International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-4, Issue-4)

Research Article | **April 2015**

Data Mining: A Tool for Banking Industry

M. P. Thapliyal
Department of Computer Science, HNB Garhwal University Srinagar
Garhwal, Uttarakhand, India

Abstract:

Data mining allows to extract diamonds of knowledge from the historical data, and predict outcomes of future situations. It helps optimize business decisions, identifying `loyal' customers, detecting fraudulent transactions, .increase the value of each customer and communication, and improve customer satisfaction. The financial institutions, such as banks and investment companies, are the pioneers in taking advantage of data mining.

Key word: Data mining, Banks.

I. INTRODUCTION

The growth of information resources along with the accelerating rate of technological change has produced huge amounts of information that often exceed the ability of managers and employees to assimilate and use it productively. Data must be categorized in some manner if it is to be accessed, re-used, organized, or synthesized to build a picture of the company's competitive environment or solve a specific business problem (Pearlson, 2001, p.196). In recent years, the need to extract knowledge automatically from very large databases has grown. In response, the closely related fields of knowledge discovery in databases (KDD) and data mining have developed processes and algorithms that attempt to *intelligently extract interesting and useful information from vast amounts of raw data.* (Fayyad, Grinstein & Wierse, 2002). For example, Wal-Mart has one of the world's largest databases of customer transactions, with over 20 million transactions being handled per day [Babcock, 1994]. Wal-Mart just wants to know to whom they should mail their next advertising circular; they are not trying to prove a hypothesis. Intelligent data mining, according to Edelstein (1996), discovers information within data warehouses that queries and reports cannot reveal. That is why a data-mining project requires the best selection of hardware, software and human resources. The market offers software tools that require expertise and high level of knowledge. Usually, because of the large amount of data, data mining tools run in sophisticated computers with high-speed processors and large storage capabilities. However, the technology is not everything and the role of the analysts who work with data mining is really important. For instance, database administrators should understand data mining requirements and try to design databases the most accessible for mining as possible. One of the biggest problems in data mining projects is to prepare data for running the algorithms. Data is not ready to mine. Preparing data could be easier if databases were previously designed taking in account mining purposes. There is a big variety of data mining software and methodologies in the market. Hence, it is important to choose the right methodology and the right tool. Not only identifying where and how to run data-mining models is important and also interpreting the results that will ultimately help making strategic decisions in organizations. Specifically, because data mining attempts to discover patterns, trends, and correlations hidden in the data, those results can give a company a strategic business advantage (O'Brien, 2001, p.363). Data mining can help managers to make decisions and apply more effective strategies in the organizations.

II. WHAT IS DATA MINING?

"Data Mining is a non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns from data" (Srikant & Agrawal, 1996).
That is a very brief definition that implies the purposes of doing mining and extracting new information from data:

- It is "valid" because it looks to be well grounded in logic patterns. In order to be valid the processes can be automatic or semiautomatic and there are many tools that are used to make the used algorithms and the resulting patterns as valid as possible;
- It is a "novel" because data mining a lot of research needs to be done yet;
- It is "potentially useful" because the results can be used in the decision making process of any organization, such as health, education, marketing, etc.
- "Understandable" patterns because the results should be capable of being understood or interpreted by users from different backgrounds and not only for researchers.
- "Patterns" from previous data because a perceptual structure has been created as a model that can be applied to new data.
- "Data" refers to the digitalized information in databases first and data warehouses later that can be accessed by data mining tools.

According to a Data Mining Glossary from Two Crows, Data Mining is an information extraction activity whose goal is to discover hidden facts contained in databases. Using a combination of machine learning, statistical analysis, modeling

techniques and database technology, data mining finds patterns and subtle relationships in data and infers rules that allow the prediction of future results (2002). The main points of this definition are the term "facts" because data mining works with "real" data and "hidden facts" because data mining shows the behavior and performance that is not easily discovered. Commercial databases are growing at unprecedented rates. In the evolution from business data to business information, each new step has built upon the previous one. According to Kurt Thearling (~1996), these are the steps in the evolution of data mining:

Table 1. Steps in the Evolution of Data Mining.

Evolutionary Step	**Business Question**	**Enabling Technologies**	**Product Providers**	**Characteristics**
Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks	IBM, CDC	Retrospective, static data delivery
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases (RDBMS), Structured Query Language (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, dynamic data delivery at record level
Data Warehousing & Decision Support (1990s)	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses	Pilot, Comshare, Arbor, Cognos, Microstrategy	Retrospective, dynamic data delivery at multiple levels
Data Mining (Emerging Today-1996)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive databases	Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry)	Prospective, proactive information delivery

Although in table 1 Thearling indicates that Data mining was emerging today, in 1996, it is still emerging today in 2008. The evolution of data mining is still going on, but what happened in the past has created the basis for organizational mining.

What data mining can do?

Data mining can do basically six tasks. The first three are all examples of **directed data mining**, where the goal is to use the available data to build a model that describes one particular variable of interest in terms of the rest of the available data. For example, analyzing bankruptcy, the target variable is a binary variable that describes if a client was declared on bankruptcy or not. In directed data mining, we try to find patterns that will make that variable have that value: 0 or 1. The next three tasks are examples of **undirected data mining** where no variable is singled out as a target and the goal is to establish some relationship among all the variables. In the previous bankruptcy example, data mining tries to identify patters of the behavior of customers without indicating that those customers are in bankruptcy or not.These are the types of information can be obtained by data mining, summarizing from two different sources: Turban & Aronson (2001) and Berry & Linoff (1999):

- **Classification**: consists of examining the features of a newly presented object and assigning to it a predefined class or group. The task is to build a model that can be applied to unclassified data in order to classify it using the defined characteristics of a certain group (e.g., classifying credit applicants as low, medium of high risk).
- **Estimation**: Given some input data, we use estimation to come up with a value for some unknown continuous variable such as income, height, or credit card balance. (e.g. a bank trying to decide to whom they should offer a home equity loan based on the probability that the person will respond positively to an offer).
- **Prediction**: records are classified according to some predicted future behavior or estimated future values based on patterns within large sets of data (e.g. demand forecasting or predicting which customers will leave within the next six months).
- **Association**: identifies relationships between events that occur at one time, determines which things go together (e.g., the contents of a shopping basket: beer with cigarettes)
- **Clustering**: identifies groups of items that share a particular characteristic segmenting a diverse group into a number of more similar subgroups or clusters. Clustering differs from classification in that it does not rely on predefined classes or characteristics for each group. (e.g. as a first step in a market segmentation effort, we can divide the customer base into clusters of people with similar buying habits, and ten ask what kind of promotion works best for each cluster or group).

- **Description and Visualization**: the purpose is to describe what is going on in a complicated database in a way that increases our understanding of the people, products, or processes that produced the data in the firs place. A good description suggests where to start looking for an explanation. (e.g., repeat visits to a supermarket).
- **Data mining techniques**
  The most commonly used techniques in data mining are (Thearling, 1995):
- **Artificial neural networks**: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- **Decision trees**: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).
- **Genetic algorithms**: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.
- **Nearest neighbor method**: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k ³ 1). Sometimes called the k-nearest neighbor technique.
- **Rule induction**: The extraction of useful if-then rules from data based on statistical significance.
- **Data mining tools**

Data mining software is based in mathematical algorithms and statistics. Developers have been working in data mining software tools to make them more user friendly and the different products available on the market have advantages and disadvantages basically related to their interface and available techniques. Today, the market offers a variety of products. According to Kdnuggets poll done between June 27 ad July 17, 2000, with the question: ***Which tool do you plan to try or by next?*** [242 votes total] the most five preferred tools are: Clementine (32), Megaputer (32), SAS EM (Enterprise Miner) (31), GainSmarts (28), and EasyMiner(15). Gartner places six tool suites in the generic market: Angoss Knowledge Suite, IBM's Intelligent Miner for Data, Oracle's Darwin, SAS' EnterpriseMiner, SGI's MineSet, and SPSS' Clementine. Some of them are very expensive - data mining products range in price from $1,000 on a PC to more than $100,000 for algorithms that run on mainframes (Foley & Russell, 1998) - and not very user friendly. However, the best product for a company would be the one that is easy to run with current corporate data and easy to interpret for the managerial purposes of the company.

III. DM IN THE BANKING INDUSTRY

The banking industry across the world has under gone tremendous changes in the way the business is conducted. With the recent implementation, greater acceptance and usage of 'electronic' banking, the capturing of transactional data has become easier and, simultaneously, the volume of such data has grown considerably. It is beyond human capability to analyze this huge amount of raw data and to effectively transform the data into useful knowledge for the organization. The enormous amount of data that banks have been collecting over the years can greatly influence the success of data mining efforts. By using data mining to analyze patterns and trends, bank executives can predict, with increased accuracy, how customers will react to adjustments in interest rates, which customers will be likely to accept new product offers, which customers will be at a higher risk for defaulting on a loan, and how to make customer relationships more profitable. The banking industry is widely recognizing the importance of the information it has about t its customers. Undoubtedly, it has among the richest and largest pool of customer information, covering customer demographics, transactional data, credit cards usage pattern, and so on. As banking is in the service industry, the task of maintaining a strong and effective CRM is a critical issue. To do this, banks need to invest their resources to better understand their existing and prospective customers. By using suitable data mining tools, banks can subsequently offer 'tailor-made' products and services to those customers. There are numerous areas in which data mining can be used in the banking industry, which include customer segmentation and profitability, credit scoring and approval, predicting payment default, marketing, detecting fraudulent transactions, cash management and forecasting operations, optimizing stock portfolios, and ranking investments. In addition, banks may use data mining to identify their most profitable credit card customers or high-risk loan applicants. There is, therefore, a need to build an analytical capability to address the above-stated issues and data mining attempts to provide the answer. Following are some examples of how the banking industry has been effectively utilizing data mining in these areas.

Marketing: Bank analysts can also analyze the past trends, determine the present demand and forecast the customer behavior of various products and services in order to grab more business opportunities and anticipate behavior patterns. Data mining technique also helps to identify profitable customers from non-profitable ones. **'Cross-selling'** is another marketing area where data mining can be extensively used. Here, a service provider makes it attractive for a customer to buy additional products or services with the same business. The more products and services a bank can provide for customers, the more likely the bank is to retain those customers.

Risk Management: Data mining is widely used for risk management in the banking industry. Bank executives need to know whether the customers they are dealing with are reliable or not. Offering new customers credit cards, extending existing customers lines of credit, and approving loans can be risky decisions for banks if they do not know anything about their customers. Data mining, however, can be used to reduce the risk of banks that issue credit cards by determining those customers who are likely to default on their accounts. An example was reported in the press of a bank discovering that cardholders who withdrew money at casinos had higher rates of delinquency and bankruptcy. It is a

common practice on the part of banks to analyze customers' transaction behaviors in their deposit accounts to determine their probability of default in their loan accounts. Credit scoring, in fact, was one of the earliest financial risk management tools developed. Credit scoring can be valuable to lenders in the banking industry when making lending decisions. Lenders would not have expanded the number of loans they give out without having an accurate, objective, and controllable risk assessment tool. The examples of both a 'good' and 'bad' loan applicant's histories can be used to develop a profile for a good and bad 'new' loan applicant. Data mining can also derive the credit behavior of individual borrowers with installment, mortgage and credit card loans, using characteristics such as credit history, length of employment and length of residency. A score is thus produced that allows a lender to evaluate the customer and decide whether the person is a good candidate for a loan, or if there is a high risk of default. Customers who have been with the bank for longer periods of time, remained in good standing, and have higher salaries/wages, are more likely to receive a loan than a new customer who has no history with the bank, or who earns low salaries/wages. By knowing what the chances of default are for a customer, the bank is in a better position to reduce the risks.

Fraud Detection: Another popular area where data mining can be used in the banking industry is in fraud detection. Being able to detect fraudulent actions is an increasing concern for many businesses; and with the help of data mining more fraudulent actions are being detected and reported. Two different approaches have been developed by financial institutions to detect fraud patterns. In the first approach, a bank taps the data warehouse of a third party (potentially containing transaction information from many companies) and uses data mining programs to identify fraud patterns. The bank can then cross-reference those patterns with its own database for signs of internal trouble. In the second approach, fraud pattern identification is based strictly on the bank's own internal information. Most of the banks are using a 'hybrid' approach. One system that has been successful in detecting fraud is Falcon's 'fraud assessment system'. It is used by nine of the top ten credit card issuing banks, where it examines the transactions of 80 per cent of cards held in the US. Mellon Bank also uses data mining for fraud detection and is able to better protect itself and its customers' funds from potential credit card fraud.

Customer Acquisition and Retention: Not only can data mining help the banking industry to gain new customers, it can also help retain existing customers. Customer acquisition and retention are very important concerns for any industry, especially the banking industry. Today, customers have so many opinions with regard to where they can choose to do their business. Executives in the banking industry, therefore, must be aware that if they are not giving each customer their full attention, the customer can simply find another bank that will. Data mining can also help in targeting 'new' customers for products and services and in discovering a customer's previous purchasing patterns so that the bank will be able to retain existing customers by offering incentives that are individually tailored to each customer's needs.

IV. CONCLUSION

Data mining is a tool used to extract important information from existing data and enable better decision-making throughout the banking industries. They use data warehousing to combine various data from databases into an acceptable format so that the data can be mined. The data is then analyzed and the information that is captured is used throughout the organization to support decision-making. It is universally accepted that many industries (including banking, retail and telecom) are using data mining effectively. Undoubtedly, data mining has many uses in industries. Its practical applications in such areas as analyzing medical outcomes, detecting credit card fraud, predicting customer purchase behavior, predicting the personal interests of Web users, optimizing manufacturing processes etc. have been very successful. It has also led to a set of fascinating scientific questions about how computers might automatically learn from past experience. The retail industry is also realizing that data mining could give them a competitive advantage. A majority of the banks in developing countries (particularly in the public sector) are not usually known to exploit their information 'asset' for deriving business value through data mining and gain competitive advantage. But with progressive liberalization of rules on entry for private and foreign multinational banks, under the GATS framework of WTO, competitive pressure on domestic banks is increasing. Thus, customer retention and acquisition will be an important determinant of the banks' bottom lines. Those banks and retailers that have realized the utility of data mining and are in the process of building a data mining environment for their decision-making process will reap immense benefit and derive considerable competitive advantage to withstand competition in future.

REFERENCES

[1] Babcock C. (1994) *Parallel Processing Mines Retail Data*. Computer World.
[2] Berry M.J.A. & Linoff G. (1999) *Mastering Data Mining: The Art and Science of Customer Relationship Management*. John Wiley and Sons, Inc.
[3] Davenport T.H. & Prusak L. (2000). *Working Knowledge: How organizations manage what they know.* Boston, Massachusetts; Harvard Business School Press.
[4] Deering B.J. (2002) Chapter 11: KM for competitive advantage: mining diverse sources for marketing intelligence. *Knowledge Management Strategy and Technology*. Bellaver R.F. & Lusa J.M. Editors. Artech House.
[5] Edelstein, H. (1996, Jan.8) . *Mining Data Warehouses" Information Week.*
[6] Fayyad U., Grinstein G.E. & Wierse A. (2002). *Information Visualization in Data Mining and Knowledge Discovery*. Morgan Kaufmann Publishers. Academic Press.
[7] Foley J. & Russell J.D. (1998) *Mining your Own Business*. Information Week.com March 6, 1998. Retrieved May 4, 2002 from: http://www.informationweek.com/673/73iudat.htm

[8] Gartner Group Sized up Workbench Market: Data Mining News. Retrieved May 5, 2002 from: http://www.idagroup.com/v3n0101.htm

[9] Kdnuggets poll (June 27 - July 17, 2000) with the question: ***Which tool do you plan to try or by next?***. Retrieved April 24, 2002 from http://www.kdnuggets.com/polls/next_dm_tool-2000-07-17.htm

[10] Laudon K.C. & Laudon J.P. (2000). *Management Information Systems: Organization and Technology in the Networked Enterprise*. Sixth Edition; Prentice Hall.

[11] Meltzer M. (2001) E-Mining Myth & Magic: Using Data Mining Successfully. Retrieved May 1, 2002, from http://www.crm-forum.com/library/art/art-048/art-048.htm.

[12] O'Brien J.A. (2001) *Introduction to Information Systems: Essentials for the Internetworked E-Business Enterprise*. Tenth Edition; McGraw-Hill Irwin.

[13] Pearlson, K.E. *Managing and Using Information Systems: A Strategic Approach*. New York, Wiley, 2001: John Wiley & Sons, Inc.

[14] Srikant, R and Agrawal, R (1996) "*Mining sequential patterns : Generalizations and performance improvements*", Proc. of the 5th International Conf. on Extending Database Technology, France (March).

[15] Thearling K. Retrieved May 6, 2002 from: http://www3.shore.net/~kht/text/dmwhite/dmwhite.htm

[16] Turban E. & Aronson J.E. (2001). *Decision Support Systems and Intelligent Systems.* Sixth Ed. New Jersey; Prentice Hall.

[17] Two Crows: *Data Mining Glossary*. Retrieved April 28, 2002 from: http://www.twocrows.com/glossary.htm.

Author's Biography

Dr.M.P.Thapliyal is working as Reader in the Department of Computer Science, HNB Garhwal University Srinagar(Garhwal) Uttarakhand.His major research interests are in the field of Software Engineering ,Human-Computer Interaction, educational research and the role of Information and communication technologies for improving teaching and learning process. Author has been involved in the design and experimentation of educational software . He has published more than 30 papers in International and National Journals/ Conferences. Author is also Reviewer of full papers of various International Conferences and visited China, USA, Itlay, France and Singapore as an expert to present his Talks/papers. Author is Editorial Board Member of various International Journals like EJEL etc and Experts of various Indian Universities. Author has 15 years of experience in the field of software development, Human-Computer Interaction and Information System Design. He is member of SIGAPP (Special Interest Group on Applied Computing), Society for Information Science and Senior Life Member of Computer Society of India..Author is also Indian Representative of the International Federation for Information Processing (IFIP),TC-13 (Human-Computer Interaction) Group. Author has written books on C Programming and System Analysis and Design.

