Chapter II

Data Mining Defined

Over the years, the term data mining has been connected to various types of analytical approaches. In fact, just a few years ago, let's say prior to 1995, many individuals in the software industry and business users as well, often referred to OLAP as a main component of data mining technology. More recently however, this term has taken on a new meaning and one which will most likely prevail for years to come. As we mentioned in the previous chapter, data mining technology encompasses such methodologies as clustering, classification and segmentation, association, neural networks and regression as the main players in this space. Other analytical processes which are related to mining, as defined in this work, include such methodologies as Linear Programming, Monte Carlo analysis and Bayesian methodologies. In fact, depending on who you ask, these techniques may actually be considered part of the data mining spectrum since they are grounded in mathematical techniques applied to historical data. The focus of this work however, revolves around the former more core approaches.

Regardless of the type of methodology, data mining has taken its roots from traditional analytical techniques. Enhancements in computer processing, (e.g., speed and processing power) has enabled a wider diffusion of more complex techniques to become more automated and user friendly and have evolved to the state of our current data mining.

THE ROOTS OF DATA MINING

The term *data mining* has become a loosely used reference to some well established analytical methodologies used to validate business and economic theory. The purpose of this chapter is to remind the user population that data mining is not some "black box" computational magic or crystal ball that provides flawless insights for decision makers but is an approach that is

grounded in traditional analysis that can be used to gain a greater understanding of business processes.

Data mining incorporates analytical procedures grounded in traditional statistics, mathematics and business and economic theory. Much of mining methodology takes its roots from what is referred to as econometrics. Econometrics is an analytical methodology that involves the application of mathematics and quantitative methods to historical data in conjunction with traditional statistics with the focus of testing an established economic or business theory (for more details see Gujarati, 1988). The following phrase puts this issue in the proper context:

> "In reviewing the development of concepts for the statistical analysis of econometric models, it is very easy to forget that in the opening decades of this century a major issue was whether a statistical approach is appropriate for the analysis of economic phenomena. Fortunately, the recognition of the scientific value of sophisticated statistical methods in economics and business has buried this issue. To use statistics in a sophisticated way required much research on basic concepts of econometric modeling that we take for granted today."[1]

The evolution of modeling and mining dates back to the use of calculators in the processing of data to identify statistical relationships between variables. The estimation of such rudimentary measures as means, medians, max & mins, variances and standard deviations provided the building blocks to today's heavy duty computer processing of large volumes of data with mathematical equations, computer algorithms and corresponding statistical measures.

A Closer Look at the Mining Process (The Traditional Method)

The traditional methodology referred to as econometrics above involves the following procedure:

1) Gather data that includes information relevant to the theory to be tested.
2) Specify quantitative (mathematical functional forms) which depict the relationships between explanatory and target variables. (e.g., A single equation or system of equations).
3) Apply corresponding statistical tests to measure the robustness or correctness of the quantitative model in supporting the economic or business theory.

Related Content

A Dual-Database Trusted Broker System for Resolving, Protecting, and Utilizing Multi-Sourced Data
Neal Gibsonand Greg Holland (2014). *Information Quality and Governance for Business Intelligence (pp. 352-362).*
www.irma-international.org/chapter/a-dual-database-trusted-broker-system-for-resolving-protecting-and-utilizing-multi-sourced-data/96159

Machine Learning Approach and Model Performance Evaluation for Tele-Marketing Success Classification
Fatma Önay Koçoluand akir Esnaf (2022). *International Journal of Business Analytics (pp. 1-18).*
www.irma-international.org/article/machine-learning-approach-and-model-performance-evaluation-for-tele-marketing-success-classification/298014

Business Intelligence and Agile Methodology for Risk Management in Knowledge-Based Organizations
Muhammad Mazen Almustafaand Dania Alkhaldi (2012). *Business Intelligence and Agile Methodologies for Knowledge-Based Organizations: Cross-Disciplinary Applications (pp. 240-267).*
www.irma-international.org/chapter/business-intelligence-agile-methodology-risk/58574

Data to Analytics to Insight: Role of rtDashboard at St. Joseph Mercy Health
Mohan Tanniru, Matt Nawrocki, David Bobrykand Anupam A. Sule (2020). *Theory and Practice of Business Intelligence in Healthcare (pp. 133-150).*
www.irma-international.org/chapter/data-to-analytics-to-insight/243353

Time Series Data Mining: A Retail Application
Daniel Hebert, Billie Anderson, Alan Olinskyand J. Michael Hardin (2014). *International Journal of Business Analytics (pp. 51-68).*
www.irma-international.org/article/time-series-data-mining/119497