the skills of our users (now and in the future) match the skills required to use the tool?
  - What skills are required to get useful results with the data mining tool? Is it a tool that is designed for the most technically savvy users, data mining beginners or can it be used by both groups and users at skill levels in between? What training costs will I have to incur to get everyone up to speed? Don't forget to consider the skill sets of potential users in the future.

- Can the tool be customized for our users?
  - How can the tool be customized for different uses? Can common processes be saved for reuse? Can the tool automate tasks? Are services available to customize interfaces or provide other help?

Additional considerations: vendor

Finally, consider the strength of the vendor in your data mining tool assessment. In many ways, purchasing a data mining solution is an investment in your future and you'll want a reliable partner for the road ahead.

- Does the vendor offer other tools and services that can help solve similar problems?
  - Does the vendor offer other data mining or data analysis tools? Are consulting, training, technical support and other services available? Are these services available on a worldwide basis?

- Will the vendor be a reliable source that can update software and services that meet my needs?
  - Is the vendor a leader in providing data mining solutions? Does the vendor have the resources required to continue to provide a high level of service in the future?