Why Human Expertise is Critical for Data Mining

Aziz Kaddouri, PhD
IT Consultant
Arlington, Virginia, USA
azizk03@yahoo.com

Shardul Y. Pandya. PhD
School of Business & Technology
Capella University
Richmond, Virginia USA
shardul.pandya@capella.edu

Abstract—**Current data mining (DM) technology is not domain-specific and therefore rarely generates reliable, business actionable knowledge that can be used to improve the effectiveness of the decision-making process in the banking industry. Despite this shortcoming, banks continue to rely on DM, hoping to gain a competitive edge in the face of rising global competition and eroding profits. Using PCA and CHAID algorithm to analyze and evaluate a survey of 1,000 DM analysts revealed that the relationship between the level of human expertise and the enhancement of the value of DM are statistically significant. However, the current level of integrating expertise in the process of knowledge discovery appears less-than satisfactory given the low level of perceived success (25%) in using DM technology.**

Keywords-actionable knowledge; banking industry; data mining; decision making; domain specific; human expertise

I. INTRODUCTION

The sharp decline in the effectiveness of conventional marketing strategies and rise of global competition are two key reasons some banks are unable to compete in the new global marketplace [24]. To gain a competitive edge, the banking industry has become increasingly dependent on data mining (DM) technology [20]. However, because DM is not domain-specific, it tends to generate large amount of patterns or relationships with very little business-actionable knowledge [2] [4] [27] [36] [40]. If used effectively, however, DM offers the banking/financial industry an opportunity to boost profitability by discovering hidden knowledge within the wealth of electronic data generated by customer transactions [10]. This knowledge can be used to understand and predict customer base behavior, assess the needs of current clients more accurately, and identify new key client bases more efficiently, while reducing costs and increasing market share [22].

DM, also known as knowledge discovery in databases (KDD), provides powerful search architecture and uses highly sophisticated analytical tools to process large datasets for the purpose of discovering, detecting and/or predicting patterns and behavior [7]. According to Fayyad, Piatetsky-Shapiro, and Smyth in [14], these analyses aim at identifying “valid, novel, potentially useful, and ultimately understandable patterns in data" to help management make well-informed decisions (p. 40).

Until recently, conventional statistical methods have been the main analytical tools available to decision makers. However, the advent of online banking and subsequent exponential increase in financial electronic data has tested the limits of traditional methods for handling large volumes of high dimensional data [5]. In addition, the convergence of many enabling factors (e.g., low cost of storage and computing technology, increasing ease of data collection, and the development of sophisticated machine-learning algorithms with robust computational power) have facilitated the adoption of DM technology by the banking/financial industry [7]. This technological progress has been welcomed by the business community as an opportunity to acquire an advantage in a highly competitive market [1].

However, while most competitive corporations are already using DM to discover new and useful knowledge, the lack of domain specific practicable research has significantly hampered the utility of DM for many knowledge-based industries [39] [29]. In fact, a survey of companies using DM technology showed that over 53% of companies have reported no direct improvement to their bottom line. About 20% have noticed a very little improvement, while only a small portion (about 8%) have noticed some substantial increase in their business profitability [38]. This low level of profitability shows that current DM technology needs to overcome its limitations before it can reliably generate insights that can be used to support the decision making process.

In the remainder of this paper Section 2 describes related work with an emphasis on the role of human expertise in data mining. Section 3 describes the proposed work in detail. Results with illustrations are presented in Section 4. Finally, Section 5 provides a conclusion followed by a list of references used in this paper.

II. Related work

Given the tremendous increase in the amount of electronic data generated over the past few decades, researchers and practitioners have identified DM technology as a means to uncover hidden and useful knowledge [15]. DM became particularly necessary when traditional methods to turn data to knowledge were overwhelmed by the magnitude of accumulated data [14] [21].

Over the last decades, DM technology has made tremendous technical progress. In fact, the strength of DM is its ability to synthesize multiple disciplines, including artificial intelligence, databases, machine learning, pattern recognition, and statistics, in one powerful platform to discover hidden knowledge using large datasets [29]. However, most efforts to enhance DM have focused on improving the technical performances of the algorithm, which does serve the needs of researchers. Little effort has been made to test DM's capabilities or limitations from a business viewpoint [11] [18]. The predominant use of technical metrics rather than domain-specific factors to evaluate the quality of newly discovered patterns is hampering the business application of DM [17]. In fact, many researchers suggest that business performances cannot be improved without integrating domain expertise [3] [28].

Because DM is mainly a data-driven process, Dybowski, Laskey, Myers, and Parsons in [13] suggested that human expertise could play a major role in improving the value of current DM technology, particularly in incorporating the necessary knowledge to build models of the domain. While using a case study method to investigate success factors of business applications, Dastidar in [8] found that domain expertise is one of the leading factors. In their study of the relevance of DM tools to the banking industry, Chye and Gerry in [6] concluded that DM is highly useful, but fully capturing the advantages offered by these tools requires user expertise in both the application domain and DM tools. Likewise, from their five case studies, Shortland and Scarfe in [35] concluded that a combination of human expertise and DM technology was highly effective, particularly in identifying and predicting fraud. The authors added that a domain expert is needed to relate the data to the domain problem under study.

Gur-Ali and Wallace in [19] emphasized that the function of DM should not be limited to discovering knowledge, but extended to provide insights to help businesses make better decisions. The authors highlighted the importance of relating DM technology to business goals. This could be initiated by defining "a mapping from managerial goals to the performance measures of the algorithm" (p. 3). In their study of 20 major companies, Davenport, Harris, De Long, and Jacobson in [9] pointed out that one of the reasons many organizations are facing difficulties transforming data into actionable knowledge is that those organizations have relied heavily on technology while neglecting the critical role of human expertise in providing actionable insights based on analysis and interpretation of the data. The authors developed a set of five core competency skills considered critical for any organization planning to build solid analytic capability. The core competency skills include technology, statistical modeling and analysis, knowledge of the data, knowledge of business, and communication/partnering. As each requires human expertise, one can therefore argue that human expertise, applied during two key stages of DM, could play a critical role in the success/failure of a DM project, ultimately affecting the quality of choices made by the decision-makers.

- Incorporating domain expertise during the knowledge discovery process and while evaluating mined results from a business standpoint is likely to generate higher business-actionable knowledge.
- Interpreting discovered knowledge and integrating it in a business framework with understandable business language is likely to improve the decision making outcome.

As Larose in [23] pointed out, even if DM is capable of generating patterns, it is the task of the user to ultimately determine the causes. In other words, DM alone is less likely to benefit the business [35]. Therefore, we hypothesize that without integrating human expertise, current (data-driven) DM technology is unlikely to generate substantial actionable knowledge, and consequently is less capable of improving the decision-making outcome.

III. Proposed work

Using a quantitative method (e.g. survey), this study investigated the potential role of human expertise in enhancing the value of current DM technology to improve the quality of business decisions in the banking industry. The study evaluated the relationship between the levels of expertise in five core competency skills mentioned above and the enhancement of the value of current DM technology, by asking the following questions:

- *Question 1: What is the relationship between the user's level of domain knowledge/expertise and the value of data mining?*
- *Question 2: What is the relationship between the user's level of communication expertise and the value of data mining?*

A. Conceptual framework

To date, DM is lacking a general theoretical framework and a commonly acceptable methodology to test and generalize its findings [4] [16] [39] [37] [41]. Despite this shortcoming, there is a methodology or model that is used to guide the process of knowledge discovery. This model is known as CRISP-DM [Cross Industry Standard Process for Data mining], Fig. 1. According to Shearer in [34], the CRISP-DM model, which was developed as a result of collaboration between industry leaders, DM users and service providers, organizes DM process into six phases in order to help organizations understand the process of knowledge discovery, and successfully carry out DM projects.

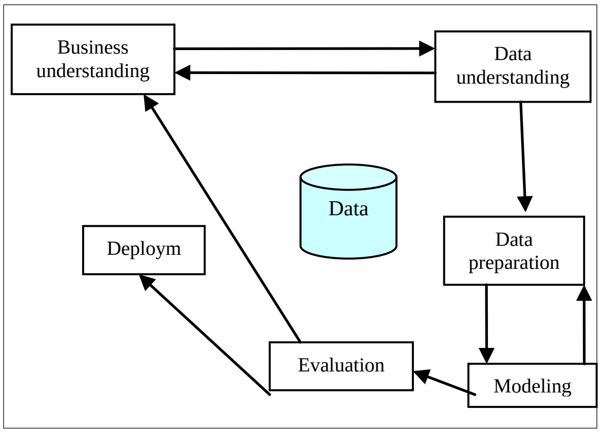

Figure 1. CRISP-DM process model. Adapted from (http://www.crisp-dm.org)

Fig. 1 shows a graphical representation of the CRISP-DM model, which illustrates the life cycle of a DM project. The model is organized into six phases, which are represented by the boxes and the relationships between different phases are indicated by arrows. The model is cyclical in nature, and going back and forth between different phases illustrates the iterative character of the process of knowledge discovery.

In this study, the CRISP-DM model is used as a guide to investigate the role of human expertise in improving the value of DM technology. To reach this goal, key areas where human expertise can contribute to the success of DM were evaluated using the concept of critical success factors (CSFs) as a framework. As defined by Rockart in [31], CSFs are "the limited numbers of areas in which results, if they are satisfactory, will ensure successful competitive performance for the organization" (p. 85). The author added that failure to get satisfactory results in those limited areas would prevent the organization from reaching its desired goals.

Human expertise as defined in this study consists of a set of five factors or competency skills developed by Davenport et al. in [9]. The relationship between the levels of expertise in these five core competency skills and the enhancement of the value of current DM technology was investigated, by surveying a random sample of DM analysts working for the banking industry. Results obtained from the survey were analyzed using two statistical analyses. First, a Principal Components Analysis (PCA) was conducted to determine the most important factors used by the banking industry to cope with DM limitations. Second, a series of Chi-Square tests was conducted to evaluate the relationship between the level of domain expertise and communication and the improvement of the value of DM.

B. Sample

Commercial banks and financial institutions were the target population. A random sample of a 1,000 DM analysts working for the banking industry was selected. A list of e-mail addresses of individuals working for different banks and financial institutions was selected from a panel of roughly 2.5 million of qualified survey respondents. The list belongs to Zoomerang, a full service online market research firm. In this case, selecting respondents was based on their function (i.e., data mining specialists working for the banking/financial industry). When the list of e-mails was generated, a random sample (i.e., every tenth number of the database) was selected to be surveyed. The survey was administrated online using the e-mails of a 1,000 qualified DM analysts. The researcher collected 200 completed questionnaires (about 20% response rate) to conduct the analysis.

C. Instrumentation

The study used a non-experimental design (e.g., survey). One of the strengths of survey research is that it can be used to obtain information from large samples of the population, making generalization relatively easy. This information can be used to describe quantitatively certain aspects of a population under study. Information about behaviors, attitudes, and opinions of the target population are frequently collected using a survey method [26]. However, as Salant and Dillman in [32] cautioned the information obtained from the population is only estimates, which are different from exact measurements.

Unlike experimental and quasi-experimental designs, under non-experimental designs researchers cannot manipulate or control the independent variable. Therefore, under non-experimental designs the goal is not establishing causality but determining relationships (correlations). However, independent and dependent variables can be used to define the scope of the study. The number and types of variables to be included in survey research are more comprehensive than under experimental designs, where costs and logistical limitations prohibit such choices.

This study used a ten-question instrument survey based on a matrix developed by Davenport et al. in [9]. The survey instrument was used to evaluate the relationship between two levels of human expertise in a set of five competency skills (technology, statistical modeling and analysis, data knowledge, domain knowledge, and communication/partnering) and the enhancement of the value of current data mining. The levels of human expertise were developed by H. Dreyfus and Dreyfus in [12] and include *Novice* and *Expert*. To increase the response rate and make sure respondents were knowledgeable about DM technology, a professional online survey service (zoomerang.com) was used for data collection purposes instead of a traditional bulk mail. The live questionnaire has a statement of informed consent offering participants the option of opting out of the survey if they wish to. In addition, this study was approved for research by the Institutional Review Board (IRB), approval number 216060-1. The research was conducted in compliance with the IRB standards for ethical research, confidentiality, informed consent, Internet research, data storage, retention, and destruction.

D. Data Analysis

The data collected using the survey was used to conduct two types of statistical analyses. First, a Principal component analysis (PCA) was used to evaluate the most important factors in domain expertise/knowledge and communication/partnering expertise that were used by the banking industry to cope with data mining limitations.

Second, a series of Chi-Square tests was conducted: (a) testing the relationship between the level of domain expertise and the perceived improvement of the value of DM; and (b) testing the relationship between the level of communication expertise and the perceived improvement of the value of DM. Domain expertise and communication are the independent variables, while perceived success in using DM is the dependent variable.

1) Variables.

In this study domain expertise and communication expertise, respectively skills 4 and 5 in [9], were the two independent variables. And perceived success in using DM was the dependent variable.

2) Measurement.

A two-point Likert scale was used to measure the independent variables. The levels, which reflect the level of human expertise perceived to improve the value of current data mining, are as follows: 1 = *Novice*; 2 = *Expert*. The dependent variable had a two-point Likert scale. The levels, which were used to evaluate the perceived success in using data mining in the banking industry, are as follows: 1 = Low; 2 = High

This section discusses the proposed work to be carried out in this study. As mentioned above, the study used a survey research method to investigate the relationship between human expertise and the enhancement of the value of data mining in the banking industry. Results obtained from this study are fully analyzed in the next section.

IV. RESULTS

A. Introduction

Using a survey, the practicality of the banking industry's current methods for coping with DM limitations was investigated with a particular focus on the role of human expertise in improving the value of DM. For this purpose, two research questions were developed and their corresponding hypotheses were tested. While research question 1 evaluated the relationship between domain expertise and perceived success in DM, research question 2 investigated the relationship between communication expertise and the perceived success in DM projects. For the purpose of testing the two research questions, the following null hypotheses were formulated:

> *Null Hypothesis 1* (H01): There is no relationship between the user's level of domain expertise and the value of data mining.
>
> *Null Hypothesis 2* (H02): There is no relationship between the user's level of communication expertise and the value of data mining.

B. Web Survey

There were a total number of 200 responses out of 1,000 qualified respondents selected. Therefore, the total sample size for this survey is 200 respondents, which represents just 20% of the population tested. The survey was limited to individuals with over three years of experience working with DM projects in a banking/financial environment. This study used a preexisting survey instrument designed by Davenport et al. [9].

Responses to the survey were gathered by Zoomerang.com, an Online Survey & Polls Service. Data were collected and stored in a secure server and made available only to the researcher. No manual data entry was involved, which ensured the accuracy of data. Then data were downloaded into Excel 2003 for coding and conducting some summary statistics. Finally, data were imported into SPSS 17.0 for analysis and testing the two hypotheses mentioned above.

1) Principal components analysis

Before testing the hypotheses, a PCA was conducted to evaluate the most important factors used in both domain and communication expertise used in the banking industry to cope with the limitations of DM. The aim of PCA is to extract the variables, or factors, that explain the pattern of correlations within a set of observed variables. PCA consists of four steps: (a) running a correlation matrix to determine the groups of variables correlated with each other; (b) estimating the number of factors or components; (c) using rotation to make factors easier to interpret; and (d) calculating factor scores. In this case, a direct Oblimin rotation procedure was specified. A total of 20 variables were included in domain expertise factors and 13 variables for the communication expertise factors. The analysis was conducted on data gathered from 200 participants using a survey questionnaire. For the domain expertise factors, an examination of the Kaiser-Meyer Olkin measure of sampling adequacy suggested that the sample was factorable (KMO = 0.924). The analysis yielded a two-factor solution. Thirteen items loaded onto component or factor 1 and 7 factors onto component, or factor 2, as shown in (Tables 1 & 4).

TABLE 1. THIRTEEN FACTORS LOADED ONTO COMPONENT 1 FOR DOMAIN EXPERTISE

Factors	Loadings
Business Critical Issues	1.025
Business Process	1.024
Business Constraint	0.893
Business Strategy	0.890
Business Drivers	0.888
Business Objectives	0.827
Business Options	0.819
Business Models	0.776
Technology Implementation Issues	0.727
Business Environment	0.688
Competitive Factors	0.667
Technology Acceptance Issues	0.611
Financial Flow Analysis	0.517

All items on Table 1 seem to be related to business management. The remaining seven items loaded onto component 2. As shown in Table 2, all items appear to be related to business expertise.

TABLE 2. SEVEN FACTORS LOADED ONTO COMPONENT 2 DOMAIN EXPERTISE

Factors	Loadings
Pricing Techniques Mechanism	0.907
Sales and Service Improvement	0.903
Fraud Prevention	0.817
Customer Acquisition and Retention	0.798
Return on Investment	0.747
Risk Management	0.722
Financial Evaluation and Forecasting	0.545

The principal component analysis for communication expertise factors yielded also a two-factor solution. Seven variables loaded onto component or factor 1 and six variables onto component, or factor 2, as shown in (Tables 3 & 4).

TABLE 3. SEVEN FACTORS LOADED ONTO COMPONENT 1 COMMUNICATION EXPERTISE

Factors	Loadings
Ability to interpret and integrate results in the business strategy	0.91
Ability to manage others	0.81
Ability to conceptualize problems	0.69
Ability to effectively present results to management	0.67
Ability to coach/teach others	0.62
Ability to clearly interpret findings	0.57
Ability to analyze issues	0.55

All items reported on Table 3 appear to be related to interpretative skills in communication expertise.

TABLE 4. SIX FACTORS LOADED ONTO COMPONENT 2 COMMUNICATION EXPERTISE

Factors	Loadings
Ability to write clearly	0.95
Ability to be a team player	0.90
Ability to listen carefully	0.83
Ability to communicate orally	0.82
Ability to communicate clearly	0.80
Understanding of customer relationship	0.76

All variables in Table 4 appear to be related to verbal skills in communication expertise.

To evaluate the relationship between the user's level of domain and communication expertise and the perceived level of success, a series of Chi-square tests was conducted. The aim of Chi-Square test is to evaluate the null hypothesis that two categorical variables are independent. The null hypothesis is rejected when the probability value (p-value) is less than or equal to the significance level. A 5% level is used across this study.

2) *Research Question 1*

What is the relationship between the user's level of domain expertise and the value of data mining? In attempting to answer this question, a cross-tabulation was conducted as illustrated in Table 5. The cross-tabulation was conducted using the responses to the survey questions 4 and 10. Question 4 was related to the level of domain expertise (1 = *Novice*, 2 = *Expert*) that a DM analyst ought to have to improve the value of current DM technology and question 10 was related to the perceived level of success (1 = Low, 2= High) in DM projects.

TABLE 5. CROSS-TABULATION DOMAIN EXPERTISE AND PERCEIVED SUCCESS

		Perceived Level of Success		Total
		Low	High	
Domain Expertise: Business Management	*Novice*	90	20	110
		45.0%	10.0%	55.0%
	Expert	60	30	90
		30.0%	15.0%	45.0%
	Total	150	50	200
		75.0%	25.0%	100.0%

A Chi-square test was conducted to test whether a relationship exists between the domain expertise (business management) and perceived success of DM. As shown in Table 6, the null hypothesis H01 was rejected with a p-value of 0.014 at the p =0.05 significance level. These results suggest that the relationship between the level of user's domain management expertise and perceived level of success in DM was statistically significant.

TABLE 6. CHI-SQUARE TEST DOMAIN EXPERTISE AND PERCEIVED SUCCESS

Chi-Square Tests					
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	6.061^a	1	0.014		
Continuity Correction^b	5.279	1	0.022		
Likelihood Ratio	6.051	1	0.014		
Fisher's Exact Test				0.021	0.011
Linear-by-Linear Association	6.03	1	0.014		
N of Valid Cases	200				
a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 22.50.					
b. Computed only for a 2x2 table					

Using a Chi-square test, the relationship between the user's level of domain expertise (business expertise) and perceived level of success was also statistically significant. With a p-value of 0.037 at the p =0.05 significance level. These results suggest that the relationship between the level of user's business expertise and perceived level of success in DM was statistically significant.

3) Research Question 2

What is the relationship between the user's level of communication expertise and the value of data mining? To answer this question, a cross-tabulation was conducted as illustrated in Table 7. The cross-tabulation was conducted using the responses to the survey questions 5 related to the level of communication expertise (1 = *Novice*, 2 = *Expert*) that a DM analyst ought to have to improve the value of current DM technology and question 10 related to the perceived level of success (1 = *Low*, 2 = *High*) in a DM project.

TABLE 7. CROSS-TABULATION COMMUNICATION EXPERTISE AND PERCEIVED SUCCESS

		Perceived Level of Success		Total
		Low	High	
Communication Expertise:	*Novice*	51	8	59
Interpretative Skills		25.5%	4.0%	29.5%
	Expert	99	42	141
		49.5%	21.0%	70.5%
	Total	150	50	200
		75.0%	25.0%	100.0%

A Chi-square test was conducted to test whether a relationship exists between the user's level of communication expertise (Interpretative Skills) and perceived level of success in DM. As shown in Table 8, the null hypothesis was rejected with a p-value of 0.016 at the p =0.05 significance level. These results suggest that the relationship between the user's level of communication expertise and perceived level of success in DM was statistically significant.

TABLE 8. CHI-SQUARE TEST COMMUNICATION AND PERCEIVED SUCCESS

Chi-Square Tests					
	Value	Df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	5.842^a	1	0.016		
Continuity Correction^b	5.009	1	0.025		
Likelihood Ratio	6.35	1	0.012		
Fisher's Exact Test				0.019	0.011
Linear-by-Linear Association	5.813	1	0.016		
N of Valid Cases	200				
a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 14.75.					
b. Computed only for a 2x2 table					

The relationship between the user's level of communication expertise (verbal skills) and the perceived level of success was not statistically significant, using a Chi-Square test. With a p-value of 0.425 at the p = 0.05 significance level. Therefore, the null hypothesis was accepted stating that the relationship between the level of user's communication expertise (verbal skills) and perceived level of success in DM was not statistically significant.

Table 9 synthesizes respondents' level of expertise needed to improve the value of DM in the banking industry.

TABLE 9. CORE COMPETENCY SKILLS AND LEVELS OF EXPERTISE

	Level of Expertise		
Core Competency	Novice	Expert	Missing values
Technology Expertise			
Respondents	125	74	1
Percent	62.8%	37.2%	
Statistical Modeling and Analysis			
Respondents	115	85	0
Percent	57.5%	42.5%	
Knowledge of Data			
Respondents	107	91	2
Percent	53.5%	45.5%	
Domain Expertise			
Respondents	115	85	0
Percent	57.5%	42.5%	
Communication/Partnering			
Respondents	60	140	0
Percent	30.0%	70.0%	

As shown in Table 9, the majority of respondents (n = 200) appeared to indicate that a data mining analyst ought to have only a novice level of expertise in technology (63%), statistical modeling and analysis (58%), knowledge of data (54%), domain expertise (58%) to improve the value of current data mining technology. However, on the other hand respondents believed that an analyst ought to be an expert in communication (70%) to successfully contribute to the value of current data mining technology.

A series of Chi-Square tests was also conducted for the other three core competency skills, namely technology expertise, statistical modeling and analysis, and knowledge of data. In addition, the set of skills respondents believe that an ideal DM ought to have improve the value of current DM are summarized in the next section.

Technology expertise**.** A Chi-square test showed that there was an association between the user's level of expertise in technology and perceived level of success in DM. With a value of $X^2 = 4.694$, $df = 1$, p-value = 0.030 and p =0.05 significance level, the relationship between the two variables was statistically significant.

Statistical modeling and analysis**.** A Chi-square test showed that there was an association between the user's level of expertise in statistical modeling and analysis and perceived level of success in DM. With a value of X^2 = 6.554, df = 1, p-value = 0.010 and p = 0.05 significance level, the relationship between the two variables was statistically highly significant.

Knowledge of data**.** A Chi-square test showed that there was an association between the user's level of expertise in knowledge of data and perceived level of success in DM. With a value of $X^2 = 10.797$, $df = 1$, p-value = 0.001 and p = 0.05 significance level, the relationship between the two variables was statistically highly significant.

Perceived level of success in data mining**.** Fig. 2 illustrates the respondents' perceived level of success dealing with current DM technology. Only a quarter of respondents have a positive experience using data mining, while the large majority (75%) reported an unsuccessful experience. In 2003, similar results were obtained by Wang and Oppeheim in [38], demonstrating little progress has been made in improving the bottom line profits, particularly in the banking industry.

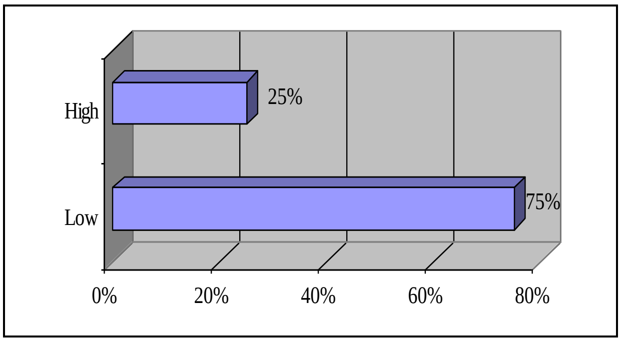

Figure 2. Perceived level of success in data mining

Areas of specialization**.** As shown in Fig. 3, respondents believed that the top four specializations that a DM analyst ought to have to improve the value of DM in the banking industry were: economics/finance (49%), computer science (45%), statistics (39.5%), and research & methodology (36%).

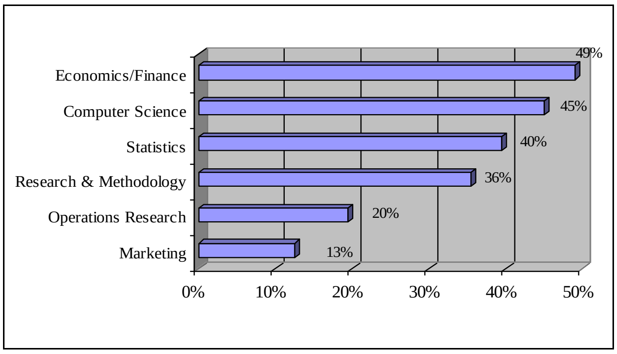

Figure 3. Areas of specialization for an ideal data mining analyst

Number of years of experience**.** As Fig. 4 illustrates, the large majority of respondents (91%) believed that a successful data mining analyst should have at least one year of experience, while 70% believe three years or more were necessary.

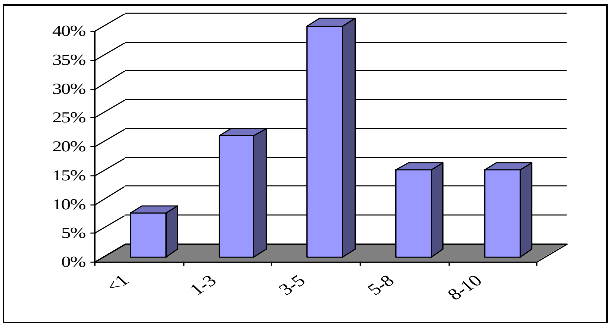

Figure 4. Number of years of experience to improve the value of data mining

Importance of data mining in decision making**.** Respondents indicated that data mining technology played an important role in the decision making process within the banking industry. In fact, the majority (75%) believed that DM played at least a somewhat important role. In addition, 52% valued the technology as being important to extremely important in the decision making, as shown in Fig. 5.

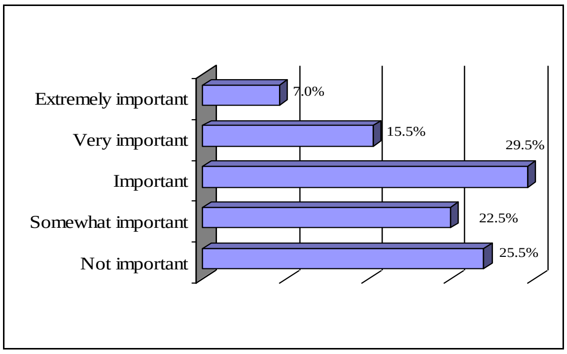

Figure 5. Importance of data mining in decision making in the banking industry

Level of education. As shown in Fig. 6, respondents viewed education as an important element in improving the value of data mining. Ninety four percent of respondents indicated that a data mining analyst should have at least four years of college. However, only (3%) believed that a doctorate degree was necessary and (4%) suggested that an analyst with an associate degree could also improve the value of a data mining project.

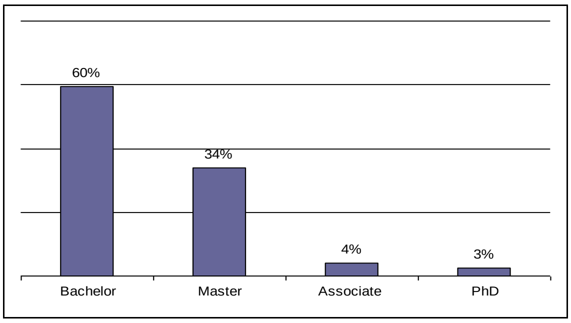

Figure 6. Level of education improving the value of data mining

C. Summary

This study conducted a Principal component analysis (PCA) and a series of Chi-square tests to determine the relationship between core competency skills and perceived level of success in DM projects. Results revealed that the relationship between a user's level of expertise in five competency skills, namely technology, statistical modeling and analysis, knowledge of data, knowledge of domain/business, and communication/partnering and the level of perceived success in data mining expertise was statistically significant. These findings would be of interest to the business community, practitioners/researchers and curricula developers, who are directly involved in DM industry.

D. Study implications

The study revealed that the banking industry's current use of human expertise is far from being satisfactory, which makes the industry ill-prepared to cope with data mining limitations, namely low level of business actionable knowledge and lack of domain-specific metrics. In fact, results showed that the banking industry was neglecting the importance of human expertise, despite research suggesting that human expertise plays a critical role in enhancing the value of DM [9].

E. Study limitations

Use of a non-experimental design (survey) can be a limitation in this study. A survey as an instrument of data collection relies mainly on self-reporting of respondents and their perception towards the issue under study not on accurate measurements. This could limit the chance of this study to generalize its findings.

V. Conclusion

The findings from the research supported the overall hypothesis that the user's level of expertise and perceived level of success in data mining projects are associated. To our knowledge, this is the first time that the hypothesis has been tested and the practicality of the banking industry's current methods for coping with DM limitations has been evaluated. However, it is important to note that the study only suggested that the level of expertise and perceived success were associated or correlated, but did not infer a cause-and-effect relationship between the two variables solely based on these correlations. These results were also consistent with the findings of Davenport et al. in [9], who postulated that the five core competencies skills are critical success factors for any organization trying to transform data into useful knowledge.

The results also supported the domain-driven data mining approach that advocates the use of human expertise in all stages of knowledge discovery to improve the value of current data mining technology [4] [39] [23] [25] [30] [33] [38]. However, the current level of integrating human expertise in the process of knowledge discovery appears less-than satisfactory given the low level of perceived success (25%) in using DM technology.

In fact, while the five core competencies are not seen as important, the very respondents who undervalue them admit that (a) their data mining efforts are not successful, (b) they are not trying to add staff with those competencies, and (c) they will continue to rely on DM. Clearly that increased reliance is only justified if they find a way to make it work for them-and this study implies that effective use of those core competency skills could address that problem.

VI. REFERENCES

[1] Apte, C., Bing, L., Pednault, E., & Smyth, P. (2002). Business applications of data mining. *Communications of the ACM*, *45*(8), 49-53. Retrieved from Business Source Complete database.

[2] Cao, L. (2009). Introduction to domain driven data mining. In L. Cao, P. S. Yu, C. Zhang, & H. Zhang (Eds.), *Data mining for business applications* (pp. 3-10). New York, NY: Springer.

[3] Finlay, D. D., Nugent, C. D., Haiying, W., Donnelly, M. P., & McCullagh, P. J. (2010). Mining, knowledge and decision support. Technology & Health Care, 18(6), 429-441. doi:10.3233/THC-2010-0603.

[4] Cao, L., & Zhang, C. (2007). The evolution of KDD: Towards domain-driven data mining. *International Journal of Pattern Recognition & Artificial Intelligence*, *21*(4), 677-692. Retrieved from Academic Search Premier database.

[5] Chopoorian, J., Witherell, R., Khalil, O., & Ahmed, M. (2001). Mind your business by mining your data. *SAM Advanced Management Journal* , *66*(2), 45-51. Retrieved from Business Source Complete database.

[6] Chye, K., & Gerry, C. (2002). Data mining and customer relationship marketing in the banking industry. *Singapore Management Review*, *24*(2), 1-27. Retrieved from Business Source Complete database.

[7] Chye, K. H., Chin, T. W., & Peng, G. C. (2004). Credit scoring using data mining techniques. *Singapore Management Review*, *26*(2), 25-47. Retrieved from ABI/INFORM Global database.

[8] Dastidar, S. (2009). Strategic elements of software product business. *ICFAI Journal of Business Strategy*, *6*(1), 17-23. Retrieved from Business Source Complete database.

[9] Davenport, T., Harris, J., De Long, D., & Jacobson, A. (2001). Data to knowledge to results: Building an analytic capability. *California Management Review*, *43*(2), 117-138. Retrieved from Business Source Complete database.

[10] Datta, R. P. (2008). Data mining applications and infrastructural issues: An Indian perspective. *ICFAI Journal of Infrastructure*, *6*(3), 42-50. Retrieved from Business Source Complete database.

[11] Debuse, J. (2007). Extending data mining methodologies to encompass organizational factors. *Systems Research & Behavioral Science*, *24*(2), 183-190. doi:10.1002 /sres.823

[12] Dreyfus, H., & Dreyfus, S. (1985). *Mind over machine: The power of human intuition and expertise in the era of the computer*. NY: Free Press.

[13] Dybowski, R., Laskey, K., Myers, J., & Parsons, S. (2004). Introduction to the special issue on the fusion of domain knowledge with data for decision support. *Journal of Machine Learning Research*, *4*(3), 293-294. doi:10.1162/153244304773633825

[14] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996a). From data mining to knowledge discovery in datasets. *American Association for Artificial Intelligence*, 17, 37-54.

[15] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996b). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*(11), 27-34.

[16] Fayyad, U, Piatetsky-Shapiro, G., & Uthurusamy, R. (2003). Summary from the KDD-03 panel: Data mining: The next 10 years. *SIGKDD Explorations Newsletter*, *5*(2), 191-196. doi:10.1145/980972.981004

[17] Freitas, A. A. (2006, Autumn). Are we really discovering "interesting" knowledge from data? *BCS-SGAI Magazine*, *9*(1), 41-47.

[18] Ghani, R., & Soares, C. (2006). Data mining for business applications: KDD-2006 workshop. *ACM SIGKDD Explorations Newsletter*, *8*(2), 79-81. doi:10.1145 /1233321.1233332

[19] Gur-Ali, O. F., & Wallace, W. A. (1997). Bridging the gap between business objectives and parameters of data mining algorithms. *Decision Support Systems*, *21*(1), 3-15.

[20] Hormozi, A., & Giles, S. (2004). Data mining: A competitive weapon for banking and retail industries. *Information Systems Management*, *21*(2), 62-71. Retrieved from ABI/INFORM Global database.

[21] Kriegel, H. P., Borgwardt, K. M., Kröger, P. , Pryakhin, A., Schubert, M., & Zimek, A. (2007). Future trends in data mining. *Data Mining and Knowledge Discovery*, *15*(1), 87-97. Retrieved from ABI/INFORM Global.

[22] kumar, J., Tejaswi, A. A., Srinivas, G. G., & kumar, A. (2010). CRM system using UI-AKD approach of D3M. *International Journal On Computer Science & Engineering*, 1(2), 159-163. Retrieved from Business Source Complete database.

[23] Larose, D. T. (2005). *Discovering knowledge in data: An introduction to data mining*. Hoboken, NJ: John Wiley & Sons.

[24] Lau, K., Chow, H., & Liu, C. (2004). A database approach to cross selling in the banking industry: Practices, strategies and challenges. *Journal of Database Marketing & Customer Strategy Management*, *11*(3), 216-234. Retrieved from ABI/INFORM Global database.

[25] Finlay, D. D., Nugent, C. D., Haiying, W., Donnelly, M. P., & McCullagh, P. J. (2010). *Mining, knowledge and decision support. Technology & Health Care*, 18(6), 429-441. doi:10.3233/THC-2010-0603

[26] Marczyk, G., DeMatteo, D., & Festinger, D. (2005). *Essentials of research design and methodology.* Hoboken, NJ: John Wiley & Sons.

[27] Padmanabhan B., Tuzhilin A. (1999). Unexpectedness as a measure of interestingness in knowledge discovery.

Decision Support Systems, *27*(3), 303-318. doi:10.1016/S0 167-9236(99)00053-6

[28] Pechnizkiy, M., Puuronen, S., & Tsymbal, A. (2005). Why data mining research does not contribute to business? Retrieved from http://www.win.tue.nl/~mpechen/ publications/DMBiz05_Pechenizkiy_etal_camera.pdf

[29] Pechenizkiy, M., Puuronen, S., & Tsymbal, A. (2008). Towards more relevance-oriented data mining research. *Intelligent Data Analysis*, *12*(2), 237-249. Retrieved from Business Source Complete database.

[30] Yong Seog Kim (2011). Multi-objective clustering with data- and human-driven metrics. *Journal of Computer Information Systems*, 51(4), 64-73.

[31] Rockart, J. (1979). Chief executives define their own data needs. *Harvard Business Review*, *57*(2), 81-93. Retrieved from Business Source Complete database.

[32] Salant, P., & Dillman, D. A. (1994). *How to conduct your own survey*. NY: John Wiley & Sons.

[33] Sharma, S., & Osei-Bryson, K. M. (2009). Role of human intelligence in domain driven data mining. In L. Cao, P. S. Yu, C. Zhang, & H. Zhang (Eds.), *Data mining for business applications* (pp. 53-61). New York, NY: Springer.

[34] Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing*, *5*(4). Retrieved from http://www.crisp-dm.org/News /86605.pdf

[35] Shortland, R., & Scarfe, R. (2007). Data mining applications in BT. *BT Technology Journal*, *25*(3-4), 272-277. Retrieved from ABI/INFORM Global database.

[36] Silberschatz, A., & Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems. *EEE Transactions on Knowledge and Data Engineering*, *8*(6), 970-974.

[37] Vityaev, E.E., & Kovalerchuk, B.Y. (2008). Relational methodology for data mining and knowledge discovery. *Intelligent Data Analysis*, 12, 189-210. Retrieved from Business Source Complete database.

[38] Wang, J., & Oppenheim, A. (2003). The pitfalls of knowledge discovery in databases and data mining. In Wang, J. (Eds.). *Data mining opportunities and challenges* (pp. 220-238). Hershey, PA: IRM Press.

[39] Wang, G., & Wang, Y. (2009). 3DM: Domain-oriented data-driven data mining. *Fundamenta Informaticae*, *90*(4), 395-426. doi:10.3233/FI-2009-0026

[40] Yang, Q. (2009). Post-processing data mining models for actionability. In L. Cao, P. S. Yu, C. Zhang, & H. Zhang (Eds.), *Data mining for business applications* (pp. 11-30). New York, NY: Springer.

[41] Yang, Q., & Wu, X. (2006). 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making*, *5*(4), 597-604. Retrieved from Library, Information Science & Technology Abstracts database.