

Improvement of performance of Association Rule mining using Apriori Algorithm and Logistic Regression

Ahasan-ul Habib Basunia
Student Id: 012111010

A Thesis
in
The Department
of
Computer Science and Engineering



Presented in Partial Fulfillment of the Requirements
For the Degree of Master of Science in Computer Science and Engineering
United International University
Dhaka, Bangladesh
December 2017
Ahasan-ul Habib Basunia, 2017

Approval Certificate

This thesis titled "**Improvement of performance of Association Rule mining using Apriori Algorithm and Logistic Regression**" submitted by **Ahasan-ul Habib Basunia**, Student ID: **012111010**, has been accepted as Satisfactory in fulfillment of the requirement for the degree of Master of Science in Computer Science and Engineering on 31st January 2018.

Board of Examiners

1.

Dr. Mohammad Nurul Huda
Professor and MSCSE Coordinator
Dept of CSE, UIU

Supervisor

2.

Dr. Swakkhar Shatabda
Associate Professor, CSE, UIU

Head Examiner

3.

Dr. Dewan Md. Farid
Associate Professor, CSE, UIU

Examiner-I

4.

Mr. Suman Ahmmed
Assistant Professor, CSE, UIU

Examiner-II

5.

Dr. Salekul Islam
Head and Associate Professor
Dept of CSE, UIU

Ex-Officio

Declaration

This is to certify that the work entitled “**Improvement of performance of Association Rule mining using Apriori Algorithm and Logistic Regression**“ is the outcome of the research carried out by me under the supervision of Professor Dr. Mohammad Nurul Huda.

Ahasan-ul Habib Basunia
Student Id: 012111010
Computer Science and Engineering

In my capacity as supervisor of the candidate’s thesis, I certify that the above statements are true to the best of my knowledge.

Professor Dr. Mohammad Nurul Huda

Abstract

In this research we like to apply popular data mining techniques e.g. Apriori algorithm on Health and Demographic Surveillance System , Matlab data of International Center for Diarrhoeal Disease Research, Bangladesh (ICDDR,B) to extract association rule for learning the knowledge considering existing population according to their socioeconomic, educational, migration, birth, death, marriage and divorce condition. Apriori algorithm has some limitations which are wasting of time for rule mining by scanning the enter database and generated some unexpected association rules. Now we present an Apriori with Logistic Regression model which is save that wasted time depending on scanning only selected transactions and stopped to generated unexpected association rules. The thesis shows by investigational observations with different values of minimum support that applied on only the Apriori model and newly implemented Apriori with Logistic Regression model that implemented new model (Apriori with Logistic Regression model) saves the time consumed by 82.09% in contrast with the only Apriori model as well as reduces the unexpected rule mining 37.93 %, and makes the association rule mining more efficient and minimum time spend. The models will develop in this research could be helpful during sample selection in any Demographic Surveillance System (DSS) in the ICDDR,B and over the world when a research project runs for a long time duration.

Acknowledgement

Firstly I want to show gratitude my supervisor **Professor Dr. Mohammad Nurul Huda** and **Professor Dr. S M Monzurur Rahman**, who did an enormous work. Although as a professor their schedule was full, they had time to inspire, support, and help me all the moment. They always motivated me to keep closely and provided new ideas on our discussion. They provided helpful suggestion when I run into problems, whatever they were technical, scientific or organizational.

Many thanks go as well to the whole Computer Science and Engineering Department of United International University, Dhaka, Bangladesh for fruitful deliberations and ideas. Thanks to all of them, the working setting is excellent.

Furthermore I am grateful to **Dr. Peter Kim Streatfield**, Ex Director, Centre for Population, Urban& Climate Change (CPUCC), ICDDR,B, Dhaka, Bangladesh for offering me the opportunity to get the data set for this thesis. I also wish to thank Mr. Sajal Kumar Shaha, Senior Programmer of HDSU for providing me a very clean dataset.

Special thanks to the staffs of UIU and HDSU, ICDDR,B whose support and co-operation make my life easy. My stay in UIU was a rewarding and unique experience both academically and personally.

Table of Contents

APPROVAL CERTIFICATE.....	i
DECLARATION.....	ii
ABSTRACT.....	iii
ACKNOWLEDGEMENT.....	iv
ABBREVIATIONS AND DEFINITIONS.....	vii
LIST OF TABLES.....	ix
LIST OF FIGURES.....	x

Chapter 1: Introduction

Preamble	1
1.1 Rule Mining Methods.....	2
1.2 Motivation and Scope.....	3
1.3 Research Questions and Contribution.....	4
1.4 Arrangement of the Thesis paper.....	6

Chapter 2: Data Mining Background and Introductions

2.1 Introduction.....	7
2.2 Data Mining	7
2.2.1 Relationship of Data Mining to alternating Disciplines	9
2.2.2 Data Mining Application	11
2.3 Data Mining Methods.....	11
2.3.1 Classification	11
2.3.2 Clustering	12
2.3.3 Regression.....	13
2.3.4 Association Rule.....	14

Chapter 3: Association Rule Mining and Statistics

Introduction.....	15
3.1 Generating frequent item sets.....	15

3.2	The Apriori Algorithm.....	16
3.2.1	Algorithm.....	17
3.2.2	Limitations of Apriori Algorithm	18
3.3	Logistic regression.....	19

Chapter 4: Case Study: Mining Association Rules Using Data Mining and Statistics

	Introduction.....	22
4.1	Datasets and Methodology.....	22
4.2	Data Preprocessing.....	24
4.3	Data modeling.....	26
4.3.1	Dataset 1.....	26
4.3.2	Dataset 2.....	26
4.4	Evaluations and deployment.....	27
4.4.1	Model 1.....	27
4.4.2	Model 2.....	28
4.4.3	Compare the output of Model 1 vs. Model 2.....	29

Chapter 5: Conclusion and Future Research

5.1	Conclusion.....	32
5.2	Future Works.....	32
	References	33
	Appendix A.....	36

ABBREVIATIONS

ANN: Artificial Neural Network.

ASSOM: Adaptive-Subspace Self-organization Map, an unsupervised neural network.

BP: Back Propagation, a learning algorithm for the supervised multi-layered neural network.

CF: Confidence, a data mining measure.

CPAR: Classification based on Predictive Association Rule, a data mining classification method.

DM: Data Mining:

DSS: Health and Demographic Surveillance System.

DT: Decision Tree, a data mining method for classification.

Err: Error, a data mining measure.

FCM: Fuzzy C-Means, a clustering algorithm based on fuzzy logic.

FL: Fuzzy Logic, a soft-computing method.

GA: Genetic Algorithm, a soft-computing method.

GHSOM: Growing Hierarchical Self-Organizing Map.

GRLVQ: Generalized Relevance Learning Vector Quantization, an unsupervised neural network.

GSOM: Growing Self-Organization Map, a dynamic unsupervised neural network

HFM: Hierarchical Feature Map, an unsupervised neural network.

HDSS: Health and Demographic Surveillance System.

IBL: Instance-Based Learning.

ICDDR,B : International Center for Diarrhoeal Disease Research, Bangladesh

IGG: Incremental Growing Grid, a dynamic unsupervised neural network

KDD: Knowledge Discovery in Databases, data mining is a sub-area of KDD.

KNN: Kohonen Neural Network

LVQ: Learning Vector Quantization, an unsupervised neural network.

MQE: Mean Quantization Error, a performance measure of unsupervised neural network learning algorithms.

NN: Neural Network.

SBA: Scoring Based Association, a data mining predictive method based on the association rule.

SSNN: Supervised Single-layered Neural Network.

SSSNN: Set of Supervised Single-layered Neural Networks.

SOM: Self-Organization Map, an unsupervised neural network.

Supp: Support, a data mining measure.

TS-SOM: Tree Structured Self-Organization Map.

UIU: United International University

VIA: Validity Interval Analysis, a rule mining method using multi-layered neural networks.

VQ: Vector Quantization, an unsupervised neural network.

LIST OF TABLES

Table 1: Number of rules and process time required for different support (Apriori
Without Regression).....27

Table 2: Number of rules and process time required for different support (Apriori with
Regression).....28

Table 3: Time required for different support.....29

Table 4: Number of rules for different support.....30

LIST OF FIGURES

Figure 2.1 Steps involved in Knowledge Discovery and Data Mining.....	8
Figure 2.2 Machine learning framework	10
Figure 2.3 Data Mining framework.....	10
Figure 2.4 Classifications Example.....	12
Figure 2.5 Simple linier regression of a loan data set.....	13
Figure 3.1: The Apriori algorithm.....	18
Figure 3.2: Regression of Sex on Height.....	20
Figure 4.1: Flow chart of Experimental Design.....	24
Figure 4.2: Required time vs support	30
Figure 4.3: Numbers of rules vs support.....	31

Chapter 1

Introduction

Preamble

A very popular branch of modern science is computer science, where data process is done in a systematic way. In our daily lives whatever we do it generate data, which are stored in many large databases in different ways. It is really difficult to analyze those data by statistical methods. At the present time, a new topic in computer science is data mining (DM), which is referred as knowledge discovery in databases (KDD)[19]. The DM is a method, which extracts implicit, unknown and potentiality helpful information from huge volume database. The strip-mined information is additionally referred as knowledge of the shape rules, constraints, and regularities. Rule mining is one in every of vital tasks in DM since rules give an epigrammatic statement of doubtless important information that's simply understood by the end users. Researchers have been using several techniques such as statistics, decision tree, Artificial Intelligence, database, psychological feature etc. for rule mining.

Association rule mining is a recognized technique in data mining. It is trying to show all interesting relationship in a database, but sometimes it generated some unexpected association rules that are not appropriate. In this thesis, we will try to compare two different association rule mining models and evaluate them in new way.

For an example, we do consider departmental shops sales data. And take into account the concerning customers' personal data and credit card information. The businessman can provide some attractive offer for the client by analyzing those data.

1.1 Rule Mining Methods

For rule mining aims, all data are stored in a database using DM algorithm. The main challenge of rule mining is to extract appropriate information from stored data which reflects the characteristic of the data. This knowledge is called rules. These rules are implemented to make a decision for the whole dataset.

Significant numbers of Data mining algorithms that are used for mine knowledge with numeric data. Most of the algorithms are developed by mathematics. But, association rule mining algorithm is mostly used on non-numeric and categorical data. It's counting also involves very easy way. These types of algorithm really efficient and sometimes they extract very impressive inventions.

Most of the time, Association rule mining are used to find numerous co-occurring relationships from an items collection. It is also called “Market Basket Analysis” because it is the base of association mining sector. If the target to discover associations of items from a sample set, which are happened more frequent. We try to present the famous example "Beer and Diapers association" which is commonly available in data mining books. The story is: people want to buy diapers from a store they also have a chance to buy beer together. Now we can explain it with a simple example. Consider that the store sales database shows this information:

- There are 1200,000 transactions in total.
- 15,000 transactions contain breads (1.25 percent)
- 120,000 transactions contain milk (10 percent)
- 12,000 transactions contain each bread and milk (1.0 percent)

We consider that beer and diaper both are independent variable (i.e. there was no association), so we expect 10% of diaper buyers can buy beer. Because the number of beer customer is only 10%. But, we find out that beer purchases among the diaper shopper of 80 %

(=12000/15000). The factor is increased 8 times of the expectation. It is called Lift. It is the proportion of the observed frequency and the expected frequency. It was a simple calculation of database transactions. That's why the rule of association can express; buyers of the diaper may buy beer based on the Lift factor 8. Statistically, Lift is calculated by the proportion of the associated chance of two items x and y , divided by the product of their individual probabilities: $\text{Lift} = P(x,y)/[P(x)P(y)]$. If both items are independent, then $P(x,y)=P(x)P(y)$, corresponding to $\text{Lift} = 1$ in that case. It is interesting findings where Lift value below 1 for the output of anti-correlation - similar to mutually distinct items, which are sometimes co-occurring. In 2004, Walmart has unfolded this event when one by one hurricane crossed the state of Florida.

1.2 Motivation and Scope

Nowadays association rule mining is common practice in DM. We can extract all attractive relationships with efficiently and without any doubt from a big database using Association rule. The massive amount of data captured within the set of association rules will be used not just for narrating the relationships within the info, however additionally for discriminatory between totally different sorts or categories of info instances.

On the other hand, complexity is the main difficulty of association rule mining. Even it is not possible to find all the expected relation from a usual database. For this reason to finding all interestingness need to apply a prune strategy on the search area. So we can say that any association rule mining algorithm is not able to extract all relationship, but the finding is interesting ones. It is really difficult to make a difference for output of different association rule mining algorithms. So, we can say a deficiency is available to assessment procedures for the excellence of association rule mining algorithms and their interestingness events.

Firstly, during this thesis, we try to find an interesting affinity within the comparison of the standard of mining two models. Hence, we have used Apriori Algorithm to mining association

rules. Secondly, we have used statistical logistic regression then applying Apriori to mining association rules that prefer correct rules. During this thesis, we have used to compare totally different rule mining process to mining association rules.

We illustrate the performance of different mining processes on the ensuing association rules. This thesis provides a comparative study of using association rules mining and using the statistical method with association rule mining.

1.3 Research Questions and Contribution

The rationale for this study is to investigate causes and association of out-migration for Matlab Health and Demographic Surveillance System (HDSS). The out-migration has been designated as an important public health problem and a major cause of appropriate population sample selection for a research program in Demographic Surveillance System (DSS). That's why research staff cannot able to enroll expected subject in due time with fixed manpower. Therefore they need to more sample selection and extend the project time and cost. If software will develop could be helpful here to measure the out-migration and select the appropriate sample size with the help of the HDSS data which are stored in a secure historical oracle database. This software has an option to set some preselected parameters considering the database and accordingly generated the expected sample dataset by considering the out-migration.

In this research, we would like to apply popular Association rule mining techniques e.g. Apriori algorithm on HDSS, ICDDR,B, Matlab data to finding association rules existing population according to their socioeconomic, educational, migration, birth, death, marriage and divorce condition. We will use two types of dataset models to evaluated, and compared with different ways. Finally, we will consider an efficient model to mining association rules and make decisions. The model will be considered in this research could be helpful during

sample selection in any DSS in the ICDDR,B or over the world, when a research program will run for a long time duration.

Strategies to address ethical issues

1. Out-migration is common characteristics of a Demographic Surveillance System. Our proposed will includes participants in Matlab Health and Demographic Surveillance System (HDSS) conducted by the ICDDR,B at the Matlab field site in Bangladesh. The subjects' data are residing in the Matlab HDSS area (both Intervention and Comparison area). We would like to use secondary data of Matlab HDSS and consider randomize 10 village population out of 142 villages and all age groups and sex.
2. We are using only pre collected dataset. We are committed to not shear any personal detail in publicly of the participant in this study. We are observing their socioeconomic, educational, migration, birth, death, marriage and divorce condition.
3. Matlab HDSS has already collected Surveillance data with the standardized procedures.
4. The identification of all study participants will remain confidential. Every effort is made to keep the records confidential as possible. Only authorized personnel have access to participant identification information. Data are analyzed and published without reference to any name or identity. Data will be shared with collaborating institute maintaining ICDDR,B Data policy.
5. Matlab HDSS staff's have already obtain written consent from the participant or the parent/guardian of the participant. This research involves participant previously enrolled in the Matlab HDSS and collect Surveillance data schedule basis.
6. The research participants no further visits required. We use the data of Matlab HDSS database, which are already collected and maintained by HDSS personnel.

7. There are no expected direct benefits to participants in this research. However, If this research finds that the efficient association with out-migration and others variables using Matlab HDSS database, it may help to develop an efficient sample selection Software for HDSS of ICDDR,B, and other developing countries HDSS.

8. This study requires using of Matlab HDSS database records for identification causes and classifying of out migration. The parts of the records that will be accessed are socioeconomic, educational, migration, birth, death, marriage and divorce condition.

1.4 Arrangement of the Thesis paper

The thesis paper is organized as follows: in Chapter 1 we discuss rule mining method, motivation, scope, and research questions of the thesis. We try to introduce a short introduction on data mining and address association rules main parts. In Chapter 2 we have discussed most of the data mining algorithms. In Chapter 3 we discuss association rule mining algorithms and Logistic Regression which are used in this thesis. In Chapter 4 construct for data collection and process, presents our experimental setup and explain the results. And last one Chapter 5 for calculation and future works.

Chapter 2

Data Mining Background and Introductions

2.1 Introduction

Nowadays, database growth is very fast in everywhere, which is the notice about information explosion. Every day increases the amount of data in databases. Most of the database volume is calculated in terra bytes previously it was in gigabytes. This types of data are not simply illustrated and analyzed by traditional manual methods like MS Excel and others ways. We can only know some general status of the data. It cannot find out important internal information from data. Some hidden relationships may very important for the organizational context, but it cannot show. In present techniques, can analyze the very large database and find out automatically undiscovered knowledge. Not solely that, they will lose some info arising from hidden relationships between data which can have an important result on structure prospects. This ends up in the requirement for new generation of techniques and tools with the interestingness to help in analyzing the mountains of data intelligently and automatically with undiscovered knowledge. In 1995, a new deftness proposed at the First International Conference on "Knowledge Discovery in Databases". We need to apply many algorithms to extract knowledge from a giant database. This is the process of DM. The DM has changed the present world analysis. Now it is popular in business, research, and scientific analysis sectors.

2.2 Data Mining

Using statistical methods we can only summarize the data and produce reports, it is called manual or traditional data analysis. But, considering the huge amount of data growth, the statistical technique is not applicable. Though sometimes it possible that time needs more

memory and time. It is not possible any people to analyze millions of records with different kinds of dimensions. For this reason, researcher community introduced a newly automated technique of data analysis is called Data Mining (DM) or Knowledge Discovery in Databases (KDD). It is the main duty of DM how to discover knowledge efficiently and easy way from a large and complex database. Many modern organizations have solved their problem using DM technique. This is a difficult job to realize, examine and appropriately use the huge database. The extracted knowledge from the large database will serve as a reliable way and this knowledge can be utilized in an application for decision-making, information management, and different purpose.

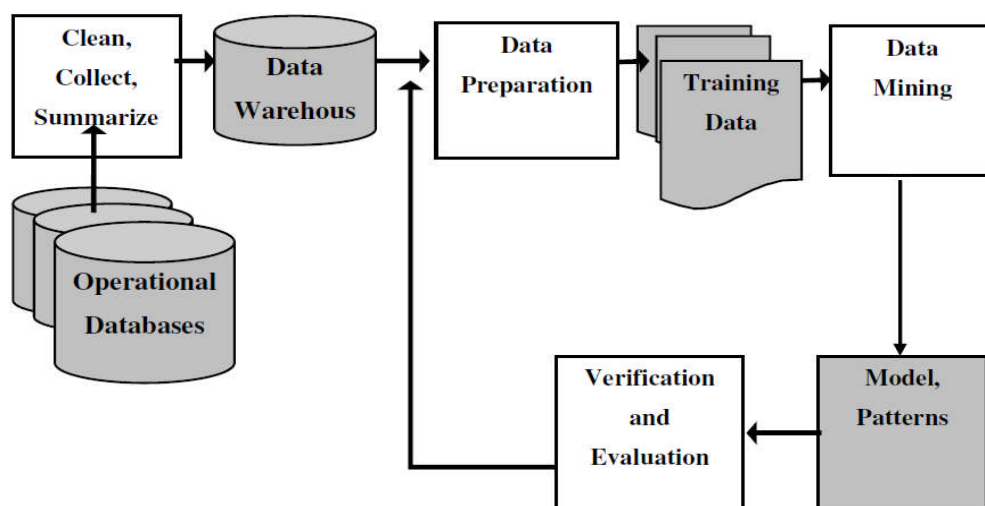


Figure 2.1 Steps involved in Knowledge Discovery and Data Mining

There are many complexes affinities are hidden in a large database; the DM algorithms can explore the relationships. The primary difference between the traditional databases and the DM systems is that the DM systems extract hidden relations, but the traditional databases show the simple pieces of information. The researchers mainly interested in DM for artificial intelligence, pattern recognition, machine learning, statistics, association rule, data visualization and knowledge acquisition.

On the basis of Hand et al. [7] Every DM algorithm could be divided into four parts:

1. Pattern structure or Model
2. Interestingness measure (score function)
3. Method of Searching
4. Data management plan

Overall, using different ways like classification rules or trees, regression, clustering, association rule, sequence modeling, dependency, and link analysis DM explore the results.

2.2.1 Relationship of Data mining to alternating Disciplines

The scientists come from the various department in KDD that's why their approaches also different. Considering the basic approaches of different researchers KDD can be classified by are five categories: machine learning approaches, applied mathematical approaches, integrated approaches, database-oriented approaches and other approaches.

For a long period, data analysis mostly depends on statistical tools. Till now it is very popular with different backgrounds professionals. Basically, a mathematical model is created to use statistical approaches, and the rules, patterns, and regularities are pulled from the model. For an example, using the training data set a Bayesian network developed and the network explore the relationships within the component. Bayesian provided a rational idea which is “the most extensively studied statistical method for knowledge discovery “[Box 1989]. It is a very powerful tool to analyze the relationships between databases. All Applied mathematics models are developed based on a theoretical foundation. Like an applied mathematics model has created based on the Bayesian distribution theorem. But, the real world data are not stored in databases based on applied mathematics model. As a result, it makes a difficulty for statisticians.

It has a very strong relationship with database and DM. Some database does not store the data in an organized way to apply DM techniques. And some are stored in semi-structured data. So, it is very important to construct a database of structured and high-level collections of resources. Where DM techniques can apply easily to access and analyze the data. For an example, if super store databases are scanned iteratively to extract information for customer purchasing habits.

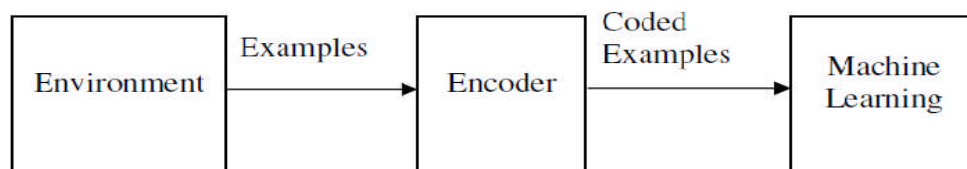


Figure 2.2 Machine learning framework

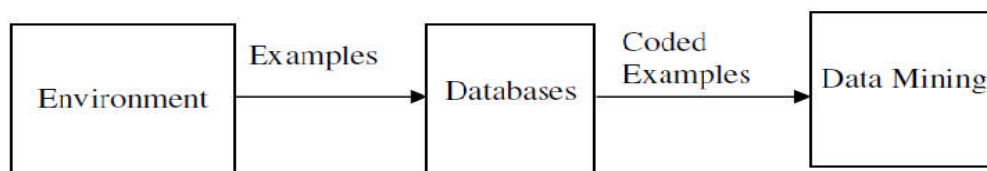


Figure 2.3 Data Mining framework

It has a significant uniformity both data mining and machine learning. We try to express a small general framework for machine learning in figure 2.2. For the machine learning, the data encoded by the encoder to understood the machine. The milieu illustrates practical field, which is learned from environment. A significant number of observations, which is the encoder encoded as the format of the machine understandable. We have shown the encoding process in Figure 2.3, where DM find the needful patterns from the database. This is the variation of all machines learning framework. When database models implemented that time

encoder replaced by the database; every stage reflects a stage of the database from the environment and every stage transition reflects another stage transition from the environment.

2.2.2 Data Mining Application

Nowadays is the massive development of information science, we also observe the much useful applications of DM. DM has successfully implemented on market analysis for product sale strategy, accurately calculate the property costing to present in the competitive market, providing the best service to the customer for better relation, fraud detection of the credit card, provides the different type of price and service in the telecommunication sector, finding the disease burden in health sector, and etc[9].

2.3 Data Mining Methods

There are different types of methods are available to achieve the DM goals. These different types of methods are used for different types of functions according to the nature of needs. We will try to discuss some main methods in briefly in this section.

2.3.1 Classification

A commonly used function of data mining, which is split fresh data into one to many categories. These categorical classes are predefined. We can define a class C_i (ith Class) in mathematically as follows:

$$C_i = \{o \in S | \text{Cond}_i(o)\} \quad (2.1)$$

Where training data set is S and o object has removed from this set, when o is a member of the category C_i .

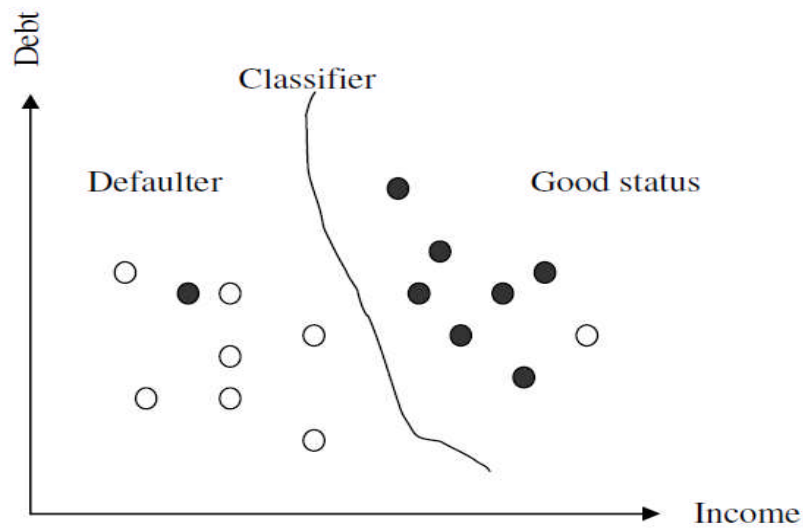


Figure 2.4 Classification example

As of some examples for classification, we can say fraud discovery on credit card, bank loan deception etc. A two-dimensional simulated dataset with 16 events is shown in figure 2.4. In the past, a person has taken a loan at the different time from a specific bank, which is shown in the figure by a circle. Here the data divided into two classes: some default their loan and some recover their loan in due time. Observing the previous information of clients, the bank can be interested to provide a loan again a person. Most of the time, this type of decision has taken by the classification technique of DM.

After analyzing the information a classifier constructs the classification method. The classifier also is known as decision function[5]. This classifier can be detected a new case the future class member. The Regression, NN, decision rules, etc are good example of a classifier.

2.3.2 Clustering

Considering some similar metric each set of data create a partition, this process is called Clustering. The similarity or proximity is the quality of a cluster and the set of objects grouped together[4]. Most of the objects are forward into a complete or reciprocally

monopolistic cluster sets. The classification classes are not similar it should be settled in data for the clustering. If data generate a few clusters means some excellent particulars are lost, but it has an entire commutation. The numbers of way are available to create clusters by data. Clustering has a great contribution in DM applications like medical diagnostics, scientific data exploration, spatial database applications, text mining, CRM, computational biology, web analysis, direct marketing and others.

2.3.3 Regression

Regression models basically used to predict one dependent variable from one or more independent variables. In statistics, many regression models are available: these are linear, multi-linear, non-linear, logistic regression. Considering previous and present events data, a

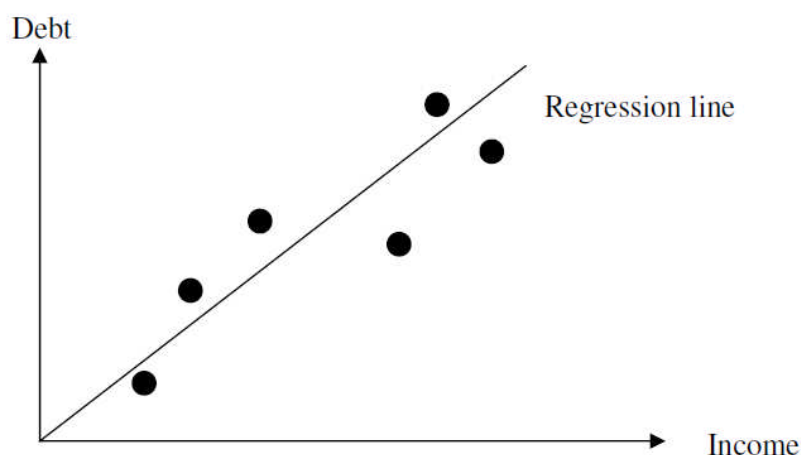


Figure 2.5 A simple linear regression of a loan data set

data miner can able to predict the present, future and even past events using regression models. The data analyst prefers the model because it is not costly or not time-consuming to gather data to make a prediction[8]. Basically, the regression model is used in everywhere. The regression model is also used in the time series prediction. An output of a simple linear regression model is shown in Figure 2.5, where the loan amount is reconciled as a linear function of income.

2.3.4 Association Rule

In 1993, Agarwal with others introduced a new data mining technique is Association rule. After that, it is made a preferring consideration particularly in market basket analysis, where retail sales transactions provided the customer buying patterns. Association rules generated an idea by the market basket analysis “When a customer buys a set of products what is the probability that he buys another product?” Formally, an association rule identifies a mixture of boolean conditions created from the attribute values of a dataset that occur along with bigger frequency than may be expected if the conditions were independent of one-another. Mathematically, an association rule is outlined as $YX \Rightarrow$, wherever X (antecedent) and Y (consequent) are logical predicates created by boolean predicates. A logical predicate in an association rule consists of 1 or a lot of Boolean conditions and that they are connected by the AND ($\wedge$) logical operator. In a very transactional information set (e.g. Sale database of a supermarket), an association rule seems as

$$(\text{item} = \text{milk}) \wedge (\text{item} = \text{sugar}) \Rightarrow (\text{item} = \text{bread})$$

That represents if someone buys milk and sugar, it's presumably that he additionally buys bread. The chance of an association rule is measured by several values e.g. support, confidence, coverage, leverage, lift etc.

Chapter 3

Association Rule Mining and Statistics

Introduction

In this chapter, we will discuss association rule mining using Apriori Algorithm. Our main target is mining association rules using the algorithmic approach. For this, we will implement two different models. We will calculate the interestingness of an association rule using the two models and their meaningful differences. Within the rule mining, we tend to have an interest in valuable time save and an extremely accurate rule set. It has no drought the Apriori algorithm [6] extract a quality set of association rules. The second methodology, implementing Logistic Regression tends to choose only significant variables after that apply Apriori Algorithm to mining association rules. Each of them turns out frequent itemsets in the same approach. An itemset is termed frequent if the support is a lot of a predefined minimum support.

3.1 Generating frequent item sets

It is an easy search problem to extract association rules. But, it is refractory for an exhaustive search; cause of the expected number of association rules is exponential within the context of a large number of attributes. If the binary attributes are n then the rules number are $O(2^n - 1)$. In context to discrete-valued attributes, the situation is worse. Considering there are n attributes and each can take m values, there would be possible rules $O(m^n)$.

In spite of that, it can be possible within a reasonable time to make a search consider support for frequent itemsets. An item set X of length k is frequent if and only if all subsets of X with length $k-1$ are frequent. This property allows the search space to be pruned substantially.

Now we start the size 1 all frequent itemsets algorithmically. All individual items are available in this frequent itemsets set which has a bigger support than the minimum support given by the user. It can be done with a linear time one pass on the data. The size 2 frequent itemsets have two steps: Firstly, the size 1 frequent itemsets generated candidate itemsets of size 2 in every possible way, then, in another pass over the data, the candidate itemsets are confirmed that they are really frequent. All infrequent ones are deleted.

The termination is mandatory, it can be empty of frequent itemsets for a distinct k , or the number of attributes is same to k . When n is the number of attributes then n linear passes over the data. Both algorithms produce association rules in various ways using a separate dimension for interestingness.

3.2 The Apriori Algorithm

The apriori algorithm [6] is the very first algorithms for association rules mining with a standard way in the data mining. To seek for association rules are guided by two parameters: support and confidence.

Apriori returns an association rule based on support and confidence values. If the user-defined threshold values are more than its support and confidence values that time no rule generates[3]. The confidence has a control over output ordering. If the confidence is same for many rules then it depends on support. That's why the Apriori algorithm indulgence more confident rules and characterizes these rules as more interesting.

There are two main factors contributes to the Apriori mining process. The first one is generating frequent itemsets. This step dependent on support based pruning, because there are itemsets considered minimum support. The second step is the generation of rules out of the frequent itemsets. In this step confidence based pruning is applied.

Rule discovery is straightforward. For every frequent itemset f and every non-empty subset s of f , apriori outputs a rule of the form $s \Rightarrow (f - s)$ if and only if the confidence of that rule is above the user-specified threshold.

3.2.1 Algorithm

The important features of the apriori algorithm are presented in the last section. There were assess of the interestingness and strategy of the pruning for parasitic the investigation area. Now we want to present the algorithm. We consider the association rules number within n , which is the only input parameter. The output is a list containing the n best rules. We will consider this list from now on as best $[n]$. It is perform to using a dominance rank.

The theme is a custom-made version from Scheffer [12]. The primary necessary step within the algorithm (see Figure 2.1) we tend to use a dynamically increasing support threshold for the support of the rule body s_{body_min} beginning with threshold 1. We tend to loop over the length of the frequent item sets, whereas they're non-empty (step 6). So within the first iteration all frequent item sets of length 1 with $s_{body_min} = 1$ area unit made. In resultant passes the frequent item sets that have a minimum of s_{body_min} are made (step 6(a)). For all those item sets we tend to decision the rule generation procedure that is explained in additional detail below. If best $[n]$ changes throughout the rule generation step we tend to increase s_{body_min} . we tend to accustomed confirm the minimum support an ideal assured rule (a rule confidence 1) should have to be compelled to get into best $[n]$ (step 6(c)). This is often our new threshold for s_{body_min} . If the new minimum support is larger than the amount of instances within the data set, the algorithm terminates. If the support threshold has increased, all item sets from the frequent item sets that have a support below minimum support are deleted. In step 6(b) we tend to decision the rule generation procedure that receives as input one frequent itemset.

```

1. Input: number of desired association rules  $n$ , database  $D$  with items  $a_1, \dots, a_k$ 
2. Set the support threshold of the rule body  $s_{body\_min} = 1$ 
3. For  $i = 1, \dots, k$  DO:
    Construct a number of associations rules of length  $i$  at random and measure
    their confidence  $\hat{c}$  provided  $s(X) > 0$ .
    Let  $\pi_i(c)$  be the distribution of confidences.
4. For all  $c$ , compute  $\_ (c)$  using equation (2.3)
5. Let  $F_0 = \{\theta\}$  be the set of frequent item sets of length 0.
6. For  $i = 1, \dots, k - 1$  Do:
    While  $(i = 1 || F_{i-1} \neq \theta$ 
        (a) Determine all frequent item sets  $X$  of length  $i$  with  $s(X) \geq s_{body\_min}$ 
        (b) For all  $X \in F_i$  call RuleGen( $X$ )
        (c) If  $best[n]$  has changed in RuleGen Then
            Increase  $s_{body\_min}$  so that,
             $E(c|1, s_{body\_min}) > E(c(best[n]) | \hat{c}(best[n]), s(best[n]))$ .
        (d) If  $s_{body\_min} > \text{size of database } D$  Then Exit.
        (e) If  $s_{body\_min}$  has been increased in step 6(d) Then
            Delete all item sets  $X$  from  $F_i$  with  $s(X) < s_{body\_min}$ .
7. Output  $best[n]$ 

```

Figure 3.1: The Apriori algorithm.

3.2.2 Limitations of Apriori Algorithm

We know that Apriori algorithm is simple and clear, but it has some weak points. Wasting of valuable time is one of the main problem when working with a massive range of candidate sets with a lot of frequent itemsets, low minimum support or huge itemsets [1]. Apriori are

going to be terribly low and inefficiency once memory capability is proscribed with sizable amount of transactions.

In this thesis, we tend to propose reducing the valuable time spent for looking out in data transactions for frequent itemsets and stop to mining unexpected rules.

3.3 Logistic Regression

In many surveys we observed that the logistic regression may applied only ordinal data. That means variables with over two ordered classes. Normally it is possible a binary dependent variable (DV) either 0 or 1. For an instance, we have enter yes for 1 and no for 0. If we use to code like this, then the mean of the distribution is adequate to the proportion of 1s within the distribution. Another example if there are 100 folks within the distribution and 30 of them are coded 1, then the mean of the distribution is .30, that is that the proportion of 1s. The mean of the distribution is additionally the probability of drawing someone labeled as 1 arbitrarily from the distribution. That is, if we tend to take someone arbitrarily from our sample of 100 that I simply represented, the probability that the person is going to be a 1 is .30. Therefore, proportion and probability of 1 are constant in such cases. The mean of a binary distribution thus coded is denoted as P , the proportion of 1s. The proportion of zeros is $(1-P)$ that is typically denoted as Q . The variance of such a distribution is PQ , and also the standard deviation is $\text{Sqrt}(PQ)$.

Suppose we wish to predict whether or not somebody is male or female (DV, $M=1$, $F=0$) using height in inches (IV). We tend to may plot the relations between the two variables as we tend to typically neutralize regression. The plot would possibly look something like this:

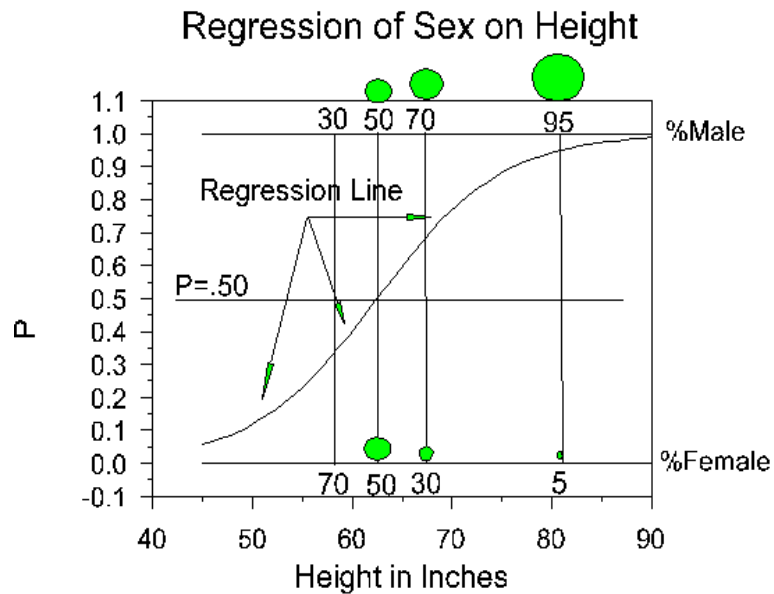


Figure 3.2: Regression of Sex on Height

Now we explain the points of the graph (data are fictional):

- The regression curve is a rolling average, even as in linear regression. The Y-axis is P that indicates the proportion of 1s at any given value of height. (review graph)
- The regression curve is nonlinear. (review graph)
- None of the observations --the raw data points-- truly fall on the regression curve. All of them fall on zero or one. (review graph)

The Logistic Curve

The logistic curve relates the independent variable, X, to the rolling mean of the DV, P ($\bar{Y}$).

$$P = \frac{e^{a+bX}}{1 + e^{a+bX}} \quad (3.1)$$

Or

$$P = \frac{1}{1 + e^{-(a+bX)}} \quad (3.2)$$

The formula to do this is also written either wherever P is that the probability of a 1 (the proportion of 1s, the mean of Y), e is that the base of the logarithm and a and b are the parameters of the model. The value of a yields P when X is zero, and b adjusts however quickly the probability changes with dynamic X a single unit (we will have standardized and unstandardized b weights in logistic regression, even as in standard linear regression). As a result of the relation between X and P is nonlinear, b doesn't have an easy interpretation during this model because it will in standard linear regression.

Chapter 4

Case Study: Mining Association Rules Using Data Mining and Statistics

Introduction

In this section, we will mine association rules using both models and analyze their performance metrics in different ways. We will also evaluate the mining processes as well as compare the performance of different rule mining models. In this research, our main focus is on analyzing the rule mining processes. Basically, the different mining approaches can be compared by the mechanism of those models. In this chapter, we compare the performance of only Apriori and Apriori with Logistic Regression. We are able to explore which model is more efficient for association rules mining.

4.1 Datasets and Methodology

In 1966, ICDDR,B set up a Health and Demographic Surveillance System (HDSS) at Matlab in Chandpur district of Bangladesh. At the beginning, HDSS collected data on births, deaths, and migrations, as well as collected the censuses data also. After that in 1975, it has been started the collection of marital data. In 1982 and 1993, HDSS has been collected internal movement and household headship data respectively. Later in 1998, the Geographical Information System (GIS) and Record Keeping System (RKS) both are merged with HDSS data depository. All HDSS data are stored in Oracle Data server. The Community Health Research Workers (CHRWs) make a contact with each household every bi-month and collect any new or changed information in their allocated area. Different level of Officers and Managers supervise the activities of CHRWs and a monitoring team monitors the quality of

collected data independently. We have select randomize 10 village's populations with all out-migration people from the Health and Demographic Surveillance System (HDSS) area leaving in the Matlab HDSS are considered for our research. We also collected the information on socioeconomic and demographic characteristics, housing, environmental conditions, food and sanitation, educational and work-related information for the selected out migration population.

We have collected the dataset in SPSS format from HDSS data management team:

- 44678 records are available in the dataset
- Those data are considered for 10 years (from January 2001 to December 2010).
- Total 128 variables come from different parts of the dataset.

For the domain knowledge, we talked to HDSS experts and accordingly we consider some variables are unnecessary for our study, so we have removed those variables from our dataset. Rest of the variables we consider for preprocessing, cleaning, transformation and modeling. Some records and fields have not relevant and important for the research. In the data preprocessing phase we further remove those records and fields. When we analyzed these data using statistical tools some of these variables has no contribution in our research so we further deleted these information from dataset. After the above activity, the variable count reduced from 128 to 80.

Applying frequency distribution and statistical information related to the variables in our dataset is taken to have a quick understanding of the data content and quality of the data. We observe that there are missing values in the data and there are skewness and sparsity in some variables, which are dealt in the data preprocessing section.

The dataset's records are contained in numerical format. But, 75 variables data are categorical in nature i.e. they have distinct integer values, which is represented the corresponding to some categorical values, like a variable "SCORE" it has 5 distinct integer values and each integer value representing the corresponding categorical value. We have separated these values and

created a new variable. Those variables contained the value (1 = Yes, 0 = No). Remaining 5 attributes have continuous values.

Outline of Experimental Design:

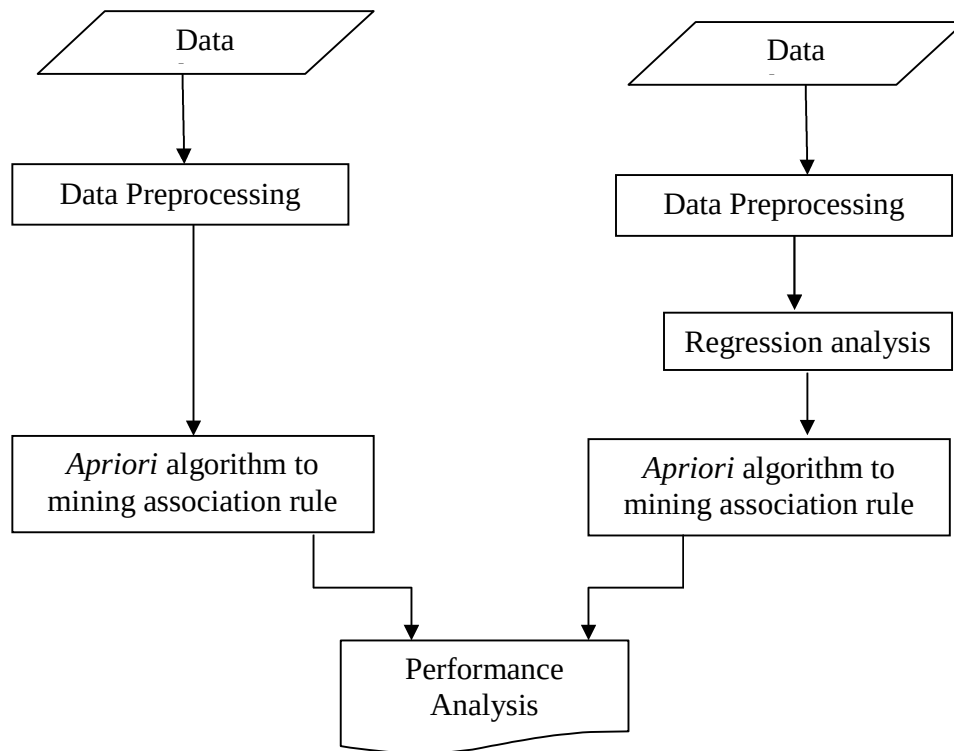


Figure 4.1 : Flow chart of Experimental Design

4.2 Data Preprocessing

Data Mining works on data. Basically, real-world data is not stored in data mining format. So, we need to organize the dataset and understandable as of data mining tools[16].

Data preparation process cannot be done blindly. Till now, there is no tool developed which is automatically process data and make a perfect dataset[17]. Data preparation is an art in science. That's why; we need to apply the knowledge of their effect on the data being prepared.

Oracle, SPSS and Excel software are used to find out missing values, empty values, misclassification error, field transformation, dimensionality reduction i.e., overall data cleaning activities are carried out in the following steps using Oracle, SPSS and Excel.

Step 1: Data filtration has done on the basis of internal movement split family and head change. Only record consider for out migration and stay in the area is taken. Size of the dataset reduces. Record count is 33,408.

This is done because by observation it is realized that in cases of records with internal movement, split family and head change has no impact on out migration.

Step 2: Some families came to the HDSS area few months ago, so they have no socioeconomic survey information. That's why we have deleted those records. The number of records in the dataset becomes 32,260.

Step 3: We create some variable by calculating age on in migration, age on out migration, duration of stay at the HDSS area. So the attributes increases from 80 to 83.

Step 4: We split these variables (AGE_IM, AGE_OM, STAY_OM, MAR_STAT, MAIN_INCOME, EDUYR, OCCP, P_MI_CAUSE, P_MO_CAUSE, R_CAUSE, R_RSHIP, LAND_A, LAND_R, LAND_T, TOT_HOME, WALL, FLOOR, ROOF, DWATER_DRY, LATRINE_FEMALE and SCORE) to make categorical. That's why now the total numbers of attributes 516.

Step 5: We have converted values 1 and 0 (1=yes and 0=no) for all variables. These are needed for statistical and data mining analysis.

Step 6: We have taken frequency distribution table for all variables. Remove all missing or empty value from data set.

All variables are not necessary to solve a specific data mining activity. Some variables may contribute interference for an appropriate data mining mission. The participation of all variables sometimes may create difficulty or decline the efficiency of an algorithm. If we applied all attributes, It may create a difficulty to determine which ones is appropriate for the mining. It is also the curse of dimensionality. We can solve this high dimensionality problem by minimizing the number of variables, which is known as dimensionality reduction.

4.3 Data modeling

In this study, we have created two dataset in .cvs format using SPSS and Excel software to implement SPSS Clementine Software version 11 for mining association rules.

4.3.1 Dataset 1

Step 1: Firstly we extract the dataset considering all variables from oracle database in .cvs format. In this case we have 516 attributes are available for data mining.

Step 2: We have converted all '1' in 'T' and '0' in 'F' for all variables. These are mandatory for Apriori algorithm in SPSS Clementine Software.

4.3.2 Dataset 2

Step 1: Firstly we extract the dataset from oracle database in SPSS format.

Step 2: Then we have implemented Logistic Regression for 'OUT_MIGRATION' as dependent variable and all others variables are considered as independent.

Step 3: We have selected only those variables in dataset which has a strong relationship with 'OUT_MIGRATION' considering the significant value ≤ 0.05 . And remove all others variables from the dataset. After that we have 345 attributes are available for data mining and transferred it in .cvs format.

Step 4: We have converted all '1' in 'T' and '0' in 'F' for all variables. These are mandatory for Apriori algorithm in SPSS Clementine Software.

4.4 Evaluations and deployment

In the previous stage we have processed the dataset and now we evaluated them. We extract association rules from those two datasets. Both datasets are not same as the number of variables. First one is contained all variables which are generated from the database and the second one only contain less number of variables which have a significant relation with 'OUT_MIGRATION'.

4.4.1. Model 1 (all instances from Dataset 1 used in association rules mining)

Firstly, we implement the dataset 1 which is created in data modeling section. Now we try to put different Minimum antecedent support as well as observe the outputs. At first we use the Minimum antecedent support 4 %, Minimum rule confidence 90 % and Maximum number

Table 1: Number of rules and process time required for different support (Apriori without Regression)

Support (%)	Apriori without Regression (Time in Second)	Apriori without Regression (Number of rules)
5	2753	16882
6	1709	5228
7	1017	1592
8	860	450
9	595	140
10	428	41
11	303	12

of antecedents 10 in Apriori model of SPSS Clementine Software. It shows "No rules found. Thresholds too high?" . When we use the Minimum antecedent support 5 %, Minimum rule confidence 90 % and Maximum number of antecedents 10 as a result we get the number of Association Rules are 16882 and the process completion time required 2753 seconds. After

that we change Minimum antecedent support from 6 % to 12. We always maintain the minimum rule confidence 90 % and maximum number of antecedents 10. Performance metrics for model 1 is presented in Table 1. When we put the minimum antecedent support 12% that time also it shows “No rules found. Thresholds too high?”. The table shows when support is low at that moment process time required more as well as rules number is too high.

4.4.2. Model 2 (all instances from Dataset 2 used for association rules mining)

Secondly, we use the dataset 2 which is created in data modeling section. Now we implement the same process again for the dataset 2 as well as observe the outputs. At first we use the Minimum antecedent support 3 %, Minimum rule confidence 90 % and Maximum number of

Table 2: Number of rules and process time required for different support (Apriori with Regression)

Support (%)	Apriori with Regression (Time in Second)	Apriori with Regression (Number of rules)
5	512	6302
6	344	2076
7	220	777
8	148	250
9	99	88
10	67	37
11	47	12

antecedents 10 in Apriori model of SPSS Clementine Software. It shows “No rules found. Thresholds too high?”. When we use the Minimum antecedent support 5 %, Minimum rule confidence 90 % and Maximum number of antecedents 10 as a result we get the number of Association Rules are 6302 and the process completion time required 512 seconds. After that we change Minimum antecedent support from 6 % to 12. We always maintain the minimum rule confidence 90 % and maximum number of antecedents 10. Performance metrics for model 2 is presented in Table 2. When we put the minimum antecedent support 12% that time also it shows “No rules found. Thresholds too high?”. The table shows when support is low

just then process time required more as well as rules number is high. But the performance is better than model 1.

4.4.3. Compare the output of Model 1 vs. Model 2

The first experiment we try to find association rules using only applying the Apriori algorithm and we called it Apriori without Logistic Regression experiment. On the other hand in second experiment, firstly we applying the Logistic Regression and finding the significant variables considering the significant value ≤ 0.05 . As a result the number of variables reduces for data mining .After that we have applied the Apriori algorithm to mining association rules and we called it Apriori with Logistic Regression experiment.

Table 3: Time required for different support

Support (%)	Apriori with Regression (Time in Second)	Apriori without Regression (Time in Second)	Time reducing (%)
5	512	2753	81.40
6	344	1709	79.87
7	220	1017	78.37
8	148	860	82.79
9	99	595	83.36
10	67	428	84.35
11	47	303	84.49

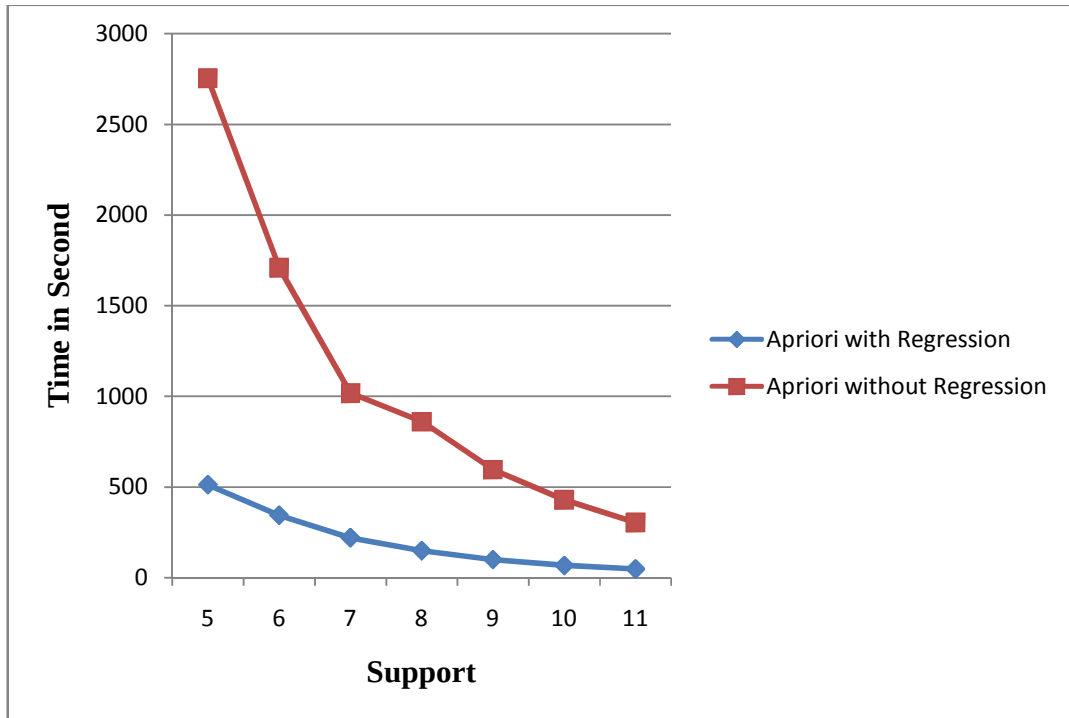


Figure 4.2: Required time vs support

Now we try to explain the differences of two experiments. Considering the table 3 and figure 4.1 shows the performance of second experiment too high. It saves average time 82.09 %, which is very much cost effective for the large dataset.

Table 4: Number of rules for different support

Support (%)	Apriori with Regression (Number of rules)	Apriori without Regression (Number of rules)	Rules reducing (%)
5	6302	16882	62.67
6	2076	5228	60.29
7	777	1592	51.19
8	250	450	44.44
9	88	140	37.14
10	37	41	9.76
11	12	12	0.00

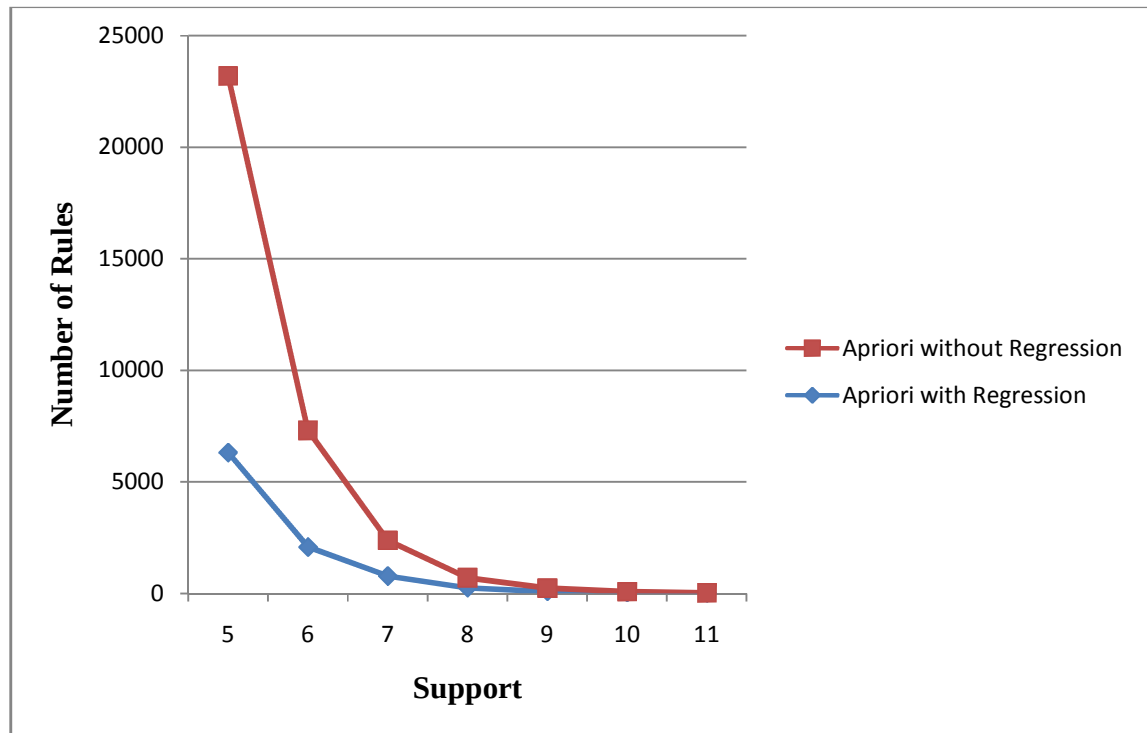


Figure 4.3: Numbers of rules vs support

Further, we consider the table 4 and figure 4.2. These also show the second experiment reduce the unexpected association rules. It protected average unexpected rules 37.93 %, which is also very much effective for the large dataset for association rule mining. When the support is high both experiments show near to same performance, but the support is low that moment second experiment performs efficiently.

Chapter 5

Conclusion and Future Research

5.1 Conclusion

In our research, an Apriori with Logistic Regression is proposed through reducing the time consumed in transactions scanning for candidate itemsets by increasing the minimum support. When we decrease the value of support, the difference of time consumed between our Apriori with Logistic Regression and only the Apriori increases. The time consumed to generate association rules in our Apriori with Logistic Regression is less than the time consumed in only the Apriori; our Apriori with Logistic Regression reduces the time consuming by 82.09 %. It also shows the same performance to reducing 37.97 % unexpected association rules mining by our Apriori with Logistic Regression model. As this is proved and validated by the experiments and observable in figure 4.1, figure 4.2, table 3 and table 4.

5.2 Future Works

This thesis we recommend a new way of measuring the quality of association rule mining by using Apriori with Logistic Regression model. Using only Apriori algorithm for mining association rule involves mining a high quality rule set in an efficient way. Apriori with Logistic Regression model involves building highly accurate and competitive association rules with a minimum time. If we follow our Apriori with Logistic Regression model for sample selection, we can save our valuable time and prepared an appropriate sample data set from a HDSS.

References:

- [1] R.Z.Inamul Hussain , etl : “A Study of Different Association Rule Mining Techniques” International Journal of Computer Applications (0975 – 8887) Volume 108 – No 16, December 2014

- [2] Enrique García, Cristóbal Romero, etl :“Drawbacks and solutions of applying association rule mining in learning management systems “.Córdoba University, Campus Universitario de Rabanales, 14071, Córdoba, Spain

- [3] Gang Yang , Hong Zhao , etl : “An implementation of improved apriori algorithm”. Machine Learning and Cybernetics, 2009 International Conference on, 12-15 July 2009, IEEE.

- [4] S. Sumathi , S.N. Sivanandam: “*Introduction to Data Mining Principles*”, Studies in Computational Intelligence (SCI) 29, 1–20 (2006), www.springerlink.com, Springer-Verlag Berlin Heidelberg 2006.

- [5] Stefan Mutter : “Classification using Association Rules”. Hamilton, New Zealand, Aotearoa, 11th March 2004.

- [6] Agrawal R. , Srikant R.” Fast Algorithms for Mining Association Rules”. *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB’94)*, pages 475–486, Santiago de Chile, Chile, September 1994. Morgan Kaufmann.

- [7] Hand D., Mannila H. , etl “*Principles of Data Mining*”. MIT Press, Cambridge, Massachusetts, USA, 2001.

- [8] S.M. Monzurur Rahman : “ Data Mining Using Neural Networks”. RMIT University July 2006.

- [9] Alex, B., Stephen. etl. (1999), “*Building Data Mining Applications for CRM*”, McGraw-Hill.
- [10] Alsabti, K., Ranka, S. and Singh, V. (1998), “CLOUDS: “A Decision Tree Classifier for Large Datasets”, *Proceedings of 4th International Conference on Knowledge Discovery and Data Mining*, pp. 2-8. New York, USA.
- [11] Liu B., Hsu W. and Ma Y. “Integrating Classification and Association Rule Mining”. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (KDD’98)*, pages 80–86, New York, USA, August 1998. The AAAI Press.
- [12] Scheffer T. “Finding Association Rules That Trade Support Optimally against Confidence”. Unpublished manuscript, more detailed version, algorithm’s pruning method changed.
- [13] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From data mining to knowledge discovery in databases,” vol. 17, no. 3, AI magazine, 1996, pp. 37.
- [14] Bereznicki, Bonnie J., Peterson, Gregory M., Jackson, Shane L., Haydn Walters, E., Fitzmaurice, Kimbra D., & Gee, Peter R. (2008). “Data-mining of medication records to improve asthma management”. MJA, 189(1), 21–25.
- [15] Wei-Chou Chen, Shian-Shyong Tseng, Ching-Yao Wang: “A novel manufacturing defect detection method using association rule mining techniques”. *Expert Systems with Applications* 29 (2005) 807–815
- [16] Rashedur M. Rahman□, Fazle Rabbi Md. Hasan: “Using and comparing different decision tree classification techniques for mining ICDDR,B Hospital Surveillance data”., *Expert Systems with Applications* 38 (2011) 11421–11436

[17] Pyle Dorian, “Data Preparation for Data Mining, Morgan Kaufman”, Morgan Kaufmann; Book & CD-ROM 1st edition (1999) ISBN-10: 1558605290, ISBN-13: 978-1558605299.

[18] Houston, Andrea L., Chen, Hsinchun, Hubbard, Susan M., Schatz, Bruce R., Ng, Tobun D., Sewell, Robin R., et al. (1999). “Tolle medical data mining on the internet: research on a cancer information system.” Artificial Intelligence Review, 13(5–6), 437–466. 30.ICDDR,B, (2008)

[19] Max, Bramer. “Principles of Data Mining”, Springer London Ltd, Published, (2007), ISBN 13: 9781846287657.

Appendix A

SPSS Variable List:

DM_ALL_2001_10_FINAL_NUMBER.sav [DataSet1] - IBM SPSS Statistics Data Editor

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	AGE_IM1	Numeric	1	0		None	None	8	Right	Nominal	Input
2	AGE_IM2	Numeric	1	0		None	None	8	Right	Nominal	Input
3	AGE_IM3	Numeric	1	0		None	None	8	Right	Nominal	Input
4	AGE_IM4	Numeric	1	0		None	None	8	Right	Nominal	Input
5	AGE_IM5	Numeric	1	0		None	None	8	Right	Nominal	Input
6	AGE_IM6	Numeric	1	0		None	None	8	Right	Nominal	Input
7	AGE_OM1	Numeric	1	0		None	None	8	Right	Nominal	Input
8	AGE_OM2	Numeric	1	0		None	None	8	Right	Nominal	Input
9	AGE_OM3	Numeric	1	0		None	None	8	Right	Nominal	Input
10	AGE_OM4	Numeric	1	0		None	None	8	Right	Nominal	Input
11	AGE_OM5	Numeric	1	0		None	None	8	Right	Nominal	Input
12	AGE_OM6	Numeric	1	0		None	None	8	Right	Nominal	Input
13	AGE_OM7	Numeric	1	0		None	None	8	Right	Nominal	Input
14	ASST01	Numeric	1	0		None	None	8	Right	Nominal	Input
15	ASST02	Numeric	1	0		None	None	8	Right	Nominal	Input
16	ASST03	Numeric	1	0		None	None	8	Right	Nominal	Input
17	ASST04	Numeric	1	0		None	None	8	Right	Nominal	Input
18	ASST05	Numeric	1	0		None	None	8	Right	Nominal	Input
19	ASST06	Numeric	1	0		None	None	8	Right	Nominal	Input
20	ASST07	Numeric	1	0		None	None	8	Right	Nominal	Input
21	ASST08	Numeric	1	0		None	None	8	Right	Nominal	Input
22	ASST09	Numeric	1	0		None	None	8	Right	Nominal	Input
23	ASST10	Numeric	1	0		None	None	8	Right	Nominal	Input
24	ASST11	Numeric	1	0		None	None	8	Right	Nominal	Input
25	ASST12	Numeric	1	0		None	None	8	Right	Nominal	Input
26	ASST13	Numeric	1	0		None	None	8	Right	Nominal	Input
27	ASST14	Numeric	1	0		None	None	8	Right	Nominal	Input
28	ASST15	Numeric	1	0		None	None	8	Right	Nominal	Input
29	ASST16	Numeric	1	0		None	None	8	Right	Nominal	Input
30	ASST17	Numeric	1	0		None	None	8	Right	Nominal	Input
31	ASST18	Numeric	1	0		None	None	8	Right	Nominal	Input
32	ASST19	Numeric	1	0		None	None	8	Right	Nominal	Input
33	ASST20	Numeric	1	0		None	None	8	Right	Nominal	Input
34	ASST21	Numeric	1	0		None	None	8	Right	Nominal	Input
35	ASST22	Numeric	1	0		None	None	8	Right	Nominal	Input
36	ASST23	Numeric	1	0		None	None	8	Right	Nominal	Input
37	ASST24	Numeric	1	0		None	None	8	Right	Nominal	Input
38	ASST25	Numeric	1	0		None	None	8	Right	Nominal	Input

Data View Variable View

SPSS data in numeric format:

DM_ALL_2001_10_FINAL_NUMBER.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

Visible: 517 of 517 Variables

	AGE_IM1	AGE_IM2	AGE_IM3	AGE_IM4	AGE_IM5	AGE_IM6	AGE_OM1	AGE_OM2	AGE_OM3	AGE_OM4	AGE_OM5	AGE_OM6	AGE_OM7	ASST01	ASST02	ASST03
1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
2	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	
3	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	
4	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
5	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	
6	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
7	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
8	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
9	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
10	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
11	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	
12	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	
13	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	
14	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	
15	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	
16	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	
17	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
18	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	
19	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
20	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	
21	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	
22	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	
23	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	
24	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	
25	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
26	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
27	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
28	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	
29	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	
30	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	
31	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	
32	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
33	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	
34	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	
35	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	
36	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	
37	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	

Data ViewVariable View

Some output of Screens Logistic Regression:

Screen 1:

LogisticRegressionOutput2001_2010.spv [Document2] - IBM SPSS Statistics Viewer

File Edit View Data Transform Insert Format Analyze Direct Marketing Graphs Utilities Add-ons Window Help

Input

- Log
- Logistic Regression
 - Title
 - Notes
 - Active Dataset
 - Warnings
 - Case Processing Summary
 - Dependent Variable Encoding
 - Block 0: Beginning Block
 - Title
 - Classification Table
 - Variables in the Equation
 - Variables not in the Equation
 - Block 1: Method = Enter
 - Title
 - Omnibus Tests of Model Coefficients
 - Model Summary

Warnings

Due to redundancies, degrees of freedom have been reduced for one or more variables.
The parameter covariance matrix cannot be computed.
Remaining statistics will be omitted.

Case Processing Summary

Unweighted Cases ^a		N	Percent
Selected Cases	Included in Analysis	32260	100.0
	Missing Cases	0	.0
	Total	32260	100.0
Unselected Cases		0	.0
	Total	32260	100.0

a. If weight is in effect, see classification table for the total number of cases.
b. The variable R_RSHIP00Y is constant for the selected cases. Since a constant term was specified, the variable will be removed from the analysis.

Dependent Variable Encoding

Original Value	Internal Value
0	0
1	1

Block 0: Beginning Block

Classification Table^{a,b}

Observed		Predicted		Percentage Correct
		OUT_MIGRATION	0	
Step 0	OUT_MIGRATION 0	19639	0	100.0
	1	12621	0	.0
Overall Percentage				60.9

a. Constant is included in the model.
b. The cut value is .500

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)	
Step 0	Constant	-.442	.011	1502.099	1	.000	.643

Variables not in the Equation^a

Screen 2:

LogisticRegressionOutput2001_2010.spv [Document2] - IBM SPSS Statistics Viewer

File Edit View Data Transform Insert Format Analyze Direct Marketing Graphs Utilities Add-ons Window Help

Output

- Log
- Logistic Regression
 - Title
 - Notes
 - Active Dataset
 - Warnings
 - Case Processing Summary
 - Dependent Variable Encoding
 - Block 0: Beginning Block
 - Title
 - Classification Table
 - Variables in the Equation
 - Variables not in the Equation
 - Block 1: Method = Enter
 - Title
 - Omnibus Tests of Model Coefficients
 - Model Summary

Block 0: Beginning Block

Classification Table^{a,b}

		Predicted		Percentage Correct
		OUT_MIGRATION		
Step 0	OUT_MIGRATION 0	19639	0	100.0
	1	12621	0	.0
Overall Percentage				60.9

a. Constant is included in the model.
b. The cut value is .500

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
Step 0 Constant	-.442	.011	1502.099	1	.000	.643

Variables not in the Equation^a

Step 0 Variables	Score	df	Sig.
AGE_IM1	22.582	1	.000
AGE_IM2	524.379	1	.000
AGE_IM3	3.210	1	.073
AGE_IM4	78.976	1	.000
AGE_IM5	118.338	1	.000
AGE_IM6	113.713	1	.000
AGE_OM1	595.163	1	.000
AGE_OM2	3080.113	1	.000
AGE_OM3	6461.634	1	.000
AGE_OM4	6054.220	1	.000
AGE_OM5	2987.373	1	.000
AGE_OM6	2302.829	1	.000
AGE_OM7	32260.000	1	.000
ASST01	63.155	1	.000
ASST02	31.235	1	.000
ASST03	36.701	1	.000
ASST04	1.274	1	.259
ASST05	27.102	1	.000
ASST06	31.504	1	.000
ASST07	.702	1	.402
ASST08	12.761	1	.000

Screen 3:

LogisticRegressionOutput2001_2010.spv [Document2] - IBM SPSS Statistics Viewer

File Edit View Data Transform Insert Format Analyze Direct Marketing Graphs Utilities Add-ons Window Help

Logistic Regression

- Title
- Notes
- Active Dataset
- Warnings
- Case Processing Summary
- Dependent Variable Encoding
- Block 0: Beginning Block
 - Title
 - Classification Table
 - Variables in the Equation
 - Variables not in the Equation
- Block 1: Method = Enter
 - Title
 - Omnibus Tests of Model Coefficients
 - Model Summary

Variable	B	SE	Wald	Df	Sig.
SCORE4	1.282	.258	1	1	.258
SCORE5	79.989	.000	1	1	.000
STAY_OM1	844.921	.000	1	1	.000
STAY_OM2	6802.671	.000	1	1	.000
STAY_OM3	2749.136	.000	1	1	.000
STAY_OM4	6422.013	.000	1	1	.000
STAY_OM5	4963.926	.000	1	1	.000
STAY_OM7	32260.000	.000	1	1	.000
TOT_HOME1Y	76.376	.000	1	1	.000
TOT_HOME2Y	65.407	.000	1	1	.000
TOT_HOME3Y	2.293	.130	1	1	.130
TOT_HOME4Y	4.780	.029	1	1	.029
TOT_HOME5Y	4.633	.031	1	1	.031
TOT_HOME8Y	.029	.864	1	1	.864
WALL1Y	21.360	.000	1	1	.000
WALL2Y	9.130	.003	1	1	.003
WALL3Y	.013	.909	1	1	.909
WALL4Y	14.207	.000	1	1	.000
WALL5Y	33.929	.000	1	1	.000
WALL7Y	12.390	.000	1	1	.000
WALL8Y	3.846	.050	1	1	.050
ENOUGH_FOOD	.000	.998	1	1	.998
NGO_MEMBER	46.905	.000	1	1	.000

a. Residual Chi-Squares are not computed because of redundancies.

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

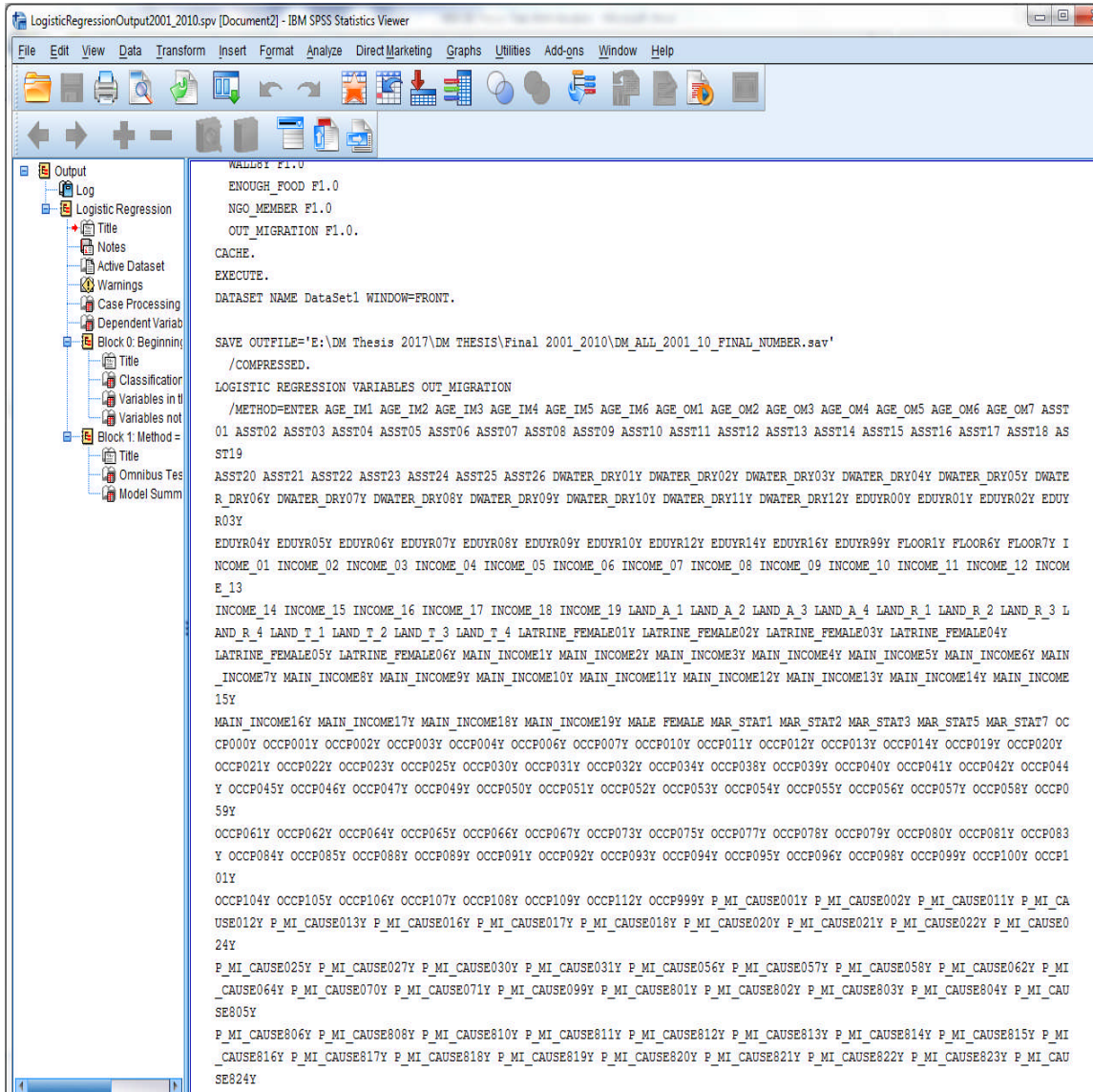
Step	Chi-square	df	Sig.
Step 1	43182.849	495	.000
Block	43182.849	495	.000
Model	43182.849	495	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	.000 ^a	.738	1.000

a. Estimation terminated at iteration number 18 because a perfect fit is detected. This solution is not unique.

Screen 4:



Data transferred into .csv format:

The screenshot shows a Microsoft Excel spreadsheet titled "DM_ALL_2001_10_FINAL_TXT.csv - Microsoft Excel". The ribbon includes Home, Insert, Page Layout, Formulas, Data, Review, View, Acrobat, and Team. The active cell is I16, containing the value "F".

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	AGE_IM1	AGE_IM2	AGE_IM3	AGE_IM4	AGE_IM5	AGE_IM6	AGE_OM1	AGE_OM2	AGE_OM3	AGE_OM4	AGE_OM5	AGE_OM6	AGE_OM7	ASST01	ASST02	ASST03	ASST04	ASST05	ASST06
2	T	F	F	F	F	F	F	F	F	F	F	T	F	F	T	F	F	T	F
3	F	F	F	T	F	F	F	F	F	F	F	T	T	F	T	F	F	F	F
4	T	F	F	F	F	F	F	F	F	F	F	T	F	T	T	F	F	F	F
5	T	F	F	F	F	F	F	F	F	F	F	F	T	F	F	F	F	F	F
6	F	F	F	T	F	F	F	F	F	F	F	T	F	F	F	F	F	F	F
7	T	F	F	F	F	F	F	F	F	F	F	T	F	F	F	T	F	F	F
8	T	F	F	F	F	F	F	F	F	F	F	T	F	F	T	F	T	F	F
9	T	F	F	F	F	F	F	F	F	F	F	T	F	F	T	F	F	F	F
10	T	F	F	F	F	F	F	F	F	F	F	T	F	F	T	F	F	F	F
11	T	F	F	F	F	F	F	F	F	F	F	T	F	F	F	F	T	F	F
12	T	F	F	F	F	F	F	F	F	F	F	T	T	T	T	F	F	T	F
13	F	F	F	T	F	F	F	F	F	F	F	T	T	T	F	F	F	F	F
14	T	F	F	F	F	F	F	F	F	F	F	T	T	T	T	F	F	T	F
15	T	F	F	F	F	F	F	F	F	F	F	T	T	F	F	F	F	F	F
16	T	F	F	F	F	F	F	F	F	F	F	T	T	F	T	F	F	F	F
17	T	F	F	F	F	F	F	F	F	F	F	T	T	F	T	F	F	T	F
18	T	F	F	F	F	F	F	F	F	F	F	T	F	F	T	F	F	F	F
19	T	F	F	F	F	F	F	F	F	F	F	T	T	T	T	F	F	F	F
20	T	F	F	F	F	F	F	F	F	F	F	T	F	F	F	F	F	F	F
21	T	F	F	F	F	F	F	F	F	F	F	T	T	F	T	F	F	F	F
22	F	F	F	T	F	F	F	F	F	F	F	T	F	F	F	F	F	F	F
23	F	T	F	F	F	F	F	F	F	F	F	T	F	F	F	F	F	F	F
24	T	F	F	F	F	F	F	F	F	F	F	T	T	T	T	F	F	T	F
25	F	F	F	T	F	F	F	F	F	F	F	T	T	F	T	F	F	F	F
26	T	F	F	F	F	F	F	F	F	F	F	T	F	F	T	F	F	F	F
27	T	F	F	F	F	F	F	F	F	F	F	T	F	F	T	F	F	F	F
28	T	F	F	F	F	F	F	F	F	F	F	T	F	F	T	F	T	F	F
29	F	F	T	F	F	F	F	F	F	F	F	T	F	F	T	F	F	F	F
30	T	F	F	F	F	F	F	F	F	F	F	T	T	T	T	F	F	F	F
31	F	T	F	F	F	F	F	F	F	F	F	T	F	F	T	F	F	F	F
32	F	F	F	T	F	F	F	F	F	F	F	T	F	F	T	F	F	F	F
33	T	F	F	F	F	F	F	F	F	F	F	T	F	F	T	F	F	F	F
34	F	F	F	F	T	F	F	F	F	F	F	T	F	F	T	F	F	F	F
35	T	F	F	F	F	F	F	F	F	F	F	T	F	T	T	F	F	F	F
36	T	F	F	F	F	F	F	F	F	F	F	T	F	T	T	F	F	F	F
37	F	F	F	T	F	F	F	F	F	F	F	T	T	F	T	F	F	F	F
38	F	F	F	T	F	F	F	F	F	F	F	T	T	F	T	F	F	F	F

