CRITICAL SUCCESS FACTORS IN DATA MINING PROJECTS

Jaesung Sim, B.P.A., M.P.A., M.S.

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

August 2003

APPROVED:

C. Steve Guynes, Major Professor
Chang E. Koh, Co-Major Professor
Alan H. Kvanli, Minor Professor
John C. Windsor, Committee Member and Chair of the Business Computer Information Systems Department
J. Wayne Spence, Coordinator of the PhD Program in Business Computer Information Systems Department
Jared E. Hazleton, Dean of the College of Business Administration
C. Neal Tate, Dean of the Robert B. Toulouse School of Graduate Studies

Sim, Jaesung, Critical success factors in data mining projects. Doctor of Philosophy (Business Computer Information Systems), August 2003, 122 pages, 37 tables, 8 illustrations, bibliography, 65 titles.

The increasing awareness of data mining technology, along with the attendant increase in the capturing, warehousing, and utilization of historical data to support evidence-based decision making, is leading many organizations to recognize that the effective use of data is the key element in the next generation of client-server enterprise information technology.

The concept of data mining is gaining acceptance in business as a means of seeking higher profits and lower costs. To deploy data mining projects successfully, organizations need to know the key factors for successful data mining. Implementing emerging information systems (IS) can be risky if the critical success factors (CSFs) have been researched insufficiently or documented inadequately. While numerous studies have listed the advantages and described the data mining process, there is little research on the success factors of data mining.

This dissertation identifies CSFs in data mining projects. Chapter 1 introduces the history of the data mining process and states the problems, purposes, and significances of this dissertation. Chapter 2 reviews the literature, discusses general concepts of data mining and data mining project contexts, and reviews general concepts of CSF methodologies. It also describes the identification process for the various CSFs used to develop the research framework. Chapter 3 describes the research framework and methodology, detailing how the CSFs were identified and validated from more than 1,300 articles published on data mining and related topics.

The validated CSFs, organized into a research framework using 7 factors, generate the research questions and hypotheses. Chapter 4 presents analysis and results, along with the chain of evidence for each research question, the quantitative instrument and survey results. In addition, it discusses how the data were collected and analyzed to answer the research questions. Chapter 5 concludes with a summary of the findings, describing assumptions and limitations and suggesting future research.

ACKNOWLEDGEMENTS

The author wishes to thank the members of his dissertation committee (Steve Guynes, Chang Koh, Alan Kvanli, and John Windsor) for all their guidance. Special thanks to go to his family for all their support and patience in this long process.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

Overview of Data Mining Issue

Over the past decade, an increasing number of organizations started routinely capturing a large volume of data describing their operations, products, and customers. In addition, the wide use of supermarket point-of-sale scanners, automatic teller machines, credit and debit cards, pay-per-view television, home shopping, electronic funds transfer, automated order processing, electronic ticketing and the like makes it possible to collect and process a massive amount of data at an unprecedented rate. Digitized information is easy to capture and inexpensive to store with the development of the latest data processing and storage technologies. People store data because doing so is easy and convenient, and because they think some valuable assets are implicitly coded within the data. Raw data, however, is rarely of direct benefit. Its true value is predicated on the ability to extract information useful for decision support or exploration and understanding of the phenomena governing the data source.

Organizational data are still largely unrecognized, inaccessible, and underutilized. The traditional method of turning data into knowledge relies on manual analysis and interpretation (Fayyad and Uthurusamy, 1996, Fayyad, Piatetsky-Shapiro, and Smyth, 1996). Our ability to analyze and understand massive datasets lags far behind our ability to gather and store the data.

Businesses in today's environment increasingly focus on gaining competitive advantages. Organizations have recognized that the effective use of data is the key element in the next generation of client-server enterprise information technology. The

technology for accessing, updating, organizing, and managing a large volume of data has matured over the past twenty years. However, many organizations have had difficulties processing these massive ores into valuable information until the boom of data mining techniques, which are the next logical steps to the discovery of usable intelligence. There is an increasing awareness of data mining technology within many organizations and an attendant increase in capturing, warehousing, and utilizing historical data to support evidence-based decision making (Mitchell, 1999).

The field of data mining addresses the question of how best to use this vast amount of historical data to discover general regularities and improve the process of making decisions.

Purpose, Problems, and Significance

The concept of data mining is gaining acceptance in business as a means of seeking higher profits and lower costs. To deploy data mining projects successfully, organizations need to know the key factors for successful data mining. The key factors can be critical success factors (CSFs), which is "the limited number of areas in which results, if they are satisfactory, will ensure successful competitive performance for the organization" (Rockart, 1979). CSFs commonly exist for industries, corporations, departments within corporations, or for individual managers fulfilling the corporate strategy. Implementing emerging IS can be risky if the critical success factors (CSFs) have been researched insufficiently or documented inadequately.

CSFs are critical to an organization's current operating activities and to its future success (Guynes and Vanecek, 1996). Extending CSFs under operating activities, we can state that there are certain critical factors in IS development projects (including data

mining projects) that if not met will lead to the failure of those IS development projects. This extension of the CSF approach has double significance: "If these factors are crucial for the success of the system, then their failure would lead to the failure of the information system" (Zahedi, 1987).

While numerous studies have listed the advantages and described the data mining process, there is little research on the success factors of data mining. First, the field of data mining is new. Second, the benefits of a data mining project are typically difficult to quantify in advance, given their exploratory nature, whereas most business cases suffer from real costs matched by intangible benefits. In addition, details on successful projects are hard to come by as few are willing to share their secrets.

Organization of the Dissertation

Chapter 2 discusses general concepts of data mining and data mining project contexts and reviews general concepts of CSF methodologies. It also describes the identification process for the various data mining CSFs used to develop the research framework.

Chapter 3 describes the validation of the CSFs identified from over 1,300 articles published on data mining and related topics. The validated CSFs are organized into a research framework using 7 factors, from which the research questions and hypotheses were generated.

Chapter 4 presents the chain of evidence for each research question, the quantitative instrument, and the survey results. In addition, it discusses how the data were collected and analyzed to answer the research questions.

Chapter 5 concludes with a summary of the findings, describing assumptions and limitations and suggesting future research.

CHAPTER 2
LITERATURE REVIEW

Finding critical success factors (CSFs) in data mining projects is very important, but there are no pre-defined CSFs in the literature of data mining projects. How, then, can we define CSFs for data mining projects? To address this problem, I will first define general concepts of data mining and data mining project contexts. I will then review general concepts of CSF methodologies, collecting CSFs from data mining contexts and reviewing them. Finally, I will use these collected CSFs to build a research framework.

Data Mining

This section defines data mining, reviews general concepts of data mining, and discusses other relevant topics related to data mining.

Definitions

Data mining is used today in a wide range of applications, from tracking down criminals to brokering information for supermarkets, from developing community knowledge for a business to cross-selling, routing warranty claims, holding on to good customers, and weeding out bad customers (Berry and Linoff, 1997). Some applications include marketing, financial investment, fraud detection, manufacturing and production, and network management (Brachman et al., 1996). Data mining is not limited to the business environment. Data mining is also useful for data analysis of sky survey cataloging, mapping the datasets of Venus, biosequence databases, and geosciences systems (Fayyad, Haussler, and Stolorz, 1996).

Data mining is the process of extracting previously unknown, valid and actionable information from large databases and then using the information to make crucial business decisions. In essence, data mining is distinguished by the fact that it is aimed at the discovery of information, without a previously formulated hypothesis (Cabena et al., 1997). The field of data mining addresses the question of how best to use the historical data to discover general regularities and improve the process of making decisions (Mitchell, 1999).

Purposes

Organizations use data mining because data mining delivers value to industry. Data mining increases customer profitability, reduces customer/product churn, reduces costs through target marketing, uncovers new markets, detects credit abuse and fraud, performs sales and trend analysis, and carries out inventory management and control.

The success of the data mining approach comes from recent advances in three different fields. In mathematics, new efficient and quick algorithms have been developed. In database technologies, new research has allowed the improvement of databases. And as for computers, new powerful architectures have made possible the elaboration of huge data volumes (Zanasi, 1998).

Data mining tools simplify analyses by employing filters based upon specific user-defined, qualifying criteria (such as a list of employees who have held specific job titles), percentile and rank filtering (such as the top 10% of their raw materials used). Users can specify information to be found regarding a particular business unit and compare it to that of multiple business units or to the company as a whole. Sometimes scanning all relevant data can help decision-makers extract similarities among events

and hence inspire hypotheses. Data mining tools often use artificial intelligence. If these tools work well, they can identify relevant data and patterns in the data, which should spur intuition or help test intuitive-based hypotheses. The patterns and rules might provide guidelines for decision-making, or they might identify the issues upon which the decision-maker should focus during the choice process (Sauter, 1999).

Introduction-based data mining software uses machine-learning algorithms to analyze records in a firm's internal and customer databases, discovering patterns, transactional relationships, and rules that can predict future trends and indicate competitive opportunities. Raw data thus transformed is maintained in a data warehouse, providing support for a variety of analytical tasks and competitive decisions. As a result, questions that traditionally required extensive trial-and-error queries or statistical segmenting can be answered automatically (Mena, 1996).

DSS can help decision-makers by prompting them to consider important issues, such as those associated with data mining tools. For example, one system used neural networks to analyze credit card data and provide hypotheses to decision-makers about credit card theft. The system returned with a unique insight; credit card thieves were charging low amounts to a card, such as $1 at a gas pump, to test the cards before using them for higher-cost purchases. This insight was complementary to those provided by humans, which tended to focus on large, unusual purchases (Sauter, 1999).

What is attracting many analysts to data mining is the relative ease with which new insights can be gained in comparison to traditional statistical approaches. Data mining is always a hypothesis-free approach, whereas most popular statistical techniques require the development of a hypothesis in advance. Data mining

algorithms typically produce a much wider set of data types and make fewer assumptions about their distribution – or no assumptions at all (Cabena et al., 1997).

Applications

Data mining, much like data warehousing, is driven by applications. The majority of these applications are aimed at understanding behavior; more specifically, comprehending customer behavior: how to acquire, retain, and increase the profitability and lifetime value of a customer.

Data mining applications include market management, risk management, fraud management, and other emerging application areas, such as text mining and web analytics. Market management includes database marketing, determining customer purchasing behavior patterns over time, and cross-selling campaigns. Risk management usually involves forecasting, customer retention, improved underwriting, quality control, and competitive analysis. Fraud management applications seek to detect fraud.

Other data mining applications include predicting customer purchase behavior, customer retention, and the quality of goods produced by a particular manufacturing line. All are applications for which data mining has been applied successfully and in which further research promises even more effective techniques (Mitchell, 1999).

Tasks

Tasks determine what we can expect from data mining. The outcomes of data mining are also referred to as the data mining tasks. These are the results that one can expect to see as a result of data mining. Data mining outcomes include classification, clustering, prediction, estimation, and affinity grouping (Thuraisingham, 1998). Peacock

(1998) proposed that five foundation-level analysis tasks are the "reasons why" of data mining: summarization, predictive modeling, clustering/segmentation, classification, and link analysis.

Brachman et al. (1996) distinguished two types of data mining methods: verification, in which the system is limited to verifying a user's hypothesis; and discovery, in which the system finds new patterns. Discovery includes prediction, through which the system finds patterns to help predict the future behavior of some entities; and description, through which the system finds patterns in order to present the patterns to users in an understandable form. Task examples of key predictive methods include regression and classification (learning a function that maps a new example into one of a set of discrete classes). Key description methods include clustering, summarization, visualization, and change and deviation detection.

Chung and Gray (1999) use the term *task* for showing examples of associations, sequences, classifications, clusters, and forecasting. Berry and Linoff (1997) used the term *task* in their book for classification, estimation, prediction, affinity grouping, clustering, and description.

However, Groth (2000) uses the term *strategies* in his book for classification, clustering, visualization, association, assortment optimization, prediction, and estimation.

Fayyad, Piatetsky-Shapiro, and Smyth (1996) suggested that more common *model functions* in current data mining practice include the following:

- Classification: maps (or classifies) a data item into one of several predefined categorical classes.
- Regression: maps a data item to a real-value prediction variable.

- Clustering: maps a data item into one of several categorical classes (or clusters) in which the classes must be determined from the data-unlike classification in which the classes are predefined. Clusters are defined by finding natural groupings of data items based on similarity metrics or probability density models.
- Summarization: provides a compact description for a subset of data. A simple example would be the mean and standard deviations for all fields. More sophisticated functions involve summary rules, multivariate visualization techniques, and functional relationships between variables. Summarization functions are often used in interactive exploratory data analysis and automated report generation.
- Dependency modeling: describes significant dependencies among variables. Dependency models exist at two levels: structured and quantitative. The structural level of the model specifies (often in graphical form) which variables are locally dependent; the quantitative level specifies the strengths of the dependencies using some numerical scale.
- Link analysis: determines relations between fields in the database (e.g., association rules to describe which items are commonly purchased with other items in grocery stores). The focus is on deriving multi-field correlations satisfying support and confidence thresholds.
- Sequence analysis: models sequential patterns (e.g., in data with time dependence, such as time-series analysis). The goal is to model the states of the process generating the sequence or to extract and report deviation and

trends over time. Data mining tasks determine what we can expect from data mining.

Techniques

As discussed in the previous section, there is no standard terminology for data mining techniques. The most common terms for this topic are *algorithm* and *technique*. One can argue that while *techniques* describe a broad class of procedures to carry out mining, *algorithms* go into more details. However, these two terms are both used for the same concept.

Groth, writing in 1998, used the term *modeling techniques* in his book, discussing "decision trees, genetic algorithm, neural nets, agent network technology, hybrid models, and statistics" as examples. Writing in 2000, Groth used the term *algorithm* when discussing "genetic algorithm, neural networks, Bayesian belief networks, statistics, advanced algorithm for association, and algorithms for assortment optimization."

Chung and Gray (1999) used the term *algorithm* with statistical analysis, neural networks, data visualization, and decision trees. Fayyad, Piatetsky-Shapiro, and Smyth (1996) used *search algorithm* as a specialized term.

Weiss and Indurkhya (1988) divide data mining algorithms into three groups: math-based methods, distance-based methods, and logic-based methods. For example, *linear discriminant analysis* is the most common math-based method, as well as the most common classification technique in general, while *tree and rule induction* is the most common logic-based method. *Neural networks* are an increasingly popular nonlinear math-based method.

Berry and Linoff (1997) used the term *techniques* for market based analysis, memory-based reasoning, cluster detection, link analysis, decision trees and rule induction, artificial neural networks, genetic algorithms, and on-line analytic processing (OLAP).

Thuraisingham (1998) did not distinguish between *techniques* and *algorithms*, but preferred the term *techniques* for examples of market basket analysis, decision trees, neural networks, inductive logic programming, and link analysis techniques.

Peacock (1998) used the term *tool* for the instrument that accomplishes a data mining technique. He suggested that the data miner's tool kit would probably include nine tools or tool sets. These include query tools, descriptive statistics, visualization tools, regression-type models, association rules, decision trees, case-based reasoning, neural network, and genetic algorithms. Decision trees, association rules, case-based learning tools, neural networks, and genetic algorithms are categorized as "machine-learning" methods, whereas the others can be classified as "machine-assisted" aids to support human learning.

Explanations for some of the more important techniques follows:

- Market based analysis: a form of clustering used for finding groups of items that tend to occur together in a transaction.
- Memory-based reasoning: making predictions about unknown instances by using known instances as a model.
- Cluster detection: the building of models that find data records that are similar to each other.

- Link analysis: a technique for describing relationships among fields; in particular, how they interconnect.
- Decision trees: a tree-like way of representing a collection of hierarchical rules that lead to a class or value.
- Neural networks: a complex nonlinear modeling technique based on a model of a human neuron.
- Genetic algorithm: a computer-based method of generating and testing combinations of possible input parameters to find the optimal output. It uses processes based on natural evolution concepts such as genetic combination, mutation and natural selection.

Services

With data warehousing, data house holding, and data syndication as related services, data mining services include consultancy, implementation and education services.

Consultancy services address data mining in the context of general business management: for example, as part of a program for business process reengineering of customer relationship management. Implementation services focus on how to implement a specific data mining solution. All vendors offer product-specific education and training; some provide more generic data mining education.

The providers of data warehousing services offer a proprietary data warehouse blueprint or architecture and combine that with an implementation methodology. Data house holding services are effectively a way of outsourcing much of the data

preparation step of the data mining process. Data syndication provides sources of external data to the data mining projects.

Procedures

When an organization embarks on a data mining project, it usually goes through a number of stages: (1) Orientation, in which management becomes aware of problems, opportunities, and issues associated with the project and determines its readiness; (2) Problem definition, in which the organization identifies the domain, sponsors and relevant applications; (3) Execution, in which decisions are made as to sourcing (in-house vs. outsource), project process and expected results; and (4) Integration, in which actions are taken based on the results. When successful, the project is made as a repeatable process and the integration step is fed back into the problem definition step iteratively if needed.

Procedures for using data mining involve learning the application domain, creating a target dataset, data cleaning and preprocessing, data reduction and projection, choosing the function of data mining, choosing the data mining algorithm(s), data mining, interpretation, and using discovered knowledge (Fayyad, Piatetsky-Shapiro, and Smyth, 1996).

Related Areas

There are several areas similar to or related to data mining. The finding of useful patterns in data is known by different names (including data mining) in different communities (e.g., knowledge extraction, information discovery, information harvesting, data archeology, and data pattern processing) (Fayyad, Piatetsky-Shapiro, and Smyth, 1996). Sometimes, because the term has become so popular, *data mining* now refers

to other types of data analysis, such as query and reporting, On-line Analytical Processing (OLAP), and statistical analysis.

Knowledge Discovery in Databases (KDD), also referred to as data mining, is the search for usable intelligence in large volumes of raw data. KDD is an umbrella term describing a variety of activities for making sense of data (Brachman et al., 1996). Some people view data mining as simply an essential step in the process of KDD.

Data warehousing is the process of collecting and cleaning transactional data and making it available for on-line retrieval. Data is extracted from operational data sources, then transformed, cleansed, reconciled, aggregated, and summarized in preparation for warehouse processing (Bontempo and Zagelow, 1998). A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decisions (Inmon, 1992).

Data warehousing aims to help improve the effectiveness of data-driven business decision making. The concept is based fundamentally on the distinction between operational data (used to run the organization) and informational data (used to manage the organization). The data warehouse is designed to be a neutral holding area for informational data and is intended to be the sole source of quality company data for decision making.

A symbiotic relationship exists between data mining and data warehousing. Having a good data warehouse can facilitate effective data mining. Although data mining is possible without a data warehouse, the data warehouse greatly improves the chances of success in data mining (Inmon, 1996)

The evaluation of *statistics* has little to offer to the understanding of search architectures of data mining but a great deal to offer to the evaluation of hypotheses in the course of a search, to the evaluation of the results of a search, and to the understanding of the appropriate uses of the results (Glymour et al., 1996).

Of all the various traditional techniques of data analysis, statistics is closest to data mining. In fact, it is fair to say that statistics traditionally has been used for many of the analyses that are now done with data mining, such as building predictive models or discovering associations in databases. Indeed, for each of the major areas of data mining endeavor, there is a broadly equivalent statistical approach and it is probably true that much, if not all, of what is done in data mining could be done with the traditional practice of statistical analysis–eventually. Data mining attracts many analysts because of the relative ease with which new insights can be gained in comparison to traditional statistical approaches.

Data mining is always a hypothesis-free approach, whereas most popular statistical techniques require the development of a hypothesis in advance. Furthermore, statisticians typically have to develop manually the equations that match the hypotheses. In contrast, data mining algorithms can develop these equations automatically.

Whereas statistical techniques usually handle only numeric data and need to make strong assumptions about its distributions, data mining algorithms typically produce a much wider set of data types and make fewer assumptions about their distribution – or no assumptions at all.

Contexts of Data Mining Projects

The previous section reviewed general concepts and related areas of data mining. Because CSFs in data mining projects are not readily available, I will investigate potential CSFs in the contexts of data mining projects so as to build a research framework. In this section, I will recognize relevant contexts within which data mining projects are planned and implemented. Later, I will use these contexts as a framework in which to recognize potential CSFs in a systematic and organized manner.

How can the contexts of data mining and data mining projects be defined? Thorough investigation of the steps in data mining process makes it possible to find contexts of data mining projects because the steps of the data mining process contain contexts, components, and related fields of data mining projects. Now I will investigate different steps or procedures of data mining projects and contexts of data mining projects.

Chung and Gray (1999) suggest utilizing the following steps in data mining. These steps are iterative, with the process moving backward whenever needed.

1. Develop an understanding of the application, of the relevant prior knowledge, and of the end user's goals.
2. Create a target data set to be used for discovery.
3. Clean and preprocess data (including handling missing data fields, noise in the data, accounting for time series, and known changes).
4. Reduce the number of variables and find invariant representation of data if possible.
5. Choose the data mining task (classification, regression, clustering, etc.)

6. Choose the data mining algorithm
7. Search for patterns of interest (this is actual data mining).
8. Interpret the pattern mined. If necessary, iterate through any of steps described above.
9. Consolidate the discovered knowledge and prepare a report.

Fayyad, Piatetsky-Shapiro, and Smyth (1996) suggested processes for successful data mining projects. They emphasized the following procedures:

1. Understanding the application domain;
2. Data cleaning and preprocessing;
3. Data reduction and projection;
4. Choosing the function of data mining;
5. Choosing the data mining algorithm(s);
6. Interpretation, and;
7. Using discovered knowledge.

Han and Kamber (2001) arranged data mining steps as followed:

1. Data cleaning
2. Data integration
3. Data selection
4. Data transformation
5. Data mining
6. Pattern evaluation
7. Knowledge presentation

Based on these steps, the architecture of a typical data mining process can have the following two major contexts:

1. Data mining: data mining engine, techniques, applications, and interpretation and using discovered knowledge.
2. Data management: database, knowledge base, data warehouse, or other information repository, data cleaning and data integration steps.

Additional contexts are *project* and *information system (IS) project* because this study deals with data mining projects based on IS environment.

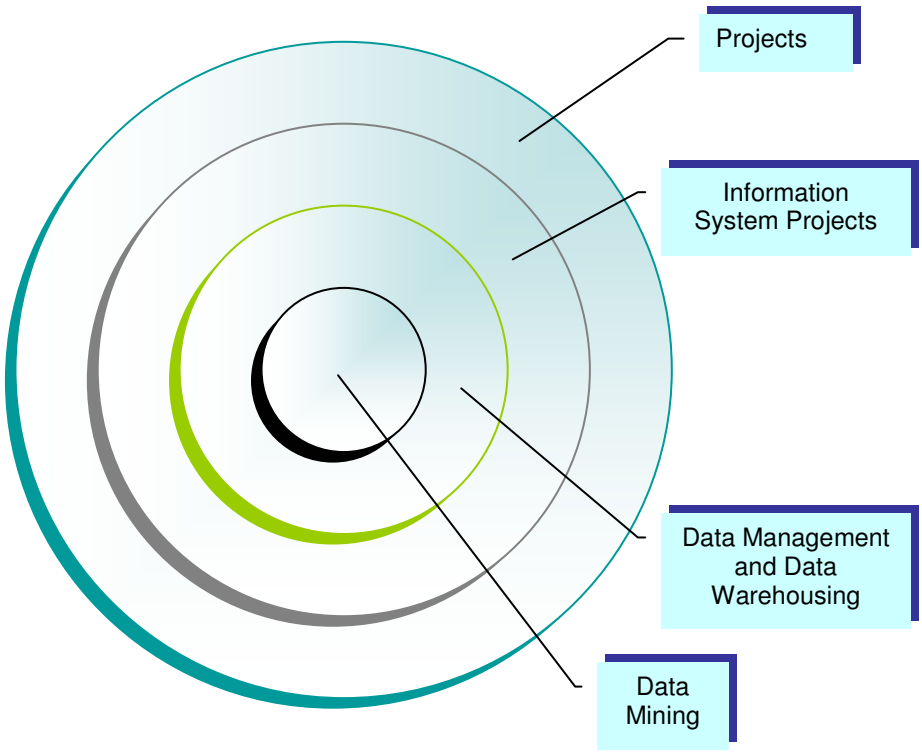

Figure 2-1: Contexts of Data Mining Project

The major contexts of data mining projects are data management, database or data warehouse, data mining module (task and technique), IS projects, and projects (Figure 2-1).

Critical Success Factor Approach

The critical success factor approach to defining and measuring an organization's performance is well-established (Rockart, 1979; Bullen and Rockart, 1981; Rockart and Crescenzi, 1984). Rockart introduced the concept in the late 1970's and defined critical success factors (CSFs) as "the limited number of areas in which results, if they are satisfactory, will ensure successful competitive performance for the organization. They are the few key areas where 'things must go right' for the business to flourish.... The status of performance in each area should be continually measured, and that information should be made available." CSFs commonly exist for an industry, a corporation, a department within the corporation, or for individual managers fulfilling the corporate strategy.

The CSF methodology is an effective way to identify corporate information needs and facilitate management information systems planning (Shank, Boynton, and Zmud, 1985). The CSF method directs managers to determine those things that must go right in order to succeed in achieving goals and objectives. The ultimate value that the CSF method brings is the ability to focus management attention on the tasks and activities that need to be done well to achieve success (Bullen, 1995). The CSF approach has proven beneficial in generating high level user involvement and support (Byers and Blume, 1994). CSFs are critical to an organization's current operating activities and to its future success (Guynes and Vanecek, 1996). Extending CSFs under operating

activities, we can state that there are certain critical factors in IS development projects (including data mining projects) that if not met will lead to the failure of those IS development projects. The factors in CSF taxonomy are classified by four levels: factors linked to success by a known causal mechanism, factors necessary and sufficient for success, factors necessary for success, and factors associated with success (Williams and Ramaprasad, 1996).

This extension of the CSF approach is supported by Zahedi, who argued that CSFs may have double significance: "If these factors are crucial for the success of the system, then their failure would lead to the failure of the information system" (Zahedi, 1987).

CSFs for Data Mining Projects

To build a research framework, I collected CSFs from major contexts of various data mining projects.

Success Factors in the Project Context

In addition to identifying CSFs, defining project success can prove to be a difficult problem. Pinto and Slevin (1988) wrote of the diversity of reported project successes in the information technology area. Many IT projects have been considered successful even though they were late and over budget. Likewise, other projects coming in on time and under budget were considered unsuccessful, as defined by a poor reception by clients and low usage levels (Pinto and Slevin, 1988). Organizations must be able to clearly define and measure the desired state called 'success' to know when it has been achieved, and do so prior to identifying critical success factors for its accomplishment.

While no strong consensus exists in the project management literature on how to define success, a number of models and techniques have been developed to aid in the definition and measurement process (Slevin and Pinto, 1986; Pinto and Slevin, 1988; Beale and Freeman, 1991). There is also a link between choosing appropriate CSFs to engage management's attention, and their support of the project activity (Rockart and Crescenzi, 1984).

Slevin and Pinto (1986) developed supporting criteria for ten factors of project success (Figure 2-2) identified in a prior survey. Their work builds on the process model and provides a framework for the necessary conditions to be present for project success. They perceive this critical path layout to be important and relevant to successful implementation. A clear mission, top management support, a defined

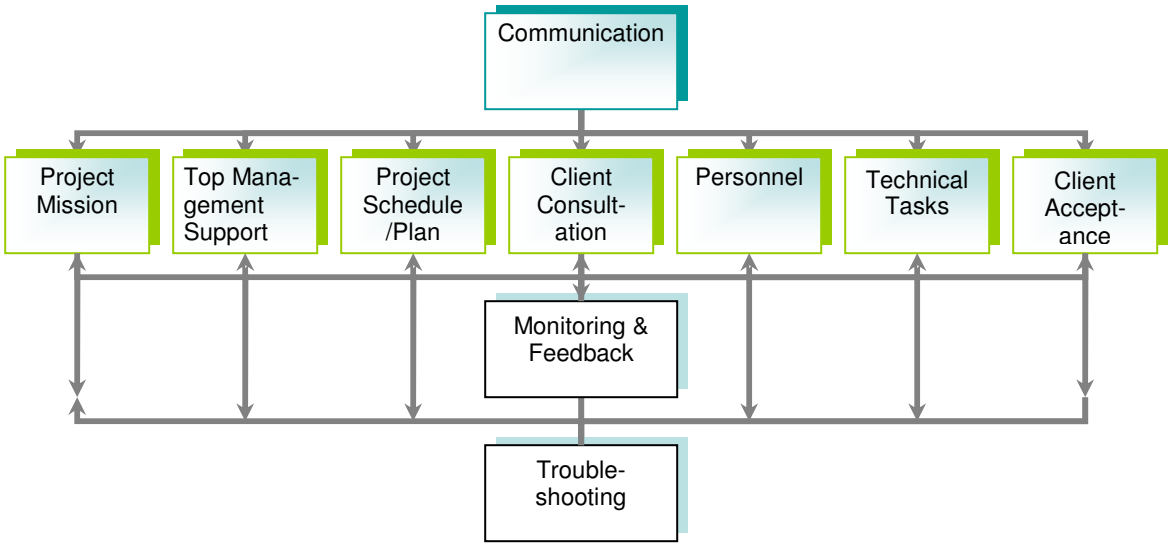

Figure 2-2: Slevin and Pinto's Factors of Project Success.

schedule or plan, client consultation, right personnel, appropriate technology, and client acceptance are defined as CSFs. Communication, monitoring, and feedback are simultaneous process steps, with troubleshooting occurring as needed throughout the implementation.

A number of principles emerged from the case study that may be relevant for applications of the CSF process in other contexts: (1) Top management support is essential; (2) Substantial group process is required; (3) Data collection is a significant problem; (4) Managers need to be continuously encouraged to consider the relevance of CSF for them; (5) The use of the factors must be pushed down through the organization; and (6) Survey development will most likely be needed (Slevin, Stieman, and Boone, 1991).

However, before applying the concept of CSFs at the project level, it may be useful to consider the generally accepted characteristics that the majority of projects exhibit. They apply to technical and non-technical project efforts and include:

- Well defined beginning and end (specified time to completion)
- A specific, preordained goal or set of goals (performance expectations)
- A series of complex or interrelated activities
- A limited budget (Pinto and Slevin, 1988).

Success will be tied to factors impacting the goals and objectives of each project as they relate to a specific or overall business strategy, and typically deal with the elements of time, budget, and performance. These four project success factors appear to be generic and are critical elements of any type of project.

Other recommendations for CSFs resulting from the project analysis include: (1) Develop realistic project schedules; (2) Match skills to needs at the proper time; and (3) Establish and control the performance of contractors (Walsh and Kanter, 1988).

From this review of the literature, CSFs in the project context can be summarized in Table 2-1.

Table 2-1: CSFs in the Project Context

CSF	Description
Goal	A clear mission; a specific, preordained goal or set of goals performance expectations)
Support	Top management support
Schedule	A defined schedule or plan; a defined beginning and end (specified time to completion); realistic project schedule
User consultation	Client consultation; client acceptance
Personnel	The right personnel
Technology	Appropriate technology; matched skills to needs
Communication	Communication
Feedback	Monitoring, and feedback are simultaneous process steps; a series of complex or interrelated activities; established and controlled performance
Troubleshooting	Troubleshooting occurring as needed throughout the implementation
Budget	A limited budget

Success Factors in Information Systems

Although there are no specific studies on CSFs for data mining, many researchers have been attempting to identify success factors that make an information system (IS) successful. Among the various studies, the one reported by Bailey and Pearson (1983) has received the most attention. This study (Table 2-2) identified 39 distinct factors that influence a user's IS satisfaction and proposed an instrument to measure them. DeLone and McLean (1992) suggested the taxonomy composed of six

Table 2-2: Dimensions and factors of IS success

covered by Bailey and Pearson's ISS instrument

Dimension	Factor
System Quality:	13. Response/turnaround time 15. Convenience of access 23. Features of computer language used 25. Realization of user requirements 26. Correction of errors 27. Security of data and models 28. Documentation of systems and procedures 38. Flexibility of the systems 39. Integration of the systems
Information Quality:	16. Accuracy of output 17. Timeliness of output 18. Precision of output 19. Reliability of output 20. Currency of output 21. Completeness of output 22. Format of output
Conflict Resolution:	2. Competition between CBIS and non-CBIS units 3. Allocation priorities for CBIS resources 5. Relationship between users and the CBIS staff 6. Communications between users and the CBIS staff 34. Personal control over the CBIS 37. Organizational position of the CBIS unit
Information Use:	24. Volume of output
User Satisfaction:	1. Top management involvement 4. Chargeback method of payment for services 32. User's confidence in the systems 33. User's participation
Individual Impact:	29. User's expectation of computer-based support 36. Job effects of computer-based support 31. Perceived utility
Service Quality:	7. Technical competence of the CBIS staff 8. Attitude of the CBIS staff 9. Scheduling of CBIS products and services 10. Time required for systems development 11. Processing of requests for system changes 12. Vendor's maintenance support 14. Means of input/output with CBIS center 30. User's understanding of the systems 35. Training provided to users

dimensions or categories (system quality, information quality, use, user satisfaction, individual impact, and organizational impact) to measure project success.

Li (1997) added seven new factors (Table 2-3) to Bailey and Pearson's 39 factors. According to Li (1997), the top five important IS success factors indicated by the IS managers are accuracy of output, reliability of output, relationship between users and the IS staff, user's confidence in the system, and the timeliness of output.

Table 2-3: The seven additional factors of IS success.

Dimension	Factor
Conflict Resolution:	*40. Users attitude toward using the CBIS:* The willingness and commitment of the user to achieve organizational goals by utilizing the CBIS capability.
User Satisfaction:	*41. Support of productivity tools:* The quality and the quantity of available computer hardware and software as well as peripheral devices which support organization's functions.
Information Quality:	*42. Clarity of output:* The degree to which output information is meaningful and unambiguous. *43. Instructiveness of output:* The capacity of output information to indicate possible corrected actions when problem occurs.
Organizational Impact:	*44. Productivity improved by the CBIS:* The ability of computer-based information systems to help user's organization produce more or better quality output per dollar of resource input. *45. Efficiency of the systems:* The ability of computer-based information systems to help the user's organization obtain the greatest possible return from the resources consumed. *46. Effectiveness of the systems:* The capacity of computer-based information systems to assist user's organization in identifying what should be done to better resolve problems.

From these literature reviews, Table 2-4 CSFs in information systems can be summarized.

Table 2-4: CSFs in Information Systems

CSF	Description
Accuracy	Accuracy of output
Reliability	Reliability of output
Relationship	Relationship between users and the IS staff
Confidence	User's confidence in the system
Timeliness	The timeliness of output.

Success Factors in Data Management

Guynes and Vanecek (1996) surveyed and interviewed 16 corporations to find CSFs for data management by using 49 factors (Table 2-5). The top three factors for administration focus on educating analysts in the use of data analysis methodology, integrating the applications logical data models into an overall enterprise data model, and developing logical data models. The top four factors for database administration

Table 2-5. Guynes and Vanecek's Factors for Data Management

Area	Factor
Data administration factors	1. Documenting all data items 2. Entering data items in the data dictionary 3. Preparing and maintaining database documentation for the data administrator 4. Preparing and maintaining database documentation for the user 5. Reviewing/approving data dictionary items 6. Performing audits of dictionary contents 7. Developing data definitions 8. Developing logical data models 9. Integrating the applications logical data models into an overall enterprise data model 10. Evaluating data analysis/data modeling software 11. Educating analysts in data analysis methodology 12. Maintaining data dictionary standards 13. Enforcing data dictionary standards

Area	Factor
Database administration factors	1. Designing integrity controls for database applications 2. Implementing approved database modifications 3. Setting database retention and backup policy 4. Loading database 5. Reorganizing database 6. Setting policies for database backup and recovery 7. Recovering database 8. Monitoring database usage 9. Enforcing database retention policy 10. Monitoring database performance 11. Tuning database to meet operational goals 12. Evaluating database related hardware 13. Selecting database related hardware 14. Determining data structures 15. Determining physical descriptions 16. Generating database descriptions 17. Evaluating database software 18. Selecting database software 19. Defining implementation strategy for databases 20. Evaluating new DBMS features 21. Developing database utilities 22. Disseminating DBMS documentation 23. Installing new DBMS features 24. Maintaining DBMS standards
Overlapping areas of responsibility involving both data administration and database administration	1. Forecasting database growth 2. Reviewing user/data administration requests for database modifications 3. Specifying database access policy 4. Selecting and designing security techniques 5. Maintaining adequate security controls 6. Educating analysts in the use of the database

Used with permission from “Information and Management” Journal.

focus on recovering databases, tuning databases to meet operational needs, determining data structures, and defining implementation strategies for databases.

From these literature reviews, Table 2-6 CSFs in data management can be summarized.

Table 2-6: CSFs in Data Management

CSF	Description
Education	Educating analysts in the use of data analysis methodology
Integration	Integrating the applications logical data models into an overall enterprise data model
Data Model	Developing logical data models
Database Recovery	Recovering databases
Operational Need	Tuning databases to meet operational needs
Data Structure	Determining data structures
Implementation Strategy	Defining implementation strategies for databases

Success Factors in Data Warehousing

The data warehouse sets the stage for successful and efficient exploration of the world of data, and the miner lucky enough to work on a foundation of a data warehouse enjoys success that comes from exploitation of this data as a resource (Inmon, 1996).

Some success factors have been defined in the literature for data warehousing. Successful data warehouse implementations have five common characteristics, or CSFs, that will help ensure a successful data warehouse implementation. They focus on the real problem; they appoint a data warehouse champion; they use detailed historical data; they apply technology to the business; and they trust that the data - history does not lie (MacDonald, 1998). Sim and Cutshall (2000) suggested the following data warehousing CSFs: (1) Identifying the business problems and data, (2) Scalability, (3) Partnership between the IT department and business users, (4) Experienced personnel, (5) Data quality, (6) Metadata management, (7) Not confusing data warehouses with data marts, and (8) Pilot testing the data warehouse with a subset of the actual data to be used in the data warehouse. David and Steinbart (1999)

suggest ensuring that the data warehousing team has a game plan. To increase their chances of success, the data warehousing team should identify a business need, maintain availability, and plan for growth. Sauls (1996) emphasizes the identification of the organization's critical data requirements. Political skills or size of data warehouse are other success factors.

Organizations that ignore these critical success factors will learn that a data warehousing project can quickly turn into a very expensive failure.

CSFs in other related areas include the following. CSFs dictate that query languages be easy to learn, easy to understand, and easy to use (Yen and Scamell, 1993). A CSF for data marts requires that the data mart be refreshed periodically and constantly updated with new information from all end users, including the field and customer service representatives (Henon and Gauches, 1999).

From these literature reviews, Table 2-7 CSFs in data warehousing can be summarized.

Table 2-7: CSFs in Data Warehousing

CSF	Description
Identification of business problem	Focus on the real problem; Identifying the business problems and data; Identify a business need
Appointment	Appoint a data warehouse champion
Scalability	Scalability; Plan for growth; Size of data warehouse
Data Requirement	Identify the organization's critical data requirements; Identifying the business problems and data
Data Quality	Data quality, metadata management; Use detailed historical data; Trust that the data - history does not lie
Communication	Partnership between the IT department and business users; Political skills
Education	Experienced personnel
Technology	Apply technology to the business
Availability	Maintain availability

Success Factors in Data Mining

Very little research on the success factors of data mining exists because the field of data mining is new. Thus, I collected possible factors from related studies to build the research framework to find CSFs for data mining projects. Some other factors can be generated in data mining steps because consequently identifying the application and following a well-articulated data mining process consistently leads to successful projects (Cabena, et al., 1997). I will explore each construct and use the information to generate a model defining successful data mining projects.

Chung and Gray (1999) suggested a data mining research framework as shown in Figure 2-3. In generating a data-mining model and in selecting the appropriate model assessment methods, data-mining research needs to incorporate the characteristics of a given task domain, the quality and composition of a dataset that represents a domain, the decision-making environment, human factors, and potential interaction among them. Each part in this figure can be a factor in defining a successful data mining project.

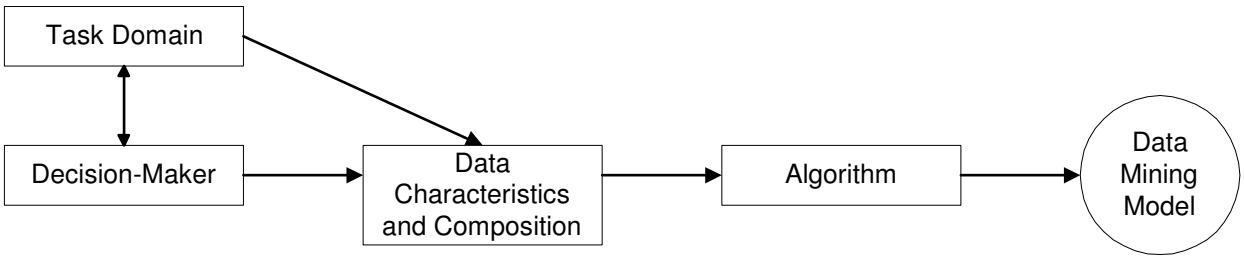

Figure 2-3: Data Mining Research Framework

Chung and Gray (1999) showed examples of the factors shown above while they defined data mining steps. For the task domain factor, we need to develop an

understanding of the application, of the relevant prior knowledge, and of the end user's goals. For the dataset factor, datasets should be cleaned and preprocessed to handle missing data fields, noise in the data, and to account for time series and known changes. The number of variables and invariant representations of data are other examples of the dataset factor. In addition to the factors shown above, the data-mining task and the data mining algorithm can be applied for researching factors because these two are important data mining steps.

When Fayyad, Piatetsky-Shapiro, and Smyth (1996) suggested processes for successful data mining projects, they emphasized the following key steps:

- Understanding the application domain (including relevant prior knowledge and the goals of the application);
- Data cleaning and preprocessing (includes basic operations, such as removing noise or outliers if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, and accounting for time sequence information and known changes, as well as deciding DBMS issues, such as data types, schema, and mapping of missing and unknown values);
- Data reduction and projection (including finding useful features to represent the data, depending on the goal of the task, and using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data);
- Choosing the function of data mining;
- Choosing the data mining algorithm(s);

- Interpretation (including interpreting the discovered patterns and possibly returning to any of the previous steps, as well as possible visualization of the extracted patterns, removing redundant or irrelevant patterns, and translating the useful ones into terms understandable by users); and
- Using discovered knowledge (including incorporating this knowledge into the performance system, taking actions based on the knowledge, or simply documenting it and reporting it to interested parties, as well as checking for and resolving potential conflicts with previously believed (or extracted) knowledge).

According to Brachman, et al. (1996), the practical application of data mining in an industry is affected by a number of issues including training, tool support, data availability, availability of patterns, changing and time-oriented data, spatially oriented data, complexity of data, and scalability. They also suggested that given a suitable domain, the costs and benefits of a potential application are affected by the following factors: alternatives, relevance, volume, complexity, quality, accessibility, change, and expertise. Several other authors mentioned other success factors, such as application, people, company, and data models.

I have identified six major categories of CSFs from the literature on data mining and related areas: task domain, human factor, dataset, tool, interpretation, and using discovered knowledge. Each of these categories is described below.

Task domain (Application). The task domain is the first factor for successful data mining projects. The developer(s) should understand the application itself, the goal of the application, the relevant prior knowledge, and the end user's goals. In addition, the developer(s) need to understand the business objectives clearly. They need to perform

solid cost-benefit analysis, to investigate the impact on the business problems or opportunities, and to determine that the project can be achieved in a reasonable time.

The application/task factor consists of five categories: classification, estimation, prediction, affinity grouping, and description.

Human Factor. As with every project in IS, data mining has the related human factor. Thus, the data mining project must consider the human factors before the data mining project can be considered a success. People in the different levels of organization who should participate in the project include the following:

- Sponsor: influential, focused on business value, enthusiastic;
- User group: owners of the business value definition of the project and evaluators of the project's success;
- Business analyst: experienced in the domain and application are;
- Data analyst: experienced in exploratory data analysis (EDA) and data mining;
- Data management specialist: experienced in database administration, has access to the physical data (including the relevant metadata); and
- The project manager: experienced in project management (Cabena, et al., 1997).

The people listed above should be sufficiently trained in the process of data mining.

The organization itself can be another factor. Today, organizations that have embraced data mining are themselves often leaders in their respective industries. Some examples are Bank of America, and AT&T. The industries where data-driven solutions to business intelligence tend to be best established are typically those where there are large volumes of data and/or a high degree of organizational complexity with

any number of complicating factors, such as international operations or a wide variety of markets served. These industries are characterized by long traditions of computerization and a focus on data-driven decision making. Hence, many of the leading examples of data warehousing and data mining today are in the banking and insurance industries. The retail industry is also a strong proponent, largely because of the increasingly competitive environment, low profit margins, and the ready supply of relatively clean data captured automatically at POS terminals. More recently, because of a worldwide movement toward the deregulation of the formerly monopolistic utility services, there is an increasing number of telecommunications and other public utility companies entering into the data mining fray. The effects of data mining may be different for various company types. Company size or revenues could be additional factors.

Dataset. The characteristics of the dataset are closely related to the data mining project. These characteristics are:

- Quality: Datasets should be cleaned and preprocessed to handle missing data fields, noise or outliers in the data, accounting for time series, and known changes. DBMS issues such as data types, schema, and mapping of missing and unknown values should be decided, too. Error rates should be relatively low.
- Accessibility: Data should be easily accessible; accessing data or merging data from different sources increases the cost of an application. While datasets can be extracted from legacy data systems, flat files, data marts, or data warehouses, ideally the data warehouse is the best source for data mining. Data mining can

be done without a data warehouse, but the data warehouse greatly improves the chances of success in data mining (Inmon, 1996).

- Complexity: Data reduction and projection include finding useful features to represent the data, depending on the goal of the task, and using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data. The more variables (fields) there are, the more complex the application. Complexity is also increased for time-series data.
- Volume: There should be a sufficient number of cases (several thousand at least). On the other hand, extremely large databases may be a problem when the results are needed quickly.

Tool. In the data mining process, choosing the technique of data mining is important because the technique is closely related to the data mining model and, thus, it provides the data mining results. The data mining technique is also referred as *modeling technique* (Groth, 2000), *algorithm* (Chung and Gray, 1999), *search algorithm* (Fayyad, Piatetsky-Shapiro, and Smyth, 1996), and sometimes *tool* (Peacock, 1998). Because there is a terminology war in data mining (Groth, 1998), different papers and texts have used different terms sometimes to mean the same concept. This research will use the term *tool* to include *technique* and *algorithm*, and add to the meaning that of the software providing for the technique and the algorithm.

The technique or algorithm used may be market-based analysis, memory-based reasoning, cluster detection, link analysis, decision trees and rule induction, artificial neural networks, or genetic algorithms.

Most available data mining tools support only one of the core discovery techniques. The tools must also support the complete data mining process and provide a user interface suitable for business users rather than for IS professionals.

Scalability is another factor for the tool. Current tools cannot handle truly vast quantities of data. Sizes of data warehouses are huge nowadays. For example, data warehouses starting at 200GB are no longer rare.

The software used for data mining projects is another concern for successful projects. The effect will be different by software category, such as commercial, public domain, and research prototype.

The type of data mining service will have different effects. The potential advantage from any data mining solution increases exponentially as the solution moves from a data mining tool to an application to a data mining service. In other words, we can expect to get a lot more leverage for our dollar investment from, say, a data mining consultancy engagement than from buying an off-the-shelf product and trying it out ourselves (Cabena, et al., 1997).

Interpretation. Interpretation includes interpreting the discovered patterns and possibly returning to any of the previous steps, as well as possible visualization of the extracted patterns, removing redundant or irrelevant patterns, and translating the useful ones into terms understandable to the users.

Using discovered knowledge. Using discovered knowledge includes incorporating this knowledge into the performance system, taking actions based on the knowledge, or simply documenting it and reporting it to interested parties, as well as

Table 2-8: Possible Success Factors in Data Mining

CSF	Description
Business Objective	Understand the business objectives clearly; perform a solid cost-benefit analysis; investigate the impact on the business problems or opportunities; understand the end user's goals; understand the relevant prior knowledge.
Schedule	Determine that the project can be achieved in a short time.
Application	Understand the application itself, particularly the goal of application.
Education	Sponsor: influential, focused on business value, enthusiastic; User group: owners of the business value definition of the project and evaluators of the project's success; Business analyst: experienced in the domain and application are; Data analyst: experienced in exploratory data analysis (EDA) and data mining; Data management specialist: experienced in database administration, has access to the physical data (including the relevant metadata); and The project manager: experienced in project management .
Quality	Datasets should be cleaned and preprocessed to handle missing data fields, noise or outliers in the data, accounting for time series, and known changes.
Accessibility	Data should be easily accessible.
Complexity	The more variables (fields) there are, the more complex the application.
Volume	There should be a sufficient number of cases.
Tool (or Technique)	Choosing the technique of data mining is important because the technique is closely related to the data mining model and, thus, it provides the data mining results; The tools must also support the complete data mining process and provide a user interface suitable for business users rather than for IS professionals.
Scalability	Scalability is the other factor for the tool.
Software	The software used for data mining projects is another concern for successful projects; The effect will be different by software category, such as commercial, public domain, and research prototype.
Service	The type of data mining service will have different effects from a data mining tool, to an application, to a data mining service.
Interpretation	Interpretation includes interpreting the discovered patterns and possibly returning to any of the previous steps, as well as possible visualization of the extracted patterns, removing redundant or irrelevant patterns, and translating the useful ones into terms understandable to the users.
Using Discovered Knowledge	Using discovered knowledge includes incorporating this knowledge into the performance system, taking actions based on the knowledge, or simply documenting it and reporting it to interested parties, as well as checking for and resolving potential conflicts with previously believed (or extracted) knowledge.

checking for and resolving potential conflicts with previously believed (or extracted) knowledge.

From the preceding literature and research reviews, I have extracted a number of possible success factors for data mining projects. Since no specific journal article revealed the CSFs for data mining projects, these suggested possible success factors remain to be verified.

Potential CSFs for Data Mining Projects

I have summarized these success factor items from all the CSF tables, generating from the major context of data mining projects a research framework to find CSFs in data mining projects.

Table 2-9: Potential Success Factor Items for Data Mining Projects

Construct	Factors	CSF Table
Business Mission	A clear mission, a specific goal	2-1
Business Objective	Identifying the business problems and data	2-6
	Understand the business objectives clearly	2-8
Budget	A limited budget	2-1
Time	A defined schedule or plan	2-1
	The timeliness of output.	2-4
	Achievable project in a short time	2-8
User consultation	Client consultation, client acceptance	2-1
	User's confidence in the system	2-4
	Relationship between users and the IS staff	2-4
	Political skills	2-6
Personnel	The right personnel	2-1
	Educating analysts in the use of data analysis methodology	2-7
	Experienced personnel	2-6

	Educated sponsor, user group, business analyst, data analyst, data management specialist, and the project manager	2-8
Technology	Appropriate technology	2-1
	Applying technology to the business	2-6
	Appointing a data warehouse champion	2-6
Communication	Communication	2-1
	Top management support	2-1
	Troubleshooting	2-1
	Monitoring, and feedback	2-1
Output	Accuracy of output	2-4
	Reliability of output	2-4
Data Management	Integrating into an enterprise data model	2-7
	Developing logical data models	2-7
	Determining data structures	2-7
	Identify the organization's critical data requirements	2-6
Database Management	Tuning databases to meet operational needs	2-7
	Defining implementation strategies for databases	2-7
	Recovering databases	2-7
Data Quality	Cleaned and preprocessed dataset	2-8
	Easily accessible data	2-8
	Not complex data structure	2-8
	A sufficient number of cases	2-8
	Data quality, metadata management	2-6
	Scalability	2-6
	Maintaining availability	2-6
Task	Data mining application	2-8
	Data mining technique	2-8
	The software for data mining project	2-8
	The type of data mining service	2-8
Action	Interpretation about results	2-8
	Using discovered knowledge	2-8

CHAPTER 3
RESEARCH FRAMEWORK AND METHODOLOGY

In this chapter, I will validate the collected success factor items discussed in the previous chapter. I will describe the initial research framework I built with information acquired through a search for specific key words in data mining articles. Following that, I will consolidate all the collected success factor items into 7 significant factors. Using these 7 factors, I will then modify the initial research framework into a new research model. Finally, from the modified research framework, I will define the research questions and hypotheses.

Initial Research Framework

Validation of Potential Critical Success Factors (CSFs) in Data Mining Projects

To verify all possible success factor items, I performed key word frequency analysis in one of the most common Internet articles repository, *EBSCOHost.* I used *Academic Search Premier* and *Business Source Premier* from the business area for this analysis. First, I generated a table containing all the key words from the above table. Then I counted the frequency of each key word in the data mining articles contained in the records of EBSCOHost. At the same time, I added new key words from some of the journal articles to verify other possible success factors. I used 1,310 articles containing the phrase “data mining” for this analysis. The results are shown in the following table: In Table 3-1, possible success factors for data mining projects shown in Table 2-8 are typed in bold font style. Most factors (except tool scalability, data quality, and data volume) have frequencies of more than one percent.

Table 3-1 Frequencies of Possible Success Factors

Percentage	Factors
More than 10%	Company, **Software**, **Business**, Information, Computer, Customer, System, **Service, Management**, **Database, Application**, **Tool, Report**, Marketing, Analysis, Process, Internet
More than 5% to 10 %	Warehousing, Development, Intelligence, Strategy, Decision, Online, Knowledge, Support, Network, **Technique,** Warehouse, Web, Research
1% to 5%	**Discovery**, Method, Sale, Site, **Access**, Cost, Problem, Result, **Pattern, Time**, Model, Organization, Machine, Prediction, **Manager**, Automation, Value, Role, Insurance, Challenge, Algorithm, Design, Experience, Building, Approach, Generation, **Visualization**, Computing, Effective, OLAP, Set, Professional, **Goal**, Level, Privacy, Competitive, **Complexity,** Corporate, Retrieval, Statistical, Usage, Interface, **Objective**, Science, Interactive, Review, Implication, Libraries, Telecommunication, Architecture, **Document**, Dynamic, Personalization, Rule, Sophisticated**, Data Management,** Existing, Loyalty, Measure, **Analyst**, Asset, Economic, Evaluation, Improving, **Education**, Selection, **Task**

Only three new factors were added from the frequency analysis. They are underlined in Table 3-1: customer, marketing, and Internet. From these three factors, I generated the following Table 3-2:

Table 3-2: Added Potential Success Factor Items in Data Mining

CSF	Description
Customer	Successful data mining projects may have close relationship with customers, and may be easy to understand for customers.
Marketing	Successful data mining projects may increase market share and profit. Successful data mining projects have close relationship with customer relationship marketing (CRM).
Internet	Successful data mining projects may be combined with Internet business easily.

By adding three more possible success factor items from the above table into Table 2-9, I generated a slightly modified Table 3-3:

Table 3-3: Potential Success Factor Items for Data Mining Projects after Validation

Construct	Factors	CSF Table
Business Mission	A clear mission, a specific goal	2-1
	Set of goals performance expectations: market share, profit, CRM	3-2
Business Objective	Identifying the business problems and data	2-6
	Understand the business objectives clearly	2-8
Budget	A limited budget	2-1
Time	A defined schedule or plan	2-1
	The timeliness of output.	2-4
	Achievable project in a short time	2-8
User consultation	Client consultation, client acceptance	2-1
	User's confidence in the system	2-4
	Relationship between users and the IS staff	2-4
	Political skills	2-6
	Close relationship with customers	3-2
Personnel	The right personnel	2-1
	Educating analysts in the use of data analysis methodology	2-7
	Experienced personnel	2-6
	Educated sponsor, user group, business analyst, data analyst, data management specialist, and the project manager	2-8
Technology	Appropriate technology	2-1
	Applying technology to the business	2-6
	Appointing a data warehouse champion	2-6
	New technology – Internet	3-2
Communication	Communication	2-1
	Top management support	2-1
	Troubleshooting	2-1
	Monitoring, and feedback	2-1
Output	Accuracy of output	2-4
	Reliability of output	2-4
Data Management	Integrating into an enterprise data model	2-7
	Developing logical data models	2-7
	Determining data structures	2-7
	Identify the organization's critical data requirements	2-6

Database Management	Tuning databases to meet operational needs	2-7
	Defining implementation strategies for databases	2-7
	Recovering databases	2-7
Data Quality	Cleaned and preprocessed dataset	2-8
	Easily accessible data	2-8
	Not complex data structure	2-8
	A sufficient number of cases	2-8
	Data quality, metadata management	2-6
	Scalability	2-6
	Maintaining availability	2-6
Task	Data mining application	2-8
	Data mining technique	2-8
	The software for data mining project	2-8
	The type of data mining service	2-8
Action	Interpretation about results	2-8
	Using discovered knowledge	2-8

Initial Research Framework

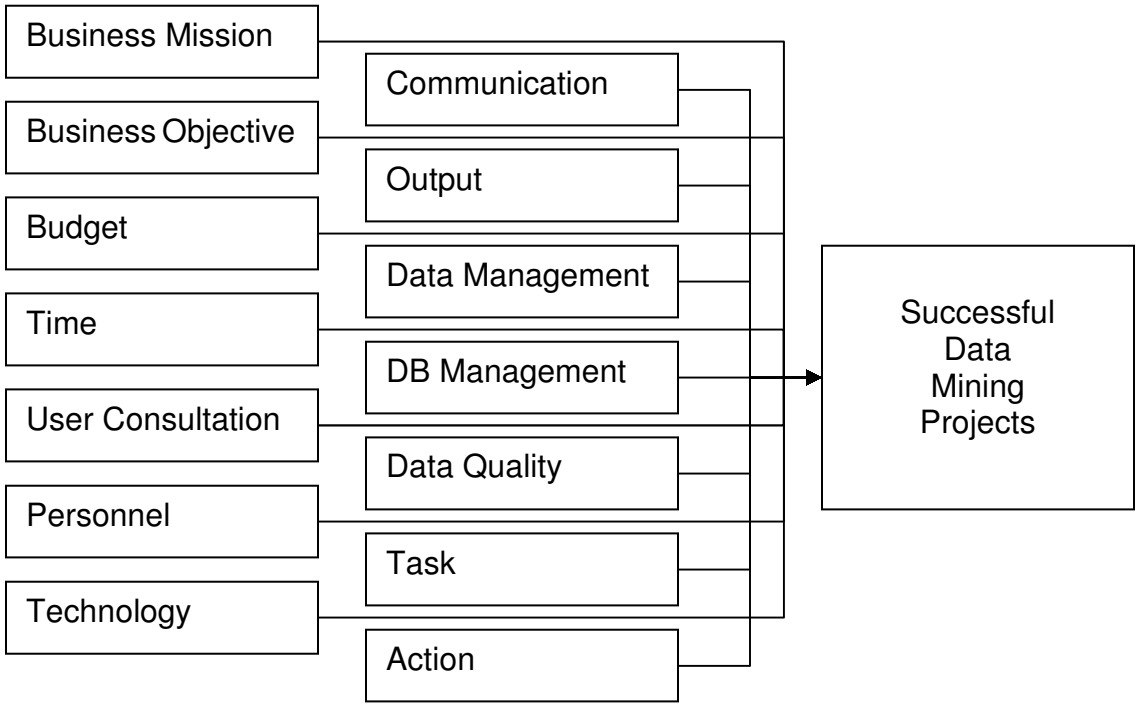

Figure 3-1: Initial Research Framework:

A Consolidation and Mapping of the Constructs to Successful Data Mining Projects

From the previous table, the success factor items could be categorized into 14 constructs. All 14 constructs may be important in successful data mining projects. Thus I have generated an initial research model, as shown in Figure 3-1.

Variables

How do we know that the application of data mining projects has met with success? This is a hard question whose answer is often domain- and/or organization-dependent. Many organizations estimate success through user's perception where quantitative results are hard to measure.

DeLone and McLean (1992) suggested six major dependent variables to measure information systems success. Their comprehensive taxonomy for six dimensions or categories includes *system quality, information quality, use, user satisfaction, individual impact,* and *organizational impact*. Because investigating the success of a data mining project means finding effects upon the organization, one of the

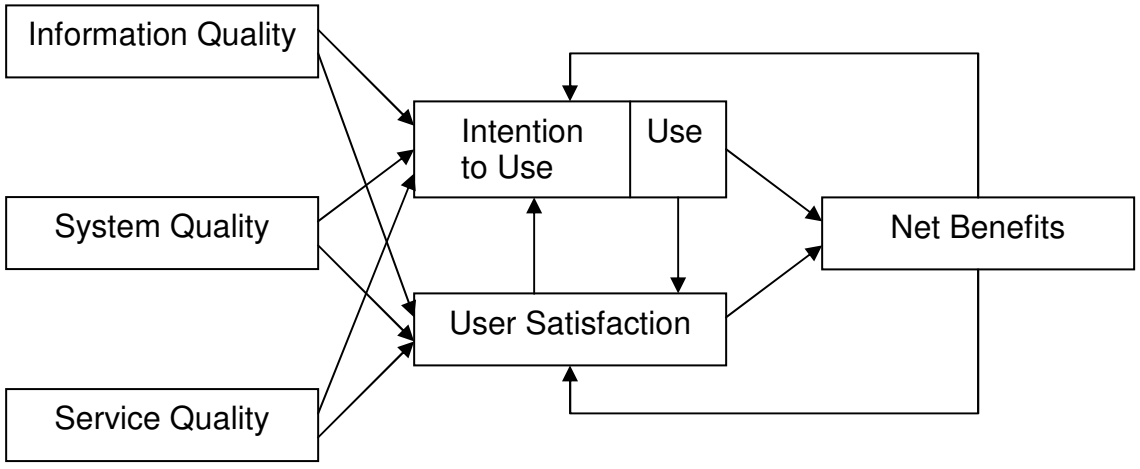

Figure 3-2: Updated DeLone &McLean IS Success Model

measures in the "organization impact" category from DeLone and McLean (1992) taxonomy may be used. User satisfaction or another measure for organization impact for a project will be used for this research. Figure 3-2 depicts DeLone and McLean's IS success model that has been updated with minor refinements to their original 1992 model (DeLone and McLean, 2003).

The following three tables list the constructs of interest, variables, types and surrogates, where needed. The independent constructs are "business mission, business objective, budget, time, user consultation, personnel, technology, communication, output, data management, database management, data quality, task, and action." Each construct contains several sub-constructs or variables. I will analyze these independent constructs with dependent constructs so as to define success factors for a data mining project.

Table 3-4: Independent Constructs

Construct #	Constructs	Variable Type	Survey Question or Surrogates
C1	Business Mission	Continuous	1, 2, 3
C2	Business Objective	Continuous	4, 5, 6, 7
C3	Budget	Continuous	8
C4	Time	Continuous	9, 10
C5	User consultation	Continuous	11, 12, 13, 14
C6	Personnel	Continuous	15, 16, 17, 18, 19, 20 ,21
C7	Technology	Continuous	22, 23
C8	Communication	Continuous	24, 25, 26, 27
C9	Output	Continuous	28, 29
C10	Data Management	Continuous	30, 31, 32, 33,
C11	Database Management	Continuous	34, 35, 36
C12	Data Quality	Continuous	37, 38, 39, 40, 41, 42
C13	Task	Continuous	43, 44, 45, 46, 47
C14	Action	Continuous	48, 49, 50, 51, 52, 53, 54, 55, 56, 57

Table 3-5: Independent Variables

V#	Variable
v01.	A clear mission
v02.	Realistic mission related to marketing strategy
v03.	Focusing on the real problem
v04.	Identifying the business problems and data
v05.	Identifying the business needs
v06.	Investigating the impact on the business problems
v07.	Investigating the impact on the business opportunities
v08.	A limited budget
v09.	Realistic project schedule
v10.	Providing information in a timely manner
v11.	Client consultation
v12.	Client acceptance
v13.	User's confidence in the system
v14.	Good relationship between users and the IS staff
v15.	Well educated team members in the project
v16.	Experienced sponsor in the project
v17.	Experienced user-group in the project
v18.	Experienced business analyst in the project
v19.	Experienced data analyst in the project
v20.	Experienced data management specialist in the project
v21.	Experienced project manager in the project
v22.	Applying relevant technology to the business needs
v23.	Applying state-of-the-art technology to the business
v24.	Good communication in the project team
v25.	Top management support to the project
v26.	Troubleshooting throughout the implementation
v27.	Monitoring, and feedback to the project
v28.	Accuracy of output
v29.	Reliability of output
v30.	Determining data structures
v31.	Developing logical data models
v32.	Integrating logical data models into an enterprise data model
v33.	Identifying the organization's critical data requirements
v34.	Tuning databases to meet operational needs
v35.	Defining implementation strategies for databases
v36.	Recovering databases
v37.	Cleaned dataset
v38.	Easily accessible data
v39.	NOT complex dataset
v40.	Sufficient number of dataset
v41.	Scalability of database
v42.	Maintaining availability of dataset

v43. Selecting the right type of data source
v44. Determining the data mining task
v45. Selecting the right data mining technique for the project
v46. Selecting the right data mining software for the project
v47. Understanding the service type of the project.
v48. Interpreting the discovered patterns
v49. Possibly returning to any of the previous steps after interpreting the discovered patterns
v50. Visualization of the extracted patterns
v51. Removing redundant or irrelevant patterns
v52. Translating the useful ones into terms understandable to the users
v53. Incorporating the knowledge into the performance system
v54. Taking actions based on the extracted knowledge
v55. Documenting the extracted knowledge
v56. Reporting the extracted knowledge to interested parties
v57. Checking for and resolving potential conflicts with previously believed (or extracted) knowledge

The dependent construct is the user's perceived success of data mining projects.

Table 3-6: Dependent Constructs

Constructs	Variable Type	Survey Question or Surrogates
Perceived Success of Data Mining Projects	Continuous	2.3.6

Consolidation of Factors

I performed a factor analysis to find major factors (or a number of groups) from 57 variables, using a principal component analysis and Varimax with Kaiser Normalization rotation.

Table 3-7: The Summary of Factor Analysis

Round	Number of Components	Number of Removed Variables	Number of Remaining Variables
1	14	12	45
2	12	7	38
3	10	4	34
4	10	2	32
5	9	3	29
6	8	1	28
7	8	1	27
8	8	1	26
9	7	0	26

Table 3-7 shows seven components after 9 rounds of factor analysis. After carefully reviewing the variables included in each group, I labeled the seven components as Action, Dataset, Communication, Output, Business Mission, Consultation, and Business Environment as shown in Table 3-8.

Component 1 contains Variables 31, 32, 50, 51, 53, 54, 55, and 57. These variables are classified into *Data management* and *Action* constructs according to Table 3-4, *Independent Constructs*. Action is the dominant construct for Component 1, and thus, Component 1 is called Factor 1, or "*Action"*.

Component 2 contains Variables 36, 38, 39, 41, and 42. These variables are classified into *Database management* and *Data Quality* constructs. Because all variables except one are part of *Data Quality* construct and all variables contain the characteristic of high quality dataset, Component 2 is called Factor 2, or "*Dataset"*.

The other factors are called Communication, Output, Business Mission, Consultation, and Business Environment, applying the same procedures as above.

Table 3-8: The Result of Final Round Factor Analysis

Component Variable	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
v01.					0.7594		
v02.					0.8558		
v05.							0.7317
v06.							0.7100
v09.							0.7611
v11.						0.8679	
v16.			0.7695				
v19.						0.8962	
v25.			0.7414				
v26.			0.7766				
v27.			0.6521				
v28.				0.8449			
v29.				0.8058			
v31.	0.8119						
v32.	0.7785						
v36.		0.8858					
v38.		0.8146					
v39.		0.7152					
v41.		0.8424					
v42.		0.7680					
v50.	0.5184						
v51.	0.5772						
v53.	0.7545						
v54.	0.8755						
v55.	0.7341						
v57.	0.6056						
Factor Name	Action	Dataset	Commu-nication	Output	Business Mission	Consult-ation	Business Environ-ment

Table 3-9 shows the relationship between constructs and consolidated factors. Two factors, F4 and F5, are exactly matched with defined constructs in the temporary research model. Other factors have close relationships with constructs. For example, factor 1, action, is the subset of construct 14, action.

Table 3-9: Relationship between Factors and Constructs

	Construct
C1	Business Mission
C2	Business Objective
C3	Budget
C4	Time
C5	User consultation
C6	Personnel
C7	Technology
C8	Communication
C9	Output
C10	Data Management
C11	Database Management
C12	Data Quality
C13	Task
C14	Action

	Factor
F5	**Business Mission**
F7	Business Env.
F6	Consultation
F3	Communication
F4	**Output**
F2	Dataset
F1	Action

I performed reliability analysis to check that each factor has a high level of reliability. The purpose of reliability analysis is to find those items that contribute to internal consistency and to eliminate those items that do not. Internal consistency is a measurable property that reflects the extent to which items intercorrelate and implies they measure the same construct. "Failure to intercorrelate is an indication that the items do not represent a common underlying construct" (Spector, 1992). "The most common and powerful method used today for calculating internal consistency reliability is coefficient alpha" (Rubin and Babbie, 1997). Coefficient alpha (Cronbach, 1951) is a direct function of both the number of items and their magnitude of intercorrelation. A widely accepted rule of thumb is that alpha should be at least 0.70 for a scale to demonstrate internal consistency (Spector, 1992).

The first four factors have higher alpha values (0.89, 0.90, 0.79, and 0.82 respectively) while the remaining three factors have lower alpha values (0.65, 0.55, and

0.67 respectively). The results are shown in Table 3-10. While Factor 5, Factor 6, and Factor 7 have low levels of reliability, I keep three factors for further analyses because alpha values are close to 0.70. The alpha value of Factor 1 can be increased from 0.8858 to 0.8894 if V50 item is deleted. To optimize the internal consistency of Factor 1, V50 could be removed.

Table 3-10: Reliability Analysis

Factor	V #	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Alpha if Item Deleted	Alpha	Stand-ardized item alpha
Factor 1	V31	37.9000	67.1692	0.6018	0.6060	0.8768	0.8858	0.8858
	V32	37.8750	65.2917	0.6454	0.6102	0.8726		
	V50	37.6750	69.7635	0.4746	0.3752	**0.8894**		
	V51	37.5250	65.3327	0.7052	0.6353	0.8667		
	V53	37.3500	61.3103	0.8130	0.7687	0.8546		
	V54	36.8750	62.5224	0.7875	0.6979	0.8577		
	V55	37.3500	67.8231	0.6632	0.4800	0.8713		
	V57	37.2250	68.4353	0.5679	0.5264	0.8799		
Factor 2	V36	18.1111	31.8283	0.8314	0.7460	0.8599	0.8998	0.9037
	V38	16.9111	33.7646	0.6920	0.5143	0.8910		
	V39	18.1556	33.2253	0.6832	0.4774	0.8941		
	V41	17.8444	31.1343	0.8069	0.7444	0.8653		
	V42	17.1556	36.1343	0.7795	0.6355	0.8774		
Factor 3	V16	16.1087	9.3879	0.5911	0.3702	0.7524	0.7879	0.8049
	V25	15.3043	11.0609	0.6618	0.4397	0.7068		
	V26	15.8478	10.0874	0.6072	0.3968	0.7319		
	V27	15.4130	12.9589	0.6038	0.3781	0.7518		
Factor 4	V28	6.0400	0.9780	0.7096	0.5035	.	0.8175	0.8301
	V29	5.6200	1.5465	0.7096	0.5035	.		
Factor 5	V01	5.6731	1.8322	0.4796	0.2300	.	0.6462	0.6483
	V02	6.2115	1.5034	0.4796	0.2300	.		
Factor 6	V11	5.7600	1.4922	0.3791	0.1437	.	0.5491	0.5498
	V19	5.8400	1.3208	0.3791	0.1437	.		
Factor 7	V05	10.4694	5.5459	0.4296	0.1848	0.6550	0.6733	0.6752
	V06	11.0612	3.9337	0.5318	0.2849	0.5162		
	V09	11.6531	4.0230	0.5189	0.2735	0.5348		

Modified Research Framework

Even though the result of factor analysis is matched with most constructs, only seven factors out of 14 constructs from the temporary framework I consider to be significant. Now, the previous preliminary research model is re-generated with these seven significant factors in Figure 3-3:

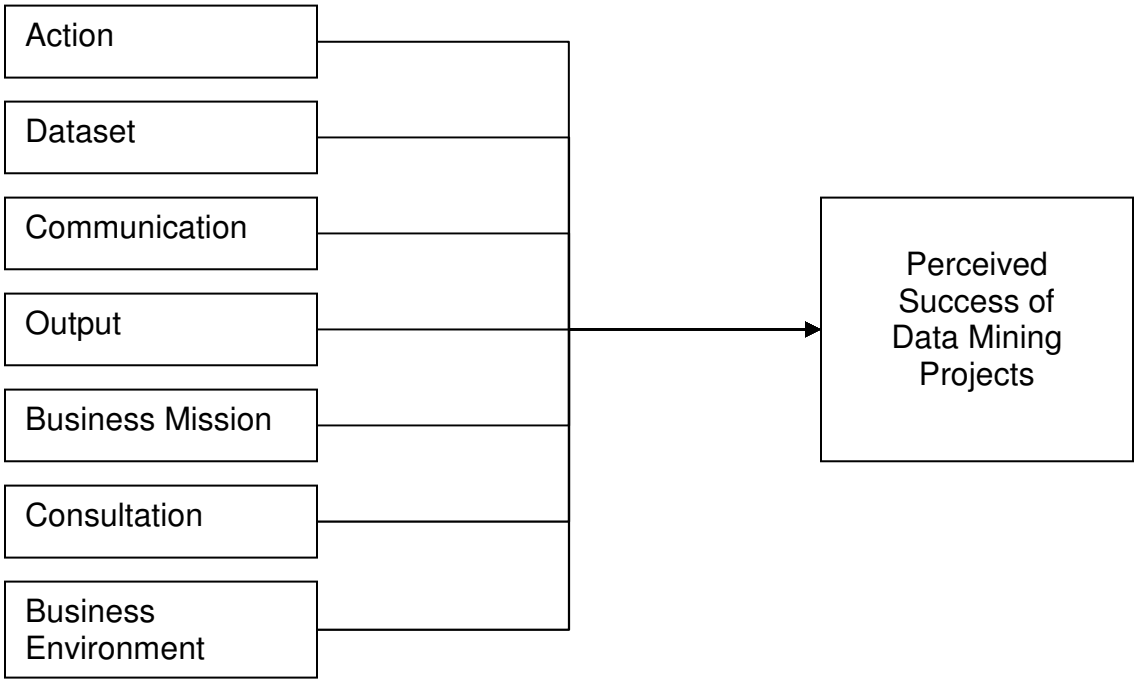

Figure 3-3: Modified Research Framework:

A Consolidation and Mapping of the Factors to Successful Data Mining Projects

Research Questions

This research addresses the overall question: What factors contribute to the successful evaluation of data mining projects? The specific research questions

assessing the contribution to the successful or unsuccessful evaluation of data mining projects are as follows:

- What factors contribute to the success of a data mining project?
- Does action factor contribute to the successful data mining projects?
- Does dataset factor contribute to the successful data mining projects?
- Does communication factor contribute to the successful data mining projects?
- Does output factor contribute to the successful data mining projects?
- Does business mission factor contribute to the successful data mining projects?
- Does consultation factor contribute to the successful data mining projects?
- Does business environment factor contribute to the successful data mining projects?
- What are the discriminating indicators for the successful data mining project?

Hypotheses

For testing purposes, I have reworded the research questions into 8 hypotheses as listed below. Hypotheses one through seven represent the various factors that the previous factor analysis indicated as important factors from all 57 possible success factor items:

H1: Action factor from the research framework is a critical success factor that contributes to the successful data mining projects.

H2: Dataset factor from the research framework is a critical success factor that contributes to the successful data mining projects.

H3: Communication factor from the research framework is a critical success factor that contributes to the successful data mining projects.

H4: Output factor from the research framework is a critical success factor that contributes to the successful data mining projects.

H5: Business mission factor from the research framework is a critical success factor that contributes to the successful data mining projects.

H6: Consultation factor from the research framework is a critical success factor that contributes to the successful data mining projects.

H7: Business environment factor from the research framework is a critical success factor that contributes to the successful data mining projects.

H8: A reduced set of indicators of data mining projects provide a better classification of successful or unsuccessful evaluation than the naive prediction rule.

Research Methodology

This research is a field survey because the researcher does not control the treatment but measures its impact. Data collection is done in a natural setting. Thus, this research provides a realistic description of existing systems.

I conducted this research in several phases: criteria development, data collection, and analysis. Each phase is explained below.

Instrument Development

The first phase resulted in the development of instruments that measure the independent and dependent variables as well as other demographic information. I derived the instrument for the independent constructs and variables through the consolidation of a “best set” of criteria, resulting in 14 constructs. This resulted in 57 questions representing the questions for the 14 constructs. The constructs and their

associated questions numbers, from which the questionnaire was developed, are located in Table 3-4 and Table 3-5. Questions in the survey instrument are arranged by their related constructs, so order bias may be present.

The dependent construct, data mining project success or failure, and variables are developed and represented in Table 3-6.

A panel of four experts critiqued and revised the surveys. The panel was instructed to address the topics of readability, clarity, consistency, reliability and construct validity. I revised the instruments based on the feedback from the panel regarding readability, clarity, consistency and reliability, as well as the organization and layout of the instruments.

Measurement

In this section I explain how the survey results are numerically computed and prepared for further statistical analysis. To measure independent variables, the following 7-point Likert scale that reflects the level of compliances to the question by the respondent:

0 = don't know or not applicable

1 = strongly disagree (that the item in this question is important for successful data mining projects)

2 = disagree

3 = somewhat disagree

4 = neutral

5 = somewhat agree

6 = agree

7 = strongly agree

Similarly, I also used a 7-point Likert scale to measure the dependent variable that assesses the level of success with a data mining project as reported by the respondent:

0 = don't know or not applicable

1 = very unsuccessful

2 = unsuccessful

3 = somewhat unsuccessful

4 = neutral

5 = somewhat successful

6 = successful

7 = very successful

Variables

The results for the dependent variables I consolidate into a discrete success or failure variable, that is, 2 for success and 1 for failure. The scores for the dependent variables for each observation are combined in order to obtain a single discreet value of success or failure. Observations scoring less than 4 are classified as failed. Those resulting in a score greater than 4 are classified as successful. A score of 4 represents a neutral response. Scores below 4 indicates noncompliance and are thus considered unsuccessful. Scores above 4 indicate compliance and are thus considered to be successful.

The groups of independent variables represent a predictor of success or failure at the aggregate level. Constructs are represented by several independent variables

making up their criteria grouping. All independent variables receive a score from zero to seven as indicated in measurement. The numerical score for the constructs consists of summing the responses to the questions that represent the construct. Based on the previously discussed numerical interpretation of the independent and dependent variable results, the survey is now in a valid format for t-test analysis, factor analysis, and constructs testing via an ANOVA procedure.

Sampling

Many data miners are currently involved in a data mining pilot project or have already deployed one or more data mining production systems. Data mining professionals often form user groups or e-mail forums on the Internet and share their knowledge and experience with each other. Some examples of such e-mail forums are Yahoo! Group, Nautilus Systems Datamine-L Discussion List, and Data Mining Group, etc.

For this research, I used the Nautilus Systems Datamine-L Discussion List, an unmoderated e-mail forum for discussing practical applications of data mining, data warehousing, and knowledge discovery. Nautilus Systems, Inc. is a business and computer consulting firm focusing on data mining and data warehousing and sponsors the e-mail forum for data miners.

Several participants in this user group agreed to participate in this research. The population is made up of a sample to find success factors in data mining projects.

Data Collection

I conducted the survey on-line and captured data over the Internet. I scored each criterion on a scale of 0-7 as depicted in Table 3-11. I loaded completed

responses from each participant into a Microsoft® Excel spreadsheet and imported to SPSS® for analysis. I calculated the dependent variables in Excel before I imported them into SPSS as a discrete 1 or 2. For example, if I used only three independent variable criteria versus the actual 57, the format of the data collected may appear as follows in Table 3-8.

Table 3-11: Sample Data Layout.

	Independent Variables			Dependent Variables	
	Criterion 1	Criterion 2	Criterion 3	User's Evaluation	User's Perception
Participant 1	5	1	4	1	1
Participant 2	4	2	4	2	1
Participant 3	3	7	4	2	2

Analyses

I used One-Sample t-Test, One-Way ANOVA, Factor Analysis, Decision Trees and Neural Network in this study.

The One-Sample t-Test procedure tests whether the mean of a single variable differs from a specified constant. This test may be used in the following situations. For example, a researcher might want to test whether the average IQ score for a group of students differs from 100. Or a cereal manufacturer might take a sample of boxes from the production line and check whether the mean weight of the samples differs from 1.3 pounds at the 95% confidence level.

The One-Way ANOVA procedure produces a one-way analysis of variance for a quantitative dependent variable by a single factor (independent) variable. Analysis of variance is used to test the hypothesis that several means are equal. This technique is

an extension of the two-sample t test. In addition to determining that differences exist among the means, a researcher may want to know which means differ. There are two types of tests for comparing means: *a priori* contrasts and *post hoc* tests. Contrasts are tests set up before running the experiment and *post hoc* tests are run after the experiment has been conducted. The researcher can also test for trends across categories. For example, doughnuts absorb fat in various amounts when they are cooked. An experiment is set up involving three types of fat: peanut oil, corn oil, and lard. Peanut oil and corn oil are unsaturated fats, and lard is a saturated fat. Along with determining whether the amount of fat absorbed depends on the type of fat used, the researcher could set up an *a priori* contrast to determine whether the amount of fat absorption differs for saturated and unsaturated fats.

Factor analysis attempts to identify underlying variables, or factors, that explain the pattern of correlations within a set of observed variables. Factor analysis is often used in data reduction by identifying a small number of factors which explain most of the variance observed in a much larger number of manifest variables. Factor analysis can also be used to generate hypotheses regarding causal mechanisms or to screen variables for subsequent analysis. The factor analysis procedure offers a high degree of flexibility. Seven methods of factor extraction are available. Five methods of rotation are available, including direct oblimin and promax for non-orthogonal rotations. Three methods of computing factor scores are available, and scores can be saved as variables for further analysis.

Decision trees and neural networks are two of numerous data mining analysis techniques. Decision trees are powerful and popular tools for classification and

prediction. Decision trees are used for directed data mining, particularly classification. They divide the records in the training set into disjoint subsets, each of which is described by a simple rule on one or more fields. One of the chief advantages of decision trees is the ability to effectively explain the model since it takes the form of explicit rules. This allows people to evaluate the results, identifying key attributes in the process. This is also useful where the incoming data is of uncertain quality – spurious results are obvious in the explicit rules. The rules themselves can be expressed easily as logic statements, in a language such as SQL, so they can be applied directly to new records (Berry and Linoff 1997).

Neural networks are simple models of neural interconnections in brains, adapted for use in digital computers. In their most common incarnation, they learn from a training set, generalizing patterns inside it for classification and prediction. One of the chief advantages of neural networks is their wide applicability. Neural networks are applicable “because: they can handle a wide range of problems, they produce good results even in complicated domains, they can handle both categorical and continuous variables, and they are available in many off-the-shelf packages” (Berry and Linoff 1997). To apply the neural network method in data mining technique, the input values should be massaged so they all lie between 0 and 1 because neural networks only handle values between 0 and 1. After training the model with sample data, this trained model will predict the success level of any new data

CHAPTER 4
ANALYSIS AND RESULTS

In this chapter, I present the chain of evidence for each research question and the quantitative instrument and survey results. In addition, I discuss how the data were analyzed to support research questions and conclusions, as well as the results of the hypothesis testing.

Web Survey

The surveys were completed via an electronic form on the Internet by 56 participants over three months, from December 1, 2001 through February 28, 2002. E-mail requests were circulated to the discussants in one of the data mining e-mail forums, the Nautilus Systems Datamine-L Discussion List. The response rate could not be assessed because number of discussants in the e-mail forum was not available.

IP address information in the message footer in the electronic form allowed the identification and elimination of duplicate responses. The mechanics of the survey included Microsoft® Front Page Form function to return the user's response to the web page upon completion and selection of the submit button.

I accumulated data from the web survey in a text database format during the survey collection process, and then exported from the text database to a Microsoft® Excel. I confirmed 57 variables, 12 demographics, two dependent variables, and comments. I converted data in the Excel spreadsheet to SPSS® statistical package format. Type, missing value, and column format were defined in SPSS data file. Once completed, I compiled descriptive statistics on key demographic factors, as well as summary data for user's evaluation and perception of data mining projects.

Demographics

I received and included a total of 56 responses in the results analysis. As with any survey tool, individual elements of the survey were completed at the participant's preference, resulting in varying numbers of responses (n) for the different demographic items and questions. 12 participants (21.4%) completed the free-form comment section.

General Demographics

100% of respondents reported gender, resulting in a return of 23.2% of female and 76.8% male.

As shown in the table below, 21 to 34 is the largest single age group represented, with 44.6%. 35 to 44 is the next largest single age group represented, with 30.36%. Thus, about 75% of all respondents are categorized into these two age groups, implicating the age characteristics of data miners at the Internet community.

Survey participants are well educated, with 96.4% reporting at least some college or college degrees. Many participants (71.4%) report master's degree and doctor's degree.

Most participants (82.1%) report that their level of management is higher or equal to middle level. More than half of participant (57.1%) report that their level of management is middle level. Only 17.9% of participants are categorized in operating level of management. I conclude data mining projects are conducted at the higher levels of management.

Table 4-1: Respondents' General Demographics

Variable	Number of Subjects	Percent
Gender		
Male	43	76.79
Female	13	23.21
Total	56	100.00
Age		
Under 21	0	0.00
21 to 34	25	44.64
35 to 44	17	30.36
45 to 54	11	19.64
55 to 64	3	5.36
65 or older	0	0.00
Total	56	100.00
Level of Education		
High School	1	1.79
Some College	1	1.79
College Graduate (4 yr)	13	23.21
Master's Degree	27	48.21
Doctorate Degree	13	23.21
Professional Degree (MD, JD, etc.)	0	0.00
Vocational/Technical School (2 yr)	0	0.00
Other	1	1.79
Total	56	100.00
Level of Management		
Top level	14	25.00
Middle level	32	57.14
Operating level	10	17.86
Total	56	100.00

Company Demographics

About a third (28.6%) of participants work in a company in which annual sales are less than 1 million dollars, and a fourth (25.0%) of participants work in a company in which annual sales are less than 5 million dollars and more than 1 million dollars. Thus, more than half (53.6%) of participants work in small companies with annual sales of less than 5 million dollars.

The number of employees in the respondents' companies is generally small. Half (50.0%) of respondents are working in companies of fewer than 50 employees.

About a fifth (21.43%) of respondents work in service company type. About half (48.2%) of respondents work in companies which are not categorized.

Table 4-2 summarizes the results for company demographics.

Table 4-2: Respondents' Company Demographics

Variable	Number of Subjects	Percent
Company's Annual Sales		
Less than $1 million	16	28.57
$1 million to less than $5 million	14	25.00
$5 million to less than $10 million	4	7.14
$10 million to less than $50 million	4	7.14
$50 million to less than $100 million	4	7.14
$100 million to less than $200 million	6	10.71
$200 million to less than $500 million	3	5.36
$500 million or more	5	8.93
Total	56	100
Number of Employees in Company.		
50 or less	28	50.00
51 to 100	0	0.00
101 to 500	5	8.93
501 to 1,000	5	8.93
1,000 to 5,000	13	23.21
Over 5,000	5	8.93
Total	56	100
Company Type		
Manufacturing	5	8.93%
Construction	0	0.00%
Service	12	21.43%
Finance/Insurance/Real Estate	8	14.29%
Sales	2	3.57%
Healthcare	2	3.57%
Transport/Communication/Utility	0	0.00%
Other	27	48.21%
Total	56	100

Data Mining Projects Demographics

14.3% of respondents have less than 1 year experience in data mining. 48.2% of respondents have more than 1 year but less than 5 years of experience. The remaining 37.5 % of respondents have more than 5 year experience.

The majority (32.1%) of respondents have completed more than 10 but fewer than 50 data mining projects. Only 5 of 56 respondents did not complete any data mining projects.

When respondents work on data mining projects, they report the need for an average of 3 to 5 people (48.2%) to complete a data mining project. An average of 1 to 2 people teams has seven occurrences, resulting in a value of 12.50%. Thus, about 60% of data mining projects need 1 to 5 members per data mining project. I conclude, from this information, that data mining projects are usually carried out by small groups of professionals.

These respondents work on various types of data mining projects, ranging from transportation/communication/utility to manufacturing. Most data mining projects were completed in the service area (26.8%), the finance/insurance/real estate area (30.4%), and the sales area (16.1%). A few data mining projects have been completed in the transportation/ communication/ utility area (8.9%) and healthcare area (5.4%). Only 1 respondent reported data mining projects in the manufacturing area, and none reported projects in the construction area.

Most datamine-l user group respondents (78.6%) reported projects that were completed in North America and Europe.

Table 4-3 summarizes results for data mining projects demographics.

Table 4-3: Respondents' Data Mining Projects Demographics

Variable	Number of Subjects	Percent
Working Years in Data Mining Area		
N/A	4	7.14
Less than 6 month	4	7.14
6 months to less than 1 yr	4	7.14
1 yr to less than 3 yr	10	17.86
3 yr less than to 5 yr	13	23.21
5 yr less than to 10 yr	12	21.43
10 yr or more	9	16.07
Total	56	100
Number of Data Mining Projects Involved		
N/A or none	5	8.93
1	3	5.36
2 to 3	11	19.64
4 to 5	4	7.14
6 to 10	9	16.07
11 to 50	18	32.14
51 or more	6	10.71
Total	56	100
Number Of Average People Working for Data Mining Projects		
No project done	6	10.71
1 to 2	7	12.50
3 to 5	27	48.21
6 to 10	8	14.29
11 to 50	6	10.71
51 to 100	2	3.57
101 to 500	0	0.00
501 or more	0	0.00
Total	56	100
Company Type for Data Mining Projects		
No project done	6	10.71
Manufacturing	1	1.79
Construction	0	0.00
Service	15	26.79
Finance/Insurance/Real Estate	17	30.36
Sales	9	16.07
Healthcare	3	5.36
Transport/Communication/Utility	5	8.93
Total	56	100
Location of Data Mining Projects Done		
No project done	6	10.71

in Africa	0	0.00
in Asia	3	5.36
in Europe	19	33.93
in North America	25	44.64
in South America	3	5.36
in Australia	0	0.00
Total	56	100

Evaluation of Data Mining Projects

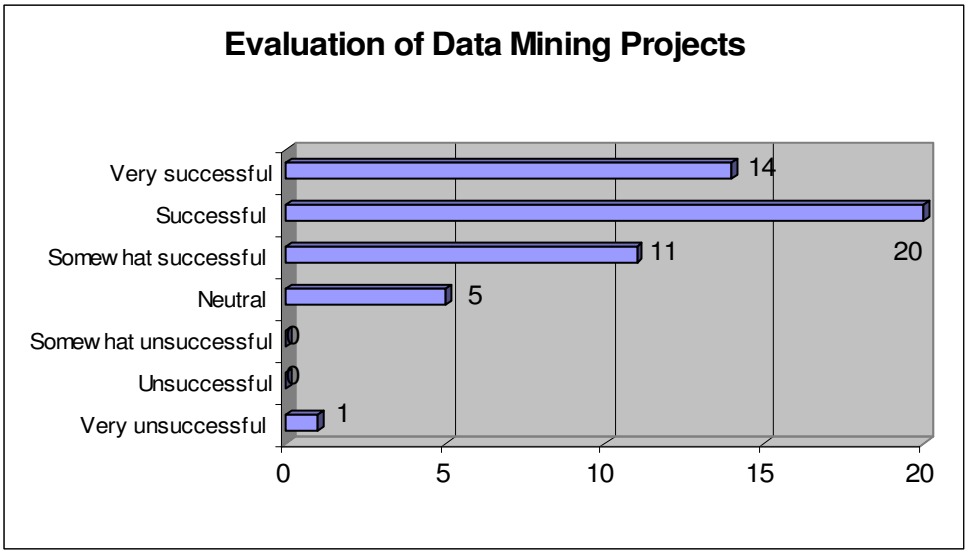

Figure 4-1: Evaluation of Data Mining Projects Chart

Figure 4-1 shows the respondents' evaluation of the success of their data mining projects. Only one respondent reported a very unsuccessful completion of data mining project(s). Five respondents (8.9%) reported that their projects were neither successful nor unsuccessful. All other respondents (80.4%) reported that their projects were completed successfully: 14 respondents reported that their projects were completed very successfully, 20 reported successfully, and 11 reported somewhat successfully.

Familiarity with Data Mining Area

The last question in the web survey regarded respondent familiarity with the data mining area as shown in Figure 4-2. Three respondents (5.4%) answered that they are very unfamiliar to the data mining area. The remaining 53 respondents answered that they were familiar with the data mining area.

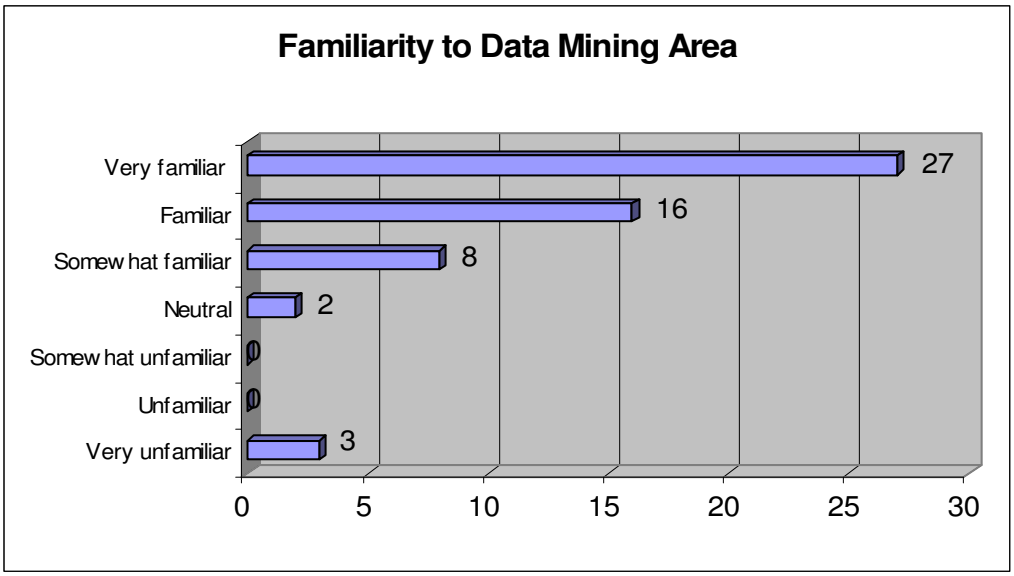

Figure 4-2: Familiarity to Data Mining Area Chart

The three respondents who were unfamiliar to data mining area I have dropped from the next analyses, leaving a sample size of 53 instead of 56.

Descriptive Analysis

In this section, I analyzed each variable by the one-sample t-test to determine whether the mean response of each variable is significantly higher than the test value.

Table 4-4: The Result of One-Sample t-Test with 57 Variables (Test Value = 4)

Variable	Value of t	Df	p-value	Mean Response
V01	13.2955	52	0.0000	6.2264
V02	8.9131	51	0.0000	5.6731
V03	10.0217	49	0.0000	5.6600
V04	19.684	51	0.0000	6.4423
V05	13.9043	52	0.0000	6.0943
V06	7.8158	48	0.0000	5.5306
V07	5.2162	49	0.0000	5.1800
V08	-0.3477	49	**0.6352**	**3.9200**
V09	5.3116	51	0.0000	5.0000
V10	5.4444	51	0.0000	5.1154
V11	11.575	50	0.0000	5.8627
V12	11.8298	51	0.0000	5.8846
V13	4.6463	48	0.0000	5.1224
V14	6.2655	50	0.0000	5.0588
V15	6.3416	51	0.0000	5.2692
V16	2.8392	50	0.0033	4.6863
V17	2.0814	48	0.0214	4.4898
V18	9.0559	50	0.0000	5.5686
V19	10.4173	50	0.0000	5.7647
V20	7.141	52	0.0000	5.2453
V21	5.4283	51	0.0000	5.0000
V22	6.7674	48	0.0000	5.4082
V23	0.4546	48	**0.3257**	4.1020
V24	14.6665	52	0.0000	5.9057
V25	9.2594	52	0.0000	5.6038
V26	5.1309	47	0.0000	5.1042
V27	11.0663	50	0.0000	5.5294
V28	8.7204	50	0.0000	5.5686
V29	14.8754	50	0.0000	6.0392
V30	6.9464	51	0.0000	5.2115
V31	4.2838	48	0.0000	4.9592
V32	2.9056	50	0.0027	4.7451
V33	5.6258	49	0.0000	5.3200
V34	-0.0728	50	**0.5289**	**3.9804**
V35	0.4592	49	**0.3240**	4.1200
V36	-0.5583	48	**0.7104**	**3.8571**
V37	7.8514	51	0.0000	5.7500
V38	4.803	49	0.0000	5.1400
V39	-0.7147	48	**0.7609**	**3.8163**
V40	4.3137	49	0.0000	5.0200

V41	0.4007	48	**0.3452**	4.1020
V42	4.398	50	0.0000	4.8235
V43	-0.0678	50	**0.5269**	3.9804
V44	11.5643	50	0.0000	6.0588
V45	7.8286	49	0.0000	5.6400
V46	5.4807	50	0.0000	5.2549
V47	5.3663	45	0.0000	5.2391
V48	14.9278	50	0.0000	6.2353
V49	15	48	0.0000	6.1429
V50	4.7821	50	0.0000	5.0784
V51	5.2853	47	0.0000	5.2083
V52	11.6587	51	0.0000	6.1154
V53	5.9934	48	0.0000	5.4082
V54	10.1014	50	0.0000	6.0392
V55	7.5796	51	0.0000	5.3654
V56	12.3997	51	0.0000	5.9038
V57	6.7556	51	0.0000	5.4038

Because this study uses the 7-point Likert scale in independent variables, a sample mean of a certain variable significantly higher than a neutral value of 4 I regard as successful.

If a certain independent variable is not significantly higher than the test value, in other words, if the p-value is higher than 0.05, then I consider that variable unimportant to achieving success in data mining projects. Five variables have lower mean responses, and eight variables, including five variables (bold font style) of lower mean responses, have insignificant p-value. Thus, I consider these eight variables as less successful than the other variables. These eight variables are **V8,** V23**, V34,** V35**, V36, V39,** V41**, and V43**. All other variables are significant and, thus, I consider them as successful variables for data mining projects.

From this analysis, I found that respondents of the datamine-l user group consider the following variables as unimportant to the success of a data mining project:

Table 4-5: Code Interpretation of Demographics.

Code	Demographic Information (Measurement)
R1	Level of management (Top level, Middle level, Operating level)
R2	Gender (Male, Female)
R3	Age (Under 21, 21 to 34, 35 to 44, 45 to 54, 55 to 64, 65 or older)
R4	Level of education (High School, Some College, College Graduate (4 yr), Master's Degree, Doctorate Degree, Professional Degree (MD, JD, etc.), Vocational/Technical School (2 yr), Other)
O1	Company's annual sales (Less than $1 million, $1 million to less than $5 million, $5 million to less than $10 million, $10 million to less than $50 million, $50 million to less than $100 million, $100 million to less than $200 million, $200 million to less than $500 million, $500 million or more)
O2	Number of employees in your company (50 or less, 51 to 100, 101 to 500, 501 to 1,000, ,000 to 5,000, Over 5, 000)
O3	Company type (Manufacturing, Construction, Service, Finance/Insurance/Real Estate, Sales, Healthcare, Transport/Communication/Utility, Other)
P1	Working years in data mining area (N/A, Less than 6 month, 6 months to less than 1 yr, 1 yr to less than 3 yr, 3 yr less than to 5 yr, 5 yr less than to 10 yr, 10 yr or more)
P2	The number of data mining project(s) involved (N/A or none, 1, 2 to 3, 4 to 5, 6 to 10, 11 to 50, 51 or more)
P3	The number of average people working for data mining projects (No project done, 1 to 2, 3 to 5, 6 to 10, 11 to 50, 51 to 100, 101 to 500, 501 or more)
P4	The company type of data mining projects developed (No project done, Manufacturing, Construction, Service, Finance/Insurance/Real Estate, Sales, Healthcare, Transport/Communication/Utility)
P5	Area of data mining projects completed (No project done, in Africa, in Asia, in Europe, in North America, in South America, in Australia)
P6	Evaluation of data mining projects (No project done, Very unsuccessful, Unsuccessful, Somewhat unsuccessful, Neutral, Somewhat successful, Successful, Very successful)
Perc.	Familiarity to data mining area (Very unfamiliar, Unfamiliar, Somewhat unfamiliar, Neutral, Somewhat familiar, Familiar, Very familiar)
Eval.	Evaluation of data mining projects in discrete number (Unsuccessful, Successful)

a limited budget (V8), applying state-of-the-art technology to the business (V23), tuning databases to meet operational needs (V34), defining implementation strategies for databases (V35), recovering databases (V36), not-complex dataset (v39), scalability of database (V41), and selecting the right type of data source (V43).

Because these variables were considered insignificant, they could have been removed before further analyses. As six of eight variables were already eliminated when I consolidated these factors in the previous chapter, the effect of keeping these variables for further analyses is insignificant.

Demographics Analysis

In this section, I examine the relationship between demographics, variables, and factors.

The demographics of the web survey comprised three parts: respondent information, company information of the respondent, and project information of the respondent, shown in Table 4-5.

Factors ANOVA

Seven factors were tested with one-way ANOVA using demographic information. Table 4-6 provides a summary of p-value of all ANOVA results with factors. Bold font styles are significant.

For example, a p-value of 0.0333 in F1 and R1 indicates that the means of factor 1 are not the same in different levels of management. In other words, this p-value

Table 4-6: P-value Summary of Factors ANOVA

ANOVA	F1	F2	F3	F4	F5	F6	F7
R1	**0.0333**	0.5670	0.7515	0.7817	0.9875	0.3349	0.3374
R2	**0.0470**	0.8288	0.4914	0.6266	0.2091	0.9456	0.7742
R3	0.6463	0.9205	0.8241	0.2192	0.9478	0.6719	0.1455
R4	0.8617	0.1611	0.2264	0.5083	0.1156	0.6466	0.2870
O1	0.9200	0.5484	0.7451	0.9420	0.4437	0.0789	0.3812
O2	0.3700	0.8259	0.5924	0.8515	0.9520	**0.0291**	0.3026
O3	**0.0339**	**0.0033**	0.1540	0.4753	0.3182	0.5336	0.8418
P1	0.3488	0.5442	0.9104	0.4965	0.4832	0.5254	0.2305
P2	0.0671	0.1761	0.2092	0.7045	0.0770	0.6641	0.3197
P3	0.3340	0.9479	0.9960	0.7457	0.9022	0.7926	0.5136
P4	0.1710	0.4241	0.4800	0.8411	0.4728	0.7246	0.1882
P5	0.2192	0.0502	**0.0477**	0.0850	0.8101	0.4292	0.0982
P6	0.0637	**0.0263**	0.9498	0.7208	0.5731	0.3581	0.5213
Perc	0.3902	0.4964	**0.0087**	0.1230	0.4217	0.7125	0.1864
Eval	0.5266	**0.0081**	0.9644	0.9794	0.3846	0.2982	0.9213

signifies that statistically there are differences in the level of management with regard to Action (factor 1) as a successful major factor for data mining projects.

As shown in Table 4-7, the mean of Factor 1 in Group 1 (Top level managers) I found to be significantly lower than that in Group 2 (middle level managers). In other words, top level managers consider Factor 1 (Action) as a less important factor than the middle level counterparts. Factor 1 is the compilation of several variables including V31 (Developing logical data models), V32 (Integrating logical data models into an enterprise data model), V50 (Visualization of the extracted patterns), V51 (Removing redundant or irrelevant patterns), V53 (Incorporating the knowledge into the performance system), V54 (Taking actions based on the extracted knowledge), V55 (Documenting the extracted knowledge), and V57 (Checking for and resolving potential conflicts with previously believed knowledge).

Table 4-7: Descriptive Analysis for Factor 1 and Level of Management.

	Level of Management	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
						Lower Bound	Upper Bound
F1	1	10	36.5000	11.60699	3.67045	28.1969	44.8031
	2	26	45.2692	7.68665	1.50748	42.1645	48.3739
	3	4	42.7500	4.50000	2.25000	35.5895	49.9105
	Total	40	42.8250	9.20671	1.45571	39.8806	45.7694

Factor 1 (Action) was perceived differently by management level (R1), age (R2), and company type (O3). Factor 2 (Dataset) scores varied by company type (O3) and data mining project group (P6). Factor 3 (Communication) score differed based on the area of data mining projects completed (P5) and the level of familiarity with data mining (PERC). Factor 6 (Consultation) scores appeared to vary with the size of a company as measured by the number of employees (O2). Factors 4, 5, and 7, (Output, Business Mission, and Business Environment) did not appear to be influenced by any demographic variables. Even though these findings seem to be interesting and warrant further investigation as to their meaning and causes, I considered this to be beyond the scope of this study and left it for future studies.

Variables ANOVA

I tested all 57 variables with one-way ANOVA using demographic information. Table 4-8 provides the summary of the p-values of all ANOVA results with variables. Values in bold font styles are significant.

For example, a p-value of 0.0071 in V1 (A clear mission) and R4 (Level of education) indicates that the means of V1 are not same in different groups of R4. In

Table 4-8: P-value Summary of Variables ANOVA

ANOVA	R1	R2	R3	R4	C1	C2	C3	P1	P2	P3	P4	P5	P6	Perc
V1	0.6549	0.5440	0.9432	**0.0071**	0.2542	0.9283	0.3879	0.8396	0.1468	0.6419	0.7643	0.8341	0.2327	0.3490
V2	0.7378	0.0984	0.8728	0.2626	0.5863	0.8710	0.3190	0.1625	0.5588	0.9557	0.3335	0.8037	0.8536	0.3147
V3	0.3557	0.0946	0.4871	0.0531	0.8768	0.1455	0.1416	0.9161	0.1352	0.6341	0.1666	0.2165	0.0510	0.4897
V4	0.9007	0.2379	0.9015	0.6041	0.4918	0.5866	0.7120	0.2689	**0.0128**	0.1250	0.0787	0.2209	**0.0384**	**0.0001**
V5	0.3007	0.5286	0.5223	**0.0282**	0.2297	0.9968	0.5971	0.1260	0.0809	0.8816	0.9014	0.7639	0.7299	**0.0220**
V6	0.2896	0.9680	0.1630	0.5875	0.4137	0.1353	0.3392	0.3620	0.6795	0.5438	**0.0129**	0.4435	0.7612	0.2043
V7	0.3358	0.6709	0.3922	0.8150	0.5638	0.1521	0.4025	0.5182	0.3725	0.6237	0.2576	0.3386	0.3516	0.3657
V8	0.7611	0.4211	0.2932	0.9805	0.9745	0.5589	0.7487	0.2511	0.3714	0.6939	0.3153	**0.0296**	**0.0143**	0.1158
V9	0.8886	0.8112	**0.0375**	0.7733	0.2192	**0.0241**	0.6955	0.1358	0.2778	0.5763	0.5435	0.0821	0.0614	**0.0273**
V10	0.2661	0.0899	0.0512	0.4952	0.1874	0.0515	0.1933	0.1067	0.1579	0.9543	0.4021	0.4704	0.2892	0.2246
V11	0.3935	0.9205	0.3842	0.3488	**0.0322**	0.1036	0.5254	0.0635	0.8568	0.8348	0.7277	0.3290	0.6244	0.6464
V12	0.4569	0.3371	0.7789	0.6303	0.4355	0.1543	0.0536	**0.0009**	0.9428	0.7483	0.4472	**0.0236**	0.4647	**0.0014**
V13	0.2203	0.5965	0.2418	0.6725	0.7827	0.3573	**0.0293**	0.0607	0.3246	0.3652	0.9474	0.5624	**0.0336**	**0.0007**
V14	0.4021	0.6469	0.5162	0.1290	0.1824	0.2788	0.0590	0.4217	0.1032	0.0831	0.2293	0.1385	0.3693	0.5262
V15	0.1515	0.7821	0.4386	0.0533	0.5046	0.2919	0.6871	0.9439	0.2957	**0.0462**	0.5982	0.1061	0.3609	0.4444
V16	0.7850	0.5270	0.1692	0.0798	0.8418	0.9129	0.0657	0.6541	**0.0197**	0.3928	0.7160	**0.0080**	0.7987	**0.0199**
V17	**0.0453**	0.5345	0.0835	0.4251	0.1042	0.7946	0.1923	0.0620	**0.0387**	0.0502	0.0961	**0.0438**	**0.0162**	0.6231
V18	**0.0419**	0.1695	0.5952	0.6951	0.3856	0.5201	0.1260	**0.0163**	0.0786	0.1674	0.5824	0.0797	**0.0087**	0.2412
V19	0.6340	0.6595	0.9482	0.3467	0.6406	0.1867	0.8474	0.7466	0.1933	0.6429	0.8486	0.3808	0.1471	0.6875
V20	0.8219	0.2988	0.2873	0.2974	0.7437	0.5190	0.7887	0.5525	**0.0081**	0.7683	0.2957	0.3448	**0.0313**	0.7060
V21	0.7988	0.4628	0.4565	0.7375	0.8267	0.6935	0.9342	0.1771	0.0579	0.3302	0.8973	0.0562	0.0603	0.8673
V22	0.7882	0.1984	0.1932	0.7099	0.8746	0.6736	0.8082	0.6312	0.4981	0.8339	0.6583	0.5936	0.7669	0.0831
V23	**0.0346**	0.2035	0.1328	0.8965	0.0992	0.4542	0.9277	0.7985	0.2718	0.7580	**0.0217**	**0.0263**	0.0530	0.5263
V24	0.6513	0.2814	**0.0376**	0.4239	0.8709	0.7495	0.6022	0.3557	0.9887	0.2886	0.3453	0.5001	0.7527	0.0729
V25	0.4988	0.3332	0.6352	0.2013	0.5294	0.8071	0.1933	0.2152	**0.0144**	0.4938	0.8379	0.0918	0.3192	**0.0031**
V26	0.9741	0.8465	0.3700	0.8655	0.9300	0.7513	0.4695	0.7295	0.4796	0.9967	0.3072	0.0771	0.2178	**0.0091**
V27	0.0600	0.2261	0.8045	0.7830	0.9610	0.9378	0.1781	0.8109	0.5255	0.7550	0.7548	0.1720	0.7967	0.7890
V28	0.9889	0.4199	0.0892	0.6955	0.8206	0.8760	0.5991	0.3401	0.9585	0.7409	0.9850	0.1040	0.4311	0.0942
V29	0.3326	0.8604	0.2637	0.1893	0.9569	0.8062	0.2845	0.5407	0.5780	0.4218	0.5662	**0.0467**	0.3197	0.2223
V30	0.6627	0.7061	0.3530	0.0788	0.9073	0.2074	0.5729	0.1879	0.4544	0.4590	0.5418	0.1414	**0.0289**	0.8933

Table 4-8: P-value Summary of Variables ANOVA, Continued

ANOVA	R1	R2	R3	R4	C1	C2	C3	P1	P2	P3	P4	P5	P6	Perc
V31	0.5049	0.5980	0.6079	0.6525	0.6185	0.4531	0.3544	0.2756	0.2441	0.6601	0.9055	0.1223	0.5240	0.5075
V32	**0.0361**	0.0693	0.0840	0.7529	0.6176	0.5372	0.0803	**0.0290**	**0.0306**	0.6474	0.2214	**0.0135**	0.0833	0.3796
V33	0.1511	0.1852	0.2497	0.8753	0.8234	0.6844	0.4745	0.0688	0.6750	0.4190	0.1475	0.4110	0.0764	0.4417
V34	0.8673	0.9683	0.0771	0.1320	0.6643	0.5593	0.1048	0.5898	0.4253	0.8594	0.6617	0.1196	**0.0163**	0.1785
V35	0.9650	0.1780	0.1724	0.3807	0.5948	0.9230	**0.0231**	0.7302	0.1106	0.2747	0.1970	0.2427	0.1042	0.1399
V36	0.5974	0.5006	0.1287	0.1198	0.8256	0.8752	0.1358	0.7190	0.0728	0.6634	0.2595	0.0958	**0.0146**	0.0535
V37	0.4194	0.4181	0.4981	0.8192	0.0883	0.0752	0.1437	0.6768	0.1064	**0.0003**	0.7257	0.8228	0.5252	0.3049
V38	0.6478	0.8949	0.9545	0.3304	0.2370	0.9051	**0.0140**	0.2448	0.2657	0.9028	0.5384	0.3837	0.3573	0.8940
V39	0.2562	0.7440	0.4124	0.6631	0.7191	0.5492	0.0895	0.5018	0.0940	0.9810	0.6428	**0.0132**	0.1033	0.2742
V40	0.1629	0.8089	0.5248	0.3678	0.7886	0.6170	0.4747	0.5025	0.4077	0.2376	0.3042	0.2523	0.4078	0.0913
V41	0.3016	0.4622	0.8081	0.5403	0.3418	0.6197	**0.0340**	0.4234	**0.0474**	0.8693	0.4019	**0.0099**	**0.0255**	0.2891
V42	0.3894	0.7857	0.8864	0.2087	0.6379	0.3077	**0.0013**	0.6562	0.6887	0.6804	0.5031	0.1870	0.1660	0.8882
V43	0.0997	0.6098	0.1770	0.0563	0.4179	0.5400	0.9658	0.0509	0.0535	0.6024	0.0606	0.0623	0.1519	0.0840
V44	0.4298	0.2477	0.8997	0.8715	0.6509	0.5055	0.0602	0.2062	0.1643	0.4485	**0.0213**	0.6448	0.1340	0.4488
V45	0.5697	**0.0358**	0.4699	0.8739	0.6032	0.3522	0.2587	0.1925	0.0729	0.9474	0.0551	0.2384	**0.0027**	0.1061
V46	0.1066	0.2341	0.3060	**0.0449**	0.3566	0.8679	0.6265	0.6372	0.5047	0.4330	0.2643	0.0740	0.1647	0.3784
V47	0.7175	0.4633	0.0699	0.7468	0.8466	0.9785	0.5854	0.3278	0.3795	0.0768	**0.0405**	0.1505	**0.0110**	0.9414
V48	0.4668	0.3318	0.3618	0.6327	0.3357	0.3895	0.2353	0.6543	0.1372	0.8355	0.3195	0.5109	0.1343	0.7308
V49	0.1648	0.5745	0.5462	**0.0025**	0.1058	0.4178	0.1428	0.9828	0.2830	0.7022	0.5699	0.3352	0.8985	0.4721
V50	**0.0495**	0.0990	0.2404	0.7524	0.6294	0.8580	0.2700	0.4707	0.0822	0.4085	0.2620	0.0601	**0.0274**	0.6913
V51	**0.0284**	**0.0139**	0.1032	0.5191	0.8080	0.3735	0.4515	0.8919	0.2324	0.9965	**0.0078**	0.3660	0.0567	0.8846
V52	0.6091	0.6887	0.9440	0.4879	0.2859	0.5445	0.8281	0.6356	0.2465	0.4809	0.8197	0.7877	0.0925	0.1507
V53	0.0617	**0.0271**	0.3384	0.6535	0.5279	0.2454	0.5848	0.8973	0.4131	0.6958	**0.0349**	0.8296	**0.0470**	0.2811
V54	0.1038	0.3044	0.2394	0.6153	0.7366	0.5527	0.0700	0.5969	0.5704	0.1966	0.1466	0.2945	0.5210	0.0743
V55	0.1539	0.2460	0.3332	0.6026	0.7822	0.2748	**0.0120**	0.0561	0.1327	0.0806	0.1150	0.6581	0.1089	0.2121
V56	**0.0276**	0.5273	0.4296	0.4704	0.7552	0.3177	0.1693	0.1107	0.4874	0.1070	0.1941	0.5053	0.1851	0.4362
V57	0.1209	0.2617	0.4232	0.5589	0.6327	0.3425	0.3954	0.0830	0.1796	0.3942	0.3113	0.2004	0.2867	**0.0403**

Table 4-9: Descriptive Analysis for Variable 1 and Level of Education.

	Level of Education	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
						Lower Bound	Upper Bound
V1	1. High School	1	3.00	.	.	.	.
	2. Some College	1	7.00	.	.	.	.
	3. College Graduate (4 yr)	11	5.45	1.81	.55	4.24	6.67
	4. Master's Degree,	26	6.38	.90	.18	6.02	6.75
	5. Doctorate Degree	13	6.69	.48	.13	6.40	6.98
	8. Other	1	7.00	.	.	.	.
	Total	53	6.23	1.22	.17	5.89	6.56

other words, if the educational level of the respondents is higher, they gave higher points to a clear mission as shown in Table 4-9. Likewise, a clear mission is a more important variable to respondents with a doctoral degree than those with a master's or bachelor's degree. I speculate that managers with a higher level of education tend to hold offices at a higher level, and consequently are more concerned about the high-level corporate mission than their lower-level counterparts.

Table 4-10 summarizes Table 4-8. Each demographic category shows differences in several variables from 1 to 14 variables. Different numbers of employees in the company (C2) have different means of realistic project schedule (V09), only. However, evaluation of data mining projects (P6) has different effects on 14 variables.

Analyzing Hypotheses 1 Through 7

In this ANOVA analysis, I ensured that each factor had a different success point with regards to the three evaluation groups, with values for very successful, successful, and somewhat successful.

Table 4-10, Number of Significant Variables by Demographics

Demographics	Variables	Count
Level of management (R1)	V17, V18, V23, V32, V50, V51, V56	7
Gender (R2)	V45, V51, V53	3
Age (R3)	V09, V24	2
Level of education (R4)	V01, V05, V46, V49	4
Company's annual sales (C1)	V11	1
Number of employees in the company (C2)	V09	1
Company type (C3)	V13, V35, V38, V41, V42, V55	6
Working years in data mining area (P1)	V12, V18, V32	3
The number of data mining project(s) involved (P2)	V04, V16, V17, V20, V25, V32, V41	7
The number of average people working for data mining projects (P3)	V15, V37	2
The company type of data mining projects developed (P4)	V06, V23, V44, V47, V51, V53	6
Area of data mining projects completed (P5)	V08, V12, V16, V17, V23, V29, V32, V39, V41	9
Evaluation of data mining projects (P6)	V04, V08, V13, V17, V18, V20, V30, V34, V36, V41, V45, V47, V50, V53	14
Familiarity to data mining area(Perc)	V04, V05, V09, V12, V13, V16, V25, V26, V57	9

Table 4-11 shows a basic frequency analysis of a dependent variable (Perceived Success) that is a recoded variable of P6 (Evaluation of data mining projects). The dependent variable, P6 (Evaluation of data mining projects), uses the 7-point Likert scale ranging from very unsuccessful (1) to very successful (7). It would be ideal to have two groups -- successful group and unsuccessful group -- for comparison to see if the success with a data mining project has any bearing on other study variables. However, doing so by splitting the sample in the middle of the scale would result in serious imbalance between two groups. Instead, I classified the respondents into three more balanced groups.

Table 4-11. Frequency of Perceived Success Variable

	Perceived Success	Frequency	Percent	Valid Percent	Cumulative Percent
Somewhat Successful	5	15	28.3	31.3	31.3
Successful	6	19	35.8	39.6	70.8
Very Successful	7	14	26.4	29.2	100.0
	Total	48	90.6	100.0	
Missing	System	5	9.4		
	Total	53	100.0		

In the recoding process, I removed all unsuccessful evaluation values (1 through 3) of P6 variable, and combined the neutral and the somewhat successful evaluation values (4 and 5) into a somewhat successful evaluation value of 5. The values for successful (6) and very successful (7) were not changed.

Factors ANOVA

Table 4-12 shows the results of the ANOVA analysis. From seven factors, only one factor, Factor 2, has different values among the three groups.

I have provided a summary of each hypothesis' result. Hypotheses H1 through H7 were tested using an ANOVA procedure; Table 4-13 summarizes the results.

Hypotheses 1 through 7 test seven factors with different evaluation groups. Three evaluation groups were developed from the P6 (average evaluation of data mining projects) demographic, including group 5 (evaluating their data mining projects as neutral or somewhat successful), group 6 (evaluating their data mining projects as successful), and group 7 (evaluating their data mining projects as very successful).

Hypothesis 1. Factor 1, Action, was analyzed by three different evaluation groups. Its analyzed p-value is 0.062, which is not significant even though its p-value is

Table 4-12. ANOVA Analysis with Factor and Perceived Success Group

Factor		Sum of Squares	df	Mean Square	F	Sig.
F1	Between Groups	473.036	2	236.518	3.023	.062
	Within Groups	2738.438	35	78.241		
	Total	3211.474	37			
F2	Between Groups	426.890	2	213.445	5.196	**.010**
	Within Groups	1561.061	38	41.081		
	Total	1987.951	40			
F3	Between Groups	13.166	2	6.583	.354	.704
	Within Groups	706.931	38	18.603		
	Total	720.098	40			
F4	Between Groups	4.348	2	2.174	.478	.623
	Within Groups	190.852	42	4.544		
	Total	195.200	44			
F5	Between Groups	3.072	2	1.536	.279	.758
	Within Groups	242.162	44	5.504		
	Total	245.234	46			
F6	Between Groups	11.919	2	5.960	1.539	.226
	Within Groups	162.659	42	3.873		
	Total	174.578	44			
F7	Between Groups	17.847	2	8.923	.989	.381
	Within Groups	370.040	41	9.025		
	Total	387.886	43			

close to 0.05. Thus, hypothesis H1 is not supported. The *Action* factor does not differ across the three groups.

Hypothesis 2. Factor 2, Dataset, was analyzed by three different evaluation groups. Its analyzed p-value is 0.010, which is significant. Group 5 has a higher mean than Group 6 and Group 7; this means that Factor 2, Dataset, is a more important factor to the somewhat successful evaluation group than to the other groups. Thus, hypothesis H2 is accepted. There is a difference in the *Dataset* factor across the three groups.

Factor 2, Dataset, was found to be significantly different by the three evaluation groups. As shown in Table 4-14, the evaluation group that evaluates their data mining

Table 4-13: Summary of Results from Hypotheses 1 through 7

	Hypothesis	F Value	Significance
H1	Action factor from the research framework is a critical success factor that contributes to the successful data mining projects.	3.023	.062 Not Significant
H2	Dataset factor from the research framework is a critical success factor that contributes to the successful data mining projects.	5.196	**.010** Significant
H3	Communication factor from the research framework is a critical success factor that contributes to the successful data mining projects.	.354	.704 Not Significant
H4	Output factor from the research framework is a critical success factor that contributes to the successful data mining projects.	.478	.623 Not Significant
H5	Business mission factor from the research framework is a critical success factor that contributes to the successful data mining projects.	.279	.758 Not Significant
H6	Consultation factor from the research framework is a critical success factor that contributes to the successful data mining projects.	1.539	.226 Not Significant
H7	Business environment factor from the research framework is a critical success factor that contributes to the successful data mining projects.	.989	.381 Not Significant

projects as somewhat successfully consider Factor 2 as more important than the other two groups. The "successful" evaluation group has a higher mean than the "very successful" evaluation group.

Five variables -- V36 (Recovering databases), V38 (Easily accessible data), V39 (NOT complex dataset), V41 (Scalability of database), and V42 (Maintaining availability of dataset) -- make up Factor 2 as shown in Table 4-15. With regards to variable V36, I noticed a significant difference between Group 5 and Group 6&7. The two groups also

Table 4-14. Descriptive Analysis of Factor 2

F2	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
5	14	25.3571	6.1219	1.6362	21.8224	28.8918	14.00	32.00
6	16	20.9375	6.6178	1.6545	17.4111	24.4639	10.00	35.00
7	11	17.0909	6.4568	1.9468	12.7531	21.4287	5.00	27.00
Total	41	21.4146	7.0497	1.1010	19.1895	23.6398	5.00	35.00

differ with the variable V41. On the other hand, Group 5&6 and Group 7 showed different responses to V42.

Hypothesis 3. Factor 3, Communication, was analyzed by three different groups. Its analyzed p-value is 0.704, which is not significant. Thus, hypothesis H3 is not supported. The *Communication* factor does not differ across the three groups.

Hypothesis 4. Factor 4, Output, was analyzed by the different groups. Its analyzed p-value is 0.623, which is not significant. Thus, hypothesis H4 is not supported. The *Output* factor does not differ across the three groups.

Hypothesis 5. Factor 5, Business Mission, was analyzed by the different groups. Its analyzed p-value is 0.758, which is not significant. Thus, hypothesis H5 is not supported. The *Business Mission* factor does not differ across the three groups.

Hypothesis 6. Factor 6, Consultation, was analyzed by the different groups. Its analyzed p-value is 0.226, which is not significant. Thus, hypothesis H6 is not supported. The *Consultation* factor does not differ across the three groups.

Hypothesis 7. Factor 7, Business Environment, was analyzed by the different

Table 4-15. Descriptive Analysis of Variables in Factor 2

		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
						Lower Bound	Upper Bound		
V36. Recovering databases	5 a	14	4.79	1.672	.447	3.82	5.75	1	7
	6 b	17	3.41	1.543	.374	2.62	4.21	1	7
	7 b	13	3.15	1.772	.492	2.08	4.22	1	7
	Total	44	3.77	1.764	.266	3.24	4.31	1	7
V38. Easily accessible data	5	14	5.79	1.626	.434	4.85	6.72	1	7
	6	19	4.95	1.433	.329	4.26	5.64	2	7
	7	13	4.46	1.941	.538	3.29	5.63	1	7
	Total	46	5.07	1.692	.249	4.56	5.57	1	7
V40. Sufficient number of dataset	5	13	5.15	1.345	.373	4.34	5.97	2	7
	6	18	4.89	1.967	.464	3.91	5.87	2	7
	7	14	4.71	1.684	.450	3.74	5.69	2	7
	Total	45	4.91	1.690	.252	4.40	5.42	2	7
V41. Scalability of database	5 a	14	4.79	1.424	.381	3.96	5.61	2	7
	6 b	18	3.72	1.776	.419	2.84	4.61	2	7
	7 b	12	3.17	1.697	.490	2.09	4.24	1	6
	Total	44	3.91	1.736	.262	3.38	4.44	1	7
V42. Maintaining availability of dataset	5 a	14	5.36	1.082	.289	4.73	5.98	3	7
	6 a	19	4.84	1.425	.327	4.16	5.53	2	7
	7 b	13	4.08	1.320	.366	3.28	4.87	1	6
	Total	46	4.78	1.365	.201	4.38	5.19	1	7

groups. Its analyzed p-value is 0.381, which is not significant. Thus, hypothesis H7 is not supported. The Business Environment factor does not differ across the three groups.

Analyzing Hypothesis 8

Hypothesis 8 tests to find rules that predict successful or unsuccessful data mining projects by variables (V) and factors (F). I used two techniques, decision tree (DT) and neural network (NN), to build the rules, developing four sub-hypotheses (two

categories and two techniques) . The four sub-hypotheses are H8DTV (decision tree technique by variables), H8DTF (decision tree technique by factors), H8NNV (neural network technique by variables), and H8NNF (neural network technique by factors).

I used two data mining techniques to find an algorithm which will result in successful data mining project evaluation. I used SPSS® Clementine software for data mining analysis. While multi-task tools typically require special training and some tool customization, Clementine, in particular, reportedly has been widely used without customization by a variety of users ranging from business analysts to biochemists. Such use is made possible by a highly graphical user interface to data mining functions.

Decision Tree Technique

First, I performed the decision tree technique. Decision trees are powerful and popular tools for classification and prediction. Each decision tree represents rules. There are a variety of algorithms for building decision trees which share the desirable trait of explicability. Two of the most popular techniques are known as CART (Classification and Regression Trees) and CHAID (Chi-squared Automatic Interaction Detection). A newer algorithm, C5.0, is gaining popularity and is now available in several software packages. For this study, I used C5.0 to perform this decision tree technique. Table 4-16 shows the rule and the correction rate.

In the analysis of the decision tree by variables, I found seven sub rules. The first rule indicates that if a user thinks that the value of V39 (NOT complex dataset) is equal to or less than 4, and the value of V48 (Interpreting the discovered patterns) is equal to or less than 6, then the user will have a Mode 2 result. In other words, if a user

Table 4-16: Results of Decision Tree

By Variables	By Factors
V39 =< 4 [Mode: 2] V48 =< 6 [Mode: 2] -> 2 V48 > 6 [Mode: 2] V07 =< 3 [Mode: 1] -> 1 V07 > 3 [Mode: 2] V34 =< 1 [Mode: 1] -> 1 V34 > 1 [Mode: 2] -> 2 V39 > 4 [Mode: 1] V20 =< 3 [Mode: 2] -> 2 V20 > 3 [Mode: 1] V30 =< 6 [Mode: 1] -> 1 V30 > 6 [Mode: 2] -> 2 Results for output field EVAL Comparing $C-EVAL with EVAL Correct : 44(89.80%) Wrong : 5(10.20%) Total : 49	F2 =< 25 [Mode: 2] -> 2 F2 > 25 [Mode: 1] F6 =< 11 [Mode: 1] -> 1 F6 > 11 [Mode: 2] F1 =< 47 [Mode: 1] -> 1 F1 > 47 [Mode: 2] -> 2 Results for output field EVAL Comparing $C-EVAL with EVAL Correct : 40 (81.63%) Wrong : 9 (18.37%) Total : 49

thinks that the not complex data (V39) is not important and that interpreting the discovered patterns (V48) is not very important, then he or she will have highly successful (Mode 2) data mining projects. The last rule means that if a user thinks that the value of V39 is greater than 4, the value of V20 is greater than 3, and the value of V30 is greater than 6, then the user will have Mode 2 result. In other words, if a user thinks that the NOT complex dataset (V39) is important, that the data management specialist in the project (V20) is important, and that determining the data structure (V30) is very important, then he or she will have highly successful (Mode 2) data mining projects. I used V39 as the first criteria to build the rule.

In the analysis of the decision tree by factors, I found four sub rules: (1) if F2 is less than or equal to 25, then mode 2; (2) If F2 is greater than 25 and F6 is less than or equal to 11, then mode 1; (3) If F2 is greater than 25, F6 is greater than 11, and F1 is

equal to or less than 47, then mode 1; and (4) If F2 is greater than 25, F6 is greater than 11, and F1 is greater than 47, then mode 2. Here, F2 (dataset) is the most important factor to build the decision tree rule. In other words, if users think that the dataset is an important factor for successful data mining projects, they have high chances of achieving successful data mining projects.

The two rules indicate interesting results. H8DTV generated the rule with an accuracy of 89.80 %, and H8DTF generated the rule with an accuracy of 81.63%. Both rules can build rules with an accuracy of 50% or higher. Thus, both hypotheses using the decision tree technique are accepted, meaning that a reduced set of indicators of data mining projects provide a better classification of successful or unsuccessful evaluation than the naive prediction rule.

Neural Network Technique

To apply the neural network method, all the input values should be massaged so that they lie between 0 and 1 because neural networks only handle values between 0 and 1. After training the model with sample data, this trained model predicts the success level of any new data. Neural networks are applicable because they can handle a wide range of problems, they produce good results even in complicated - domains, they can handle both categorical and continuous variables, and they are available in many off-the-shelf packages (Berry and Linoff, 1997). Table 4-17 shows the rules and the correction rates.

The neural network technique generates two sub models by variables and factors. The first model by variables used all 57 variables to build a model. V23 (applying state-of-the-art technology to the business) is the most important variable

Table 4-17. Results of Neural Network

By Variables	By Factors
Neural Network "EVAL" architecture Input Layer : 57 neurons Hidden Layer #1 : 20 neurons Output Layer : 1 neurons Predicted Accuracy : 75.00% Relative Importance of Inputs V23 : 0.03016 V13 : 0.02729 V12 : 0.02691 V09 : 0.02426 V02 : 0.02035 V36 : 0.01905 V42 : 0.01862 V54 : 0.01842 V27 : 0.01797 V25 : 0.01629 V18 : 0.01589 V01 : 0.01552 V29 : 0.01442 V28 : 0.01428 V17 : 0.01351 V30 : 0.01328 V52 : 0.01305 V22 : 0.01279 V38 : 0.01234 V05 : 0.01215 V34 : 0.01207 V21 : 0.01196 V49 : 0.01182 V47 : 0.01177 V40 : 0.01144 V26 : 0.01135 V11 : 0.01112 V35 : 0.01111 V04 : 0.01065 V55 : 0.01023 V44 : 0.01004 V51 : 0.01004 V39 : 0.00888 V15 : 0.00873	Neural Network "EVAL" architecture Input Layer : 7 neurons Hidden Layer #1 : 20 neurons Output Layer : 2 neurons Predicted Accuracy : 73.08% Relative Importance of Inputs F2 : 0.32848 F5 : 0.14005 F1 : 0.08699 F3 : 0.05057 F7 : 0.03888 F4 : 0.02798 F6 : 0.01621 Results for output field EVAL Comparing $N-EVAL with EVAL Correct : 34 (69.39%) Wrong : 15 (30.61%) Total : 49 Results for output field EVAL Comparing $C-EVAL with EVAL Correct : 40 (81.63%) Wrong : 9 (18.37%) Total : 49

V45 : 0.00745 V03 : 0.00737 V08 : 0.00696 V19 : 0.00693 V10 : 0.00690 V31 : 0.00639 V07 : 0.00622 V53 : 0.00581 V33 : 0.00395 V43 : 0.00253 V48 : 0.00230 V56 : 0.00227 V50 : 0.00219 V46 : 0.00211 V37 : 0.00194 V32 : 0.00187 V41 : 0.00186 V16 : 0.00179 V14 : 0.00177 V57 : 0.00169 V06 : 0.00168 V20 : 0.00153 V24 : 0.00151 Results for output field EVAL Comparing $N-EVAL with EVAL Correct : 33 (67.35%) Wrong : 16 (32.65%) Total : 49	

because it has the highest relative importance of inputs (0.03016). V13 (user's confidence in the system), V12 (client acceptance), V09 (realistic project schedule), and V02 (realistic mission related to marketing strategy) are also important variables in which the values of the relative importance of inputs are higher than 0.02.

The second model by factors used all 7 factors to build the model. F2 (dataset) has the highest relative importance of inputs (0.32848), and F6 (consultation) has the lowest relative importance of inputs (0.01621). In other words, dataset is the most important factor influencing the accuracy of this model.

H8NNV generated a rule with an accuracy of 67.35 %, and H8NNF generated a rule with an accuracy of 69.39% and 81.63%. Both hypotheses can build models with an accuracy of 50% or higher. Thus, both hypotheses using the neural network are accepted. A reduced set of indicators of data mining projects provide a better classification of successful or unsuccessful evaluation than the naive prediction model.

Even though these two techniques achieved highly accurate rules and models, rules by decision tree seemed to rate higher than models by neural network.

CHAPTER 5
CONCLUSIONS

Summary of the Research Process

Conclusion

I proposed, in this study, to identify from all possible success factor items what factors are important to the success of data mining projects and what factors are not. This I accomplished by documenting the concepts of data mining, related areas of data mining, contexts of data mining, the critical success factor (CSF) approach, and all possible success factor items from all relevant areas to data mining projects. I focused on eight hypotheses, one dependent construct, 57 independent variables, and seven factors. I found evidence indicating which factors are more important to successful data mining projects.

I tested eight hypotheses and identified six factors receiving less emphasis from participants of more successful data mining projects than they did from participants of less successful data mining projects. This evidence indicates that more attention should be given to one factor (Dataset) and that a reevaluation should be done of the techniques to define and deploy the six non-significant factors.

Assumptions and Limitations

This study is based on the assumption that the criteria and success factor items from the literature contain the “best set” of indicators that contribute to the successful data mining projects. More specifically, CSFs in major components of data mining contain important factors, and these factor items are validated by using published data

mining articles. The dependent construct assumption is that the success measures selected from the DeLone and McLean model (1992) resulted in an accurate measure for successful data mining projects. Data collection assumptions are that responses of the participating respondents were as accurate as possible to the best of their knowledge.

As with any field research, only partial control is possible and there is limited ability to accommodate extraneous variables (Buckley et al. 1976). Thus, each researcher has a responsibility to ensure that the study does not seriously suffer from various internal and external validity threats. Generally, internal validity is not expected to be as high with field studies as with lab-based experimental studies. This is because the experimental studies are usually conducted in a highly controlled environment whereas most field studies are carried out in a more natural setting. Nevertheless, the researcher must recognize potential threats to the validity of the study and ensure they do not adversely affect the results of the study. Some of sources of validity threats include (Cook and Campbell, 1979): history, maturation, testing instrumentation, statistical regression, selection, mortality, interactions with selection, ambiguity about the direction of casual influence, diffusion or imitation of treatments, compensatory equalization of treatments, compensatory rivalry, and resentful demoralization. In this study, I did not consider history or maturation to be threatening because of the relatively short time span of data collection. I used the panel of experts to reduce the potential testing instrumentation validation threat. I addressed and reported statistical threats by significance testing appropriate for each test statistic. The data collected also has high internal validity, given that the data were loaded electronically directly into the testing

software package from the participant's survey response, eliminating potential data entry error. I did not consider compensatory equalization of treatments to be a significant threat even though this study suggests the results of the study to the participants.

There are three major threats to external validity because there are three ways the researcher could be wrong -- people, places or times. While external validity for place and time were considered to be strong by virtue of the nature of the study, I recognize the potential validity threat for people. Due to the relatively small sample size (n = 56), the results of this study may not be easily generalized to other studies.

With field data, the possibility and risk of extraneous variables and interactions exist. External validity could be the problem because of small sample size. The significance testing indicates that the internal and external validity of this study has no major threats.

Future Research

One of the weaknesses of this study is the small sample size, which may have adversely affected external validity. Thus, I recommend repeating the study with a larger sample size for further research. Further studies may serve to explain why certain groups (such as level of management, level of age group, company types, locations of data mining projects completed, number of employees in respondents' company, etc.) differ in their response to some factors and variables. Why, for example, do the responses to questions concerning Factor 6 (Consultation) vary among groups of organizations of different sizes (O2)?

Other issues from data mining remain to be researched. For example, privacy issues are further exacerbated now that the World Wide Web makes it easy for new data to be automatically collected and added to databases. As data mining tools and services become more widely available, privacy concerns are likely to intensify (Cranor, 1999). Data mining has the potential to equip companies with the ability to invade individual privacy. Information about a person's purchases and even which Internet sites they visit can be bought and sold without their knowledge or permission (Wreden, 1997).

The process of discovering or predicting the competitors' strategic decisions and/or understanding the characteristics of the business using quantitative analysis techniques applied to open sources (for example, online databanks) (Zanasi, 1998).

Today, more than ever, it is vital for organizations to monitor their competitors and the direction of the market. This increased focus on competition has given rise to the practice of competitive intelligence (CI), the process of collecting, analyzing, and disseminating information about industry developments or market trends to enhance a company’s competitiveness.

APPENDICES

APPENDIX A

QUESTIONNAIRE

Critical Success Factors in Data Mining Projects

Successfully implemented, data mining can improve decision making processes by analyzing massive historical data to discover patterns and regularities that are hard to detect with conventional technologies. Unfortunately, not all data mining projects succeed. To improve the chance of success with a data mining project, organizations need to understand and pay close attention to a number of key critical success factors. Despite the obvious importance of Critical Success Factors (CSFs) in data mining, there is little research-based information on the topic.

The findings from this study should be of substantial interest and value to all of us in the field. However, this would not be possible without your help. Please take a few minutes (no more than 10 minutes) to answer a few questions based on your experience and knowledge. Your participation is voluntary and your response will be kept confidential. Only aggregated data will be used in presenting the study findings to protect individual and organizational identities. If you are interested in receiving a report of the study as soon as it becomes available, please send an e-mail request to the researcher (jxsim@ualr.edu).

If you have any question regarding this survey, please contact the researcher (sim@unt.edu) or the faculty advisor (guynes@unt.edu).

This study has been approved by the University of North Texas committee for the protection of human subjects (940-565-3940).

Thank you

Instructions

Please select the number that best describes your response to each statement using a scale of 0 to 7. Select n/a (not applicable) if the question is not applicable or you do not know the answer. Think of the scale from 1 to 7 as a continuum from complete disagreement to complete agreement with the statement. For a neutral response, select 4.

n/a = This question is not applicable or I don't know the answer.
1 = strongly disagree
2 = disagree
3 = somewhat disagree
4 = neutral
5 = somewhat agree
6 = agree
7 = strongly agree

Part 1: Critical Success Factors for data mining projects.

The following factors are (were) important for successful data mining projects;

Factors	n/a	1	2	3	4	5	6	7
A clear mission.	◉	○	○	○	○	○	○	○
Realistic mission related to marketing strategy.	◉	○	○	○	○	○	○	○
Focusing on the real problem.	◉	○	○	○	○	○	○	○
Identifying the business problems and data	◉	○	○	○	○	○	○	○
Identifying the business needs.	◉	○	○	○	○	○	○	○
Investigating the impact on the business problems.	◉	○	○	○	○	○	○	○
Investigating the impact on the business opportunities.	◉	○	○	○	○	○	○	○
A limited budget.	◉	○	○	○	○	○	○	○
Realistic project schedule.	◉	○	○	○	○	○	○	○
Providing information in a timely manner.	◉	○	○	○	○	○	○	○
Client consultation.	◉	○	○	○	○	○	○	○
Client acceptance.	◉	○	○	○	○	○	○	○
User's confidence in the system.	◉	○	○	○	○	○	○	○
Good relationship between users and the IS staff.	◉	○	○	○	○	○	○	○
Well educated team members in the project.	◉	○	○	○	○	○	○	○
Experienced sponsor in the project.	◉	○	○	○	○	○	○	○
Experienced user-group in the project.	◉	○	○	○	○	○	○	○
Experienced business analyst in the project.	◉	○	○	○	○	○	○	○
Experienced data analyst in the project.	◉	○	○	○	○	○	○	○
Experienced data management specialist in the project.	◉	○	○	○	○	○	○	○

Experienced project manager in the project.	◉ ○ ○ ○ ○ ○ ○ ○
Applying relevant technology to the business needs.	◉ ○ ○ ○ ○ ○ ○ ○
Applying state-of-the-art technology to the business.	◉ ○ ○ ○ ○ ○ ○ ○
Good communication in the project team.	◉ ○ ○ ○ ○ ○ ○ ○
Top management support to the project.	◉ ○ ○ ○ ○ ○ ○ ○
Troubleshooting throughout the implementation.	◉ ○ ○ ○ ○ ○ ○ ○
Monitoring, and feedback to the project.	◉ ○ ○ ○ ○ ○ ○ ○
Accuracy of output.	◉ ○ ○ ○ ○ ○ ○ ○
Reliability of output.	◉ ○ ○ ○ ○ ○ ○ ○
Determining data structures.	◉ ○ ○ ○ ○ ○ ○ ○
Developing logical data models.	◉ ○ ○ ○ ○ ○ ○ ○
Integrating logical data models into an enterprise data model.	◉ ○ ○ ○ ○ ○ ○ ○
Identifying the organization's critical data requirements.	◉ ○ ○ ○ ○ ○ ○ ○
Tuning databases to meet operational needs.	◉ ○ ○ ○ ○ ○ ○ ○
Defining implementation strategies for databases.	◉ ○ ○ ○ ○ ○ ○ ○
Recovering databases.	◉ ○ ○ ○ ○ ○ ○ ○
Cleaned dataset.	◉ ○ ○ ○ ○ ○ ○ ○
Easily accessible data.	◉ ○ ○ ○ ○ ○ ○ ○
NOT complex dataset.	◉ ○ ○ ○ ○ ○ ○ ○
Sufficient number of dataset.	◉ ○ ○ ○ ○ ○ ○ ○
Scalability of database.	◉ ○ ○ ○ ○ ○ ○ ○
Maintaining availability of dataset.	◉ ○ ○ ○ ○ ○ ○ ○
Selecting the right type of data source (flat file, relational DB, data warehouse, etc).	◉ ○ ○ ○ ○ ○ ○ ○

Determining the data mining task.	◉ ○ ○ ○ ○ ○ ○ ○
Selecting the right data mining technique for the project.	◉ ○ ○ ○ ○ ○ ○ ○
Selecting the right data mining software for the project.	◉ ○ ○ ○ ○ ○ ○ ○
Understanding the service type (consultancy, implementation, education, or related services) of the project.	◉ ○ ○ ○ ○ ○ ○ ○
Interpreting the discovered patterns.	◉ ○ ○ ○ ○ ○ ○ ○
Possibly returning to any of the previous steps after interpreting the discovered patterns.	◉ ○ ○ ○ ○ ○ ○ ○
Visualization of the extracted patterns.	◉ ○ ○ ○ ○ ○ ○ ○
Removing redundant or irrelevant patterns.	◉ ○ ○ ○ ○ ○ ○ ○
Translating the useful ones into terms understandable to the users.	◉ ○ ○ ○ ○ ○ ○ ○
Incorporating the knowledge into the performance system.	◉ ○ ○ ○ ○ ○ ○ ○
Taking actions based on the extracted knowledge.	◉ ○ ○ ○ ○ ○ ○ ○
Documenting the extracted knowledge.	◉ ○ ○ ○ ○ ○ ○ ○
Reporting the extracted knowledge to interested parties.	◉ ○ ○ ○ ○ ○ ○ ○
Checking for and resolving potential conflicts with previously believed (or extracted) knowledge.	◉ ○ ○ ○ ○ ○ ○ ○

Part 2: Demographic Information.

2.1. Respondent Information:

2.1.1. What is your level of management? ○ Top level ◉ Middle level ○ Operating level

2.1.2. What is your gender? ◉ Male ○ Female

2.1.3. Please indicate your age group.

○ Under 21 ○ 21 to 34 ◉ 35 to 44 ○ 45 to 54 ○ 55 to 64 ○ 65 or older

2.1.4. Please indicate the highest level of education completed.

○ High School
○ Some College
◉ College Graduate (4 yr)
○ Master's Degree
○ Doctorate Degree
○ Professional Degree (MD, JD, etc.)
○ Vocational/Technical School (2 yr)
○ Other

2.2. Company Information:

2.2.1. Please indicate the level of your company's annual sales.

◉ Less than $1 million
○ $1 million to less than $5 million
○ $5 million to less than $10 million
○ $10 million to less than $50 million
○ $50 million to less than $100 million
○ $100 million to less than $200 million
○ $200 million to less than $500 million
○ $500 million or more

2.2.2. Please indicate the level of number of employees in your company.

◉ 50 or less
○ 51 to 100
○ 101 to 500
○ 501 to 1,000
○ 1,000 to 5,000
○ Over 5, 000

2.2.3. Please indicate the type of your company.

○ Manufacturing
○ Construction
○ Service
○ Finance/Insurance/Real Estate
○ Sales
○ Healthcare
○ Transport/Communication/Utility
◉ Other

2.3. Data Mining Project Information:

2.3.1. Please indicate years working in data mining area.

◉ N/A
○ Less than 6 month
○ 6 months to less than 1 yr
○ 1 yr to less than 3 yr
○ 3 yr less than to 5 yr
○ 5 yr less than to 10 yr
○ 10 yr or more

2.3.2. Please indicate the number of data mining project(s) involved.

◉ N/A or none
○ 1
○ 6 to 10
○ 11 to 50

- ○ 2 to 3
- ○ 4 to 5
- ○ 51 or more

2.3.3. Please indicate the number of average people working for data mining projects.

- ◉ No project done
- ○ 1 to 2
- ○ 3 to 5
- ○ 6 to 10
- ○ 11 to 50
- ○ 51 to 100
- ○ 101 to 500
- ○ 501 or more

2.3.4. Please indicate the company type to which most of your data mining projects are developed.

- ◉ No project done
- ○ Manufacturing
- ○ Construction
- ○ Service
- ○ Finance/Insurance/Real Estate
- ○ Sales
- ○ Healthcare
- ○ Transport/Communication/Utility

2.3.5.Where do you mostly complete your data mining projects?

- ◉ No project done
- ○ in Africa
- ○ in Asia
- ○ in Europe
- ○ in North America
- ○ in South America
- ○ in Australia

2.3.6. How do you evaluate the average success of your data mining projects ?

- ○ No project done
- ○ Very unsuccessful
- ○ Unsuccessful
- ○ Somewhat unsuccessful
- ◉ Neutral
- ○ Somewhat successful
- ○ Successful
- ○ Very successful

2.4. Personal perception to Data Mining Project

2.4.1. How familiar are you to data mining area? This includes every area such as consulting, researching and teaching.

- ○ Very unfamiliar
- ○ Unfamiliar
- ○ Somewhat unfamiliar
- ◉ Neutral
- ○ Somewhat familiar
- ○ Familiar
- ○ Very familiar

Part 3: Comments.

Please, enter additional CSFs for data mining projects or comments in the space provided below:

Submit Form Reset Form

Jaesung Sim (501-569-8853) and Dr. Steve C. Guynes (940-565-3110)

Department of Business Computer Information Systems

College of Business Administration

University of North Texas

Denton, TX 76203

Revised: May 31, 2002

APPENDIX B

STATISTICAL ANALYSIS RESULTS

Oneway

Descriptives

		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
						Lower Bound	Upper Bound		
F1	5	11	44.0000	9.79796	2.95420	37.4176	50.5824	18.00	55.00
	6	16	45.1875	7.35952	1.83988	41.2659	49.1091	30.00	56.00
	7	11	37.0000	9.82853	2.96341	30.3971	43.6029	12.00	45.00
	Total	38	42.4737	9.31647	1.51133	39.4114	45.5359	12.00	56.00
F2	5	14	25.3571	6.12193	1.63615	21.8224	28.8918	14.00	32.00
	6	16	20.9375	6.61784	1.65446	17.4111	24.4639	10.00	35.00
	7	11	17.0909	6.45685	1.94681	12.7531	21.4287	5.00	27.00
	Total	41	21.4146	7.04974	1.10098	19.1895	23.6398	5.00	35.00
F3	5	12	20.5833	4.90748	1.41667	17.4653	23.7014	9.00	26.00
	6	17	21.1176	2.84786	.69071	19.6534	22.5819	15.00	28.00
	7	12	19.7500	5.32789	1.53803	16.3648	23.1352	12.00	28.00
	Total	41	20.5610	4.24293	.66263	19.2217	21.9002	9.00	28.00
F4	5	14	11.5714	2.47182	.66062	10.1442	12.9986	6.00	14.00
	6	18	11.8333	1.75734	.41421	10.9594	12.7072	8.00	14.00
	7	13	11.0769	2.21591	.61458	9.7379	12.4160	8.00	14.00
	Total	45	11.5333	2.10627	.31398	10.9005	12.1661	6.00	14.00
F5	5	14	11.5000	3.03188	.81030	9.7494	13.2506	4.00	14.00
	6	19	11.9474	2.17239	.49838	10.9003	12.9944	5.00	14.00
	7	14	12.1429	1.70326	.45522	11.1594	13.1263	10.00	14.00
	Total	47	11.8723	2.30893	.33679	11.1944	12.5503	4.00	14.00
F6	5	13	12.0000	1.15470	.32026	11.3022	12.6978	10.00	14.00
	6	18	11.9444	1.89340	.44628	11.0029	12.8860	7.00	14.00
	7	14	10.8571	2.56776	.68626	9.3746	12.3397	5.00	14.00
	Total	45	11.6222	1.99190	.29694	11.0238	12.2207	5.00	14.00
F7	5	12	16.1667	2.69118	.77688	14.4568	17.8766	12.00	21.00
	6	18	17.0556	2.83823	.66898	15.6441	18.4670	11.00	21.00
	7	14	15.5714	3.43543	.91816	13.5879	17.5550	8.00	20.00
	Total	44	16.3409	3.00343	.45278	15.4278	17.2540	8.00	21.00

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
F1	Between Groups	473.036	2	236.518	3.023	.062
	Within Groups	2738.438	35	78.241		
	Total	3211.474	37			
F2	Between Groups	426.890	2	213.445	5.196	.010
	Within Groups	1561.061	38	41.081		
	Total	1987.951	40			
F3	Between Groups	13.166	2	6.583	.354	.704
	Within Groups	706.931	38	18.603		
	Total	720.098	40			
F4	Between Groups	4.348	2	2.174	.478	.623
	Within Groups	190.852	42	4.544		
	Total	195.200	44			
F5	Between Groups	3.072	2	1.536	.279	.758
	Within Groups	242.162	44	5.504		
	Total	245.234	46			
F6	Between Groups	11.919	2	5.960	1.539	.226
	Within Groups	162.659	42	3.873		
	Total	174.578	44			
F7	Between Groups	17.847	2	8.923	.989	.381
	Within Groups	370.040	41	9.025		
	Total	387.886	43			

BIBLIOGRAPHY

Adams, N.M., Hand, D.J., and Till, R.J. “Mining for classes and patterns in behavioural data,” Journal of the Operational Research Society 52(9), Sep 2001, pp.1017-1024

Apte, C., Liu, B., Pednault, E.P., and Smyth, P. “Business applications of data mining,” Communications of the ACM 45(8), Aug 2002, pp.49-53

Arunasalam, R. G., Richie, J. T., Egan, W., Gur-Ali, O., and Wallace, W. A. “Reengineering claims processing using probabilistic inductive learning,” IEEE Transactions on Engineering Management 46(3), August 1999, pp.335-345

Backhaus, B. “Mining your engineering data,” Computer-Aided Engineering 18(1), January 1999, pp.56-59

Bailey, J.E., and Pearson, S. W. “Development of a tool for measuring and analyzing computer user satisfaction,” Management Science 29 (5), May 1983, pp. 519-529.

Baker, S. and Baker, K. “Mine over matter,” Journal of Business Strategy 19(4), Jul/Aug 1998, pp.22-26

Balachandran, K., Buzydlowski, J., Dworman, G., Kimbrough, S. O. et al. “MOTC: An interactive aid for multidimensional hypothesis generation,” Journal of Management Information Systems 16(1), 1999, pp.17-36

Bass, T. “Intrusion detection systems and multisensor data fusion,” Communications of the ACM 43(4), April 2000, pp.99-105

Beale, P. and Freeman, M. “Successful project execution: A model,” Project Management Journal 22(4), December 1991, pp. 23-30.

Bentley, J. “Long common strings,” Unix Review 15(13), December 1997, pp.61-66

Bentley, T. “Mining for information,” Management Accounting-London 75(6), June 1997, pp.56

Berry, M. and Linoff, G. Data Mining Techniques: for marketing, sales, and customer support. John Wiley & Sons, Inc., New York, NY, 1997

Berzal, F., Blanco, I., Cubero, J.C., and Marin, N. “Component-based data mining frameworks,” Communications of the ACM 45(12), Dec 2002, pp.97-100

Bessler, D. A. and Chamberlain, P. J. “Composite Forecasting with Dirichlet Priors,” Decision Sciences 19(4), Fall 1988, pp.771-781

Bontempo, C. and Zagelow, G. “The IBM data warehouse architecture,” Communications of the ACM 41(9), September 1998, pp. 38-48.

Boone, M. E. “Why fuzzy queries are logical,” Sales & Marketing Management 151(2), February 1999, pp.77

Borok, L. S. “Data mining: Sophisticated forms of managed care modeling through artificial intelligence,” Journal of Health Care Finance 23(3), Spring 1997, pp.20-36

Bose, I., and Mahapatra, R.K. “Business data mining - A machine learning perspective,” Information & Management 39(3), Dec 20, 2001, pp.211-225

Brabazon, T. “Data mining: A new source of competitive advantage?,” Accountancy Ireland 29(3), June 1997, pp.30-31

Brachman, R.J., Khabaza, T., Kloesgen, W., Piatetsky-Shapiro, G., and Simoudis, E. “Mining business databases,” Communications of the ACM 39(11), November 1996, pp. 42-48.

Bradley, P., Gehrke, J., Ramakrishnan, R., and Srikant, R. “Scaling mining algorithms to large databases,” Communications of the ACM 45(8), Aug 2002, pp.38-43

Brown, A J. “A model for effective, customer-oriented market plans,” Direct Marketing 60(3), July 1997, pp.40-43

Buckley, J. W., Buckley, M. H., Chiang, H, Research Methodologies and Business Decisions, New York: National Association of Accountants and The Society of Industrial Accountants of Canada, 1976.

Bui, T. and Lee, J. “An agent-based framework for building decision support systems,” Decision Support Systems 25(3), April 1999, pp.225-237

Bullen, C. V., “Reexamining productivity CSFs: The knowledge worker challenge,” Information Systems Management. 12(3), Summer 1995, pp. 13-18.

Bullen, C. V., and Rockart, J.F. “A primer on critical success factors”, Center for Information Systems Research Working Paper, 69, Sloan School of Management, MIT, Cambridge, MA., June 1981.

Byers, C. R., and Blume, D., “Tying critical success factors to systems development”, Information Management 26(1), January 1994, pp. 51-61.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., and Zanasi A. Discovering Data Mining : From concepts to implementation. Prentice Hall, Upper Saddle River, NJ, 1997

Castelluccio, M. “Data warehouses, marts, metadata, OLAP/ROLAP, and data mining--a glossary,” Management Accounting 78(4), October 1996, pp.59-61

Caudill, S. B. “The Necessity of Mining Data,” Atlantic Economic Journal 16(3), September 1988, pp.11-18

Chan, C., and Lewis, B. “A basic primer on data mining,” Information Systems Management 19(4), Fall 2002, pp.56-60

Chen, A.N., Goes, P.B., and Marsden, J.R. “A query-driven approach to the design and management of flexible database systems,” Journal of Management Information Systems 19(3), Winter 2002-2003, pp.121-154

Chen, L.D., Sakaguchi, T., and Frolick, M.N. “Data mining methods, applications, and tools,” Information Systems Management 17(1), Winter 2000, pp.65-70

Cheung, K.W., Kwok; J.T., Law, M.H., and Tsui, K. “Mining customer product ratings for personalized marketing,” Decision Support Systems 35(2), May 2003, pp.231-243

Chiu, C. “Towards integrating hypermedia and information systems on the web,” Information & Management 40(3), Jan 2003, pp.165-175

Chou, D. C. and Chou, A. Y. “A manager's guide to data mining,” Information Systems Management 16(4), Fall 1999, pp.33-41

Chung, H.M., and Gray, P. “Special Section: Data Mining,” Journal of Management Information Systems, 16(1), Summer 1999 pp. 11-16.

Clark, T. “Constructing the SAN, part two,” Computer Technology Review 19(6), June 1999, pp.56-58

Cook, T. D., and Campbell, D. T., Quasi-experimentation design & analysis issues for field setting. Houghton Mifflin Company, Boston, 1979.

Cooper, L. G. and Giuffrida, G. “Turning datamining into a management science tool: New algorithms and empirical results,” Management Science 46(2), February 2000, pp.249-264

Cranor, L. F. “Internet privacy,” Communications of the ACM 42(2), February 1999, pp. 28-31.

Crocker, J. “Introduction to Neural Networks and Data Mining: For Business Applications,” Journal of the Operational Research Society 51(6), June 2000, pp.771-772

Cronbach, L. J. “Coefficient alpha and the internal structure of tests,” Psychometrika, 16, 1951 pp. 297 – 334.

Culnan, M. J. and Armstrong, P. K. “Information privacy concerns, procedural fairness, and impersonal trust: An empirical investigation,” Organization Science 10(1), Jan/Feb 1999, pp.104-115

Daskalaki, S., Kopanas, I., Goudara, M. and Avouris, N. “Data mining for decision support on customer insolvency in telecommunications business,” European Journal of Operational Research 145(2), Mar 1, 2003, pp.239-255

David, J. S., and Steinbart, P. J. “Drowning in data,” Management Accounting 81(6). December 1999, pp. 30-34.

DeLone, W. D., and McLean, E. R. “Information Systems Success: The Quest for the Dependent Variable,” Information Systems Research 3(1), March 1992, pp. 60-95.

DeLone, W. D., and McLean, E. R. “The DeLone and McLean Model of Information Systems Success: A Ten-Year Update,” Journal of Management Information Systems 19(4), Spring 2003, pp. 9-30.

DeLong, J. B. and Lang, K. “Are All Economic Hypotheses False?,” Journal of Political Economy 100(6), December 1992, pp.1257-1272

Deming, D. “Performance of wide Ultra SCSI and SSA architectures,” Computer Technology Review 18(2), February 1998, pp.34-36

Deogun, J., Choubey, S. K., Raghavan, V. V., and Sever, H. “Feature selection and effective classifiers,” Journal of the American Society for Information Science 49(5), Apr. 15 1998, pp.423-434

Drew, J. H., Mani, D. R., Betz, A. L., and Datta, P. “Targeting customers with statistical and data-mining techniques,” Journal of Service Research 3(3), February 2001, pp.205-219

Dube, N. A. “Globalsearch: The international search for missing and exploited children,” International Review of Law Computers & Technology 13(1), March 1999, pp.69-74

Eick, S.G. “Visualizing online activity,” Communications of the ACM 44(8), Aug 2001, pp.45-50

Etzioni, O. “The World Wide Web: Quagmire or gold mine?,” Communications of the ACM 39(11), November 1996, pp.65-68

Evans, G. W. “A Test for Speculative Bubbles in the Sterling-Dollar Exchange Rate: 1981-84,” American Economic Review 76(4), September 1986, pp.621-636

Ewers, A. “A review of new developments in text retrieval systems,” Journal of Information Science 20(6), 1994, pp.438-443

Exner, F. “Advances in Knowledge Discovery and Data Mining,” Journal of the American Society for Information Science 49(4), 1998, pp.386-387

Farrow, S. “Testing the Efficiency of Extraction from a Stock Resource,” Journal of Political Economy 93(3), 1985, pp.452-487

Fayyad, U. “The digital physics of data mining,” Communications of the ACM 44(3), 2001, pp.62-65

Fayyad, U., and Uthurusamy, R. “Data mining and knowledge discovery in databases,” Communications of the ACM. 39(11), November 1996, pp. 24-26.

Fayyad, U., and Uthurusamy, R. “Evolving data mining into solutions for insights,” Communications of the ACM 45(8), Aug 2002, pp.28-31

Fayyad, U., Haussler, D. and Stolorz, P. “Mining scientific data,” Communications of the ACM 39(11), November 1996, pp. 51-57.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. “The KDD process for extracting useful knowledge from volumes of data,” Communications of the ACM 39(11), November 1996, pp. 27-34.

Feelders, A., Daniels, H., and Holsheimer, M. “Methodological and practical aspects of data mining,” Information Management 37(5), 2000, pp.271-281

Fogler, H. R. “Investment analysis and new quantitative tools,” Journal of Portfolio Management 21(4), 1995, pp.39-48

France, T., Yen, D., Wang, J.C., and Chang, C.M., “Integrating search engines with data mining for customer-oriented information search,” Information Management & Computer Security 10(5), 2002, pp.242-254

Frawley, A., and Thearling, K. “Increasing customer value by integrating data mining and campaign management software,” Direct Marketing 61(10), 1999, pp.49-53

Freedman, J. “IIA announces 1997 research priorities,” Management Accounting 78(10), 1997, pp.65-66

Furash, E. E. “Data mining,” Journal of Lending & Credit Risk Management 79(11), 1997, pp.7-11

Galal, G. M., Cook, D. J., and Holder, L. B. “Exploiting parallelism in a structural scientific discovery system to improve scalability,” Journal of the American Society for Information Science 50(1), 1999, pp.65-73

Garrity, E.J. “Data Mining II,” Information Resources Management Journal 15(1), Jan-Mar 2002, pp.38-39

Glass, C. “Success sure is a lot of fun,” Sales & Marketing Management 151(1), 1999, pp.54-59

Gluck, M. “The use of sound for data exploration,” Bulletin of the American Society for Information Science 26(5), 2000, pp.26-28

Glymour, C., Madigan, D., Pregibon, D., and Smyth, P. “Statistical inference and data mining,” Communications of the ACM 39(11), November 1996, pp. 35-41.

Greenfeld, N. “Data mining,” Unix Review 14(5), 1996, pp.9-14

Grossman, R.L., Hornick, M.F., and Meyer, G. “Data mining standards initiatives,” Communications of the ACM 45(8), Aug 2002, pp.59-61

Groth, R. Data Mining: A hands-on approach for business professionals. Prentice Hall, Upper Saddle River, New Jersey, 1998

Groth, R. Data Mining: Building Competitive Advantage. Prentice Hall, Upper Saddle River, New Jersey, 2000

Grzymala-Busse, J. W., Ziarko, W. “Data mining and Rough Set Theory,” Communications of the ACM 43(4), 2000, pp.108-109

Gur-Ali, O. and Wallace, W. A. “Bridging the gap between business objectives and parameters of data mining algorithms,” Decision Support Systems 21(1), September 1997, pp.3-15

Gutwin, C., Paynter, G., Witten, I., Nevill-Manning, C., and Frank, E. “Improving browsing in digital libraries with keyphrase indexes,” Decision Support Systems 27(1,2), 1999, pp.81-104

Guynes, C. S. and Vanecek, M. T. "Critical success factors in data management," Information and Management 30(4), July 1996, pp. 201- 209.

Hall, H. “Online information sources: Tools of business intelligence?,” Journal of Information Science 26(3), 2000, pp.139-143

Hall, O.P. “Mining the store,” Journal of Business Strategy 22(2), Mar/Apr 2001, pp.24-27

Hamel, G. “The quest for value,” Executive Excellence 16(3), 1999, pp.3-4

Han, J., Altman, R.B., Kumar, V., Mannila, H. and Pregibon, D. “Emerging scientific applications in data mining,” Communications of the ACM 45(8), Aug 2002, pp.54-58

Han, J., and Kamber, M. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, San Francisco, California, 2001

Heinrichs, J.H., and Lim, J. “Integrating web-based data mining tools with business models for knowledge management,” Decision Support Systems 35(1), Apr 2003, pp.103-112

Hendry, D. F. “Achievements and challenges in econometric methodology,” Journal of Econometrics 100(1), 2001, pp.7-10

Henon, E. A., and Gauches, C., “Winning the battle of customer preference,” Limra's Marketfacts 18(5), September/October 1999, pp. 9-12.

Hirji, K.K. “Exploring data mining implementation,” Communications of the ACM 44(7), Jul 2001, pp.87-93

Hui, S.C. and Jha, G. “Data mining for customer service support,” Information & Management 38(1), Oct 2000, pp.1-14

Inmon, W H. “The data warehouse and data mining,” Communications of the ACM 39(11), November 1996, pp. 49-50.

Jacso, P. “Peter's picks & pans,” Database 22(2), 1999, pp.70-74

Jeffery, B. “Getting to know the customer through a customer information system,” Information Strategy: the Executive's Journal 13(1), 1996, pp.26-36

Jern, M. “The advent of visual data mining on the World Wide Web,” Information Systems Management 15(3), 1998, pp.66-69

Johnson, D. E. “Web-based data analysis tools help providers, MCO contain costs,” Health Care Strategic Management 19(4), 2001, pp.16-19

Jurisica, I. “Systematic knowledge management and knowledge discovery,” Bulletin of the American Society for Information Science 27(1), 2000, pp.9-12

Kahn, R. N. “What Practitioners Need to Know ,” Financial Analysts Journal 46(4), 1990, pp.17-18,20

Kautz, H., Selman, B., and Shah, M. “Referral Web: Combining social networks and collaborative filtering,” Communications of the ACM 40(3), 1997, pp.63-65

King, W. R. “IT-enhanced productivity and profitability,” Information Systems Management 15(3), 1998, pp.70-72

Kirchner, C. “Highly scalable storage for very large database (VLDB) and datawarehousing applications,” Computer Technology Review 16(2), 1996, pp.46,49

Klemettinen, M., Mannila, H., and Toivonen, H. “Interactive exploration of interesting findings in the Telecommunication Network Alarm Sequence Analyzer TASA,” Information & Software Technology 41(9), 1999, pp.557-567

Knight, K. “Mining online text,” Communications of the ACM 42(11), 1999, pp.58-61

Koczkodaj, W. W., Orlowski, M., and Marek, V. W. “Myths about rough set theory,” Communications of the ACM 41(11), 1998, pp.102-103

Kostoff, R.N., Del Rio, J.A., Humenik, J.A., and Ramirez, A.M., “Citation mining: Integrating text mining and bibliometrics for research user profiling,” Journal of the American Society for Information Science and Technology 52(13), Nov 2001, pp.1148-1156

Krieger, A. M., Green, P. E., and Umesh, U N. “Effect of level of disaggregation on conjoint cross validations: Some comparative findings,” Decision Sciences 29(4), 1998, pp.1047-1058

Lavington, S., Dewhurst, N., Wilkins, E., and Freitas, A. “Interfacing knowledge discovery algorithms to large database management systems,” Information & Software Technology 41(9), 1999, pp.605-617

Leamer, E. E. “Revisiting Tobin's 1950 study of food expenditure,” Journal of Applied Econometrics 12(5), 1997, pp.533-561

Leeds, S. “Data mining: Beware of the shaft,” Direct Marketing 62(9), 2000, pp.38-42

Leinweber, D. J., and Arnott, R. D. “Quantitative and computational innovation in investment management,” Journal of Portfolio Management 21(2), 1995, pp.8-15

Lejeune, M.A. “Measuring the impact of data mining on churn management,” Internet Research 11(5), 2001, pp.375-387

Lesser, E., Mundel, D., and Wiecha, C. “Managing customer knowledge,” Journal of Business Strategy 21(6), 2000, pp.35-37

Leung, K. S., and Cheng, J. C. “Discovering knowledge from noisy databases using genetic programming,” Journal of the American Society for Information Science 51(9), 2000, pp.870-881

Levinsohn, A. “Modern miners plumb for gold,” ABA Banking Journal 90(12), 1998, pp.52-55

Li, E.Y., “Perceived importance of information system success factors: A meta analysis of group differences,” Information and Management 32(1), February 1997, pp. 15-28.

Liddy, E. D. “Text mining,” Bulletin of the American Society for Information Science 27(1), 2000, pp.13-14

Lindsay, R. M. “Lies, damned lies and more statistics: The neglected issue of multiplicity in accounting research,” Accounting & Business Research 27(3), 1997, pp.243-258

Lingras, P J., and Yao, Y. Y. “Data mining using extensions of the rough set model,” Journal of the American Society for Information Science 49(5), 1998, pp.415-422

Linoff, G. “Which way to the mine?,” As/400 Systems Management 26(1), 1998, pp.42-44

Lisse, W. C Jr. “The economics of information and the Internet,” Competitive Intelligence Review 9(4), 1998, pp.48-55

Loro, L.. “Pacific Bell taps database to find business customers,” Business Marketing 82(9), 1997, pp.19

Loveman, G. “Diamonds in the data mine,” Harvard Business Review 81(5), May 2003, pp.109-113

Lutgens, K. “Data Warehousing and Data Mining for Telecommunications,” Database 21(3), 1998, pp.94

MacDonald, M., “Beat the odds,” CMA Magazine 72(5), June 1998, pp. 16-18.

Magrath, A. J. “Mining for new product successes,” Business Quarterly 62(2), 1997, pp.64-68

Malhotra, Y. “Tools@work: Deciphering the knowledge management hype,” Journal for Quality & Participation 21(4), 1998, pp.58-60

Markowitz, H. M., and Xu, G. L. “Data mining corrections,” Journal of Portfolio Management 21(1), 1994, pp.60-69

Mason, C. “Optimal Database Marketing: Strategy, Development, and Data Mining,” Journal of Marketing Research 39(4), Nov 2002, pp.499-501

McCarthy, J. “Phenomenal data mining,” Communications of the ACM 43(8), 2000, pp.75-79

McErlean, F. J., Bell, D. A., and Guan, J. W. “Modification of belief in evidential causal networks,” Information & Software Technology 41(9), 1999, pp.597-603

McQueen, G., and Thorley, S. “Mining fool's gold,” Financial Analysts Journal 55(2), 1999, pp.61-72

McQueen, G., Shields, K., and Thorley, S. R. “Does the "Dow-10 investment strategy" Beat the Dow statistically and economically?,” Financial Analysts Journal 53(4), 1997, pp.66-72

McSherry, D. “Knowledge discovery by inspection,” Decision Support Systems 21(1), 1997, pp.43-47

Mena, J., "Machine-learning the business: Using data mining for competitive intelligence," Competitive Intelligence Review 7(4) Winter, 1996 pp.18-25.

Menczer, F. “Complementing search engines with online Web mining agents,” Decision Support Systems 35(2), May 2003, pp.195-212

Mendonca, M. G., Basili, V. R. “Validation of an approach for improving existing measurement frameworks,” IEEE Transactions on Software Engineering 26(6), 2000, pp.484-499

Mendonca, M. G., Basili, V. R., Bhandari, I. S., and Dawson, J. “An approach to improving existing measurement frameworks,” IBM Systems Journal 37(4), 1998, pp.484-501

Miller, L. L., Honavar, V., and Barta, T. “Warehousing structured and unstructured data for data mining,” Journal of the American Society for Information Science 34, 1997, pp.215-224

Mishina, M. “Enterprisewide business intelligence gains favor among corporations,” As/400 Systems Management 26(3), 1998, pp.30-33

Mitchell, T.M. “Machine learning and data mining,” Communications of the ACM 42(11) November 1999, pp.30-36.

Mobasher, B., Cooley, R., and Srivastava, J. “Automatic personalization based on Web usage mining,” Communications of the ACM 43(8), Aug 2000, pp.142-151

Mulvenna, M.D., Anand, S.S., and Buchner, A.G. “Personalization on the Net using Web mining,” Communications of the ACM 43(8), Aug 2000, pp.122-125

Nasukawa, T., and Nagano, T. “Text analysis and knowledge mining system,” IBM Systems Journal 40(4), 2001, pp.967-984

Nazem, S. M., and Shin, B. “Data mining: New arsenal for strategic decision-making,” Journal of Database Management 10(1), 1999, pp.39-42

Nemati, H.R., Steiger, D.M., SIyer, L., and Herschel, R.T. “Knowledge warehouse: An architectural integration of knowledge management, decision support, artificial intelligence and data warehousing,” Decision Support Systems 33(2), Jun 2002, pp.143-161

Newing, R. “Data mining,” Management Accounting-London 74(9), 1996, pp.34-36

Ng, M. K., and Huang, Z. “Data-mining massive time series astronomical data: Challenges, problems and solutions,” Information & Software Technology 41(9), 1999, pp.545-556

Norton, M. J. “Knowledge discovery with a little perspective,” Bulletin of the American Society for Information Science 27(1), 2000, pp.21-23

Olson, D. O., and Jonish, J. “The Robustness of Translog Elasticity of Substitution Estimates and the Capital Energy Complementarity Controversy,” Quarterly Journal of Business & Economics 24(1), 1985, pp.21-35

Padmanabhan, B., and Tuzhilin, A. “Unexpectedness as a measure of interestingness in knowledge discovery,” Decision Support Systems 27(3), 1999, pp.303-318

Padmanabhan, B., and Tuzhilin, A. “Knowledge refinement based on the discovery of unexpected patterns in data mining,” Decision Support Systems 33(3), Jul 2002, pp.309-321

Park, S.C., Piramuthu, S., and Shaw, M.J. “Dynamic rule refinement in knowledge-based data mining systems,” Decision Support Systems 31(2), Jun 2001, pp.205-222

Peacock, P. R. “Data mining in marketing: Part 1,” Marketing Management 6(4) Winter 1998, pp. 8-18.

Peacock, P. R. “Data Mining in Marketing: Part 2,” Marketing Management 7(1), 1998, pp.14-25

Peacock, P. R. “Data warehouses and marts,” Marketing Management 6(4), 1998, pp.13

Perkowitz, M., and Etzioni, O. “Adaptive Web sites,” Communications of the ACM 43(8), 2000, pp.152-158

Pinto, J. K. and Slevin, D. P. “Project success: definitions and measurement techniques,” Project Management Journal 19(1), February 1988, pp. 67-72.

Press, S. “Fool's Gold?,” Sales & Marketing Management 150(6), 1998, pp.58-62

Rabinovitch, L. “America's "first" department store mines customer data,” Direct Marketing 62(8), 1999, pp.42-44

Rajagopalan, B., and Krovi, R. “Benchmarking data mining algorithms,” Journal of Database Management 13(1), Jan-Mar 2002, pp.25-35

Rawlings, I. “Using data mining and warehousing for knowledge discovery,” Computer Technology Review 19(9), 1999, pp.20-22

Rhodes, W. L. Jr. “Once it's in there, how do I get it out?,” Systems Management 3x/400 24(6), 1996, pp.23-26

Roccapriore, D. “Building IT for export exploration, part II,” Business & Economic Review 46(3), 2000, pp.21-23

Rockart, J.F., "Chief executives define their own needs," Harvard Business Review, 57(2), March-April 1979, pp.81 - 93.

Rockart, J.F., and Crescenzi, A.D. “Engaging top management in information technology,” Sloan Management Review 25(4), Summer, 1984 pp. 3-16.

Roussinov, D., and Zhao, J.L. “Automatic discovery of similarity relationships through Web mining,” Decision Support Systems 35(1), Apr 2003, pp.149-166

Rubin, A., and Babbie, E., Research methods for social work (3rd ed.). Pacific Grove, CA: Brooks/Cole Publishing Company. 1997

Rufat-Latre, J. “Data storage: A strategic vacuum waiting for a strategic company,” Computer Technology Review 19(12), 1999, pp.50-51

Sauls, W., “Leveraging the data warehouse,” Management Accounting 78(4), October 1996, pp. 39-43.

Sauter, V. L. “Intuitive decision-making,” Communications of the ACM 42(6), 1999, pp.109-115

Schiereck, D., De Bondt, W., and Weber, M. “Contrarian and momentum strategies in Germany,” Financial Analysts Journal 55(6), 1999, pp.104-116

Schreck, T., and Chen, Z. “Branch grafting method for R-tree implementation,” Journal of Systems & Software 53(1), 2000, pp.83-93

Schroeder, A. T. Jr. “Data Mining with Neural Networks: Solving Business Problems from Application Development to Decision Support,” Journal of the American Society for Information Science 48(9), 1997, pp.862-863

Scotney, B., and McClean, S. “Efficient knowledge discovery through the integration of heterogeneous data,” Information & Software Technology 41(9), 1999, pp.569-578

Scott, P D., and Wilkins, E. “Evaluating data mining procedures: Techniques for generating artificial data sets,” Information & Software Technology 41(9), 1999, pp.579-587

Shank, M. E., Boynton, A. C., and Zmud, R. W., "Critical success factor analysis as a methodology for MIS planning," MIS Quarterly 9(2), June 1985, pp. 121 - 129.

Shanks, G., and Darke, Peta. “Understanding corporate data models,” Information Management 35(1), 1999, pp.19-30

Shaw, M.J., Subramaniam, C., Tan, G.W., and Welge, M.E. “Knowledge management and data mining for marketing,” Decision Support Systems 31(1), May 2001, pp.127-137

Sim, J., and Cutshall, R., “The Critical Success Factors in Data Warehousing: A Literature Review”**,** Proceedings of the 31st Annual Conference of the Decision Sciences Institute, Southwest Region, March 2000.

Slater, J. “Super sleuths,” Far Eastern Economic Review 162(40), 1999, pp.90-91

Slevin, D. P. and Pinto, J. K. “The project implementation profile: New tool for project managers,” Project Management Journal 17(4), September 1986, pp. 57-65.

Slevin, D. P., Stieman, P. A. and Boone, L.W. “Critical Success Factor Analysis for Information Systems Performance Measurement and Enhancement,” Information Management 21(3), October 1991, pp. 161-174.

Smith, K. A., Gupta, J. N. “Neural networks in business: Techniques and applications for the operations researcher,” Computers & Operations Research 27(11,12), 2000, pp.1023-1044

Smith, K. A., Willis, R. J., and Brooks, M. “An analysis of customer retention and insurance claim patterns using data mining: A case study,” Journal of the Operational Research Society 51(5), May 2000, pp.532-541

Smyth, P., Pregibon, D., and Faloutsos, C. “Data-driven evolution of data mining algorithms,” Communications of the ACM 45(8), Aug 2002, pp.33-37

Snyder, C. A., McManus, D., J. and Wilson, L. T. “Corporate memory management: A knowledge management process model,” International Journal of Technology Management 20(5,6,7,8), 2000, pp.752-764

Sorby, B. “How new technologies affect RAID storage,” Computer Technology Review 18(2), 1998, pp.50

Spangler, W. E., May, J. H., and Vargas, L. G. “Choosing data-mining methods for multiple classification: Representational and performance measurement implications for decision support,” Journal of Management Information Systems 16(1), 1999, pp.37-62

Spector, P.E., Summated rating scale construction: an introduction. Newbury Park, CA.: SAGE Publications, Inc., 1992

Spiliopoulou, M. “Web usage mining for Web site evaluation,” Communications of the ACM 43(8), Aug 2000, pp.127-134

Spinellis, D., and Raptis, K. “Component mining: A process and its pattern language,” Information and Software Technology 42(9), Jun 1, 2000, pp.609-617

Srinivasan, P., Ruiz, M. E., Kraft, D. H., and Chen, J. “Vocabulary mining for information retrieval: Rough sets and fuzzy sets,” Information Processing & Management 37(1), 2001, pp.15-38

Srinivasan, U, Ngu, A.H., and Gedeon, T. “Managing heterogeneous information systems through discovery and retrieval of generic concepts,” Journal of the American Society for Information Science 51(8), Jun 2000, pp.707-723

Sung, T. K., Chang, N., Lee, G. “Dynamics of modeling in data mining: Interpretive approach to bankruptcy prediction,” Journal of Management Information Systems 16(1), 1999, pp.63-85

Szewczak, E. “My privacy and the Internet,” Information Resources Management Journal 12(4), 1999, pp.3-4

Thatcher, M. E., and Clemons, E. K. “Managing the costs of informational privacy: Pure bundling as a strategy in the individual health insurance market,” Journal of Management Information Systems 17(2), 2000, pp.29-57

Thelwall, M. “A web crawler design for data mining,” Journal of Information Science 27(5), 2001, pp.319-325

Trowbridge, D. “Database overload overwhelms database administrators,” Computer Technology Review 20(6), 2000, pp.22-24

Trybula, W. J. “Data mining and knowledge discovery,” Journal of the American Society for Information Science 32, 1997, pp.197-229

Tsuji, B. “The information gold rush,” As/400 Systems Management 25(8), 1997, pp.40-43

Tyler, G. “We've got the information somewhere, but,” Management Accounting-London 75(6), 1997, pp.58-59

Vanecko, J. J., and Russo, A. W. “Data mining and modeling as a marketing activity,” Direct Marketing 62(5), 1999, pp.52-55

Vanecko, J. J., and Russo, A. W. “Taking the mystery out of mining and modeling,” Direct Marketing 62(4), 1999, pp.42-43

Veall, M. R. “Bootstrapping the process of model selection: An econometric example,” Journal of Applied Econometrics 7(1), 1992, pp.93-99

Walsh, J. J. and Kanter, J. “Toward More Successful Project Management,” Journal of Systems Management 39(1), January 1988, pp. 16-21.

Weir, J. “Data mining: Exploring the corporate asset,” Information Systems Management 15(4), 1998, pp.68-71

Wencer, R. A. “When a business abandons a data mine: A case study,” Information Systems Management 15(4), 1998, pp.27-35

White, H. D. “Computing a curriculum: Descriptor-based domain analysis for educators,” Information Processing & Management 37(1), 2001, pp.91-117

Williams, J. J. and Ramaprasad, A. “A taxonomy of critical success factors,” European Journal of Information Systems 5(4), December 1996, pp. 250-260.

Wong, S. K. M., Butz, C. J., and Xiang, Y. “Automated database schema design using mined data dependencies,” Journal of the American Society for Information Science 49(5), 1998, pp.455-470

Wormell, I. “Informetrics: Exploring databases as analytical tools,” Database 21(5), 1998, pp.25-30

Wreden, N. “Insight or intrusion? Data mining's effect on privacy,” Communicationsweek (650), February 17, 1997, p. 44.

Wu, K., Yu, P. S., and Ballman, A. “SpeedTracer: A Web usage mining and analysis tool,” IBM Systems Journal 37(1), 1998, pp.89-105

Wu, X. “Rule induction with extension matrices,” Journal of the American Society for Information Science 49(5), 1998, pp.435-454

Xanthopulos, Z., Melachrinoudis, E., and Solomon, M. M. “Interactive multiobjective group decision making with interval parameters,” Management Science 46(12), 2000, pp.1585-1601

Yen, M. Y., and Scamell, R. W., “A human factors experimental comparison of SQL and QBE,” IEEE Transactions on Software Engineering 19(4), April 1993, pp. 390-409.

Yoon, Y. “Discovering knowledge in corporate databases,” Information Systems Management 16(2), 1999, pp.64-71

Zahedi, F. "Reliability of information systems based on the critical success factors - formulation," MIS Quarterly 11(2), June 1987, pp. 187 - 203.

Zanasi, A. “Competitive intelligence through data mining public sources,” Competitive Intelligence Review 9(1), January-March 1998, pp. 44-54.

Zhu, D., Premkumar, G., Zhang, X., and Chu, C.H. “Data mining for network intrusion detection: A comparison of alternative methods,” Decision Sciences 32(4), Fall 2001, pp.635-660

Zivot, E., Andrews, D. W K. “Further Evidence on the Great Crash, the Oil-Price Shock, and the Unit-Root Hypothesis,” Journal of Business & Economic Statistics 10(3), 1992, pp.251-270