Data mining and the process of taking decisions in e-business

Ana Maria Mihaela Tudorache
Romanian-American University

ABSTRACT

Data mining software allows users to analyze large databases to solve business decision problems. Data mining is, in some ways, an extension of statistics, with a few artificial intelligence and machine learning twists thrown in. Like statistics, data mining is not a business solution, it is just a technology. For example, consider a catalog retailer who needs to decide who should receive information about a new product. The information operated on by the data mining process is contained in a historical database of previous interactions with customers and the features associated with the customers, such as age, zip code, their responses. The data mining software would use this historical information to build a model of customer behavior that could be used to predict which customers would be likely to respond to the new product. By using this information a marketing manager can select only the customers who are most likely to respond. The operational business software can then feed the results of the decision to the appropriate touch point systems (call centers, direct mail, web servers, email systems, etc.) so that the right customers receive the right offers.

Keywords: data mining, business decisions, data analysis, cluster analysis, decision strategy

1. INTRODUCTION

Data mining identifies trends within data that go beyond simple analysis. Through the use of sophisticated algorithms, non-statistician users have the opportunity to identify key attributes of business processes and target opportunities. However, abdicating control of this process from the statistician to the machine may result in false-positives or no useful results at all.

Although data mining is a relatively new term, the technology is not. For many years, businesses have used powerful computers to sift through volumes of data such as supermarket scanner data to produce market research reports (although reporting is not considered to be data mining). Continuous innovations in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy and usefulness of data analysis.

The term data mining is often used to apply to the two separate processes of knowledge discovery and prediction. Knowledge discovery provides explicit information that has a readable form and can be understood by a user. Forecasting, or predictive modeling provides predictions of future events and may be transparent and readable in some approaches (e.g., rule-based systems) and opaque in others such as neural networks. Moreover, some data-mining systems such as neural networks are inherently geared towards prediction and pattern recognition, rather than knowledge discovery.

Metadata, or data about a given data set, are often expressed in a condensed data-minable format, or one that facilitates the practice of data mining. Common examples include executive summaries and scientific abstracts.

Data mining relies on the use of real world data. This data is extremely vulnerable to collinearity precisely because data from the real world may have unknown interrelations. An unavoidable weakness of data mining is that the critical data that may expose any relationship might have never been observed. Alternative approaches using an experiment-based approach such as Choice Modelling for human-generated data may be used. Inherent correlations are either controlled for or removed altogether through the construction of an experimental design.

2. THE ALGORITHM USED IN DATA MINING

Data Mining is an analytic process designed to explore data (usually large amounts of data - typically business or market related) in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The ultimate goal of data mining is prediction - and predictive data mining is the most common type of data mining and one that has the most direct business applications. The process of data mining consists of three stages: (1) the initial exploration, (2) model building or pattern identification with validation/verification, and (3) deployment (i.e., the application of the model to new data in order to generate predictions).

Stage 1: Exploration. This stage usually starts with data preparation which may involve cleaning data, data transformations, selecting subsets of records and - in case of data sets with large numbers of variables ("fields") - performing some preliminary feature selection operations to bring the number of variables to a manageable range (depending on the statistical methods which are being considered). Then, depending on the nature of the analytic problem, this first stage of the process of data mining may involve anywhere between a simple choice of straightforward predictors for a regression model, to elaborate exploratory analyses using a wide variety of graphical and statistical methods (see Exploratory Data Analysis (EDA)) in order to identify the most relevant variables and determine the complexity and/or the general nature of models that can be taken into account in the next stage.

Stage 2: Model building and validation. This stage involves considering various models and choosing the best one based on their predictive performance (i.e., explaining the variability in question and producing stable results across samples). This may sound like a simple operation, but in fact, it sometimes involves a very elaborate process. There are a variety of techniques developed to achieve that goal - many of which are based on so-called "competitive evaluation of models," that is, applying different models to the same data set and then comparing their performance to choose the best. These techniques - which are often considered the core of predictive data mining - include: Bagging (Voting, Averaging), Boosting, Stacking (Stacked Generalizations), and Meta-Learning.

Stage 3: Deployment. That final stage involves using the model selected as best in the previous stage and applying it to new data in order to generate predictions or estimates of the expected outcome.

The concept of Data Mining is becoming increasingly popular as a business information management tool where it is expected to reveal knowledge structures that can guide decisions in conditions of limited certainty. Recently, there has been increased interest in developing new analytic techniques specifically designed to address the issues relevant to business Data Mining (e.g., Classification Trees), but Data Mining is still based on the conceptual principles of statistics including the traditional Exploratory Data Analysis (EDA) and modeling and it shares with them both some components of its general approaches and specific techniques.

3. DATA MINING IN BUSINESS

Data mining in customer relationship management applications can contribute significantly to the bottom line. Rather than contacting a prospect or customer through a call center or sending mail, only prospects that are predicted to have a high likelihood of responding to an offer are contacted. More sophisticated methods may be used to optimize across campaigns so that we can predict which channel and which offer an individual is most likely to respond to - across all potential offers. Finally, in cases where many people will take an action without an offer, uplift modeling can be used to determine which people will have the greatest increase in responding if given an offer. Data clustering can also be used to automatically discover the segments or groups within a customer data set.

Businesses employing data mining quickly see a return on investment, but also they recognize that the number of predictive models can quickly become very large. Rather than one model to predict which customers will churn, a business could build a separate model for each region and customer type.

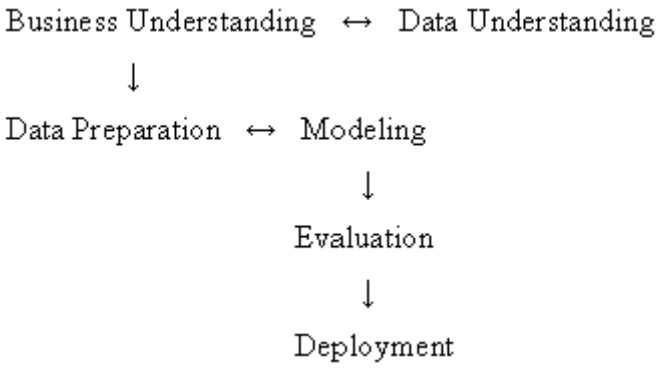
Then instead of sending an offer to all people that are likely to churn, it may only want to send offers to customers that will likely take to offer. And finally, it may also want to determine which customers are going to be profitable over a window of time and only send the offers to those that are likely to be profitable. In order to maintain this quantity of models, they need to manage model versions and move to automated data mining.

Data mining can also be helpful to human-resources departments in identifying the characteristics of their most successful employees. Information obtained, such as universities attended by highly successful employees, can help HR focus recruiting efforts accordingly. Additionally, Strategic Enterprise Management applications help a company translate corporate-level goals, such as profit and margin share targets, into operational decisions, such as production plans and workforce levels.

4. A MODEL FOR DATA MINING

In the business environment, complex data mining projects may require the coordinate efforts of various experts, stakeholders, or departments throughout an entire organization. In the data mining literature, various "general frameworks" have been proposed to serve as blueprints for how to organize the process of gathering data, analyzing data, disseminating results, implementing results, and monitoring improvements.

One such model, CRISP (Cross-Industry Standard Process for data mining) was proposed in the mid-1990s by a European consortium of companies to serve as a non-proprietary standard process model for data mining. This general approach postulates the following (perhaps not particularly controversial) general sequence of steps for data mining projects:



5. PREDICTIVE DATA MINING

The term Predictive Data Mining is usually applied to identify data mining projects with the goal to identify a statistical or neural network model or set of models that can be used to predict some response of interest. For example, a credit card company may want to engage in predictive data mining, to derive a (trained) model or set of models (e.g., neural networks, meta-learner) that can quickly identify transactions which have a high probability of being fraudulent. Other types of data mining projects may be more exploratory in nature (e.g., to identify cluster or segments of customers), in which case drill-down descriptive and exploratory methods would be applied. Data reduction is another possible objective for data mining (e.g., to aggregate or amalgamate the information in very large data sets into useful and manageable chunks).

6. CONCLUSIONS

The purpose of data visualization is to give the user an understanding of what is going on. Since data mining usually involves extracting "hidden" information from a database, this understanding process can get somewhat complicated. Because the user does not know beforehand what the data mining process has discovered, it is a much bigger leap to take the output of the system and translate it into an actionable solution to a business problem.

Data mining is the process of sorting through large amounts of data and picking out relevant information. It is usually used by business intelligence organizations, and financial analysts, but is increasingly being

used in the sciences to extract information from the enormous data sets generated by modern experimental and observational methods. It has been described as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data"and "the science of extracting useful information from large data sets or databases." Data mining in relation to enterprise resource planning is the statistical and logical analysis of large sets of transaction data, looking for patterns that can aid decision making.

REFERENCES

[1] www.wikipedia.com
[2]http://www.statsoft.com/textbook/stdatmin.html
[3] http://www.thearling.com/
[4] Berry, M., J., A., & Linoff, G., S., (2000). **Mastering data mining**. New York: Wiley.
[5] Edelstein, H., A. (1999). **Introduction to data mining and knowledge discovery** (3rd ed). Potomac, MD: Two Crows Corp.
[6] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). **Advances in knowledge discovery & data mining** Cambridge, MA: MIT Press.
[7] Han, J., Kamber, M. (2000). **Data mining: Concepts and Techniques** New York: Morgan-Kaufman.
[8] Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). **The elements of statistical learning: Data mining, inference, and prediction**. New York: Springer.