Demystifying Data Mining

By Michael Lieberman

A little knowledge can go a long way in the hands of marketers. Data mining is a powerful set of methodologies that, when successfully applied, will increase business revenue, cuts costs, or result in other actions to improve the bottom line. Data mining exercises are deployed to make decisions about marketing strategies, new product promotion, and to compare and contrast competitors.

Data mining know-how is not new, though the buzzword is relatively so. Companies have long used statistics and database systems to extract valuable information from data sets. Today, innovations in artificial intelligence, machine learning, computer processing power, storage, data warehousing, and statistical software are dramatically increasing analysis accuracy and speed while driving down costs.

The intent of data mining is to identify correlations or patterns among multiple fields in large relational databases. In simple terms, it seeks patterns in large scale data sets, attempting to find something new in the data from all parts of business, from production to management. Data mining software allows users to analyze data from different angles, categorize the data, and summarize the relationships identified for business decisions.

Typical Data Mining Methodologies

Some of the more frequently employed data mining methodologies include classification, regression, and clustering.

Classification

Classification is a popular data mining application where the variable of interest—the one we would like to predict—is categorical in nature. Categorical data distinguishes between groups such as gender or age group. It is different from continuous data such as weight or price. Examples include:

- Credit scoring, determining whether a person is a good or bad credit risk
- Medicine, determining if a given patient is at a high or low risk for diabetes
- Insurance, predicting an individual's risk category to commit fraud

Classification data mining techniques can take on descriptive or predictive aspects. For example, we can look for new categories of behavior that are strongly related to the main variable of interest, such as brands purchased most frequently. Or, we might ask what indicators, such as driving record, might be red flags for insurance fraud.

The goal of a classification data mining project is to develop predictive models able, in real-time, to implement or adjust marketing and business decisions.

Regression

Regression analysis is the most frequently used statistical method in marketing research. It can reveal the effect of one variable on another variable and thus the relative strengths of effects. In a regression data mining project, the variable of interest is continuous. Examples include:

- Estimating the effect on sales if prices are increased by 5%
- Understanding if sales are affected more by price or advertising
- Determining which promotion outperforms other promotions

For example, we might be interested in learning the amount of money donors are willing to give to a foundation. Say, a university foundation wants to determine what most motivates non-alumni donors to give to the university's foundation. These individuals did not attend the university so why would they be interested in donating? We can test several different regression techniques and perhaps we find, after running a multivariate linear regression model based on survey responses, that three variables pop out that best explain donor behavior: visits to the university hospital, desire to honor a current colleague, or a business connection.

The goals of a regression problem are like that of a classification project. We wish to find the best predictors related to the variable of interest, and to develop a predictive model to find something valuable, say, the lifetime value of a donor.

Clustering

Clustering is the third very useful data mining application but it has quite a different goal than our first two. Here, a variable of interest does not exist. Instead, we attempt to sort the data into groups of similar clusters. Some typical clustering project examples are:

- Clustering individuals for a marketing campaign
- Clustering symptoms in medical research to find relationships
- Finding clusters of products purchased based on customer survey responses.

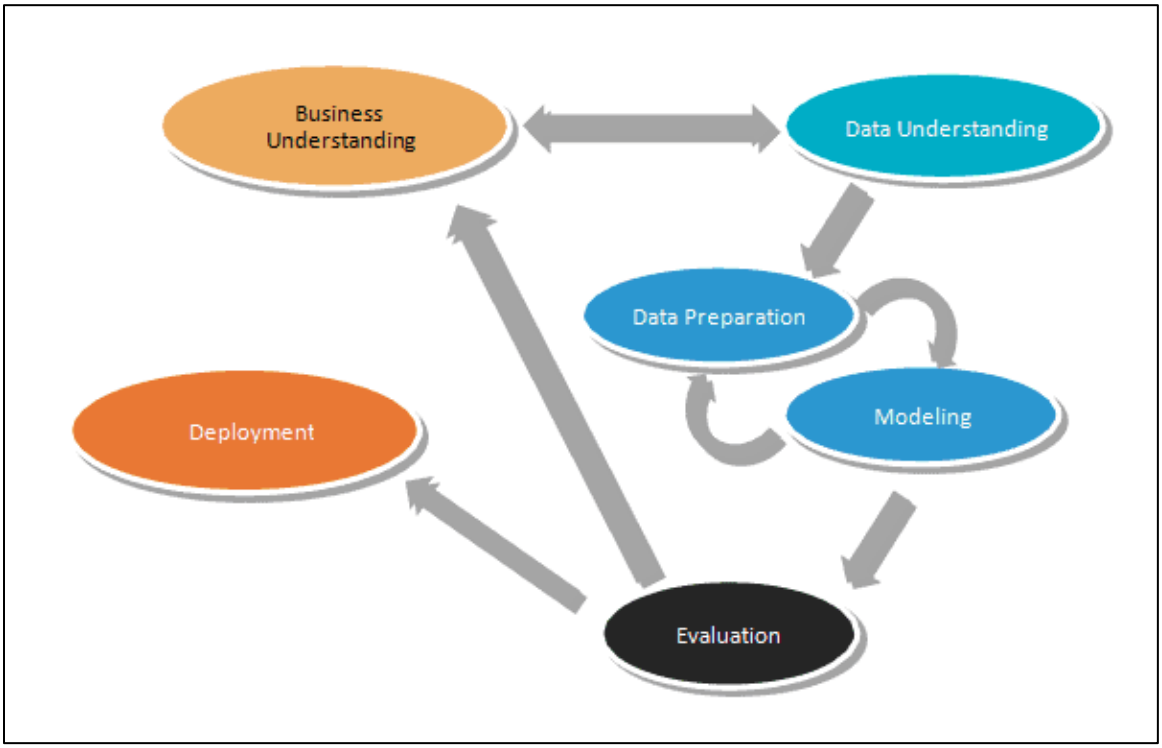
Perhaps the best example of a clustering data mining project is the ubiquitous Netflix model, which clusters customers into groups if they rent or enjoy similar movies, and makes recommendations based on their movie watching history.

The goals of a clustering data mining project are also descriptive, in that we are looking for the variables around which clusters are defined. We might also want to compare the clusters across variables of interest. Often, the most important part of cluster analysis is to assign new cases to clusters by our measure of their strengths of cluster membership.

The CRISP-DM Model

The exponential explosion of computer power and data availability over the past two decades has made it useful to organize projects using a standardized data mining process called CRISP-DM

(Cross Industry Standard Process for Data Mining). This is a process model that describes best-practice approaches that expert data miners use to tackle problems. CRISP-DM includes six major phases that are not necessarily linear as there is usually movement back and forth between key phases as early exploration yields greater understanding of the problem and solution.



Business Understanding
The project objectives and requirements are outlined from a business perspective and then converted into a data mining problem definition. Using the problem definition, a preliminary plan to achieve the business objectives is developed.

Data Understanding
Before modeling with any data set, the data mining expert must become familiar with it by identifying data quality problems, detecting preliminary insights, learning about all the fields from domain experts, and exploring the possibility of interesting data subsets in which useful information may be hidden.

Data Preparation
A final dataset is developed to be used during modeling, which may entail manipulating raw data multiple times. Actions in this phase include tabling, recording, and attribute selection as well as transformation and cleaning of data.

Modeling
Modeling techniques are dictated by the nature of the data and the goal. Typically, several will be tried. As some have specific data format requirements, returning to the data preparation phase is often required.

Evaluation
Before deploying a model it is critical to thoroughly evaluate it and review the steps executed to construct the model to be certain it meets the business objectives and doesn't make one of the Top Errors, such as employing information leaked from the future.

Deployment
The end of a data mining project is not the creation of a model. The knowledge gained must be organized and presented in a way that management can understand. Deployment can be as simple as generating a report or as complex as implementing an interactive tool useable by front-line workers.

Conclusion
Data mining and resulting knowledge management are powerful tools in the marketer's arsenal. The thought of manipulating large data sets can be daunting to some; yet, for those with vision, exploiting data and turning it into usable and actionable knowledge to inform and drive marketing strategy can provide significant impact on business profitability.

ELDER RESEARCH
DATA SCIENCE · AI · MACHINE LEARNING

Author
Michael Lieberman

Founder, Multivariate Solutions

Guest author Michael D. Lieberman has more than 30 years of experience as a researcher and statistician in the marketing and advertising research fields. He founded Multivariate Solutions in 1998 and now works with an international clientele including advertising firms, political strategy groups, and full service market research companies. Mr. Lieberman is the author of more than 90 professional articles and three marketing research books. He teaches at New York University and the University of Georgia as an adjunct professor of statistics and market research. He holds a B.A. degree in Mathematics for Rutgers University, an M.A. degree in International Affairs from George Washington University, and an M.S. degree in Statistics from Rutgers University.