



Tutorial: A Data Mining Project

Data mining is the process of applying computerized techniques to databases in order to discover new and useful relationships in the data. Data mining may be carried out using a variety of techniques and tools, and may range from the straightforward analysis of changes in profit ratios over time to complicated statistical and manipulation procedures directed toward answering a single important business question, and even extending to Distributed Data Mining and Control techniques deploying independent software agents that analyze and control from a bottom-up perspective - intelligent learning agents that react, adapt, and improve over time.

Depending on data availability and the complexity of the business question, data mining efforts can vary from simple week-long projects to more intensive multi-month programs. First projects are typically 6 to 10 weeks long. Beyond the core analysis, an important part of data mining programs is making the result operational. Achieving value from data mining requires embedding the resulting information, and sometimes the analysis technique itself, into business processes. Furthermore, implementing a MIS that flags significant deviations in current business practices from historical norms derived from the data analytics is a useful end product.

An example of the type of question that data mining might be used to answer is one proposed during an on-site assessment with the client: *What is the relationship between profitability and production volume?* Let us consider in some detail how a data mining team might go about answering this question for its clients—the individuals who proposed it for study. This discussion is relevant because each of the questions proposed by our experts for data mining study would involve the steps and considerations described below.

1. Get the relevant data

The data mining team must talk with the clients in order to understand what data in the client databases is relevant to answering the question. Frequently, data from multiple databases must be brought together to provide the most extensive support for the data mining effort. The process of stitching together data from multiple databases can take some time. Fortunately, if the data mining team is studying multiple questions, this type of effort generally takes place once, and subsequent data mining projects require less time integrating data.

2. Understand, check, and clarify the data

The data mining team will probably spend a good deal of time understanding what the different fields in the data mean. In real-world databases, there are a number of fields that may be incompletely filled in, or that may have systematically inaccurate entries. Analysis of the data in the database and discussion of the data with the clients can be a lengthy process, particularly for the first data mining project carried out on a database.

The data mining team will need to decide how to handle missing and inaccurate data in the database. Should records containing missing critical data be ignored? Should the data be filled in by manual intervention (for example, with a call to an operations manager)? Can the missing data be inferred based on other entries in the database?

The data mining team might need to settle on its approach to fields in the database that contain freely structured text. A good deal of information can be contained in such text, but extracting the information can require parsing procedures that are time-consuming to produce. The clients will probably be involved in determining whether to use text fields, and if so, how to process them for data mining purposes.

The data mining team will check to see whether there are fields that seem irrelevant to answering the question, so that they can be excluded. Project description line item information, for example, is probably too detailed for consideration when answering a higher-order question such as the question relating profitability to production volume. Furthermore, the data mining team will also need to determine whether there are fields that are alternate encodings of the answer. Such fields should be disabled during the data mining process.

3. Understand the question terms

The question posed by the clients relates production volume and profitability. Understanding the various alternative measures that can be used for these two variables requires discussions with the experts. For example, *should production volume be Net or Gross?* The interpretation of profitability also involves discussion. (Several hours during the assessment phase can be spent on just this topic.) There are two top-level definitions of profitability that were suggested by our experts: GOP (Gross Operating Profit), and Discounted Sales. These two measures of profitability are correlated, but not perfectly. The data mining team may decide to carry out its studies using both measures.

There are several ways in which these two measures of profitability can be computed. For instance, one could consider project profitability, continuing quote profitability, and account profitability over the lifetime of the account. Each of these measures could be of interest to the clients (and each may be of interest to one or more of the experts).

There are also two different ways to measure profitability itself: as a total amount, and as a percentage of project revenues. The experts felt that, in general, larger projects will show greater total profit but lower percentages of profit, since deeper levels of discounting are often associated with larger projects. Conversely, smaller projects would tend to be associated with higher percentages of profit but lower total amounts of profit. In this data mining project, and the others discussed below, both of these measures of profit are important.

Thus, the measures of profitability that could be relevant to the question under study include: GOP by project, GOP by continuing quote, GOP by account, Discounted Sales by project, Discounted Sales by continuing quote, and Discounted Sales by account. Each of these could be related to total profit, and percentage profit, yielding a total of twelve different definitions of profit to consider. The study could reveal differences between any two such interpretations of profit, and such differences could be of interest to the clients. One of the decisions to be made before the data mining process begins is which of these measures of profitability will be considered in the study. The more definitions considered, the more complex the study will be. The experts and the data mining team will make the decision based on a tradeoff between effort to be expended and the expected value of the results.

4. Apply data mining techniques

There is a wide array of techniques that can be used to mine data. Statistical techniques, neural networks, machine learning techniques, genetic algorithms, rough sets techniques, fuzzy set techniques, decision tree building procedures, k nearest neighbors techniques, and a host of other tools are available for data mining. Each of these techniques has its strengths and weaknesses, and part of the value provided to the project by the data mining team lies in understanding which techniques to use, and when. At NuTech, we use the techniques that are generally employed, such as those in SAS Institute's Enterprise Miner data mining toolkit. We also use some advanced and specialized techniques that have been developed by NuTech, and that supplement and improve on the performance of the more traditional techniques. Data mining personnel in one of our offices are often used as partners by SAS Institute when a data mining project begins and the SAS tools are not well suited to the job.

Some data mining tools can be very good for preliminary investigation of data. At the end of the Assessment Phase, it may be shown through a demonstration how decision trees can be used to understand relationships in one of the client databases. Decision trees are also excellent for discovering fields in the database that are alternate encodings of the quantity under study.

In studying the question of profitability and volume, the data mining team will use data mining tools to build “models” of the data. Each tool “learns”, based on this data, how to predict the level of profitability for a given set of input values. The way it carries out such predictions constitutes its model of the data. A good deal of expertise is required in applying data mining tools to build models. A tool that is too powerful can “memorize” each instance it sees during its training phase, producing a model that will not generalize well outside the set of instances it is trained on. A tool that is too weak will build a model with poor predictive power. One of the most important parts of the data mining process is building models at the appropriate level of generality. There are ways to verify that this has been done. Withholding part of the training data and then testing the model’s performance on the data that has not been seen is one popular approach, as is its more complex variant, multifold cross-validation.

When one or more models have been built, the data mining team can use them to predict what the level of profitability for future projects will be, based on the model constructed from projects that served as input to the study.

5. Understand the answers

Sometimes the value of a data mining project lies in uncovering unusual aspects of the data – outliers. For example, in the sample analysis given below, we have identified a subset of high volume production where the profitability is unusually high and also counter to the expectation that high production volume is associated with low profitability levels. Similarly, we also identify subsets where the profitability levels are unusually low. In general, outlier segments help you learn about your business and often represent opportunities where intervention can make a big impact.

Other times, the value of a data mining project lies in the predictions of the models themselves. Given a neural network model, for example, that relates various factors to profitability, we might want to feed the data related to a potential new asset into that model in order to see what the profit possibilities may be, before committing to a level of investment for the new asset. The pure prediction capability of the neural network might be of great value to us, even though the model itself cannot be readily interpreted by a human.

6. Interpret the answers

If the models built in the data mining process have unexpected features or unexpected outcomes, it may be useful to determine whether the correlations discovered are accidental or real. One of the most useful features of an unexpected data mining outcome is the way it can stimulate new approaches in the minds of its clients. Client reactions to unexpected outcomes of data mining projects are sometimes the greatest benefits of those projects.

For example, NuTech Solutions did a series of data mining projects for a major automotive manufacturer in 2000. The client had purchased a new database of data centered on car sales and trade-ins across a number of automobile brands, and had asked NuTech to answer ten important business questions by applying data mining techniques to the new database. The answers to six of the questions were more or less in line with client’s expectations. The answers to the other four results were counter-intuitive, and motivated changes in the client’s approach to determining the colors of new models of cars, the allocation of incentives, and assessment of customer worth based on the type of automobile being purchased.

We do not know in advance which of the questions proposed for study a client will yield answers that are unexpected. In our experience, at least two or three of them should do so. If so, the interpretation of the results may produce the greatest benefits of the data mining project.

There can also be great benefit in learning that the expected outcome is indeed supported by the data. Many of our clients can successfully predict the outcome of a data mining project, but they are not certain of their prediction. Once the study has confirmed their beliefs, they may be more prepared to make business decisions than they were when the belief was held but unsupported.

7. New data mining efforts

Most successful data mining projects lead to related projects. If the answer to a business question is interesting, the client generally thinks of a number of variations of the question that refine its value or extend the approach used to related domains. Like successful software tools, successful data mining efforts are often revised and extended. They also often lead to improved MIS that focuses on aspects of your business not previously given much attention.

8. Repeating the study

A data mining project builds models of data collected up to a certain point in time. Market conditions, economic conditions, and reservoirs change with time. A successful data mining project yields a methodology for extracting data, one or more model-building techniques that provided the result, and one or more models based on the data. After time has passed, a set of more recent data can be used to repeat the data mining process, particularly when the process has supported important business decisions. Typically, continuing data mining efforts require much lower degrees of effort than the initial effort. In many cases, they can be run in an entirely automated way to produce reports that can be reviewed by the clients. The ability to repeat a data mining project as new data comes in is another of the benefits of data mining projects.

Successful data mining projects can require interaction with the domain experts and a good deal of work with the data. They use appropriate data mining tools. When most successful, they generate unexpected results or corroborate expectations, leading to beneficial changes in business processes.