Journal of International Technology and Information Management

Volume 16 | Issue 4 Article 4

2007

Improving Credit Card Operations with Data Mining Techniques

Malini Krishnamurthi
California State University, Fullerton

Improving Credit Card Operations with Data Mining Techniques

Malini Krishnamurthi
California State University, Fullerton

ABSTRACT

Consumer credit is ubiquitous and lending poses credit risk – the risk of economic loss due to the failure of a borrower to repay according to the terms of his or her contract with the lender. And so, managing credit risk entails estimating the potential ability of borrowers to repay their debts. Researchers have sought to identify factors that contribute to consumer risk, by using quantitative models. However, the presence of data mining techniques to identify credit risk cannot be ignored. There is a paucity of research to demonstrate the use of data mining techniques in this context, and such studies could be instructive to practitioners and academicians. This study fills that void. Using a data mining tool, this study shows that consumers can be segmented by their characteristics such as education, income, years on the job, and payment habits. The study showed that the rich were highly educated and always paid in full. Delinquency experiences were more frequent in the lower income segments. Knowledge about the risk of delinquency can be useful for lenders to price for credit risk and therefore to expand the reach of credit to consumers without having to compromise on profitability.

INTRODUCTION

Scott (2007) reports that the average United States household has 8 credit cards, which are used to charge nearly $2 trillion in goods and services annually. Further the study reports that consumers use credit cards inappropriately and spend beyond their means thereby accumulating inessentials that they cannot reasonably afford. According to the Federal Reserve Board's survey of consumer finances done in 2004, 76% of U.S. families carried some form of debt. Credit use was prevalent among families of all types regardless of their age, race, ethnicity, housing status, net worth and work-force status of the household head. Further, the percentage of families holding credit cards issued by banks rose from 16% in 1970 to about 71% in 2004. Debt was carried by 90% of the families in the top income groups and 53% in the lower income groups. The amount of outstanding debt was over $900 billion on bank-type credit cards at the end of 2006 (Scott, 2007). Further, about 1.56 million households, or about 1.4% of all U.S. households, filed for bankruptcy. Job loss was reported as one of the reasons for filing for bankruptcy.

Until the late 1970s state usury laws established limits on the interest rates credit card issuers could charge on outstanding balances, which limited issuers' ability to price for credit risk. But after that, legislation relaxed the restrictions on credit card interest rates, and it allowed national banks to charge market-determined rates throughout the country. This reduction in legal impediments, along with improvements in information technology, allowed for the development of risk-based pricing nationally and contributed to the growth of revolving consumer credit.[1] Therefore, estimating the risk of potential borrowers is very important.

"Risky" customers are those who have been delinquent. Credit card companies need to identify low-risk and high-risk customers because card rates are set based on the degree of risk that consumers reflect. For instance, Citigroup claims that it added 5.99% to the prime rate (which was 6.28%) in the case of low-risk customers, and in the case of high-risk customers it added 9.99% to the prime rate (Peng, Kou, Shi, Wise & Xu, 2005; Hsing, Gibson, Lin & Wallace, 2003).

Determining and managing credit card rates are a challenge for credit card companies, and they must do well in this challenge if they are to survive in this competitive sector (Park, 2004; Hogarth, Hilgert & Kolodinsky, 2004). There

[1] In this paper the term "consumer credit" refers to credit that is used by individuals for non business purposes and that is not collateralized by real estate or specific financial assets like stocks and bonds.

is immense competition among firms, as some firms try to offer very low teaser rates to attract potential customers. Offering low rates can lead to rising charge-offs, increase in delinquency rates and declining credit quality. On the other hand, setting high credit card rates may reduce borrowing, as customers may look for variable rate credit cards or may consider home equity loans which are more responsive to market rate movements (Park, 2004; Hsing, et al., 2003). And so, determining credit risk and setting prices to match that level of risk is crucial for card issuers if their operation is to be profitable.

Bankruptcy and delinquency have been identified as important factors that contribute to high credit card rates (Hsing, et al., 2003; Brausenberger, Lucas & Roach, 2004; Park, 2004; Peng, et al., 2005). The Federal Reserve reports that credit card rates are high due to a large number of bankruptcies. In the second half of 2005 bankruptcy filings increased sharply ahead of the October enactment of the stricter bankruptcy provisions passed by Congress.

And so, given this enormous size of personal debt that people tend to carry, and the need for card companies to forecast risk and to be selective about their customers, it is understandable that researchers are keen on developing models that can identify consumer behavior and characteristics that contribute to their risk profile. Research endeavors so far have sought to classify "risky" and "safe" customers based on the use of quantitative models such as the multiple criteria linear programming method (Peng, et al., 2005) and the multi-criteria decision aid method, as well as machine learning algorithms (Matsatsinis, 2002) and other statistical models (Lee & Kwon, 2002). While these models are impressive in theory and quantitative rigor, they lack empirical validation and practical significance. Further, there is very little known about their ease of use. Therefore, the purpose of this paper is to demonstrate how a data mining technique called the "unsupervised clustering technique" can be used to identify risk among credit card users. This data mining tool employs the rigor of the hierarchical clustering algorithm and relies on MS Excel's user-friendly front-end interface.

This paper is organized as follows: Section Two presents the architecture of a Decision Support System. Section Three presents the relevant literature, Section Four contains the methods, and Section Five presents the results. Discussion of the results is presented in Section Six. Section Seven contains the implications for research and practice.

BUSINESS INTELLIGENCE SYSTEMS

Today, we are witnessing an unparalleled digital revolution brought about by the Internet, intranets, superb multimedia, and powerful and relatively inexpensive computing and storage facilities. This digital revolution has resulted in major changes in the way business is conducted and has serious implications for managerial decision making. In order to make good business decisions managers rely on Business Intelligence (BI), which used to be known as Decision Support Systems (DSS). DSS incorporate data mining (DM) techniques to extract hidden predictive information from large volumes of data and transform them to useful information and knowledge intelligently. BI systems mean a large payoff to companies (McManus & Snyder, 2003).

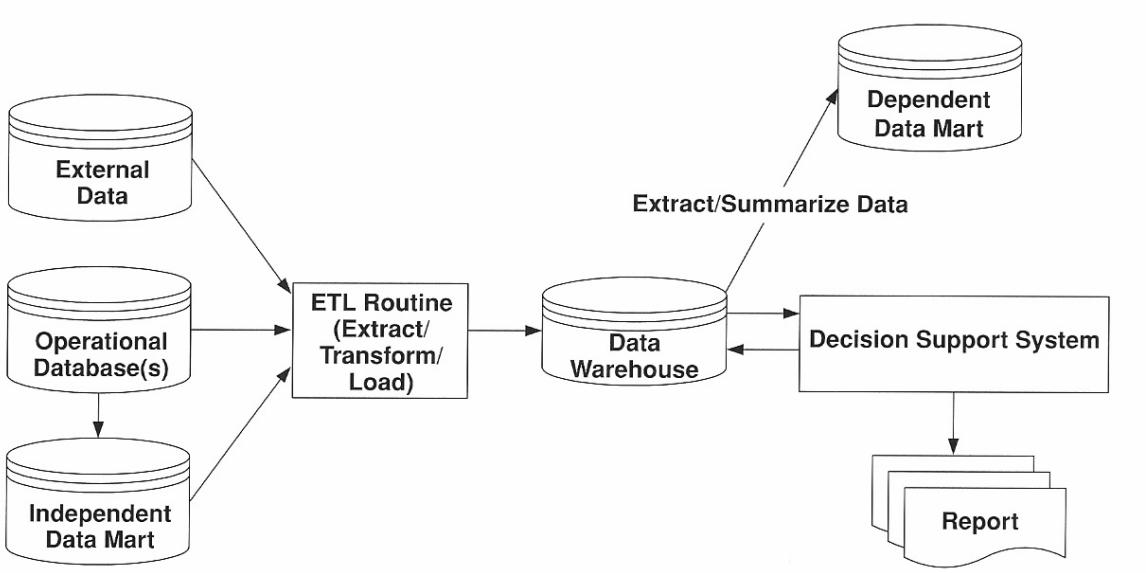


Figure 1: Information Systems Architecture for a Decision Support System.

Figure 1 represents the architecture of a Decision Support System. Operational databases and data marts contain day-to-day data that are extracted, cleansed, transformed and loaded into a data warehouse. The data warehouse is a repository of data that is used for building various decision support systems that incorporate data mining techniques to discover patterns of data. Data Warehouses can also be integrated with Knowledge warehouses that contain temporal document retrieval techniques to retrieve time sensitive information (Kalczynski, 2005).

DSS are specialized purpose information systems designed to support decision making at virtually all levels of the organization (McCarthy & Claffey, 2005). There is a high degree of cooperation in information systems practices and the business functions in organizations (Jitpaiboon, et al., 2006). A DSS consists of input, process and output components as illustrated in the above figure. A dependent data mart is a source of input to DSS. A data mart is limited in scope and contains selected data from the data warehouse, such that each separate data mart is customized for the decision support applications of a particular end-user group. For instance, Finance & Marketing groups. Within the process component of the DSS, models and data are utilized. Microsoft's Excel is the most commonly used decision analysis tool in DSS environment to either analyze or create meaningful data to support decision making (Jessup & Valacich, 2005). Excel refers to models as "templates". Besides Excel, DSS rely on many data mining tools such as drill down tools to get a more detailed view of data, statistical analysis tools to perform correlation analysis, trend analysis, forecasting (What-if) and analysis of variance. Results from the analysis are displayed in graphical and textual formats.

LITERATURE REVIEW

A review of literature was conducted in order to identify crucial factors that contribute to bankruptcy and delinquency. They are: 1. Unemployment 2. Lifestyle expectations 3. Financial illiteracy and 4. Need for borrowing in the short term. A customer pleading bankrupt is crucial to delinquency (Peng et al., 2005). Delinquent customers reflect risk, and so identifying delinquent customers is a way of differentiating risky customers from customers who have good credit standing (Peng , et al., 2005; Hsing, et al., 2003; Lee & Kwon, 2002; Stein, 2007).

Macro economic factors play a crucial role in predicting credit card behavior patterns. Prevalence of credit cards and the accumulation of consumer debt have been reported as important drivers of economic growth (Cohen, 2007). Sustainable consumption is made possible to a large extent through the possibility of the credit card payment system. On account of a negative shock to an economy, card holders can find themselves *unemployed.* In response to this, they tend to borrow for the short term and increase their credit card indebtedness to compensate for the lost source of income, in the hope of repaying the debt as soon as they are reemployed. On account of this reasoning they

become delinquent on their monthly payments either in the same month or months that follow. If they do not find a job in the next few months, they declare bankruptcy or default on their payments (Agarwal & Liu, 2003). Consumers' usage of credit therefore is a function of job position and several other variables such as income level, education, ethnicity & marital status (Lee & Kwon, 2002).

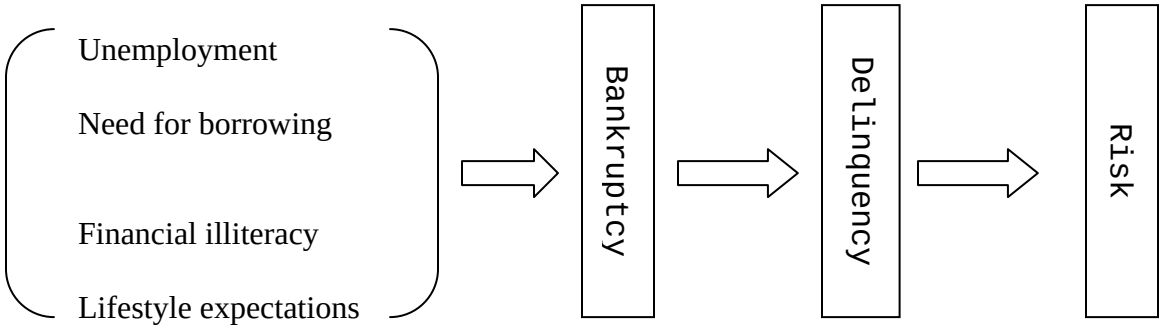
People have lifestyle expectations, and in order to maintain a desired standard of living, they borrow more than they can afford. Consumers see credit cards as *life style facilitators* (Bernthal, Crockett & Rose, 2005; Carrow & Staten, 2002; Stein, 2007). Rather than leading a frugal lifestyle based on income level, they resort to excessive borrowing and spending and eventually find themselves encumbered by debt, struggling to make even the minimum payments, and eventually becoming delinquent. Owning prestigious cards or co-branded cards and paying one card with another are also some of the common practices among credit card holders (Fock, Woo & Hui, 2005; Hogarth, et al., 2004).

Managing a lifestyle that relies on credit card spending requires knowledge of how credit cards operate. Customers should know that it is an expensive privilege, and that purchases are likely to cost more even after discounts if they plan on using credit rather than cash. And so, *financial literacy* is a key factor in credit card usage. However, knowledge of credit use is not widespread; there are many vulnerable people among the young and the old who fall prey to credit card promotions and policies and end up being indebted because they are unable to evaluate complex and competing product offerings (Scott, 2007; Stein, 2007; Braunsberger, et al., 2004).

Households use credit cards to avoid the inconvenience of paying "all cash" at once. They use credit cards as a means of payment in order to enjoy the float that is created in the meantime; or they merely use credit cards as a form of borrowing (King, 2004; Lee & Kwon, 2002; Federal Reserve, 2006 & 2007). More often than not it is people in the lower income groups who have the need to borrow (Lee & Kwon, 2002; Federal Reserve, 2007). While borrowing per se is not a problem, falling behind on payments or not being able to repay the loan is a problem. Credit use is common among all income groups (Lee & Kwon, 2002; Federal Reserve, 2006); however, it is the rich who often pay in full, while the poor rarely pay in full (Brausenberger, et al., 2004; Lee & Kwon, 2002).

This research is well informed about the several crucial factors that contribute to bankruptcy and delinquency. To date there no studies that have attempted to examine the consumer risk profile using a data mining tool. The next section illustrates the use of a data mining tool.

Proposed Model



METHODS

In the case of credit card management, classifying credit card holders' behavior into fewer than five categories is manageable and encouraged. This belief is not only based on business practice, but also comes from George Miller's classic article: "The Magical Number Seven Plus or Minus Two" (Miller, 1956). A review of the literature shows that most of the studies that have tried to explain credit card holders' behavior have applied quantitative methods such as the multiple criteria linear programming method (Peng, et al., 2005; Crook, Edleman & Thomas, 2007) or multi-criteria decision aid methods, as well as machine learning algorithms (Matsatsinis, 2002) and other statistical

models (Lee & Kwon, 2002). By applying these methods, researchers have developed classification models that are capable of separating consumer behavior by risk. These research attempts so far have provided models that are more theoretical in nature. And so, a data mining technique called the "unsupervised clustering technique" was seen as appropriate because this technique is an inductive computer learning technique that is used to create a model. The unsupervised technique identifies trends and patterns in data that do not exhibit a predefined classification. Data instances are grouped together based on a similarity scheme defined by the clustering algorithm in the data mining tool. Using the unsupervised clustering technique in this way leads to model creation.

Several trends can be identified to explain the increasing popularity of data mining. First, there is an ever-increasing volume of data available from purchasers using credit cards. Second, there is an increasing awareness of the inadequacy of the human brain to process data, particularly in situations that involve multi-factorial dependencies or correlations. Miller (1956) pointed out that the human capacity for processing information is limited to the magic number seven plus or minus two. Third, although statistical techniques are the most mature of all data mining technologies, after a certain point they tend to diverge from data mining methods that use techniques capable of accommodating multiple outliers and non-numerical data typically found in a data warehouse. Finally, data mining tools have graphical user interfaces, which make them user-friendly and which can be used by many people. Data mining techniques have wide application in business such as in e-commerce ventures (Lo & Hsieh, 2003). The data mining tool used for this study is called iData Analyzer (iDA).

iDATA ANALYZER (iDA)

iDA, a program developed by Information Acumen Corporation is an add-on to Microsoft's Excel software and therefore relies on Excel's functionality and user interface. It is installed on a user's desktop computer or a laptop computer and then it is invoked by enabling the macro option while launching MS Excel. Once the user is in the Excel program the iDA program is activated by selecting the "add-ins" option from the tools menu. If iDA is successfully installed it should appear in the "add-ins" list at which point the user can opt for it by placing a check in the check box next to iDA. iDA consists of a preprocessor, three data mining tools and a report generator. The preprocessor scans the data file for several types of errors before the data is presented to the mining engines; it corrects several types of errors such as illegal numeric values, blank lines and missing items. The preprocessor does not attempt to fix numerical data errors. It outputs a data file ready for data mining as well as a document indicating the nature and location of unresolved errors.

One of the data mining tools of iDA is called ESX, which is an exemplar-based mining tool. iDA uses a hierarchical clustering technique in which data is partitioned in a hierarchical fashion, where each level of the hierarchy is a generalization of the data at some level of abstraction. Some of the features of ESX are as follows:

1. It does not make statistical assumptions about the nature of the data to be processed.
2. It supports an automated method for dealing with missing attribute values.
3. It can be applied in domains containing both categorical and numerical data.
4. It can point out inconsistencies and unusual values in data.
5. For the supervised classification, ESX can determine those instances and attributes best able to classify new instances of unknown origin.
6. For unsupervised clustering, ESX incorporates a globally optimizing evaluation function that encourages a best instance clustering.

Results of the mining session are placed in several separate worksheets

Hypothesis: "By applying unsupervised clustering to the instances of the credit card data it will be possible to find a subset of input attributes that differentiate consumers." The choice of input attributes was limited to the ones shown in the tables that are presented in the paper. Once the clusters were well defined, a pattern of delinquency was expected to emerge in each of the clusters. Eventually, once a best set of input attributes is determined, they can be used to develop a supervised learner model for predicting future outcomes. However, such an application is not within the scope of this study.

The credit card database used in this study has data about 47 individuals. The data contains information obtained about customers through initial credit card application as well as data about their account activity. Although data can be collected through various surveys including the Internet (Bozman & Stem, 2005), the data for this study is a subset of a larger data base from a well known bank whose name is being kept confidential. Some facts have been modified to protect customer privacy and confidentiality. Although the data is small, it is useful for illustrative purposes. The data set was obtained from www.Kdnuggets.com a website that makes data sets available for Data mining and Knowledge discovery activities.

RESULTS

The data mining session was conducted on the entire dataset, keeping the default settings of the iDA analyzer for an unsupervised clustering method and using minimum rule coverage of 30 to get a better formation of clusters. Following the mining session, an overall summary of results was obtained for the session as well as for the individual clusters that were formed. The results showed three well-formed clusters with significant variables in them. The numerical attributes had significance values > .25 recommended for unsupervised clustering. Dividing by the standard deviation normalized the mean difference values, thereby making it possible to compare the attribute significance scores for all numeric attributes.

Table 1: Domain and Class Resemblance Statistics.

	Class 1	Class 2	Class 3	Domain
Res. Score:	0.724	0.698	0.623	0.50
No. of Inst.	17	14	16	47
Cluster Quality:	0.45	0.39	0.24	

We begin by analyzing the overall summary results of the data mining session given in Table 1, called domain and class resemblance statistics. The class resemblance statistics show that ESX has partitioned the 47 instances of data into 3 classes or clusters, denoted as Class 1, 2 and 3 respectively. In order to find out how well the instances in the each class fit together, we look at the class resemblance score, which is a similarity value and not a probability value. Class 1 is the best among the three well-formed classes, as indicated by the class resemblance score of .72, compared to Class 2 which has .69 and Class 3 which has .62. The domain resemblance score indicates the degree of similarity among the instances in the entire data set. As a rule it is important to focus on class resemblance scores that are higher than domain resemblance scores. Poor class resemblance scores are likely to lead to poor model performance. In this study the class resemblance scores in the case of all the three classes are higher than the domain resemblance scores and therefore are worth examining. The class resemblance score is indicative of the quality of the clusters formed and is expressed as a percentage. The cluster quality score is the percent increase or decrease of the class resemblance score relative to the domain resemblance score. A higher class resemblance score relative to the domain resemblance score results in higher values for cluster quality and therefore indicates superior cluster formation. In this case cluster 1, with a value of 45%, is the superior cluster compared to clusters 2 and 3 with their values of 39% and 24% respectively.

Table 2: Domain Summary of Categorical Attributes.

Name	Value	Frequency	Predictability
Gender	Female	26	0.55
	Male	21	0.45
Marital st	No	21	0.45
	Yes	26	0.55
Income range	0	6	0.13
	"60,000-70,000"	5	0.11
	"80,000 - 90,000"	1	0.02
	"20,000-30,000"	4	0.09
	"40,000-50,000"	4	0.09
	"30,000-40,000"	7	0.15
	"10,000-20,000"	1	0.02
	"90,000-100,000"	1	0.02
	"80,000-90,000"	4	0.09
	"100,000-110,000"	3	0.06
	"70,000-80,000"	4	0.09
	"150,000-160,000"	2	0.04
	"200,000-210,000"	1	0.02
	"110,000-120,000"	3	0.06
	"50,000-60,000"	1	0.02
Nchild	0	19	0.40
	2	15	0.32
	1	10	0.21
	3	3	0.06
Race	Caucasian	16	0.34
	Oriental	11	0.23
	Hispanic	2	0.04
	African	8	0.17
	American Ind	2	0.04
	Asian	8	0.17
Know Cred use	Yes	42	0.89
	No	5	0.11
Number of cards	1	1	0.02
	4	16	0.34
	5	20	0.43
	2	6	0.13
	3	2	0.04
	6	2	0.04
Own home	No	18	0.38
	Yes	29	0.62
Pay in full	Yes	17	0.36
	No	30	0.64
Pay min & carry bal	No	18	0.38
	Yes	29	0.62
Pay w/aNother	No	22	0.47
	Yes	25	0.53
Savings	Yes	39	0.83
	No	8	0.17
>30	No	27	0.57
	Yes	20	0.43

Next we examine the categorical attribute values for the domain summarized in Table 2. We draw several conclusions by looking at the categorical attribute values. For instance, 89% of the credit card users knew how to use a credit card. Forty-three percent of the people had at least 5 credit cards. Thirty-six percent paid their amounts in full, whereas 62% paid the minimum amount and carried a balance. Fifty-three percent paid with another credit card. Seventeen percent had no savings and 43% were delinquent. The biographic profile of the sample indicates that 55% were married women; 15% belonged to the $30-40 thousand income category; 13% had no income; 11% earned between $60-70 thousand dollars; 40% had no children and 34% were Caucasians.

Table 3a: Domain Summary of Numerical Attributes.

	Class 1	Class 2	Class 3	**Domain**	**Attribute Significance**
Age (mean)	43.88	46.86	29.31	39.81	1.22
(sd)	15.88	8.67	10.74	14.36	
Yrs in Pr job (mean)	14.82	10.57	3.19	9.60	1.12
(sd)	11.55	9.44	6.00	10.39	
yrs of ed aft High (mean)	6.47	4.29	3.88	4.94	0.83
(sd)	3.36	2.20	3.14	3.14	
BalanceDue (mean)	205.88	11,985.71	9,687.50	6,942.55	1.61
(sd)	730.13	5,727.38	7,155.13	7,299.55	

Next we examine the predictive value of each numerical attribute in the dataset by looking at the attribute significance value summarized in Table 3a. This value is a normalized value that allows the comparison of attribute significance scores for all numeric attributes. Numeric attributes with lower significance scores (less than .25) are likely to be of little value in differentiating one class from another. In this study four numerical attributes had significance scores over .25.

The next section of the domain summary in Table 3b offers insight about which categorical attributes were best able to differentiate the individual clusters. For instance, Class 1 consists of married women with no children and who had an income of over $110,000. Class 2 consists of married women with at least two children and whose income ranged from $60,000 to $70,000. Class 3 consists of single men whose income ranged from $30,000 to $40,000.

Tables 4a, 4b, and 4c summarize the results of the three classes that were formed. The tables show the number of instances within each class, and what were the most and least typical instances in each class. Two most and two least typical instances are shown. The typicality shows the average similarity of one instance compared to all other instances within each class. Examining the most and least typical instances gives us a first impression about the structure of the class.

Table 3b: Most Commonly Occurring Categorical Attribute Values.

MOST COMMONLY OCCURRING CATEGORICAL ATTRIBUTE VALUES			
	Class 1	Class 2	Class 3
Gender	Female	Female	Male
Marital st	Yes	Yes	No
Income range	"110,000-120,000"	"60,000-70,000"	"30,000-40,000"
Nchild	0	2	0
Race	Caucasian	Caucasian	Oriental
Know Cred use	Yes	Yes	Yes
Number of cards	5	5	5
Own home	Yes	Yes	No
Pay in full	Yes	No	No
Pay min & carry bal	No	Yes	Yes
Pay w/aNother	No	Yes	Yes
Savings	Yes	Yes	No
>30	No	Yes	Yes

Table 4a reflects the typical pattern in Class 1, which has a typicality score of 78%. The most commonly occurring pattern in this cluster is that of married Asian women between the ages of 45 and 58 with at least 10 years in their present job and earning an income between $110,000 and $120,000. Further, these women have two children on the average, own a home and pay their amounts in full. Women in this cluster have 6 to 8 years of education beyond the high school degree. They are knowledgeable about the use of credit cards, are far from being delinquent, and have savings. Single Caucasian or African American women between the ages of 18 to 20 who pay their amounts in full are least likely to be in this cluster.

American Indian men and Oriental women with an income between $50,000 and $70,000 and who are between the ages of 35-40 and who have 2 or 3 children are typical instances in Class 2, as shown in Table 4b. These people constitute 78% of the class. Further, the men and women in this cluster have been in their present jobs for at least 10 years and have 4 years of education beyond high school. Although they own a home and are knowledgeable about the use of a credit card, they pay only the minimum amount and carry a balance of at least $10,000. Paying one card with another and being 30 days past due are also typical characteristics in this class. However, older men and women who are between the ages of 45 to 55 and who pay the minimum amount and carry a balance are among the least typical of instances in this class.

Table 4c shows that seventy-two percent of the instances in Class 3 are unmarried men and women who have been working for 2 to 5 years on their present job, are between the ages of 24 to 35, and are in the income bracket of $30,000 to 40,000. They have on the average 2 to 4 years of education beyond high school. Although they were knowledgeable about credit cards, they paid the minimum amount and carried a balance. They paid one card with another and were often delinquent. Oriental men and Hispanic women were common in this cluster. Caucasian men or Oriental women between the ages of 30 and 60 with income between $80,000 to $100,000 and who pay the minimum amount and carry a balance are least likely to be found in this class.

Table 4a:

Class Resemblance Score:	0.72																	
Most Typical Instances:	**Age**	**Gen**	**Marst**	**Yrs in Pr job**	**Y of E aft Hi**	**Income range**	**Num child**	**Race**	**Know Cred use**	**Num of cards**	**Own home**	**Pay in full**	**Pay min & carry bal**	**Pay w/ other**	**Save**	**Bal Due**	**>30**	**Typicality**
	45	F	Yes	10	6	100,000-110,000	2	Asian	Yes	4	Yes	Yes	No	No	Yes	3000	No	0.78
	58	F	Yes	25	8	110,000-120,000	2	Asian	Yes	5	Yes	Yes	No	No	Yes	0	No	0.77
Least Typical Instances:	**Age**	**Gen**	**Marst**	**Yrs in Pr job**	**Y of E aft Hi**	**Income range**	**Num child**	**Race**	**Know Cred use**	**Num of cards**	**Own home**	**Pay in full**	**Pay min & carry bal**	**Pay w /other**	**Save**	**Bal Due**	**>30**	**Typicality**
	18	F	No	0	0	0	0	Cauc	Yes	1	No	Yes	No	No	Yes	0	No	0.64
	20	F	No	2	2	30,000-40,000	0	Afri	Yes	6	No	Yes	Yes	No	Yes	500	No	0.58

Table 4b.

	40	Male	Yes	10	4	"60,000-70,000"	3	A-Ind	Yes	5	Yes	No	Yes	Yes	Yes	12000	Yes	0.78
	35	Female	No	10	4	"50,000-60,000"	2	Orie	Yes	5	Yes	No	Yes	Yes	Yes	10000	Yes	0.76
Least Typical Instances:	**Age**	**Gender**	**Marst**	**Yrs in Pr job**	**Y of E aft Hi**	**Income range**	**Num child**	**Race**	**Know Cred use**	**Num of cards**	**Own home**	**Pay in full**	**Pay min & carry bal**	**Pay w /other**	**Save**	**Bal Due**	**>30**	**Typicality**
	56	Male	Yes	35	4	"60,000-70,000"	2	Cauc	Yes	5	Yes	No	Yes	No	Yes	800	No	0.60
	45	Female	No	1	0	"20,000-30,000"	1	Orie	No	3	Yes	No	Yes	Yes	Yes	25000	Yes	0.49

Table 4c:

Gen	Mar st	Yrs in Pr job	Y of E aft Hi	Income range	Num child	Race	Know Cred use	Num of cards	Own home	Pay in full	Pay min & carry bal	Pay w /other	Save	Bal Due	>30	Typic ality
Female	Yes	2	8	"80,000-90,000"	1	Orie	Yes	4	Yes	No	Yes	Yes	Yes	15000	No	0.52
Male	Yes	25	8	"100,000-110,000"	2	Cauc	Yes	5	Yes	No	Yes	Yes	Yes	20000	No	0.42

Tables 5a, 5b, & 5c summarize the categorical attribute values for the three different classes formed. The tables show class predictiveness and class predictability scores which are the deciding factors for class membership.

Table 5a: Categorical Attribute Summary for Class 1.

Name	Value	Frequency	Predictability	Predictiveness
Gender	Female	10	0.59	0.38
	Male	7	0.41	0.33
Marital st	No	8	0.47	0.38
	Yes	9	0.53	0.35
Income range	0	1	0.06	0.17
	"80,000 - 90,000"	1	0.06	1.00
	"40,000-50,000"	1	0.06	0.25
	"90,000-100,000"	1	0.06	1.00
	"100,000-110,000"	2	0.12	0.67
	"150,000-160,000"	2	0.12	1.00
	"200,000-210,000"	1	0.06	1.00
	"80,000-90,000"	1	0.06	0.25
	"60,000-70,000"	1	0.06	0.20
	"70,000-80,000"	1	0.06	0.25
	"110,000-120,000"	3	0.18	1.00
	"30,000-40,000"	1	0.06	0.14
	"20,000-30,000"	1	0.06	0.25
Nchild	0	8	0.47	0.42
	1	3	0.18	0.30
	2	6	0.35	0.40
Race	Caucasian	5	0.29	0.31
	Hispanic	1	0.06	0.50
	African	4	0.24	0.50
	Asian	5	0.29	0.63
	Oriental	2	0.12	0.18
Know Cred use	Yes	17	1.00	0.40
Number of cards	1	1	0.06	1.00
	5	6	0.35	0.30
	2	1	0.06	0.17
	4	6	0.35	0.38
	6	2	0.12	1.00
	3	1	0.06	0.50
Own home	No	4	0.24	0.22
	Yes	13	0.76	0.45
Pay in full	Yes	17	1.00	1.00
Pay min & carry bal	No	16	0.94	0.89
	Yes	1	0.06	0.03
Pay w/aNother	No	17	1.00	0.77
Savings	Yes	17	1.00	0.44
>30	No	17	1.00	0.63

Class predictability tells us the percent of instances in the class having a particular value for a categorical attribute. For instance, Table 5a shows the class predictability for the attribute "pay minimum and carry a balance" to be .94 for the "no" value. This means that 94% of the instances did not pay the minimum amount but rather paid the amount in full. The attribute value predictiveness score indicates the probability that a particular instance resides in a specified class. For instance, the attribute "Pay in full" has a value of 1.00, which implies that all the instances in the sample where the people paid the entire amount have been placed in this cluster.

When the class predictiveness score and class predictability score are both equal to 1 for a particular attribute, then that particular attribute value is said to be necessary and sufficient for class membership. For instance in Table 5a the attribute "Pay in Full" has a score of 1 for predictiveness and predictability. This means that all instances in Class 1 have a "Yes" value for the attribute and all instances in the sample with this value for the attribute have been placed in Class 1 or will be classified in Class 1.

When the class predictiveness score is 1 and the class predictability score is less than 1, then it can be said that all instances in the sample with the value for the attribute reside in a particular class. However, there are some instances in that class that do not have the value for that attribute in question. In such a situation we call the attribute value sufficient but not necessary for class membership. For example, in Table 5a the attribute "income range" with a value of 80,000 to 90,000 and the attribute "number of credit cards" being 1 have a predictiveness score of 1. Such values for these attributes are sufficient but not necessary for class membership.

When the class predictability score is 1 and the class predictiveness score is less than 1, we can conclude that all instances in that class have the same value for the chosen attribute. However, some instances outside of that class also have this same value for the given attribute. The attribute is said to be necessary but not sufficient for class membership. For example, in Table 5a the value of "no" for the attribute "pay with another card"; and the value of "no" for the attribute "more than 30 days late" have a predictability score of 1, implying that these values are necessary but not sufficient for class membership.

It should be noted here that the report generator for iDA analyzer lists attribute values with predictiveness scores greater than or equal to .80 as highly sufficient. In general any categorical attribute with at least one highly predictive value should be designated as an input attribute. Also, a categorical attribute with little predictive value ought to be flagged as "unused" or "for display only".

In Table 5b all the instances in Class 2 paid the minimum amount, carried a balance, and did not pay the full amount. Knowledge about the use of credit cards and owning a home were the attributes that define the necessary condition for class membership. The predictability score of these attributes was 1. Eighty percent of the sample that had the value of $60,000 and 70,000 for the attribute income range were placed in the class. All of the instances in the sample with an income range of $50,000 to $60,000 were placed in this class. The same can be said for the number of children attribute, with a value of 3. The predictiveness score of these attributes was 1, implying a sufficient condition for class membership. While this cluster indicated attribute values that were either necessary or sufficient, there were no attributes with values that were necessary and sufficient condition for class membership.(i.e.; no attribute with a predictiveness and predictability score of 1).

Table 5b: Categorical Attribute Summary for Class 2.

Name 5b	Value	Frequency	Predictability	Predictiveness
Gender	Female	9	0.64	0.35
	Male	5	0.36	0.24
Marital st	Yes	12	0.86	0.46
	No	2	0.14	0.10
Income range	"60,000-70,000"	4	0.29	0.80
	"70,000-80,000"	3	0.21	0.75
	0	2	0.14	0.33
	"20,000-30,000"	1	0.07	0.25
	"40,000-50,000"	2	0.14	0.50
	"80,000-90,000"	1	0.07	0.25
	"50,000-60,000"	1	0.07	1.00
Nchild	2	8	0.57	0.53
	3	3	0.21	1.00
	1	3	0.21	0.30
Race	Oriental	4	0.29	0.36
	American Ind	1	0.07	0.50
	Caucasian	6	0.43	0.38
	African	2	0.14	0.25
	Asian	1	0.07	0.13
Know Cred use	Yes	13	0.93	0.31
	No	1	0.07	0.20
Number of cards	4	5	0.36	0.31
	5	8	0.57	0.40
	3	1	0.07	0.50
Own home	Yes	14	1.00	0.48
Pay in full	No	14	1.00	0.47
Pay min & carry bal	Yes	14	1.00	0.48
Pay w/aNother	Yes	10	0.71	0.40
	No	4	0.29	0.18
Savings	Yes	14	1.00	0.36
>30	Yes	7	0.50	0.35
	No	7	0.50	0.26

In Table 5c all the instances in Class 3 did not pay their amount in full. Eighty-eight percent paid the minimum amount and carried a balance. Ninety-four percent paid one card with another, and 81% of the instances in this class were thirty days past due. These attributes indicate the necessary condition for membership to this class. All those instances in the sample where the income range was between $10,000 and $20,000 and had no savings were placed in this class. These conditions were sufficient for class membership. Like Class 2, this cluster reflects no attribute values that were necessary and sufficient for class membership.

DISCUSSION

This study provides evidence that data mining can be an effective tool in identifying credit card holders' behaviors and characteristics which contribute to their risk profile. More specifically, the unsupervised clustering technique was effective in partitioning the sample into segments of consumers wherein it was possible to note delinquency patterns in each of the segments. The varying experiences in delinquency is a valuable input when estimating risk and pricing for it.

Table 5c: Categorical Attribute Summary for Class 3.

Name	Value	Frequency	Predictability	Predictiveness
Gender	Female	7	0.44	0.27
	Male	9	0.56	0.43
Marital st	No	11	0.69	0.52
	Yes	5	0.31	0.19
Income range	"20,000-30,000"	2	0.13	0.50
	"30,000-40,000"	6	0.38	0.86
	"10,000-20,000"	1	0.06	**1.00**
	"80,000-90,000"	2	0.13	0.50
	"100,000-110,000"	1	0.06	0.33
	0	3	0.19	0.50
	"40,000-50,000"	1	0.06	0.25
Nchild	1	4	0.25	0.40
	0	11	0.69	0.58
	2	1	0.06	0.07
Race	African	2	0.13	0.25
	Hispanic	1	0.06	0.50
	American Ind	1	0.06	0.50
	Oriental	5	0.31	0.45
	Caucasian	5	0.31	0.31
	Asian	2	0.13	0.25
Know Cred use	Yes	12	0.75	0.29
	No	4	0.25	0.80
Number of cards	5	6	0.38	0.30
	2	5	0.31	0.83
	4	5	0.31	0.31
Own home	No	14	0.88	0.78
	Yes	2	0.13	0.07
Pay in full	No	16	**1.00**	0.53
Pay min & carry bal	Yes	14	0.88	0.48
	No	2	0.13	0.11
Pay w/aNother	Yes	15	0.94	0.60
	No	1	0.06	0.05
Savings	No	8	0.50	1.00
	Yes	8	0.50	0.21
>30	Yes	13	0.81	0.65
	No	3	0.19	0.11

Cluster one is the safest segment as it reflects low risk. It shows that matured adults with high levels of education, longer job tenure, and who pay their balances in full are less likely to default and be delinquent. Prior research by Lee and Kwon (2002), King (2004), and Peng and colleagues (2005) confirms this finding. Credit issuers can find this segment of consumers very attractive and can reward this behavior with increased credit limits. It is possible that people in this segment are less encumbered by family responsibilities since they are in a high income group.

Clusters two and three have episodes of delinquency. People pay the minimum amount and carry a balance. Between cluster two and three, the former is less risky than the latter. It is possible that people in cluster two, given their modest level of income which when coupled with family responsibilities does not go a long way, need additional resources. And so, they could use their credit cards to borrow for the short term. This finding is supported by King (2004) and Lee and Kwon (2002). People in this group know how to use credit cards and do believe in saving.

Cluster three is the riskiest; it reflects a younger population who has to prove themselves in terms of their education and profession. They are in a low income group and their financial strength is questionable. Carow and Staten

(2002), Fock and colleagues (2005), Hsing and colleagues (2003), and Brausenberger and colleagues (2004) reached similar conclusions. Credit issuers are wary of this segment and so tend to set high rates for their riskiest segment of consumers (Hsing, et al., 2003).

In summary, results from this study are supported by prior research using other methods of forecasting risk and by the Federal Reserve's report on the survey of consumer finances conducted in the year 2004 and presented in 2006. Delinquency is common among low income households, and credit use is common to all income groups. Further, credit card use is seen as a convenient means of payment among people in high income groups and a source of borrowing for those in the low income groups. Regardless of income levels, people own an average of four to five cards.

The generalizability of these findings may seem limited on account of the small sample size; however, the fact that they are supported in the literature by authors such as Lee and Kwon (2002), King (2004), Peng and colleagues (2005) and many others mentioned above lends credibility.

IMPLICATIONS FOR RESEARCH AND PRACTICE

Credit card issuers and financial and governmental institutions can use technological tools to manage credit risk more effectively. Credit issuers can segment their market based on risk and tailor their incentives and price their products accordingly, so as to reflect the underlying risk in each segment. As result they will able to expand their reach and make credit accessible to a wider population. Data mining techniques can help to prescreen the customers so that card companies can focus their promotional mailings to consumers who satisfy the established credit criteria, thereby making the solicitation process more cost-effective.

During the application process, credit card issuers decide on the customers to whom they will extend credit and on whose accounts they will set credit limits, rates and terms. In order to make these decisions, card issuers calculate certain ratios, such as debt to income and debt service to income, that can help predict repayment capacity. Data mining techniques can be used to trace trends and patterns on critical factors such as income, employment status, length of employment, and home ownership, all of which can be useful in setting rates and credit limits.

Knowledge of consumer debt and its repayment is very important for governmental institutions since it has macroeconomic implications. Its importance is evidenced by the enactment of the Bankruptcy Abuse Prevention and Consumer Protection Act of 2005. Section 1129 of this act requires the Federal Reserve to report to the Congress on the methods by which issuers of consumer credit choose the consumers to whom they solicit and extend credit. The Federal Reserve is required to report on whether the industry's practices in these matters encourage consumers to accumulate additional debt.

CONCLUSION

The growth in consumer credit raises concerns that it may sometimes be made available to consumers who are not capable of repaying, and that the accumulation of such debt may contribute to consumer insolvency. Therefore, issuers of consumer credit use increasingly sophisticated tools to identify potential customers on the basis of their expected ability and willingness to repay accumulated debt**.** This study illustrated how clustering algorithms employed by data mining techniques can be relied upon to provide valuable insights about consumer behavior and characteristics. It can be said that income, education, and fiscal discipline indicate consumers' socio-economic status, which is closely related to their credit worthiness.

REFERENCES

Agarwal, S. & Liu, C. (2003). Determinants of Credit Card Delinquency and Bankruptcy: Macroeconomic Factors. *Journal of Economics and Finance.* Spring, 27(1), 77-84.

Board of Governors of the Federal Reserve System (2007). Report to the Congress on credit scoring and its effects on the availability and affordability of credit. Submitted to the congress pursuant to section 215 of the fair and accurate credit transactions act of 2003.

Board of Governors of the Federal Reserve System (2006). Report to the Congress on Practices of the Consumer Credit Industry in Soliciting and Extending Credit and their Effects on Consumer Debt and Insolvency. June 2006. Http://www.Federalreserve.gov.

Bernthal, M. J.; Crocket, D,: Rose, R. L,; (2005). Credit Card as Life Style Facilitators. *Journal of Consumer Research*. June, 32(1), 130-145.

Braunsberger, K, Lucas, L.A, & Roach, D. (2004). The Effectiveness of Credit Card Regulation for Vulnerable Customers. *The Journal of Services Marketing,* 18(4/5), 358-370.

Bozman, C. S. & Stem, D. E. (2005). Non-response Error within Internet Surveys: A Cautionary Note. *Journal of International Technology and Information Management,* 14(2), 109-116.

Carow, K.A, & Staten, M.E. (2002). Plastic Choices: Consumer Usage of Bank Cards versus Proprietary Credit Cards. *Journal of Economics and Finance*, 26(2), 216-232.

Cohen, M, J. (2007). Consumer Credit, Household Financial Management, and Sustainable Consumption. *Journal of Consumer Studies*, 3(1), 57-82.

Crook, N. J., Edleman, B.D. & Thomas, C, L (2007). Recent Development in Consumer Credit Risk Assessment. *European Journal of Operational Research*. 183(3), 1447-1467.

Fock, H, Woo, K, & Hui, M. (2005). The Impact of a Prestigious Partner on Affinity Card Marketing. European *Journal of Marketing*, 39(1/2), 33-53.

Gaetz, W. M., & Roiger, R. J. (2003). Data Mining – A Tutorial Based Primer. Addison Wesley.

Hogarth,J.M, Hilgert, M. A, & Kolodinsky, J.M. (2004). Consumers' Resolution of Credit Card Problems and Exit Behaviors. *The Journal of Services Marketing,* 18(1), 19-34.

Hsing, Y, Gibson, J, Lin, T & Wallace, D. (2003). Determinants of credit card rates and policy implications. *International Journal of Management.* 20(3), 395-400.

Jitpaiboon, T, Nathan, T. S. R. & Vonderembse, M. A. (2006). An Empirically Derived Taxonomy of Information Systems Integration. *Journal of International Technology and Information Management,* 15(2), 17-38.

Kalczynski, P. J.(2005). Time Dimension in the Business News in the Knowledge Warehouse. *Journal of International Technology and Information Management,* 14(3), 21-32.

King, A.S (2004). Untangling the Effects of Credit Cards on Money Demand: Convenience Usage vs. Borrowing. *Quarterly Journal of Business and Economics*, 43(1/2), 57-80.

Lo, M. & Hsieh, C. (2003). Mining the Fx Electronic Inter-Dealer Market. *Journal of International Technology and Information Management*, 12(1),61-76.

Matsatsinis, N. F. (2002). CCAS: An Intelligent Decision Support System for Credit Card Assessment. *Journal of Multicriteria Decision Analysis.* 11 (4-5), 213-220.

McCarthy, R. V. & Claffey, G. F. (2005). Task-Technology Fit in Data Warehousing Environments: Analyzing the Factors that Affect Utilization. Journal of International Technology and Information Management, 14(4), 45-60.

McManus, D. J. & Snyder, C. A. (2003). Knowledge Management: The Role of Epss. *Journal of International Technology and Information Management*, 12(2), 17-28.

Jessup. L. & Valacich, Joseph (2005). Information Systems Today . Why IS Matters. Second Edition. Prentice Hall. 219.

Lee, J. & Kwon, K. (2002). Consumers' Use of Credit Cards: Store Credit Card Usage as an Alternative Payment and Financing Medium. *The Journal of Consumer Affairs*, 36(2), 239-262.

Miller, G. A. (1956). "The Magic number Seven Plus or Minus Two". *The Psychology Review*. 63 (2), 101-105.

Park, S. (2004). Consumer Rationality and Credit Card Pricing: An Explanation Based on the Option Value of Credit Lines. 25(5), 243-254.

Peng, Y., Kou, G., Shi, Y., Wise, M & Xu, W. (2005). Discovering Credit Cardholders' Behavior by Multiple Criteria Linear Programming. *Annals of Operations Research*, 135, 261-274.

WWW.Kdnuggets.com . Data set downloaded on 5/20/06.

Scott, Robert H, III (2007). Credit Card Use and Abuse: A Veblenian Analysis. *Journal of Economic Issues*, 41(2), 567-574.

Stein, K. D. (2007). Wrong Problem, Wrong Solution: How Congress Failed The American Consumer. *Emory Bankruptcy Developments Journal*, 23(2), 619-646.