ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION STUDIES FOR AFRICA

APPLICATION OF DATA MINING TECHNIQUES TO SUPPORT CUSTOMER RELATIONSHIP MANAGEMENT AT ETHIOPIAN AIRLINES

A THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES OF ADDIS ABABA UNIVERSITY IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATION SCIENCE

አዲስ አበባ ዩኒቨርሲቲ
ADDIS ABABA UNIVERSITY
SISA
የአፍሪካ የኢንፎርሜሽን
ጥናት ትምህርት ቤት

By

Henock Woubishet Tefera

July 2002

ADDIS ABABA UNIVERSITY
LIBRARIES
P.O.BOX 1176
ADDIS ABABA ETHIOPIA

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION STUDIES FOR AFRICA

APPLICATION OF DATA MINING TECHNIQUES TO SUPPORT CUSTOMER RELATIONSHIP MANAGEMENT AT ETHIOPIAN AIRLINES

By
HENOCK WOUBISHET

Name and Signature of Members of the Examining Board

Ato Getachew Jemaneh, Chairman, Examining Board

Ato Million Meshesha, Advisor

Ato Ermias Abebe, Advisor

Ato Nigussie Tadesse, Advisor

Dr. Osei Adjei, External Examiner

DEDICATION

I would like to dedicate this paper to my father and mother, Colonel Woubishet Tefera and Sister Beliyou Haile, who have always been there, and supported me all the way.

ACKNOWLEDGEMENT

I would like to thank my advisors Ato Million Meshesha, Ato Ermias Abebe and Ato Nigusse Tadesse for their constructive comments and overall guidance.

Special thanks go to Ato Ermias Alemu, who has always been there when I needed his help, and especially for his assistance with the data collection and preparation work. Ato Hailemelekot Mamo, from the Customer Loyalty Department, was very cooperative, and his ideas invaluable.

Ato Melese Tesfaye was very helpful in the initial acquisition of the software, and later Angoss Software Corporation were willing to extend the software evaluation period for this project.

Ato Daniel Woubishet was very helpful in acquiring the necessary materials for the study. W/t Lakech Haile and W/o Abebech Bekele were very supportive and caring in every possible way.

Last, but not least, I would like to thank the whole of Marketing I.S. staff, who were very understanding and extended their unreserved support whenever I asked for it.

LIST OF ABBREVIATIONS

CLD – Customer Loyalty Department

CRM – Customer Relationship Management

DCS – Departure Control System

DB – Database

DIF – Departure Information Forms

DIM – Departure Information Message

ETHIOPIAN – Ethiopian Airlines Enterprise

FFP – Frequent Flyer Program

KDD – Knowledge Discovery in Databases

SOM – Self Organizing Map

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ABSTRACT

The airline industry is highly competitive, dynamic and subject to rapid change. As a result, airlines are being pushed to understand and quickly respond to the individual needs and wants of their customers. Most airlines use frequent flyer incentive programs to win the loyalty of their customers, by awarding points that entitle customers to various travel benefits. Furthermore, these airlines maintain a database of their frequent flyer customers.

Customer relationship management (CRM) is the overall process of exploiting customer- related information and using it to enhance the revenue flow from an existing customer. As part of implementing CRM, airlines use their frequent flyer data to get a better understanding of their customer types and behavior. Data mining techniques are used to extract important customer information from available databases.

This study is aimed at testing the application of data mining techniques to support CRM activities at Ethiopian Airlines. The subject of this case study is Ethiopian Airlines' frequent flyer program's database, which contains individual flight activity and demographic information of more than 22,000 program members.

The data mining process was divided into three major phases. During the first phase, data was collected from different sources, since the frequent flyer database lacked revenue data, which was essential for the study's goal of identifying profitable customer segments. The data preparation

phase was next, where a procedure was developed to compute and fill-in for missing revenue values. Moreover, data integration and transformation activities were performed.

In the third phase, which is model building and evaluation, K-means clustering algorithm was used to segment individual customer records into clusters with similar behaviors. Different parameters were used to run the clustering algorithm before arriving at customer segments that made business sense to domain experts. Next, decision tree classification techniques were employed to generate rules that could be used to assign new customer records to the segments.

The results from this study were encouraging, which strengthened the belief that applying data mining techniques could indeed support CRM activities at Ethiopian Airlines. In the future, more segmentation studies using demographic information and employing other clustering algorithms could yield better results.

Chapter 1

Introduction

1.1 Background

Marketing is the process of planning and executing the conception, pricing, promotion, and distribution of ideas, goods, and services to create exchanges that satisfy individual and organizational objectives (McDaniel et.al., 1995). Nowadays, customers, that have real value to a company, are at the center of marketing strategies. Accordingly, businesses have found it essential to acquire new customers as well as retain those that have high value.

The airline industry is complex, dynamic and subject to rapid change and innovation. Doganis (1991) notes that the homogeneous nature of the airline product pushes airlines into making costly marketing efforts to try to differentiate their product from that of their competitors.

One way that airlines have used to win passenger loyalty is through "frequent-flyer" incentive programs (also known as customer loyalty programs or FFPs). Under such schemes, passengers are awarded points for each flight on a particular airline. As their points total builds up, passengers are entitled to increasingly attractive free flights or other travel benefits.

According to Chandler (2001), FFPs are concerned with rewarding behavior that is assumed to be loyal. Airlines later found out customers tend to fixate on the rewards. Moreover, with many such programs, one reward is generally as good as another for a customer, thus creating costs for the company with no sustainable differentiable competitive advantage.

According to most of the research that have been done with members of popular FFPs, it has been determined that only about 11 percent of active members fall into a defined category as being 'loyal' (Chandler, Ibid). The reason for this low figure is that customers usually perceive the mileage awarding airlines as companies of convenience, rather than as companies of care.

Airlines have started looking at the 'lifetime value' of each customer so they know which ones are worth investing money and effort to hold on to and which ones to let go. This change in focus from broad market segments to individual customers requires changes throughout the enterprise, but nowhere more than in marketing, sales, and customer support (Berry et.al., 1997).

Customer relationship management (CRM) is the term used for the overall process of exploiting customer-related information and using it to enhance the revenue flow from an existing customer (Bigus, 1996). Customer segmentation, according to Bounsaythip (2001), is the process of dividing customers into homogeneous groups on the basis of shared or common attributes, and is at the heart of CRM.

Segmentation describes the characteristics of the customer groups (called *segments* or *clusters*) within the data. By determining similar classes of customers, more targeted communication is possible, and marketing return on investment can be enhanced since marketing messages are accurately reaching those customers most likely to respond. Furthermore, different marketing strategies can be developed that are more appealing to members of the specified group. Segmentation requires the collection, organization and analysis of customer data.

Airlines, through interactions with their customers, generate vast amounts of customer data. Data is being extracted from the reservations, departure control, and sales information systems.

According to Fickel (2001), even though airlines have a vast store of customer information, they have not been able to put it to good use.

Machine learning and statistical techniques can be used to automatically extract information from data. *Data mining,* which is also referred to as *knowledge discovery in databases* (*KDD*), is defined as the efficient discovery of valuable, non-obvious information from a large collection of data (Trybula, 1997). Data mining centers on the automated discovery of new facts and relationships in data.

The data mining process involves the major activities of understanding the business, data collection, data preparation, model building, evaluation, and finally the deployment of results. According to Two Crows Corporation (TCC) (1999), the data collection and preparation steps may take between 80% - 90% of the time and effort of the entire data mining process.

Using data mining techniques against databases with marketing information is generally referred to as *database marketing*, and it can be used in several aspects of the customer-business relationship (Bigus, 1996). Bounsaythip (2001) notes that data mining techniques have been used to identify customer groups with high revenue potential, select criteria for mailing lists, and improve customer retention rate by identifying customers who were likely to switch to a competitor.

According to Pritscher (n.d.), data mining and data analysis are prerequisites to push CRM ahead in the airline industry. Pritscher (Ibid) further notes that knowledge about data mining techniques,

marketing strategies and airline business processes need to be integrated to successfully implement CRM. Even though CRM is not well developed in the airline business, FFPs provide a wealth of data, thus allowing to get a better understanding of customer types and their behavior.

Ethiopian Airlines (ETHIOPIAN) currently has a FFP named "ShebaMiles" with more than 22,000 members. Data pertaining to an individual member's flight activity details and personal details (like name, address, contact, date-of-birth, etc.) are stored in a database. Furthermore, the total member size increases as new members enroll and existing members' data is updated each time there is a flight activity.

The researcher chose to study the possible application of data mining techniques to support CRM activities at ETHIOPIAN. The availability of the FFP database with customers' information, willingness of ETHIOPIAN to allow the study to be conducted, and the researcher's familiarity with the airline were instrumental in deciding to proceed with the study.

1.2 Statement of the problem

Competitive pressure is very strong in the airline business. The homogeneous nature of the airline business makes product differentiation very difficult and costly. As a result, airlines have shifted their focus towards understanding their customers better that enables them to quickly respond to their individual needs and wants.

ETHIOPIAN currently has a FFP (called ShebaMiles) in order to increase and award the loyalty of its customers. The key program features are mileage accrual (where members can earn miles for air travel) and mileage redemption (members can spend miles for air travel), both on

ETHIOPIAN. Therefore, the 'currency' of the FFP is miles. Furthermore, the program is also used to identify customers with 'high value' and provide them with special benefits and services, such as access to lounges and free upgrades to an upper cabin.

Currently, 'customer value' is based on mileage. However, according to Pritscher (n.d.), mileage is not a good measure of customer profitability. Furthermore, since FF data are collected only for administrative purposes, monetary measures (which are relevant for CRM) are missing.

The Customer Loyalty Department (CLD), which is managing ETHIOPIAN's FFP with over 22,000 members, has embarked on building a CRM environment. The CLD periodically advertises special promotions to ShebaMiles' members, where members earn bonus miles over and above the usual number of base miles if they can fly on one or more of the special promotional flights. These promotions are advertised to all members, irrespective of whether these promotions apply to them or not. In addition to the additional promotional costs incurred by advertising to members for whom the promotions do not apply, members could get the impression that the airline does not know them, and that they are just getting 'junk mail.'

According to Chandler (2001), 80% of a company's revenue comes from 20% of its customers. Furthermore, it costs more to get a new customer than to retain an existing one. If members that are more valuable in terms of their revenue contribution could be identified, the CLD could design special marketing strategies that would engender their loyalty to the airline. Furthermore, a separate strategy could be devised to make the remaining members more valuable.

The CLD also wanted to know whether valuable customers were being left unrewarded while less valuable ones enjoy privileges. The program's currency being mere miles, the CLD needed new information to better manage members.

The CLD validates that correct flight activity information is uploaded into the database daily (both automated and manual), handles members' queries (telephone, e-mail, and fax), and periodically sends out statements to its members regarding their account mileage balance. With, more than 91,000 records of members' flight activity data, which is stored in its database, it has not been possible for the CLD to analyze its member data and gain better insight into the behavior and preferences of its members.

1.3 Justification

In today's competitive airline business, ETHIOPIAN needs to understand its customers better, and quickly respond to its customers' individual needs and wants. Furthermore, it wants to forge its CRM activities forward, which includes retaining customers that have high value, and acquiring new ones.

According to Pritscher et.al. (n.d.), CRM in the airline industry is rather at a starting point. Pritscher continues, a wealth of data is available from airlines' FFPs, which leads to better understanding of customer types and behavior.

ETHIOPIAN's CLD has a customer database, which contains information on the over 22,000 ShebaMiles members, that include their flight activities, demography, and their current status in

the program. These data need to be analyzed in order to find information that could help develop new business strategies and opportunities. If found, this information can increase the share of wallet for each customer and enable the CLD save costs by focusing on more targeted promotions.

Data mining, according to Bounsaythip (2001), can handle large amounts of data and 'learn' inherent structures and patterns in data as well as generate rules and models that are useful in replicating or generalizing decisions that can be applied to cases in the future. Data mining techniques are very useful in market segmentation and customer profiling.

Pritscher (n.d.) notes, by making use of data mining techniques, FFP members can be divided into similar groups on the basis of their common attributes such as their travel behavior and their revenue contribution. If valid and meaningful segments are found, they could offer the CLD an opportunity to know more about the loyalty and profitability of ShebaMiles' members.

The customer knowledge derived from these segments will enable the CLD to focus on more targeted promotions. Furthermore, knowing customers' needs better and treating them accordingly can increase their lifetime value.

In addition, the researcher believes the segmentation model, which will be the result of this study, can lead to better understanding of airline customers' behavior, especially in the African context. Most importantly, it is believed that this study will make a contribution to ETHIOPIAN's endeavors in implementing CRM.

1.4 Objectives

1.4.1 General Objective

The general objective of this research is to study the application of data mining techniques on the frequent flyer customers' database, in order to discover strategic customer segments that could support CRM activities at ETHIOPIAN.

1.4.2 Specific Objectives

The specific objectives of this study are the following:

1. Identify sources of customer data that are required for the segmentation study.
2. Collect data from the sources identified.
3. Prepare data for model building by selecting, cleaning, constructing and integrating the collected data.
4. Identify an appropriate data mining software and apply the algorithm of choice that would segment the FF members' records, based on their shared attributes.
5. Evaluate, with CLD staff, whether the discovered segments are meaningful.
6. If segments are meaningful, identify decision rules to classify new member records to the segments found.

1.5 Research Methodology

In order to achieve the objectives of this research, the researcher has adopted the CRoss-Industry Standard Process for Data Mining (CRISP-DM) model (SPSS, 2000). The primary reason for

choosing the CRISP-DM process was the fact that this approach has been widely applied for data mining studies in other industries, including similar customer data mining studies conducted for a major European airline consortium (Pritscher, n.d.). Accordingly, the following methods have been employed for this study.

a) Identifying available data sources

The primary source of data for this study was ETHIOPIAN's FF database (ShebaMiles DB), which contains flight activity and demographic data pertaining to members of ShebaMiles. Another important variable, which is the corresponding revenue value for each flight activity made by a member, was not available in ShebaMiles DB. A revenue accounting database was the source of this revenue data, where revenue information regarding individual flight activity is available.

b) Data collection and preparation for analysis

Members' flight activity data was initially collected from ShebaMiles DB. The main issue being the assignment of a revenue value for each flight activity from a revenue accounting DB, revenue data was filled for the flight activities, where a unique match existed. A procedure was developed that would fill missing revenue data for the remaining flight activities. The execution of this procedure filled in revenue values for the majority of the flight activity records. Those records, for which this procedure was not able to fill missing revenue data, were excluded.

Since the flight activities data was complete at this point, the following task was the selection of sample records from the universe of the members' data. The sampling criteria used was members' records, which had at least one flight activity within a 12 months period (which

corresponded with CLD's calendar year). The resulting flight activity data were then aggregated at member level, and new attributes were derived from the original ones.

c) Model building and evaluation of results

Clustering algorithms were applied to identify segments that were different from each other, but whose members were very similar to each other. Clustering algorithms were used since there were no predefined customer segments. Normalized counts and numeric variables were used to ensure that the influence of all variables was similar.

The segmentation results were reviewed with CLD staff to verify whether the clusters made business sense. Once meaningful segments concerning the business problem were found, the segmentation rules were applied to the validation set in order to further evaluate the results. Next, by means of decision trees, rules were identified, which could be used to classify new members to the respective segments.

1.6 Scope and limitations

The scope of this research is limited to the members of the frequent flyer program of ETHIOPIAN, where the required customer data was available. Furthermore, the study was limited to the development of a customer segmentation model, and does not include the deployment of the same.

Related literature on the application of data mining techniques in the airline industry was very limited. The acquisition of appropriate data mining software was a time taking activity.

1.7 Organization of the thesis

The thesis is organized in five chapters. The first chapter is mainly introductory, which illuminates on the problems that form the basis for this study. The general and specific objectives of this study, as well as the methodology used to achieve them are discussed.

In the second chapter, literature on the technology that is at the center of this study, namely data mining, and its application for CRM, is reviewed. Clustering techniques, such as the K-means and self-organizing map (SOM), which are commonly used for customer segmentation, are also reviewed.

The third chapter contains a business survey of CRM activities at ETHIOPIAN. The overall business process of the FFP is surveyed and problems are identified. Moreover, possible solutions to these problems are indicated.

In the fourth chapter, which is the experimentation part of this research, the different steps followed in collecting and preparing data are described. Furthermore, the model building process using K-means clustering and decision tree algorithms, including the various parameters used, are discussed in detail. Finally, summary and interpretation of the experiments' results are given.

In the fifth chapter, concluding remarks and recommendations are made.

Chapter 2

Customer Relationship Management and Data Mining

2.1 Loyalty and Customer Relationship Management

2.1.1 Overview

Loyalty is defined as a true and faithful act or behavior (Oxford Dictionary, 1997). Businesses have long known the importance of creating and maintaining customer loyalty. It is a common belief among businesses that it costs more to find a new customer than to keep and grow an existing one. However, recent studies indicate that despite heavy investments in customer satisfaction efforts and rewards programs, loyalty remains an elusive goal in almost every industry (Mc Kinsey & Company, 2001).

The primary job of a loyalty-based marketing effort is to enable the firm to find and retain the right customers. Reichheld (1995) believes that the right customers are those to whom the best value can be delivered by the firm over a sustained period of time. Companies study their customers' base and segment it into those who are highly loyal and those who are less loyal. In response to these findings, companies focus all their marketing activities on the loyal customer segment.

The recent trend in loyalty management is changing from a reward-based relationship to one that is defined through sharing information with customers. Petersen (as quoted by Chandler, 2001) believes it is about letting the customers decide that the company understands who they are,

rather than what they are. It is the understanding of 'who' the customer is that underlies what is known as *customer relationship management* (CRM).

Subject experts advocate that there is a vast difference between loyalty/reward programs, and CRM. According to Petersen (Ibid), the first is concerned with rewarding behavior that is assumed to be loyal, while the second is concerned with managing behavior to create loyalty. Petersen continues that the first deals with creating value for a customer, while the second with developing value from a customer, and that the real value of CRM is when the company earns loyalty without reward.

Market segmentation, according to DSS Research (2001), describes the division of a market into homogeneous groups, which will respond differently to promotions, communications, advertising and other marketing mix variables. Furthermore, each group or 'segment' can be targeted by a different marketing mix, because the segments are created to minimize inherent differences between respondents within each segment and maximize differences between each segment.

2.1.2 Loyalty and CRM in The Airline Industry

The competitive nature of the airline industry dictates that airlines put in a lot of effort and money to ensure that their customers remain loyal. To this effect airlines have launched loyalty programs, the earliest of which are Frequent Flyer Programs (FFPs).

According to McDonald (2001), American Airlines Inc. were pioneers in launching *AAdvantage*, the first true FFP in the airline industry. Under this program, passengers are awarded mileage

points for each flight they flew. As their points total builds up, they are entitled to increasingly attractive free flights or other travel benefits. Free flights are normally offered only on low load factor[1] services, so the airlines can claim that the cost of their program is low.

Petersen (as quoted by Chandler, 2001) tells us that, although reward programs were successful in creating a form of loyalty, many such programs found that airlines are only as valuable to their customers as the last major awards. One shortcoming of such programs is that customers tend to fixate on the rewards. Consequently, product superiority becomes less of a priority. Moreover, with many such programs, one reward is generally as good as another and creates cost for the company with no sustainable differentiable competitive advantage.

According to Chandler (2001), in most of the research that has been done with members of popular FFPs, it has been determined that only about 11 percent of active members fall into a defined category as being 'loyal'. The reason given for this rather low figure is that customers usually perceive the mileage awarding airlines as 'companies of convenience', rather than 'companies of care'.

Ever since American Airlines launched the first frequent flyer program of its kind, other airlines started to emulate it in setting-up their own frequent flyer programs. The source of the airlines' inspiration was how well the 80/20 Pareto principle applied to their business; where according to Holtz (1992), 80 percent of their business was attributable to 20 percent of their passengers, the passengers who flew regularly on business trips.

[1] Load factor is defined as the number of passengers carried as a percentage of seats available (Doganis, 1991).

FFPs, which are also known as loyalty/reward programs, are concerned with rewarding behavior that is assumed to be loyal (Chandler, 2001). The key features of the program being members earning and spending miles for air travel, mileage accrual and redemption[2] were possible for activities such as hotel stays, car rental and credit card usage. Chandler further notes that Airlines often spend 3 to 6 percent of their revenue on frequent flyer programs compared to 3 percent on advertising. However, frequency programs alone do not produce a very good return on investment if airlines' aim is to retain their top customers.

Many airline frequent flyer programs generate mass mailings to virtually every program member. According to Mammano (as quoted by Chandler, 2001) members later find out that the promotion can't possibly apply to them, thus lowering their opinion of the airline in the process. On the other hand, targeted promotions based on a customer's behavior and inclinations have a chance of working and earning loyalty in the process.

Petersen (Ibid) notes that, instead of concentrating only on rewarding behavior that is assumed to be loyal, airlines realized that they should concentrate on managing behavior to create loyalty, which is the theme of CRM. Furthermore, the miles and points which are accrued are not the measure of a good CRM program, and that the real value of CRM is when loyalty exits without reward.

It is widely shared in the loyalty and CRM industry that, loyalty programs could be an entrée into CRM, while frequency programs alone are not. Frequency programs are not loyalty programs; but legitimate loyalty programs often lead to CRM (Chandler, 2001).Chandler believes that the

[2] Redemption is the act of spending miles for air travel, hotel stays, etc.

primary focus of frequency programs is to build repeat business, while the focus for loyally programs is to build an emotional attachment to the brand.

More focused and more productive promotions are one advantage of CRM. According to Anderson (as quoted by Canaday, 1999), the big advantage starts with an airline's ability to segment its customers based on their profitability. Marketing will then be able to run more targeted promotions geared towards the different customer segments. In addition, the new customer insight can be used to improve customer services.

Another note of advise that comes from Dettman (as quoted by Canaday, 1999) is that airlines should not just focus on selling more, that they have to include service. He believes that the hardest part of CRM is integrating sales, marketing and service. Integration in the airline industry is especially hard because much of the selling is done indirectly. In air travel, 70-80 % of the sales are made through indirect channels (through travel agents). Therefore, airlines should strive to integrate their service process with direct and indirect sales channels.

According to Pritscher (n.d.), most market leaders in the airline industry orient their CRM around frequent flyer programs. The reason is that there is a wealth of data available in these frequent flyer programs, which allows to get a better understanding of customer types and customer behavior. For instance, when a FF customer calls Delta Airlines' sales office and inputs a membership number, the sales agent starts the conversation with the customer information already displayed by means of a customer management system (Donoghue, 2002).

primary focus of frequency programs is to build repeat business, while the focus for loyally programs is to build an emotional attachment to the brand.

More focused and more productive promotions are one advantage of CRM. According to Anderson (as quoted by Canaday, 1999), the big advantage starts with an airline's ability to segment its customers based on their profitability. Marketing will then be able to run more targeted promotions geared towards the different customer segments. In addition, the new customer insight can be used to improve customer services.

Another note of advise that comes from Dettman (as quoted by Canaday, 1999) is that airlines should not just focus on selling more, that they have to include service. He believes that the hardest part of CRM is integrating sales, marketing and service. Integration in the airline industry is especially hard because much of the selling is done indirectly. In air travel, 70-80 % of the sales are made through indirect channels (through travel agents). Therefore, airlines should strive to integrate their service process with direct and indirect sales channels.

According to Pritscher (n.d.), most market leaders in the airline industry orient their CRM around frequent flyer programs. The reason is that there is a wealth of data available in these frequent flyer programs, which allows to get a better understanding of customer types and customer behavior. For instance, when a FF customer calls Delta Airlines' sales office and inputs a membership number, the sales agent starts the conversation with the customer information already displayed by means of a customer management system (Donoghue, 2002).

By making use of their marketing database, organizations have been able to improve marketing results or lower their marketing costs. This general area of making use of marketing databases is known as database marketing. Holtz (1992) defines database marketing as marketing in which the approaches, strategies, methodologies, and other key marketing factors are founded on a consumer database that has this wealth of information about the customer in it.

In the airline industry, CRM is heavily dependent on IT. CRM being about appealing to the 'top-tier customers' (that section which generates the highest yield), it is very difficult to exploit top-tier data without the aid of IT. Data analysis tools can be used to extract knowledge from data. Among these tools, data mining tools can handle large amounts of data and learn inherent structures and patterns in data. Bounsaythip (2001) notes that, data mining tools can also generate rules and models that are useful in replicating or generalizing decision that can be applied to future cases.

2.2 Data Mining

2.2.1 Overview

According to Berry et.al. (2000), although the rapid pace of change in the past century was felt in nearly every area, it is hard to find examples of anything, anywhere, that has changed as fast as the quantity of stored information. Berry et.al. assert that this information explosion has created new opportunities and new headaches in every field, ranging from marketing to medicine to manufacturing.

Fayyad et.al. (1996) note, historically, the notion of finding useful patterns in data has been given a variety of names, including data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing. Furthermore, the term *data mining* has mostly been used by statisticians, data analysts, and the management information systems (MIS) communities and also gained popularity in the database field.

Many definitions of data mining could be found in the literature, Berry et.al. (1997) define it as the exploration and analysis of large quantities of data by automatic or semiautomatic means in order to discover meaningful patterns and rules. According to Bigus (1996), data mining is the efficient discovery of valuable, non-obvious information from a large collection of data. The steps in the evolution of data mining are depicted in Table 2.1, which is taken from Thearling (n.d.).

Evolutionary Step	Enabling Technologies	Characteristics
Data Collection (1960s)	Computers, tapes, disks	Retrospective, static data delivery
Data Access (1980s)	Relational databases (RDBMS), Structured Query Language (SQL), Open Database Connection (ODBC)	Retrospective, dynamic data delivery at record level
Data Warehousing & Decision Support (1990s)	On-line analytical processing (OLAP), multidimensional databases, data warehouses	Retrospective, dynamic data delivery at multiple levels
Data Mining (Emerging Today)	Advanced algorithms, multiprocessor computers, massive databases	Prospective, proactive information delivery

Table 2.1 Steps in the evolution of data mining.

The term knowledge discovery in databases (KDD) is synonymously used with data mining. The phrase *knowledge discovery in databases* was coined at the first KDD workshop in 1989 (Piatetsky-Shapiro, 2000). The purpose was to emphasize that knowledge is the end product of a

data-driven discovery and it has been popularized in the 'artificial intelligence' and 'machine-learning' fields. In Fayyad's et.al. (1996) view, KDD refers to the overall process of discovering useful knowledge from data and that data mining refers to a particular step in the process. Furthermore, data mining is considered as the application of specific algorithms for extracting patterns from data.

Trybula (1997) states that knowledge discovery (KD) is the process of transforming data into previously unknown or unsuspected relationships that can be employed as predictors of future action. Furthermore, Trybula notes that KDD is a term that has been employed to encompass both data mining and KD. Essentially, the basic tasks of data mining and KD are to extract particular information from existing databases and convert it into understandable or sensible conclusions (i.e., knowledge). In this research paper, the term data mining refers to the entire process from construction of databases through pattern identification and reporting. The KDD process is described in a graphical form in Figure 2.1, which is taken from Fayyad et.al. (1996).

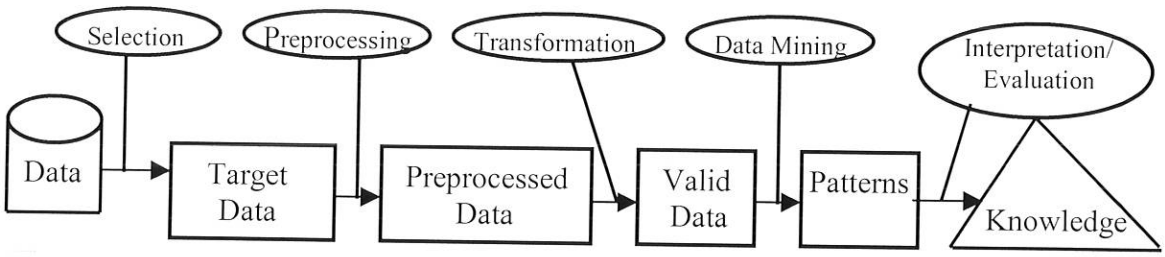


Figure 2.1 An overview of the steps that compose the KDD process.

Data mining involves more than merely applying software. According to Levin et.al. (1999), it is a process that involves a series of steps to preprocess the data prior to mining and post-processing

steps to evaluate and interpret the modeling results. Starting with the definition of the business problem, data mining is an iterative process requiring quite an important input from the user.

Data mining is a tool and its effective use requires a business to have good knowledge of its business process, comprehend its data and maintain an understanding of analytical methods. According to Berry et.al. (2000), Data mining assists business analysts with finding patterns and relationships in the data - it does not tell the value of the patterns to the organization. Furthermore, the patterns uncovered by data mining must be verified in the real world.

According to TCC (1999), the basic steps of data mining for knowledge discovery can be summarized as:

1. Understand the business problem
2. Build data mining database (includes the collection, description, selection, cleansing, and consolidation and integration of data)
3. Explore data
4. Prepare data for modeling (includes the selection of variables and rows, as well as constructing new variables and transforming them)
5. Build model
6. Evaluate model
7. Deploy model and results.

Understanding the available data before embarking on building a model(s) is a very important step. Data can be *continuous*, having any numerical value or *categorical*, fitting into discrete classes. Categorical data can be further defined as *ordinal*, having a meaningful order, or

nominal, that is unordered. According to TCC (Ibid), graphing and visualization tools are vital aids in data preparation and very important to effective data analysis.

Data mining models could either be *descriptive* or *predictive*. In descriptive models, the training is conducted using data for which all variables are *independent*, and there is no *dependent* or *target variable*. This kind of model building is known as *unsupervised learning* or *undirected data mining*.

In predictive models, the values or classes that are predicted are called *dependent* or *target variables*, while *independent variables* are used to make the prediction. This kind of model building is referred to as *supervised learning* or *directed data mining*, because unlike descriptive models the training is conducted using data for which the dependent or target variable is already known.

Two of the major descriptive data mining tasks are *clustering* and *link analysis*. Clustering divides a database into different groups, and its goal is to find groups that are very different from each other, and whose members are very similar to each other (Han et.al., 2001). Clustering is different from *classification* in that there are no predefined classes and it belongs to what is known as *unsupervised learning*. The clusters must be interpreted by someone who is knowledgeable in the business. The most common algorithms used to perform clustering include *K-means* and *Kohonen feature maps*. Table 2.2, which was taken from Bigus (1996), shows the common data mining tasks.

Data Mining Task	Techniques	Application Examples
Association discovery	Statistics, set theory	Market basket analysis
Classification	Decision trees, neural networks	Target marketing, quality control, risk assessment
Clustering	Neural networks, statistics	Marketing segmentation, design reuse
Regression	Linear and nonlinear regression, Curve fitting, neural networks	Ranking/scoring customers, pricing models, process models
Time-series forecasting	Statistics ARMA models, Box-Jenkins, neural networks	Sales forecasting, interest rate prediction, inventory control
Sequential discovery	Statistics, set theory	Market basket analysis over time

Table 2.2 Most common data mining tasks

Another descriptive data mining approach that can help identify relationships among values in a database is *link analysis.* According to TCC (1999), two of the most common approaches to link analysis are *association discovery* and *sequence discovery.* In the case of association discovery, rules about items that appear together are found. *Affinity grouping*, which is also known as *market basket analysis,* is a well known technique used in Association discovery. Sequence discovery is an association related over time.

Classification, according to Berry et.al. (1997), the most common data mining task, consists of examining the features of a newly presented object and assigning it to one of a predefined set of classes. Han et.al.(2001) state that classification and prediction are two form of data analysis that can be used to extract models describing important data classes or to predict future data trends. Classification models are created by examining already classified data and inductively finding a predictive pattern.

Other types of models that are commonly used for prediction are *regression* and *time series forecasting*. Regression uses existing values to forecast the outcome of other values. Standard statistical techniques such as linear, nonlinear, and generalized linear regression models. Time series forecasting predicts unknown future values based on a time-varying series of predictors. It uses known results to make its predictions, just like regression. The most widely used techniques for classification are decision trees, neural networks and memory-based reasoning (Berry et.al., 1997).

The application of data mining spans various industries. Telecommunications and insurance industries make use of data mining techniques to detect fraudulent activities. In medicine, data mining is used to predict the effectiveness of surgical procedures and medical tests. Companies in the financial sector use data mining to determine market and industry characteristics as well as to predict individual company and stock performance (TCC, 1999).

2.2.2 Data Mining and CRM

In describing the very important role that data mining plays in CRM, Berry et.al. (2000) note that, it is only through the application of data mining techniques that a large enterprise be able to turn the myriad records in its customer databases into some sort of coherent picture of its customers.

According to TCC (1999), many organizations are using data mining to help manage all the phases of a customer lifecycle, including acquiring new customers, increasing revenue from existing customers, and retaining good customers. Using data mining to profile its customers, a company can better treat its customers with similar characteristics.

One of the most widely used applications of data mining for CRM is in understanding customer behavior. In a customer segmentation study for a telecommunications company, Berry et.al. (2000) note that, initially they conducted an investigatory work to determine what information is interesting by talking to business experts, since it was important to use data that business users understand and find value. In the model-building phase, only four fields were chosen for the segmentation, from the many tables with customer information. Their study results showed that 'call detailed' records contain a wealth of information on customer behavior, and that detailed transaction records in other industries also provide important information about customers.

In the retail sector, data mining is used to mine point-of-sale transactions to find associations between products (Bigus, 1996). This information is used to determine product groupings and devise promotion strategies that can maximize profits. In data mining studies made on supermarkets, Berry et.al. (2000) used *K-means* algorithm to find groups of customers with similar behavior. Data mining techniques were used to improve 'shelf placement' decisions and to uncover a small, but very profitable group of customers.

According to Schultz (2001), a food store in the U.K. has used data mining to derive 157,000 different segments from the 11 million households in its database. Accordingly, each household is placed in the segments so that it would be easier to plan strategies to move customers from one segment to the next.

Churn, according to Berry et.al. (2000), is the word used in the wireless telecommunications industry to refer to customers' likelihood to defect to competitors. Churn modeling, which

predicts customers who are likely to leave in the near future, is becoming a common data mining application. Berry et.al. further note that decision tree techniques are more advantageous for churn modeling.

Levin et.al. (1999) conducted a study to increase donation amounts, by using data mining techniques to exclude people that are not likely to respond to a charity solicitation. The target marketing models identified a subset of the testing audience to solicit, which increased the net donation amount by almost 40% as compared to mailing to everybody.

Recent use of data mining is by on-line banks, where decisions as to which customer to cross-sell (sell additional services to existing customers) are supported. According to Berry et.al. (2000), decision tree models were built before arriving at the best cross-sell model.

In the hospitality business, data mining is used to support CRM. According to Sickel (as quoted by Chandler, 2001), cluster models of customer segments are supporting differential marketing activities at Six Continents Hotels. Furthermore, predictive models are built based on the segmentation results.

2.2.3 Data Mining in the airline industry

According to Pritscher (n.d.), the most obvious application of data mining in the airline business is related to frequent flyer programs. In a study conducted for a major European airlines alliance group (Qualiflier), Pritscher et.al. (n.d.) note that the objective was to explore the available databases by use of data mining methods in order to support the implementation of an efficient

CRM, in which case the first task is to identify market segments containing customers with high profit potential.

The segments, according to Pritscher (n.d.), must be explainable and the added value must be evident. Since the value of a passenger is measured in miles, a monetary value must be assigned to each passenger, which can be used to calculate profitability based on segmentation results, and allow to identify core customers.

Pritscher (n.d.) notes that the resultant segmentation, which was based on travel behavior, led to six customer segments that made sense to the business problem. Pritscher advises that, since there is no actual quantitative definition of a good segmentation, assessing the groups by investigating their revenue distribution (customer value) is important. These found segments could therefore be used for special marketing strategies.

According to Pritscher et.al. (Ibid), in an initial phase of CRM, customer segments based on individual patterns are found, describing groups of customers with distinct needs and value. These segmentation results are useful for marketing concerns and for improving customer services, and conclude that data mining is very useful to support CRM in the airline industry.

According to Harris (n.d.), British Airways analyzed customer data to discover instances where high revenue generating customers had flown one-way, but used another airline on the return. It then offered these valued customers a special incentive to use their services both ways.

According to IBM (2000), Cathay Pacific Airways (Hong Kong's national airline) has all relevant frequent flyer customer data in a data warehouse, from which various segmentation models are derived, thus enabling the airline to focus on specific customer segments.

Gobena (2000) studied the possible application of data mining techniques that could help in forecasting flight revenue information for ETHIOPIAN. Neural network algorithms, specifically multi-layer perceptron back propagation network and radial basis function architectures, were used for this study. According to Gobena, the final revenue model selected for his study had an average of 33-37% error rate, and believes that better results were possible with more training. Moreover, he believes that data mining techniques could be applied to support decision making at ETHIOPIAN.

Data mining in the airline business is not limited to customer databases. Another area where data mining has been put to use is 'airline pricing'. According to Data Warehouse Report (1998) online airline pricing employing speeded-up data mining techniques are employed to allow Reno Air to quickly track rival airlines' fare changes, and suggest competitive fare matches. These on-line, airline pricing, solutions store historical market data (including fare changes) for comparison purposes and 'what-if analysis', as well as to highlight competitors' changes by market.

EDS (2001) have used data mining to study the cause for delays at airports. The variables used were the flight schedule, weather conditions, and the types of flight delays. The study results indicated what the major causes for the delays were, and steps were taken to mitigate the problem accordingly.

2.3 Customer Segmentation

2.3.1 Overview

Bounsaythip (2001) describes customer segmentation as the process of dividing customers into homogeneous groups, where customers within each group are similar to each other than to others on the basis of shared or common attributes.

The data mining techniques mostly used for customer segmentation are *clustering* and *classification*. Saarenvirta (1998) notes that customer clustering and classification are two of the most important data mining methodologies used in marketing and CRM. Furthermore, Saarenvirta believes that businesses can use this data to divide customers into segments based on such variables as current customer profitability, a measure of the lifetime value of a customer, and retention probability, which highlight visible marketing opportunities.

2.3.2 Clustering Techniques

The ultimate goal of clustering is to find groups that are very different from each other, and whose members are very similar to each other. Clustering, according to Berry et.al. (2000), is the task of segmenting a diverse group into a number of more similar subgroups or clusters. Basically, clustering divides a database into different groups. Clustering is also the technique of choice at the beginning of a new data mining project.

This process of building models that find data that are similar to each other (clusters) belong to *undirected (unsupervised) data mining*, the goal of which is to find previously unknown

similarities in the data. There is no prior knowledge of what the clusters will be, or the attributes by which the data will be clustered. Berry et.al. (2000) state that it is up to the data miner to determine what meaning, if any, to attach to the resulting clusters.

Pritscher et.al. (n.d.) note that clustering algorithms are appropriate, if there is no predefined segmentation. According to TCC (1999), a person knowledgeable in the particular business domain must interpret the clusters. It is often necessary to modify the clustering by excluding variables that have been used to group instances, which upon examination by the domain expert have been identified as irrelevant or not meaningful.

According to Bishop (1995), as an improvement on simply choosing a subset of the data points as the basis function centers, clustering techniques can be used to find a set of centers, which more accurately reflect the distribution of the data points. The most common methods used to perform clustering are *K-means* and *Kohonen feature maps* (also known as *self-organizing maps* or *SOM*) (Bounsaythip, 2001).

In a customer segmentation study, which was conducted on a FFP database, Pritscher et.al. (n.d.) applied the K-means clustering algorithms to identify groups which are different from each other according to their product mix, but whose members are very similar to each other. According to Pritscher et.al., several runs of K-means were applied with 6 – 10 clusters before arriving at six segments, which made 'business sense'. The SOM algorithm was used to validate the cluster results from the K-means algorithm, where the clusters were separated in the SOM, and the properties of neighboring clusters were sensible, thus indicating a stable clustering solution.

In another study, which was conducted for a loyalty group that runs an 'air miles' reward program, Saarenvirta (1998) notes that the specific objectives were to create a customer segmentation using the K-means clustering algorithm. In the study, a maximum of nine clusters were chosen and a maximum five passes through the data (iteration). According to Saarenvirta, the results from the study were valid.

2.3.3 The K-Means Method

The K-means algorithm for cluster detection, according to Berry et.al. (2000), is the most widely used in practice. This method (algorithm) divides a data set into a predetermined number of clusters. That number is the "K" in the phrase K-means. Just as a mean is an average statistically, it refers to the average location of all of the members (which are records from a database) of a particular cluster. The *K-means* algorithm 'self-organizes' to create clusters. According to Bishop (1995), the algorithm involves a simple re-estimation procedure.

Supposing there are N data points x^n in total, and the intention is to find a set of K representative vectors μ_j where $j = 1,\ldots, K$, the algorithm seeks to partition the data points $\{x^n\}$ into K disjoint subsets S_j containing N_j data points. This would minimize the sum-of-squares clustering function given by

$$J = \Sigma_{j=1}^{k} \Sigma_{n \in Sj} [x^n - \mu_j]^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

Where μ_j is the mean of the data points in set S_j and is given by

$$\mu_j = 1/N_j \, \Sigma_{n \varepsilon Sj} \, x^n \qquad \ldots\ldots\ldots\ldots\ldots\ldots \qquad (2)$$

As described by Bishop (Ibid), the process begins by assigning the points at random to K sets and then computing the mean vectors of the points in each set. The algorithm assigns each of the points to the cluster to whose center it is closest in *Euclidean* distance. Next, each point is re-assigned to a new set according to which is the nearest mean vector. The means of the sets are then recomputed. This procedure is repeated until there is no further change in the grouping of the data points. At each such iteration the value of J will not increase.

Bishop (Ibid) further states that the calculation of the means can be formulated as a stochastic on-line process. In this case, the initial centers are randomly chosen from the data points, and as each data point x^n is presented, the nearest μ_j is updated using

$$\Delta\mu_j = \eta(x^n - \mu_j) \qquad \ldots\ldots\ldots\ldots\ldots\ldots \qquad (3)$$

, where η is the learning rate parameter. Once the centers of the basis functions have been found in this way, the covariance matrices of the basis functions can be set to the co-variances of the points assigned to the corresponding clusters.

In order to form clusters, each record from a database is mapped to a point in 'record space.' The number of dimensions contained in the space correspond to the number of fields in the records. The value of each field can be geometrically interpreted as a distance from the origin along the corresponding axis of the space. In addition, to ensure the usefulness of this interpretation, the

fields must all be converted into numbers and the numbers must be normalized so that a change in one dimension is comparable to a change in another.

As described by Berry et.al. (2000), records are assigned to clusters through an iterative process that starts with clusters centered at essentially random locations in the record space and moves the cluster means (*centroids*) around until each one is actually at the center of some cluster records. Though this process can best be illustrated using two dimensional diagrams, in reality the record space will have many more dimensions, because there will be a different dimension for each field in the records. This has been depicted in Figure 2.2, which is taken from Berry et.al. (Ibid).

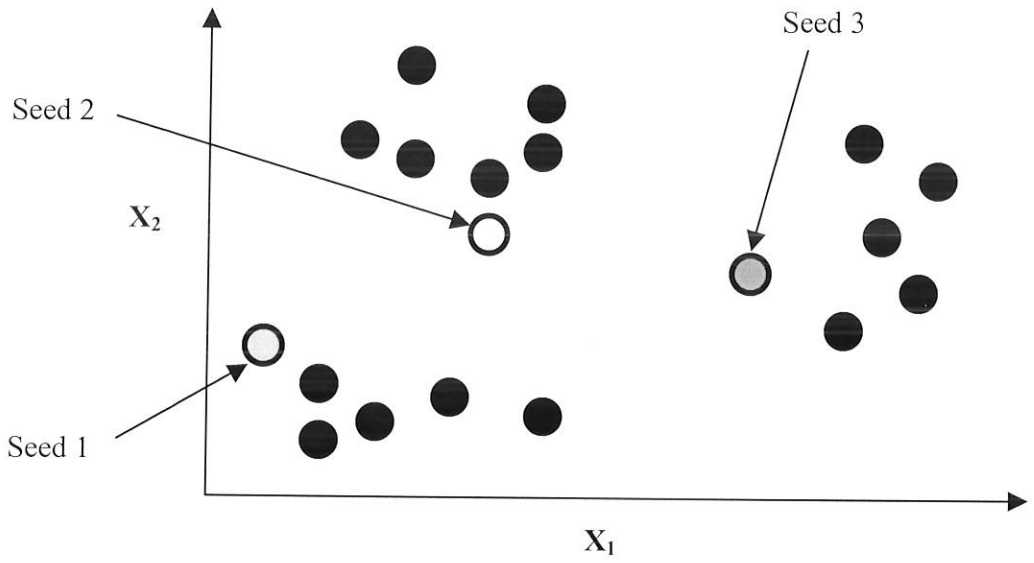


Figure 2.2 Initial Cluster Seeds

In Figure 2.3, which is taken from Berry et.al. (Ibid), the new centroids are marked with crosses. The arrows show the motion from the position of the original seeds to the new

centroids of the clusters. Once the new clusters have been found, each point is once again assigned to the cluster with the closest centroid. The process of assigning points to cluster and then re-calculating centroids continues until the cluster boundaries stop changing. The cluster boundaries are set after a handful of iterations for most data sets.

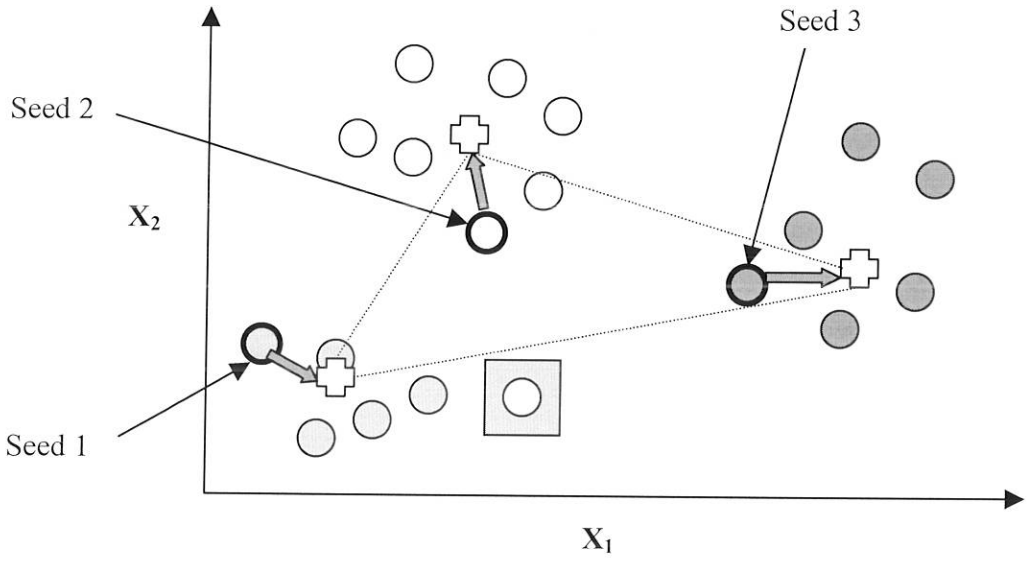


Figure 2.3 Cluster seeds after one iteration

According to Bounsyathip (2001), K-means is based on a concept of distance, which requires a metric to determine distances. Euclidean distance can be used for continuous attributes, while for categorical variables, one has to find a suitable way to calculate the distance between attributes in the data. Bounsyathip further believes that, since choosing a suitable metric is a very delicate task, a business expert is needed to help determine a good metric.

The original choice of a value for K determines the number of clusters that will be found. Furthermore, if this number does not match the natural structure of the data, the technique will not obtain good results. Unless the data miner suspects the existence of a certain number of clusters, she/he will have to experiment with different values for K.

Every set of clusters will then have to be evaluated. Berry et.al. (1997) believe that, in general, the best set of clusters is the one that does the best job of keeping the distance between members of the same cluster small and the distance between members of adjacent clusters large. They further state that, the best set of clusters in descriptive data mining may be the one showing some unexpected pattern in the data.

Once the clusters have been created, they need to be interpreted. Though there are several approaches to understanding clusters, according to Berry et.al. (2000), the three that are commonly used are:

1. Building a decision tree with the cluster label as the target variable and using it to derive rules explaining how to assign new records to the correct cluster.
2. Using visualization to see how the clusters are affected by changes in the input variables.
3. Examining the differences in the distributions of variables from cluster to cluster, one variable at a time.

Automatic cluster detection using the K-means algorithm is an undirected knowledge discovery process. According to Berry et.al. (1997), the algorithm works well with categorical, numeric, and textual data. Furthermore, it is easy to apply.

2.3.4 Self-Organizing Map (SOM)

The other popular clustering algorithm is self-organizing map (SOM), which is also known as Kohonen feature map. According to Bounsyathip (2001), SOM is a special kind of neural network architecture that provides a mapping from the multi-dimensional input space to a lower-

order regular lattice of cells, which is typically a two dimensional grid. Such a mapping is used to identify clusters of elements that are similar (in a Euclidean sense) in the original space.

A SOM tries to find clusters such that any two clusters that are close to each other in the grid space have clusters to each other in the input space. But the reverse does not hold, that is cluster centroids that are close to each other in the input space do not necessarily correspond to clusters that are close to each other in the grid. According to Bigus (1996), SOM is a feed forward neural network with no hidden layer that uses an unsupervised training algorithm and Euclidean distance measure. SOM has two layers of nodes, the input and output layers. The input layer is fully connected to one, two or multi-dimensional output layer.

In SOM algorithm, cluster centroids are assigned a location in spatially organized matrix. According to (Bounsyathip, 2001) the data is processed as follows:

1. Assign a 'neighborhood' function which, for a given centroid, identifies 'neighboring' cluster centroids.

2. For each data point:
    a. Find the cluster centroid which is 'closest' to the data point (the *winner*).
    b. Move the winner centroid towards the data point.
    c. Use the 'neighborhood' function to identify neighbor centroids and move them towards the data point .
3. Decrease the size of the neighborhood and repeat the process until the neighborhood only includes the winner centroid.

This way, the clustering starts out as a very general process and proceeds becoming more and more localized as the neighborhood decreases.

In the grid in Figure 2.4, which is taken from (Bounsyathip, Ibid), each square corresponds to a cluster. Each customer point has its distance computed from 16 cluster points. The cluster centroid is closest to this customer point is chosen to be the cluster center. Next all the other cluster centroids will move towards this chosen cluster center. This process is iterated until the cluster centroids can hardly move.

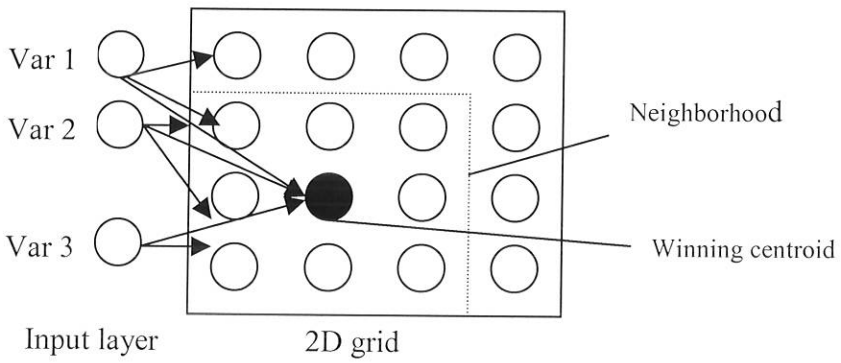


Figure 2.4 Example of SOM

2.3.5 Decision Trees

According to TCC (1999), decision trees are a way of representing a series of rules that lead to a class or value. There are two types of decision trees, namely classification trees and regression trees. Classification trees label records and assign them to the appropriate class. Classification trees can also provide the confidence that the classification is correct. In such cases, the classification tree reports the class probability, which is the confidence that a record belongs to a given class. Regression trees estimate the value of a target variable that takes on numeric values.

Decision tree learning, according to Bounsyathip (2001), is a method for approximating discrete-valued target functions, in which the learned function is represented by a decision tree. Decision trees make many tests and try to arrive at the best sequence for predicting the target. Each test creates branches that lead to more tests, until testing ends in a *leaf node*. The route or path from the root to the target leaf is the *rule* that classifies the target, and the rules are expressed in **if-then** form.

A decision tree grows from the root node, at each node the data is split to form new branches, until reaching a node that can not be split any more (leaf node). Traversing the tree from the best leaf node to the root provides the rule that classifies the target variable.

All records that arrive at a given leaf of a tree are classified the same way. Moreover, there is a unique path from the root to each leaf. That path is an expression of the rule used to classify the records. Figure 2.5, which is taken from Berry et.al. (2000), depicts a typical decision tree.

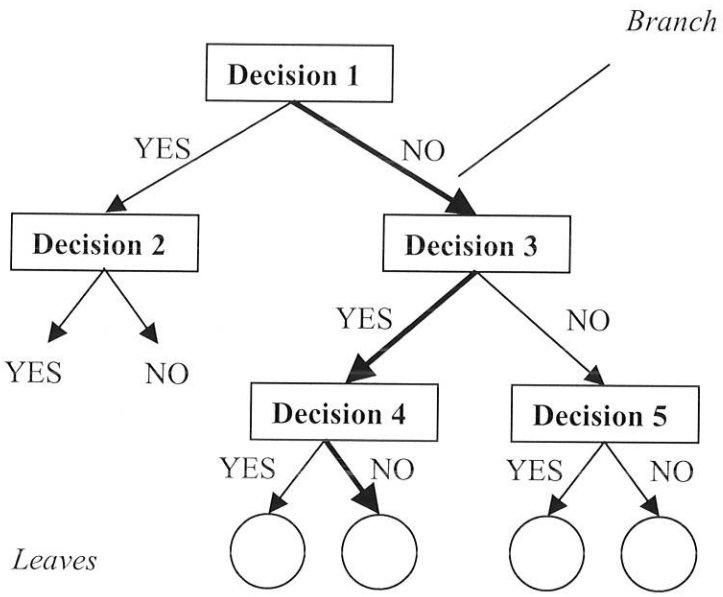


Figure 2.5 A Decision Tree.

Trees can grow in many forms. They could be binary trees of non-uniform depth, that is, each node has two children and the distance of a leaf to the root varies. In Figure 2.5, each node represents a 'yes' or 'no' question, the answer to which determines by which of two paths a record proceeds to the next level of the tree. Decision trees could have a mixture of binary and ternary nodes.

According to Berry et.al. (1997), decision-tree-building algorithms begin by trying to find the test that does the best job of splitting the data among the desired categories. At each succeeding level of the tree, the subsets created by the preceding split are further split according to rules that are appropriate at that level. The tree continues to grow until it is no longer possible to find better ways to split up incoming records.

Decision trees are built through a process known as recursive partitioning. According to Berry et.al. (2000), recursive partitioning is an iterative process of splitting the data up into partitions, and further splitting it up some more. All of the records in the training set, which are the preclassified records that are used to determine the structure of the tree, are initially together in one big box. Next, the algorithm tries breaking up the data, using every possible binary split on every field.

The decision tree algorithm chooses the 'split' that partitions the data into parts that are purer than the original. Each of the new boxes are further split or partitioned until no more splits can be found. The most important component of the algorithm is the rule that determines the initial split.

Han et.al. (2001) note that various decision tree algorithms produce trees that differ from one another in the number of splits allowed at each level of the tree, how those splits are chosen when the tree is built, and how the tree growth is limited. Among the various decision tree algorithms, the major ones are Classification and Regression Trees (CART), C4.5/C5.0, and Chi Square Automatic Interaction Detection (CHAID).

Chapter 3

A Survey of CRM at Ethiopian Airlines

3.1 General

This survey intends to assess the customer relationship management (CRM) process at Ethiopian Airlines (ETHIOPIAN). The purpose is to conduct an analysis of the current CRM process, to identify the critical functions and activities involved, and to identify and assess the availability of data sources that can support to derive a customer segmentation model that yield a reliable revenue value for each customer.

According to Toon Quee (1999), the primary function of marketing research is to utilize research abilities/facilities for gathering facts and knowledge to support marketing decision-making. Kotler (1998) describes that marketing research is the systematic design, collection, analysis, and reporting of data and findings relevant to a specific marketing situation facing the company.

In the process of conducting this survey, a series of interviews were made with staff of concerned departments. Furthermore, secondary sources of information, such as documents, departmental memos, and publications were made use of.

3.2 Ethiopian Airlines

ETHIOPIAN, which is a commercial airline, was founded in 1946 with an inaugural flight to Cairo in a war surplus airplane (Ethiopian Airlines: Bringing Africa Together, 1988). Today, ETHIOPIAN serves 40 international and 28 domestic destinations in the transport of passengers and cargo. In addition, more than half of its international destinations are in Africa (Ethiopian Airlines: Worldwide Timetable: Summer, 2002).

In 2001, ETHIOPIAN's international passenger carriage increased by 4.2% bringing up the total passenger volume to 970,000. During the same period, the airline realized a net profit of 6.5 million U.S. dollars (Ethiopian Airlines: Annual Report, 2001).

Among the airlines' service improvement strategies, the airline has finalized a study to purchase new aircraft. Furthermore, the airline has also given due emphasis to improving its information systems infrastructure. A major accomplishment in this regard has been the implementation of a wide area network (WAN) connecting its field sales offices to the head office using a virtual private network (VPN), which was leased from an international telecommunications services provider (Ethiopian Airlines: Annual Report, 2001).

3.3 The Frequent Flyer Program

ETHIOPIAN launched its frequent flyer program (FFP) named "*ShebaMiles*" in January, 1999. The name ShebaMiles is inspired by the legend of Makeda, the Queen of Sheba, who ruled Ethiopia around the 10th century B.C (ShebaMiles Membership Application, n.d.). The primary

reason for introducing Sheba Miles is to increase and award loyalty of customers. Moreover, the program is also used to identify high value customers and provide them with special services and benefits (like award tickets, upgrades, check-in and executive lounge privileges, special baggage allowances, etc.) by means of a top tier program.

According to the survey's results, over 22,000 members are currently enrolled in ShebaMiles. These members fall in one of the three Club levels of the program, which are Blue, Silver and Gold. Membership level is determined by the number of *Base Miles* flown annually. Base Miles refer to the number of miles a passenger flies on ETHIOPIAN, and is awarded for the sector flown.

There is no enrollment fee to become a member of ShebaMiles. Blue Club membership is granted when one enrolls in the program, making the member eligible to immediately start earning award miles. Members will be eligible for upgrading to Silver or Gold Club membership as soon as they have earned the required number of Base Miles.

When a member travels in excess of 25,000 Base Miles a year, she/he becomes eligible for Silver Club membership (ShebaMiles FFP Membership Guide, 2000). Some of the special privileges and benefits that Silver Club members are entitled to include: booking priority on waiting-lists, easier and more convenient check-in, excess baggage allowance, access to executive lounges, a certain percentage bonus on all Base Miles earned, advance boarding, and extended miles validity period.

The top tier of the program is the Gold Club membership which requires a member to fly more than 50,000 Base Miles in a year (Ibid). Though the types of benefits are similar as for Silver Club members, Gold Club members get to have better benefits such as: highest booking priority on waiting-list, a higher excess baggage allowance, a higher percentile bonus on all Base Miles earned, and a 24 hours hotline reservations service in Addis Ababa.

ShebaMiles members earn two types of miles, Base Miles and Bonus Miles. Bonus Miles are special member awards designed to reward frequent flyers as generously as possible. Each time a member earns Bonus Miles, they are added to her/his Base Miles to become part of her/his Award Mile balance (Base Miles + Bonus Miles = Award Miles).

There are different types of Bonus Miles that members can earn. Enrollment bonus miles are awarded to every new member across the board. Cloud Nine Class Bonus Miles entitle members to earn double the miles they would have earned in economy class. The airline runs special promotional programs on a regular basis. These include special promotional routes, where members who fly on these routes earn Promotion Bonus Miles over and above the usual number of Base Miles.

According to this survey, the Customer Loyalty Department (CLD) is responsible for running ShebaMiles. This department falls under the Market Development Division. Headed by a department manager, the department engages in developing, coordinating and directing all activities pertaining to keeping the loyalty of customers and on all matters related to the frequent flyer program. The department is responsible for ShebaMiles' objectives, policies, procedures, products, plans, programs and other related activities.

3.3.1 Business Processes of the Frequent Flyer Program

The CLD interacts with *frontline customer service* offices. These are offices that are in direct contact with the passengers of an airline. They include ticket offices, travel agencies, airport, reservations offices, in-flight services, ETHIOPIAN lounges, etc.

Reservations, Ticket Offices and Travel Agents

It is at one of these front-line customer services offices that passengers are likely to make their initial contact with the airline. When a ShebaMiles member makes a reservation, the reservations office will make sure that they include her/his program membership details in the booking profile. The booking profile of the member will be accessible to all the other offices of the airline through the airline's automated reservations system. Bilateral agreements between airlines to award frequent travelers may contain provision to exchange this information between International Air Transport Association (IATA) member airlines or from Computerized Reservation System (CRS) suppliers.

Ticket offices and travel agents also promote the program to potential passengers and solicit their enrollment in ShebaMiles.

Airport Offices

The main functions of these offices are check-in and boarding of passengers. The boarding information is provided electronically through the departure control system. The departure control system automatically generates a frequent traveler list (FTL). The FTL, a list containing

boarded passengers who have frequent traveler numbers is sent to the CLD. In addition, a departure information message (DIM), a post departure list containing the details of departure information forms (DIF) for a given flight, is composed and sent to the CLD as soon as a flight has departed. A DIF is a form to be filled out and signed by a ShebaMiles member before flight departure and verified by an airport staff.

The interactions of the frontline customer service offices with the CLD as well as the ShebaMiles members, which have been identified during the survey, are illustrated in Figure 3.1. The figure attempts to depict the overall business flow of ShebaMiles.

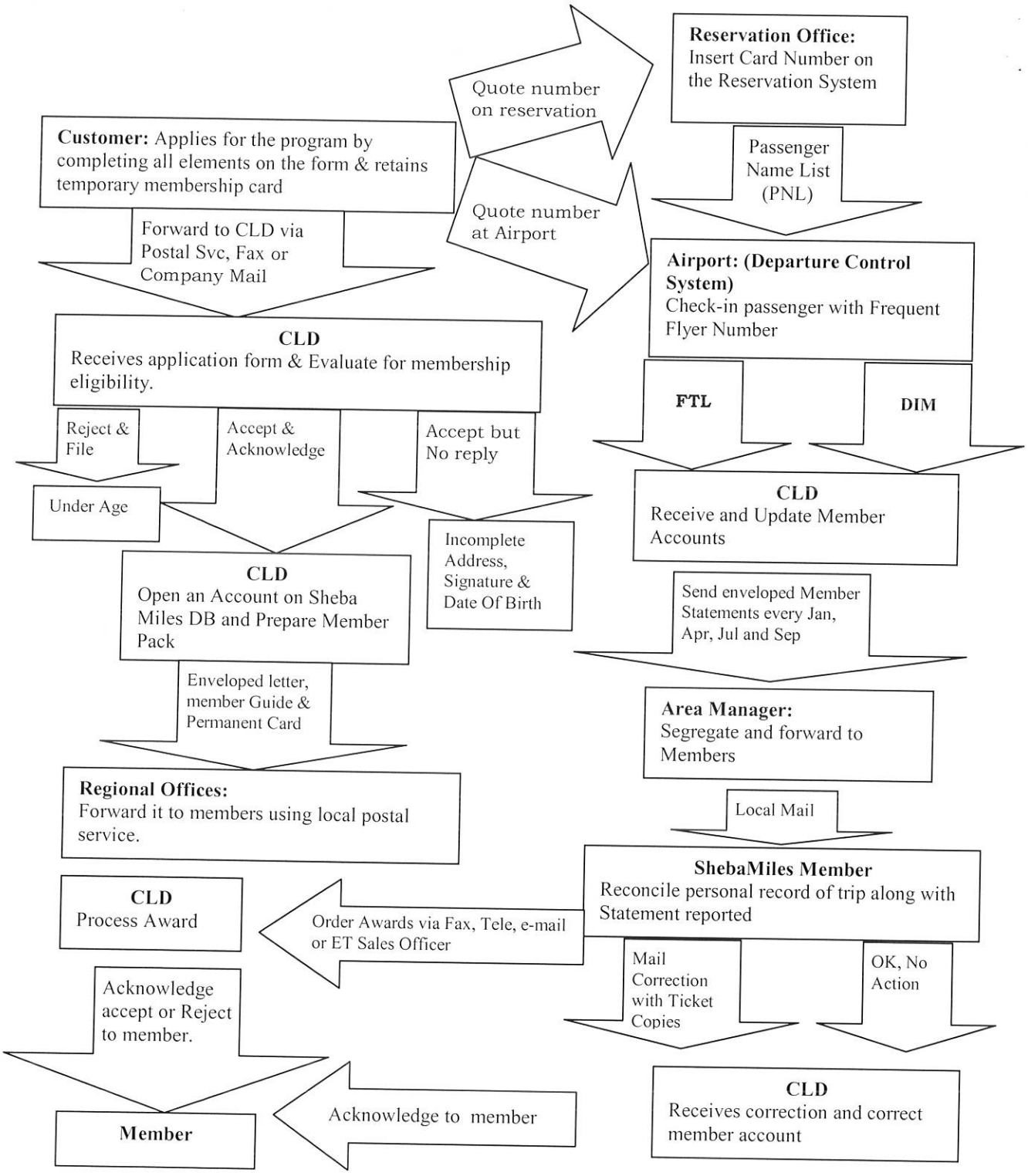


Figure 3.1 Business process of the ShebaMiles FFP Program at ETHIOPIAN

3.3.2 Overview of ShebaMiles' Database System

The CLD maintains a database (ShebaMiles DB) in order to store information and manage the program. Data pertaining to ShebaMiles members' activity is obtained through a customized interface to the airline's as well as other automated DCSs. The database consists of member account information, program requirements, and other pertinent data.

The computer programs that interact with the database regularly receive input data from users and other systems (primarily the DCS). These programs update the database with new, current information. In addition, the programs use database information to create several output files, which are used to print member materials (such as award redemption certificates and activity statements).

The main purpose of the ShebaMiles DB system is to:

- Establish and maintain the requirements of the ShebaMiles program rules.
- Establish and maintain member account data.
- Track member points within the established requirements.

In order to fulfill this purpose, the ShebaMiles DB system performs the following processes:

- Receive the input data provided and use it to retrieve the appropriate member account data from the database.
- Compare the activity-related input data to the flight segments, bonus and promotion programs.

- Perform calculations, such as comparing the member activity to the bonus and promotion requirements already established and determine the appropriate number of points to post to the member account.
- Apply the data to the member account and update the database.
- Make the database information available through reports.

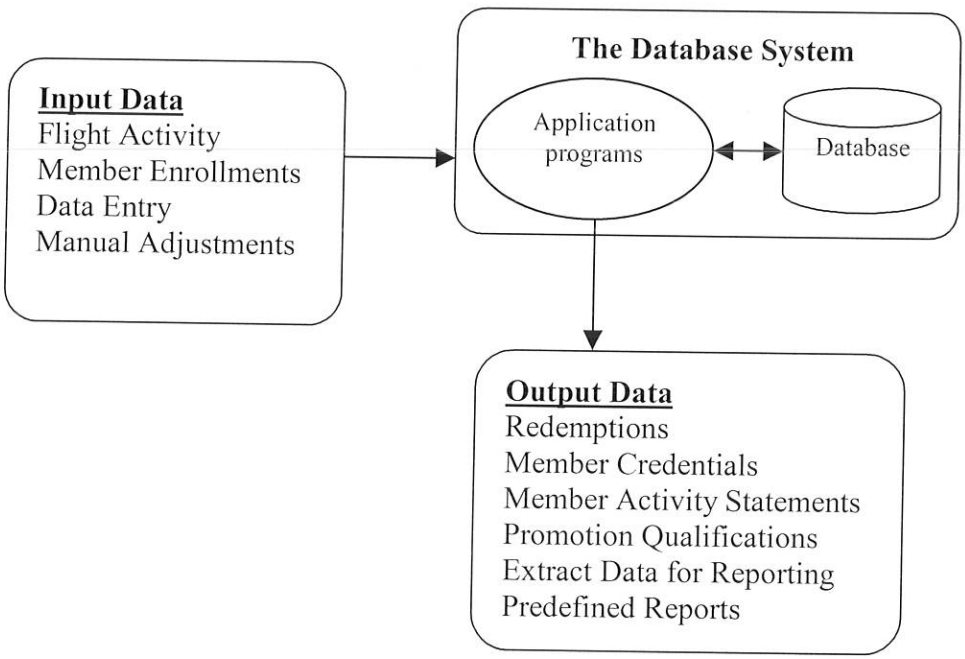


Figure 3.2 The data flow of ShebaMiles' database system.

The overall data flow of the ShebaMiles DB is illustrated in Figure 3.2. The data input functionality that the database system supports are:

- **Flight Activity:** The airlines' host departure control system creates a data file and electronically transmits the file (usually daily) to the database (DB) system. The data file consists of all member activity that occurred during the period.

- **New Member Enrollments:** Individuals wanting to join ShebaMiles should complete and mail enrollment forms to the airline. The information from the forms may be manually entered into the DB system or be electronically transmitted.

- **Data Entry:** Data entry clerks manually enter flight activity. This form of input is used when other computer systems do not electronically transmit flight activity.

- **Manual Adjustments:** Enables manual entry of adjustments for flight activity. This form of input is used when the DB system fails to automatically track points for an activity. For instance, if a flight reservation does not include a membership number, and the member fails to provide his/her number at the airport prior to flight departure, the airline host system does not include the activity in the flight activity data file; therefore, the activity data will not be entered into the member's account.

The ShebaMiles DB system creates the following output data:

- **Redemption file:** The DB system searches the database, locates instances where members redeemed points from their accounts to claim awards, and copies information for those members accounts into this file. The CLD uses the file to print letters and award certificates.

- **Member credential file:** The DB system searches the database, locates instances where members requested replacement cards or where the tier levels assigned to their accounts changed, and copies information for those member accounts into this file. The CLD uses this file to print letters, replacement cards, and packets of information for tier level changes.

- **Member activity statement file:** The DB system searches the database, locates activity records in member accounts, and copies the activity information into this file. The CLD uses the file to print member activity statements.

- **Promotion qualification file:** The DB system searches the database, locates instances where member activity qualified for a promotion, and copies information for those member accounts into this file. The CLD uses the file to print letters and promotion results (such as award certificates).

- **Extract files for reporting:** The DB system extracts requested information from the database and places the information into a file. Tools (such as Microsoft Access) can later be used to perform data queries or reporting analysis.

- **Predefined reports:** The DB system searches the database and generates predefined reports that can either be viewed or printed.

3.4 Findings of the Survey

The CLD is in the process of building a CRM environment at ETHIOPIAN. The department wants to enrich its knowledge of ShebaMiles' members so that it could run more targeted promotions, rather than the mass mail promotions it currently conducts.

The ShebaMiles DB, which runs on a Microsoft Visual Foxpro database management system (DBMS), contains over 22,000 program members that have accumulated over 90,000 flight

activities over the three years' lifetime of the program. The complete activity history of ShebaMiles members is stored to keep track of accrued and redeemed miles, and of the qualification for a tier level.

The researcher has found out that CLD currently holds a wealth of customer data, which could help to get a better understanding of customer types and customer behavior. The researcher believes that exploring this customer data may reveal new information that could help the CLD manage its customers better.

In the course of working with the domain experts in order to get an understanding of the business process, the researcher has identified some important questions, which could not be discovered by making use of conventional database query methods. It is the researcher's belief that the sheer volume of data and business requirements necessitate the use of data mining techniques to get a better customer understanding.

If 'valuable customers' could be identified among the ShebaMiles members, it could help CLD take actions that would reduce the risk of losing 'valuable customers' by rewarding them properly, and also reduce lost revenue from customers who would be more valuable if better rewarded. The CLD could also reduce costs of rewarding customers that do not deserve to be so.

Currently, customer value is based on individual mileage, which according to Pritscher (n.d.), is not a good measure of customer profitability. Although members may pay different fares for the same flight and class, the mileage points which accrues to their account is the same. To illustrate this point, for instance, if the ticket price for an 'Economy Class' round trip flight from Addis

Ababa to London ranges between 500 – 1500 US dollars, each member receives the same mileage reward of 3,686 mileage points per each flight segment irrespective of whether he/she paid 500, 1000 or 1,500 US dollars.

Based on the survey results, the next task is the exploration of the available databases by using data mining techniques, in order to determine whether the results could add value to the CRM implementation process.

Chapter 4

Experimentation

4.1 Overview

This section makes up the central and most important part of this research project. Thearling (1999) notes that in order to enable successful CRM, the initial task is to identify market segments containing high profit potential. Accordingly, the main objective of this research was to provide customer segmentation with respect to the important dimensions of customers' needs and value.

This research project incorporates all the typical stages that characterize a data mining process, and especially the CRoss-Industry Standard Process for Data Mining (CRISP-DM) process cycle, which is depicted in Figure 4.1 (SPSS, 2000).

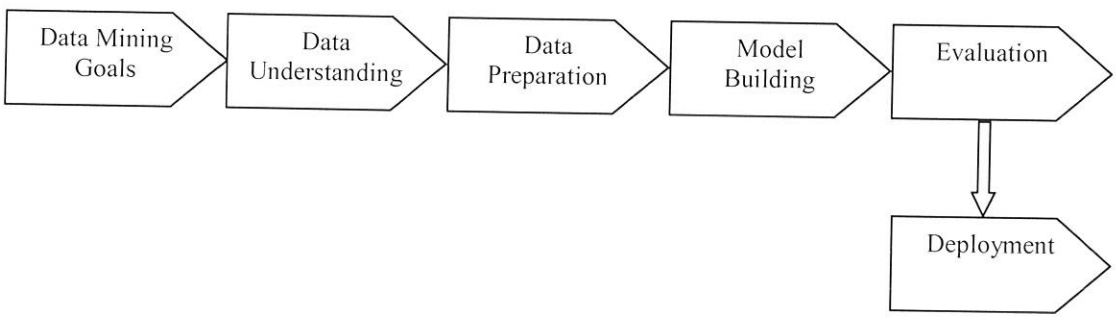


Figure 4.1 Phases of the CRISP-DM process cycle

4.2 Data Mining Goals

The business survey, which was conducted earlier, revealed that current customer value is based on individual mileage, and that mileage is an arbitrary measure of customer profitability. Thus, the first data mining goal was to combine distinct data sources and arrive at a reliable revenue value for each customer, based on individual flight activities.

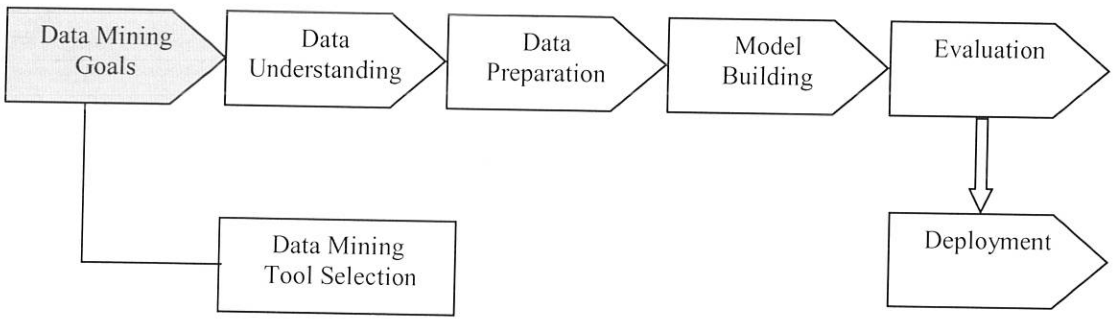


Figure 4.2 The data mining goals setting phase

The variables that determine customer value were to be used to derive customer segments, which in turn would prompt the formation of strategic marketing initiatives. The most appropriate data mining techniques, which are clustering and decision tree classification, were chosen for this purpose.

In order to provide segments that could be explained to domain experts from the CLD, emphasis was to be given to data preparation and an exploratory data analysis. This process allowed the identification of important attributes as input for the model building phase.

The success criteria for this data mining project was the discovery of customer segments with high profit potential. Provided that meaningful segments were to be discovered, the CLD could device special marketing strategies geared towards each of the segments, that would enhance their profitability as well as ensure their loyalty.

4.2.1 Data Mining Tool Selection

Among the factors considered in the selection process for an appropriate data mining tool, the important ones were:

- The data mining tasks that the tool is intended for (clustering and classification)
- The algorithms supported (K-means or SOM, and decision trees)
- Architecture and operating system: the computer architecture on which the software runs (stand alone) and a MS Windows operating system.
- Data sources: possible formats for the data that is to be analyzed (MS Access or MS Excel)
- Size: the maximum number of records the software can comfortably handle (up-to 10,000 records)
- Visualization capabilities

The identification of an appropriate data mining tool was a time taking process. Reviews of data mining tools, by Elder et.al. (1998) and Goebel et.al. (1999), were used as basis for further investigation on the Internet. The researcher finally approached the vendors of two software tools that more or less fulfilled the above factors, namely Knowledge Studio version 3.0 of Angoss Software Corporation (Angoss), and Clementine version 5.0 of SPSS Inc.

An evaluation version of Clementine was not available and the cost of the tool was beyond the budget allocated for this project. The other option was Knowledge Studio, where subsequent contacts were made with Angoss (the vendor) before downloading the software on the Internet. Though it was the sole tool to be evaluated, Knowledge Studio fulfilled most of the criteria set above, and performed well during experimentation.

4.3 Data Understanding

Having defined the data mining goals, the next step was the investigation of which data were available and useful for achieving the goals. Most of the data collected by the CLD are for administrative purposes. Therefore, a method had to be devised to close the gap between the data requirements for this experiment and the existing data situation.

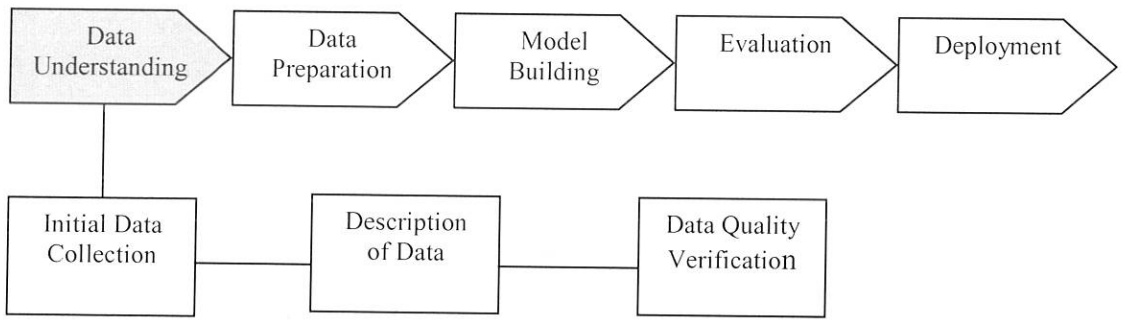


Figure 4.3 The data understanding phase

4.3.1 Initial Data Collection

The primary source of data for this research is the ShebaMiles DB. Demographic data as well as the current state of each member in the program are collected in the database. The database contains a list of single past flight activities, which contain departure and arrival airports,

information pertaining to the booking made, as well as the member's unique membership number.

Information pertaining to each passenger's revenue is not available since the ShebaMiles DB is basically used for administrative purposes. In order to complete the flight activities in the ShebaMiles DB with revenue data, corresponding revenue values were extracted from a revenue accounting database and assigned to the individual flight segments from a revenue accounting database, where revenue information of individual flight segments is available.

Further exploration of the ShebaMiles DB revealed that, a key identification variable, that is ticket number, which is necessary for a unique match to extract sales information from a sales information database concerning the flight activities, was missing.

The airlines' host departure control system creates a data file and electronically transmits the file (usually daily) as input to the ShebaMiles DB. The data file consists of all member activity that occurred during the period, including ticket number information, which is not a mandatory field for the Sheba Miles DB. These data were archived in 306 separate files.

The first task in the data collection process, which is depicted in Figure 4.4, was the extraction of the **member activity load files** from the archived 306 data files. A procedure was used that read through each record and merged the records into a single member activity data file. This exercise was very crucial as the ticket number information was contained in these separate data files. The resulting single data file was exported into MS Access and a table named **Trips Load** was

created. **Trips Load** contained 101,189 records with members' flight activity and ticket number information.

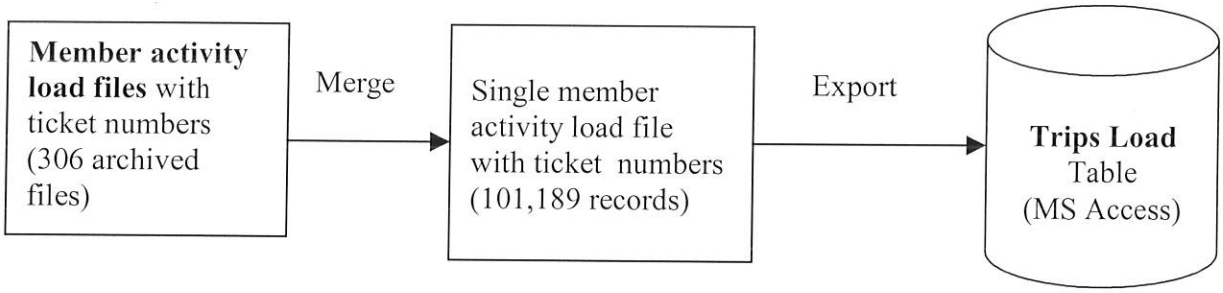


Figure 4.4 Extraction of archived member activity load files

The next task, which is illustrated in Figure 4.5, was the extraction of 90,833 records of member flight activities with the corresponding mileage awards from the ShebaMiles DB into another table named **Trips**. The **Trips** table lacked ticket number information, which was later extracted from **Trips Load** table using a matching query. The number of records in **Trips Load** being greater than **Trips** can be explained by the existence of duplicate records in the former, which were later found and excluded.

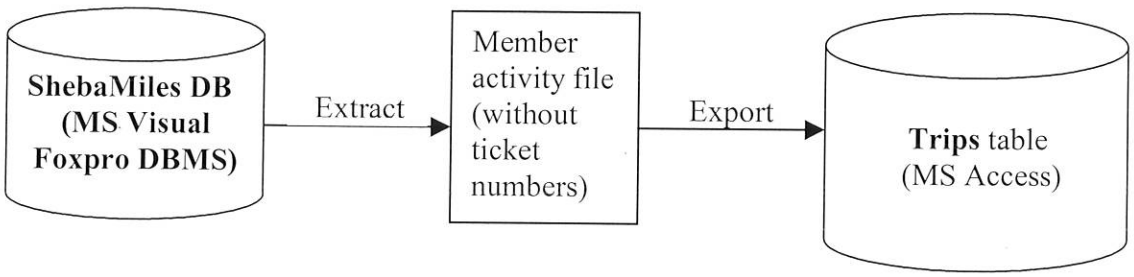


Figure 4.5 Extraction of member activity files from ShebaMiles DB

In order to match the **Trips** table with the **Trips Load** table and extract the ticket number information, records in both tables needed to have the same ShebaMiles permanent member number. Unfortunately, many records existed in **Trips Load** with temporary member numbers. Therefore, another procedure was used to read through a table in ShebaMiles DB (named **Alias**), and retrieve the corresponding permanent number for records that lacked this information in **Trips Load**.

Next, a procedure was used to match records in the **Trips** table with the **Trips Load** table, get the required *ticket number* information, and insert it as a new field in the **Trips** table. Out of the 90,833 records from the **Trips** table, corresponding valid ticket numbers were available in only 24,687 (27%) of the records in **Trips Load**. This is depicted in Figure 4.6.

The remaining records had either missing or invalid numbers. A procedure was used to fill the missing values in the '**data preparation**' phase. Once the key identification variable was available for the 24,687 records, the corresponding revenue values in U.S. dollars had to be extracted, yet from a separate revenue accounting database. Another procedure was used to extract revenue data from a revenue accounting DB.

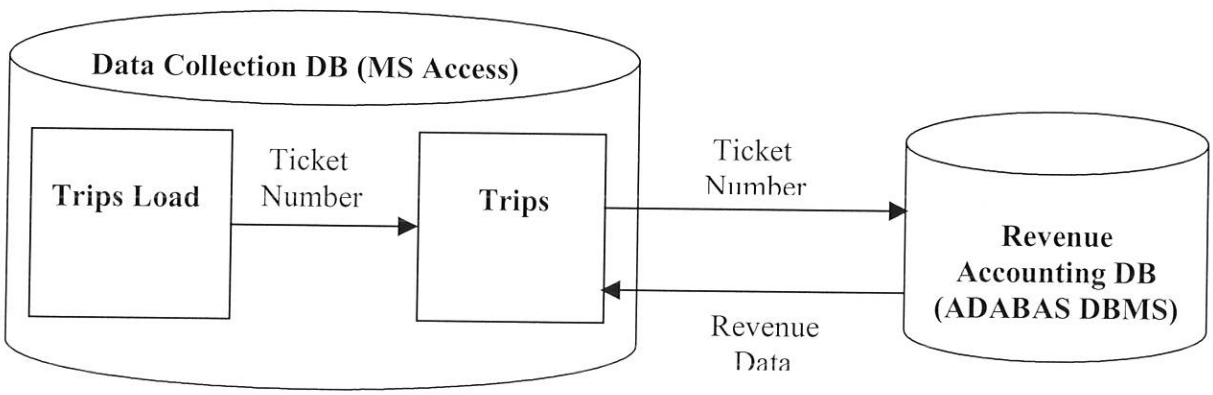


Figure 4.6 The revenue data collection process

The revenue data extraction revealed that out of the 24,687 records, a unique match was found for 20,158 (81%) of the records. A unique match was not found in 4,529 of the records. Records with matching revenue data were later integrated into the **Trips** table.

Furthermore, the **Member** and **Members Point** tables, which contain members' demographic data and their mileage point status respectively, were exported to the same data collection DB as the others.

4.3.2 Description of the Data Collected

After the initial data collection, the new database created on MS Access contained the following tables:

i) Trips

This table contains a total of 90,833 records and 10 fields. A description of the field names as well as their data types is listed as follows:

Field Name	Data Type	Description
FF_Num (Primary Key)	Text	ShebaMiles member number
Orig	Text	Origin city
Dest	Text	Destination city
Flight	Text	Flight number
Date	Date/Time	Date of flight
Class	Text	Reservations (booking) class (First/Business/Economy)
Points	Number	Total points awarded per each flight segment* traveled
Revenue	Number	Revenue in US Dollars per each flight segment traveled

* A trip from one origin city to another destination city makes up one flight segment.

Table 4.1 Attributes of the Trips table

ii) **Member**

This table contains demographic data pertaining to each of Sheba Miles member. This table contains 22,022 records and 15 fields describing the member. A description of the fields names as well as their data types is listed as follows:

Field Name	Data Type	Description
FF_Num (Primary Key)	Text	ShebaMiles member number
Lname	Text	Member's last name
Fname	Text	Member's first name
Address	Text	Member's mailing address
City	Text	Member's city of residence
Country	Text	Member's country of residence
Zip	Text	Member's zip code
Tier	Text	Member's current status in the program (Blue, Silver or Gold tier)
Enrl_date	Date/time	Member's enrollment date in the program
Lang	Text	Member's language of preference
Dob	Date/time	Member's date of birth
Phone	Text	Member's phone number
Email	Text	Member's e-mail address
Smoking	Text	Member's smoking habits (yes/no)
Seating	Text	Member's seating preference (Window or Aisle seat)

Table 4.2 Attributes of the Member table

iii) Members Point

This table contains the number of points posted to each of the 22,022 members and 5 fields that are described as follows:

Field Name	Data Type	Description
FF_Num (Primary Key)	Text	ShebaMiles member number
Exp_date	Date/Time	Points expiration date
Points	Number	Base points
Bon_points	Number	Bonus points
Rdm_points	Number	Redeemed points

Table 4.3 Attributes of the Points table

4.3.3. Data Quality Verification

The revenue data extraction revealed that out of the 90,833 records in the **Trips** table, a unique match was found for 20,158 (22%) of the records. Since it is imperative that a revenue value be assigned to each flight activity in order to provide a reliable customer valuation for the complete individual history, a procedure had to be developed to fill records whose corresponding revenue values were missing. This activity was performed in the *data preparation* phase.

4.4 Data Preparation

The main goal of this activity was the production of the dataset (datasets) used for modeling by the data mining tool of choice. The main activities during this phase included data cleaning, selection, transformation and aggregation, integration, and formatting.

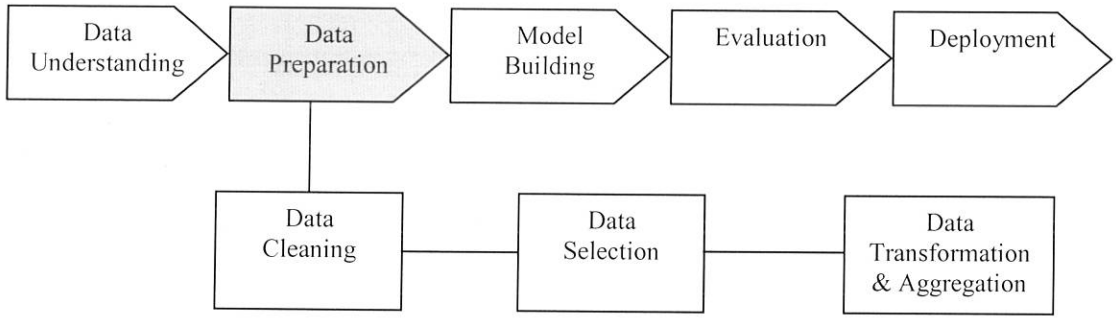


Figure 4.7 The data preparation phase

4.4.1 Data Cleaning

This phase is about raising the data quality to the level required by the project. According to CRISP-DM (2000), this may involve selection of clean subsets of the data, the insertion of suitable defaults or more ambitious techniques such as the estimation of missing data by modeling.

There are various recommendations as to how to compute missing values of key attributes, such as revenue. Saarenvirta (1998) notes that methods range from assigning the average value, to building a classification model and compute the missing values. Revenue values for a given flight segment result from several fares, which are published for this flight segment and are usually constant for a certain period. Pritscher et.al. (n.d.) believe that airline fares are no random values, and it is advisable to make use of observed values within a subgroup.

Subsequently, a procedure was used to calculate mean revenue values for the flight activities, by using historical data of similar flight segments. The results of this exercise were look-up tables by which revenue values were matched to flight activities with unknown revenue. The entry for the look-up table was the average revenue of a certain subpopulation, which was determined by the attributes' origin and destination city pair and booking class. The matching procedure was applied in steps, and at the same time the matching criteria were sequentially relaxed.

Out of 70,675 of the records, the procedure successfully filled revenue values in 53,613 of the records based on the matching criteria, while such a criteria could not be found for the remaining 17,082 records. This exercise significantly raised the number of records with revenue values to 73,751 records from the original 20,158 records. The results were
validated by domain experts as being realistic.

4.4.2 Data Selection

During this phase, selection decisions were made on the data to be used for analysis. The criteria included relevance to the data mining goals as well as quality constraints. The data mining goal being the possible identification of valuable customer groups, records with missing revenue value were excluded in order to avoid compromising the results. Accordingly, this phase saw the selection of 73,751 of the records from the **Trips** table with complete revenue data.

4.4.3 Data Transformation and Aggregation

This task, according to CRISP-DM (2000), includes constructive data preparation operations such as the production of derived attributes, complete new records or transformed values for existing attributes.

The first task performed in this case was the aggregation of the member activity records in the **Trips** table. The aggregation of the records was done by each member's records, thus reducing the total number of records from 73,751 to 11,922. The aggregation of records also necessitated the numeric representation of some attribute values. The new aggregated Trips table had the attributes shown in Table 4.4.

The key attributes from the **Trips** table that were used as inputs to Knowledge Studio's clustering algorithm were the total number of individual segments flown, the total revenue collected, and the total base points he has been awarded.

Field Name	Data Type	Description
FF_Num (Primary Key)	Text	ShebaMiles member number
Ttl_Trips	Number	Total number of segments flown by member
Ttl_Revenue	Number	Total revenue collected from member
Ttl_Points	Number	Total base mileage points awarded

Table 4.4 Attributes of the Trips table aggregated at member level

Data integration was essential at this point as information pertaining to the demographic characteristics of each member was located in a different table, that is the **Member** table.
Records in following fields pertaining to each member were first assigned a numeric code and then integrated into the Trips table.

Field Name	Data Type	Description
Tier	Number	1 = Blue; 2 = Silver; 3 = Gold
Member Tenure	Number	Number of months since member first enrolled in

		ShebaMiles
Country	Number	Member's country of residence (numerically represented from 1 to 132)

Table 4.5 Attributes from the Member table integrated into the Trips table

The demographic variables, which are shown in Figure 4.5, from the **Member** table were used to derive new variables that are shown in Figure 4.6. According to Saarenvirta (1998), data creation involves the creation of new variables by combining existing variables to form ratios, difference and so forth. The new derived attributes are the ones listed in the following table.

Field	Data Type	Description
Rev_Points	Number	Ratio of Total Revenue to Total Mileage Points
Rev_Tenure	Number	Ratio of Total Revenue to Member Tenure
Rev_Trips	Number	Ratio of Total Revenue to Total Number of Segments
Trips_Tenure	Number	Ratio of Total Number of Segments to Member's Tenure

Table 4.6 Derived attributes on the Trips table

Since the data set that Knowledge Studio accepts is a single table, all the above customer specific attributes were aggregated into a single table by making use of MS Access database query utilities. The fact that Knowledge Studio had Open Database Connection (ODBC) facilities enabled the researcher to import the data set directly from the data mart , which was constructed on MS Access. In addition, Knowledge Studio had options for choosing which of the attributes to consider in building a model from the 'single table' data set.

Figure 4.8 shows the data model of the ShebaMiles data mart, which contains the dataset that was input into the modeling tool. The Trips and Member tables were joined to create a data set with records aggregated at a member level.

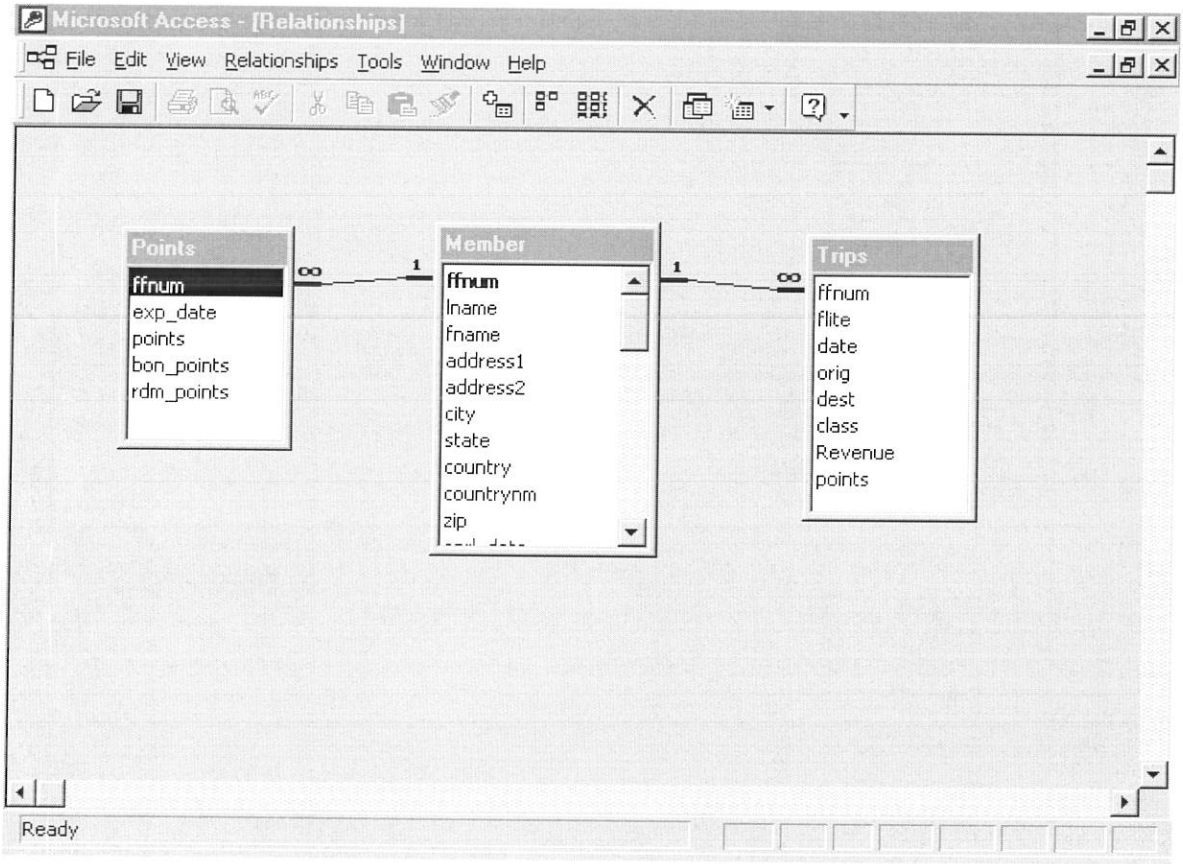


Figure 4.8 The ShebaMiles Data Mart data model

4.5 Modeling

The major tasks performed during the modeling phase were the selection of the modeling technique, laying out a test design, building a model, and the assessment of the model built.

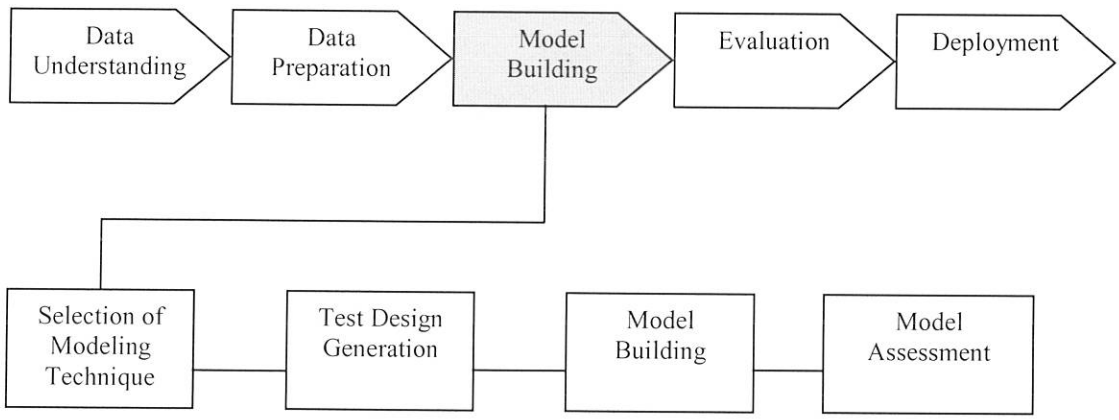


Figure 4.9 The model building phase

4.5.1 Selection of Modeling Technique

The major objective of this project being the generation of strategic customer segments among ShebaMiles members, a clustering algorithm was applied in order to identify groups which were different from each other according to their product mix as well as to their value, but whose members were very similar to each other. Since no predefined segmentation existed, employing clustering algorithms was appropriate.

Knowledge Studio supports two types of Clustering Algorithms; K-means and Expectation Maximization (EM). Unlike K-means, similarity in the Expectation-Maximization (EM) algorithm is based on the probability theory. A record being assigned to a particular cluster when it is most likely generated by the probability distribution corresponding to this cluster, with distributions being different for individual clusters. According to Bishop (1995), EM is best used when one has to deal with large amounts of missing data. Since the number of missing data was very minimal, K-means was used for this study.

The K-means clustering algorithm passes through each customer record, assigning each to the closest existing cluster center. According to Pritscher (n.d.), a critical task of using the K-means clustering algorithm is the choice of the right variables and the right scales.

The K-means algorithm requires that input variables should all be converted into numbers, and that the numbers be further normalized to ensure that the influence of all variables is similar. The task of converting the input variables to numbers has already been performed in the data preparation phase. Furthermore, data is normalized automatically in Knowledge Studio.

Once the clusters were created, the interpretation of the clusters rested with the domain experts and the researcher. The following three approaches were employed to understand the clusters using Knowledge Studio:

1. Visually analyzed how the clusters were affected by changes in the input variables
2. Examined the differences in the distributions of variables from cluster to cluster, one variable at a time

3. Finally, automatically grew a decision tree with 'cluster index' as the dependent variable, and used it to derive rules explaining how to assign new records to the correct cluster.

4.5.2 Test Design

The first activity performed in this phase of the modeling activity was coming up with a test plan for training, testing and evaluating the models. Among the available member activity data, it was decided (with the involvement of the domain experts) to consider those members records whose flight activities occurred during a 12 months period between April 01, 2001 and March 31, 2002 based on two factors.

The basic factor for the above selection criteria was that the period corresponded to ShebaMiles' calendar year, and furthermore, the number of records for this period were more than those in either of the previous years. Accordingly, this sampling criteria yielded 7,602 records out the total 11,922.

Out of the 7,602 records, 4,000 records were randomly selected to train the clustering model, while the remaining 3,602 were set aside to assess how well the model assigned them to the different clusters.

- Total number of segments flown by member (Ttl_Trips)
- Total number of segments flown by member during the 12 months between April 2001 and March 2002 (Ttl_TripsYear)
- Total revenue collected from member (Ttl_Revenue)
- Total base mileage points awarded (Ttl_Points)

- Number of months since member first enrolled in ShebaMiles (Tenure_Months)
- Ratio of Total Revenue to Total Mileage Points (RevPerPoints)
- Ratio of Total Revenue to Member's Tenure (RevPerTenure)
- Ratio of Total Revenue to Total Number of Segments(RevPerTrips)
- Ratio of Total Number of Segments to Member Tenure(TripsPerTenure)

4.5.3 Model Building

Automatic Cluster Detection

After the selection of the data set and K-means clustering algorithm in the previous phases, the next step was the process of choosing the basic run parameters for the algorithm. The basic parameters available in Knowledge Studio for K-means clustering include:

- Number of clusters: the number of clusters (k in K-means) that need to be created. This value has to be manually input into the system.

- Number of iterations: This parameter indicates the maximum number of times the algorithm will read the data. The higher this number and the lower the accuracy criterion, the longer the algorithm will run, and the more accurate the results will be. This parameter is a stopping a criterion for the algorithm. If the algorithm has not satisfied the accuracy criterion after the maximum number of passes, it will stop.

K is a user-defined number. Initially, different values of K, which ranged between 4 and 10, were randomly used. Saarenvirta (1998) advises, the number of clusters chosen should be driven by how many clusters the business can manage. The business experts were then consulted in setting values for K for the various cluster runs in order to make the whole exercise more realistic. Finally, the values chosen for the final experimentation were 4, 5 and 6.

The data overview report from Knowledge Studio was the basis for determining the threshold values (very low, low, medium, high, very high) that were used in the analysis of the results. The report provides a summary of the minimum, maximum, mean and standard deviation values for the different data sets.

The clustering experiments conducted have been broadly divided into four. In each of the experiments, different combinations of variables were used for the cluster runs. Moreover, in each of the experiments, the same variables were used for cluster runs by altering the K value and the maximum number of iterations.

Experiment 1

During the first experiment, all of the variables were input into the clustering run. Since clustering is an unsupervised data mining technique, all the variables were set as independent variables. The parameters set for the cluster runs are shown in Table 4.7.

Cluster run	No. of variables	No. of records	No. of clusters	No. of iteration
1	9	4,000	6	10,000
2	9	4,000	5	10,000

Table 4.7 Summary of clustering input parameters for the first two runs

Knowledge Studio's data overview report, shown in Figure 4.10 was used to select variables to be used for the clustering runs.

KnowledgeSTUDIO - [crm-Training set]

File Edit View Insert Tools Window Help

Data Source: c:\Experimental\Project4\crm-Training set.kdd Records: 4000 Fields: 12

Weight field: -- No Weighting -- Calc. Progress:

#	Field Name	Data Type	Cardinality	# of Missing Values	Minimum	Maximum	Mean	Standard Deviation	Unique Count
1	Ifnum	String	N/A	0	N/A	N/A			N/A
2	Ttl_Revenue	Number	3279	0	42.23	43335.75	2114.07	2973.01	3063
3	Ttl_Points	Number	2187	0	500.0	265474.0	15110.91	21330.13	1826
4	Ttl_Trips	-2,147,483,647 to 2,147,483,647	74	0	1.0	96.0	7.90	10.40	15
5	Ttl_TripsYear	-2,147,483,647 to 2,147,483,647	41	0	1.0	51.0	4.07	4.70	10
6	TierCode	-32,767 to 32,767	3	0	1.0	3.0	1.15	0.43	0
7	Tenure_months	-2,147,483,647 to 2,147,483,647	38	0	1.0	39.0	17.39	11.24	0
8	Country ID	-2,147,483,647 to 2,147,483,647	N/A	0	1.0	132.0	N/A	N/A	N/A
9	RevPerTenure	Number	3705	0	1.46	5304.51	129.31	182.09	3546
10	RevPerTrips	Number	3243	0	29.03	875.18	272.01	113.38	3025
11	TripsPerTenure	Number	548	0	0.03	25.0	0.50	0.84	200
12	RevPerPoints	Number	3241	0	0.04	0.7	0.15	0.06	3023

Calculate Calculate All Define Columns

Overview report | Data Set Chart | Segment Viewer | Cross Tabs

10 significant splits found. CAP

Figure 4.10 Overview report of the training data set

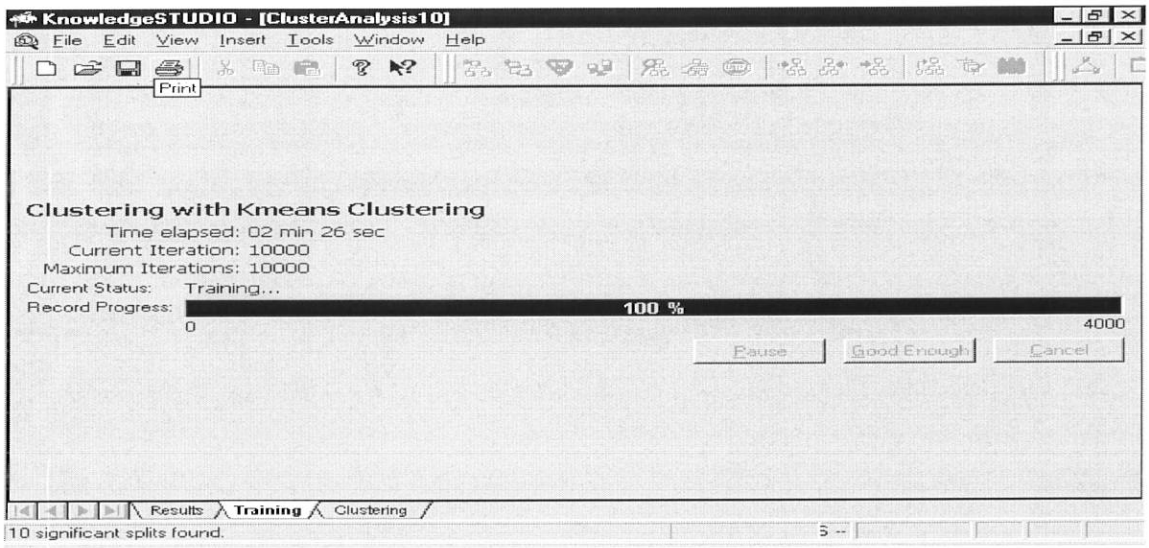


Figure 4.11 The first cluster run

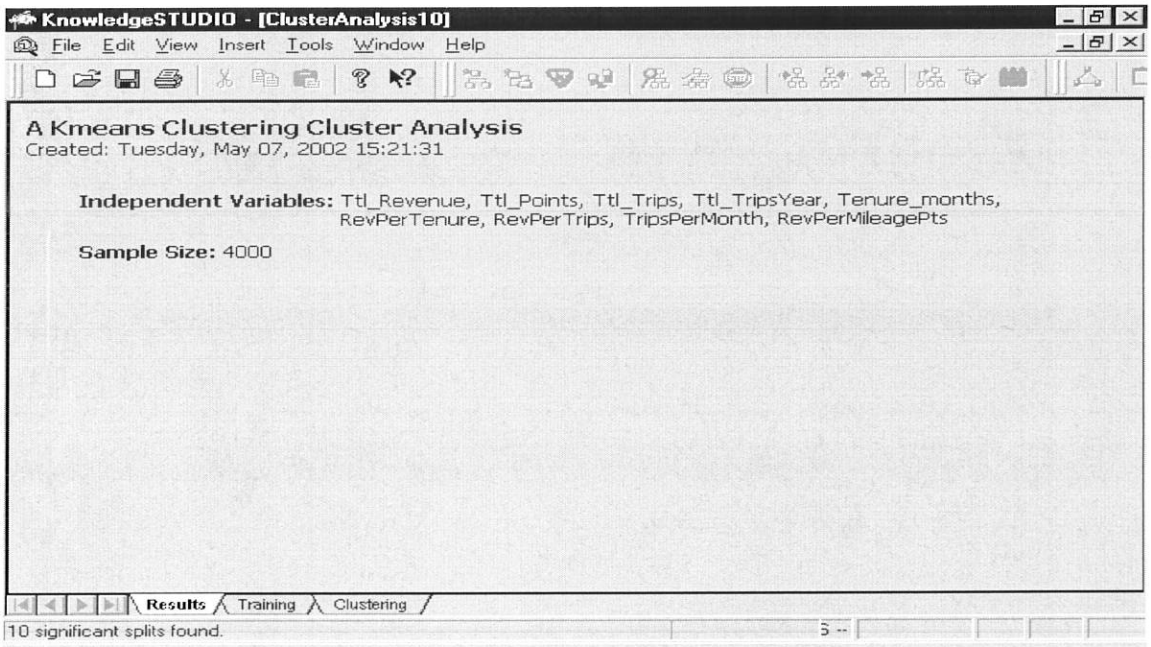


Figure 4.12 Summary of the first cluster run

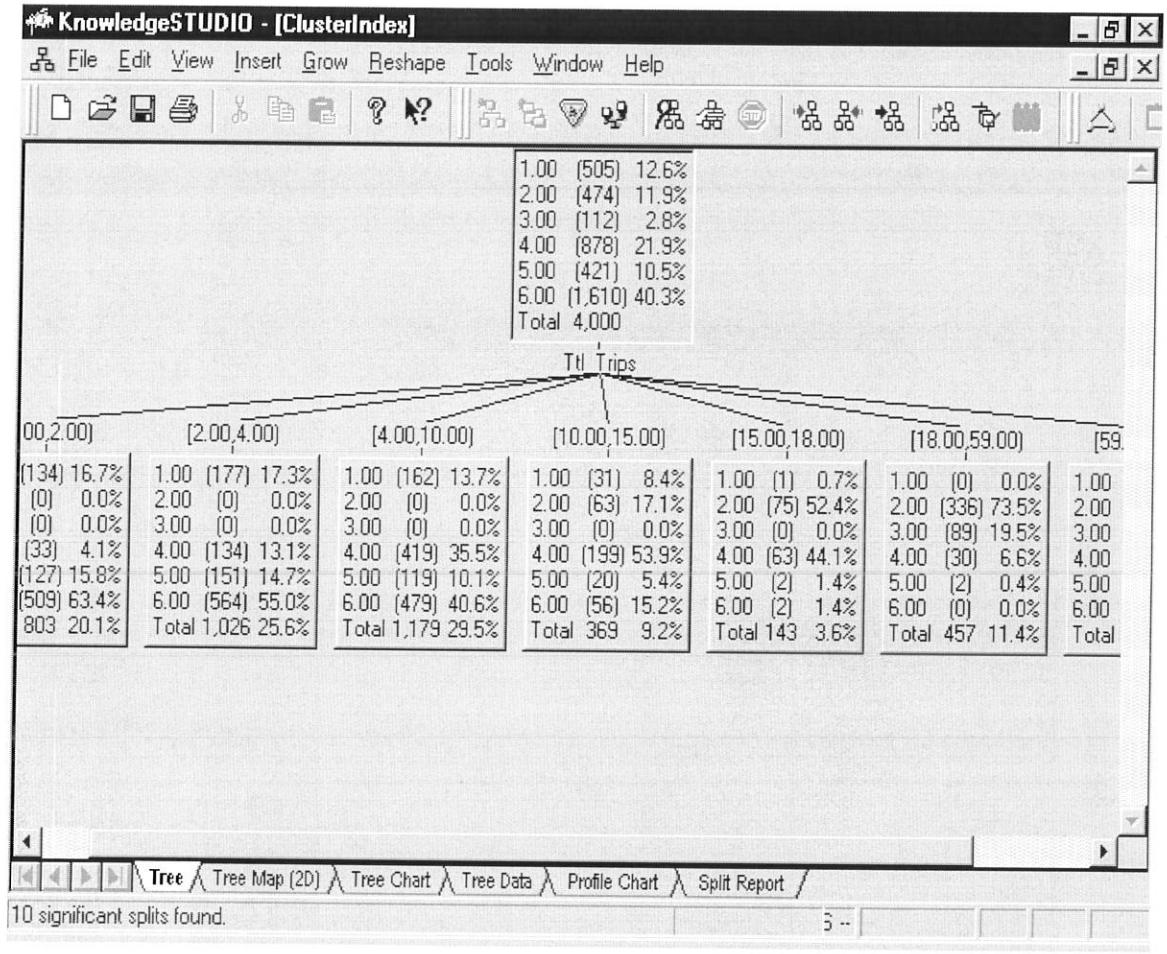


Figure 4.13 Output of the first cluster run

The output of both cluster runs was a decision tree with the Cluster Index as the dependent variable. The decision tree provided a descriptive classification model of the clusters, thus enabling exploration and detection of the characteristics of each cluster.

Analysis of the outputs of the clusters revealed that it was quite difficult to detect patterns identifying the characteristics of each cluster. Furthermore, additional 5 cluster runs with K set at 4, 7, 8,9, 10 did not yield in segments that were meaningful.

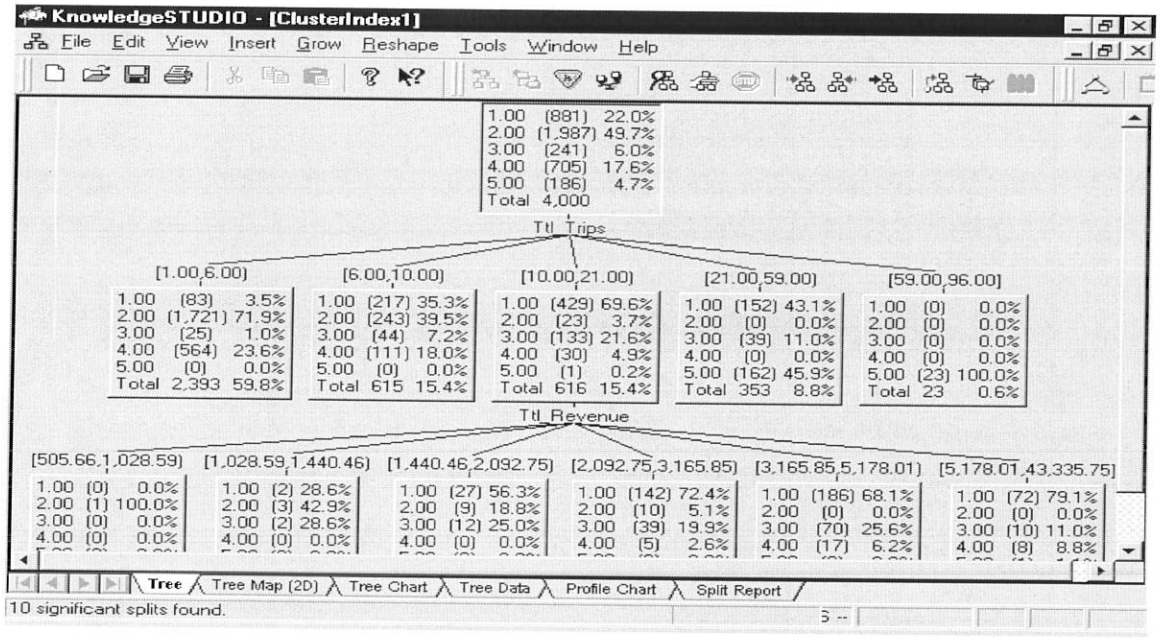


Figure 4.14 Output of the second cluster run

First Cluster Run		Second Cluster Run	
Cluster	Frequency of records	Cluster	Frequency of records
1	505 (12.6%)	1	881 (22.0%)
2	474 (11.9.%)	2	1987 (49.7%)
3	112 (2.8%)	3	241 (6.0%)
4	878 (21.9%)	4	705 (17.6%)
5	421 (10.5%)	5	186 (4.7%)
6	1610 (40.3%)		

Table 4.8 Summary of cluster results

The descriptive decision tree with the cluster index as the dependent variable indicated that among the nine variables, Ttl_Trips, Ttl_TripsYear, Ttl_Revenue and Ttl_Points were the top four variables used for classifying the clusters. These variables were chosen for the cluster runs to be conducted in the next experiment.

Experiment 2

Based on results from the previous experiment, the variables chosen for the cluster run during this phase were:

- Total number of segments flown by member (Ttl_Trips)
- Total number of segments flown by member in the last 12 months (Ttl_TripsYear)
- Total revenue collected from member (Ttl_Revenue)
- Total base mileage points awarded (Ttl_Points)

For the third cluster run, the number of clusters (k) was set at 5, and the maximum number of iterations at 1000.

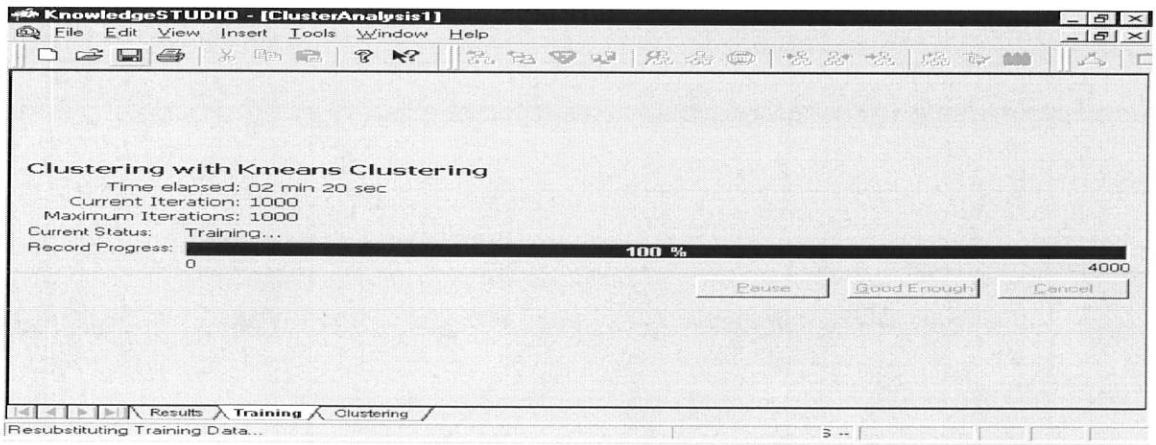


Figure 4.15 Third cluster run

The resultant model was assessed with the domain experts to decipher what the different clusters referred to, so that a segmentation metric could be developed from the member's behavior based on the input variables. Since there were no predefined customer segments that have been created to compare the results with, the researcher had to rely on the expertise of the domain experts to find patterns of customer behavior in the clusters.

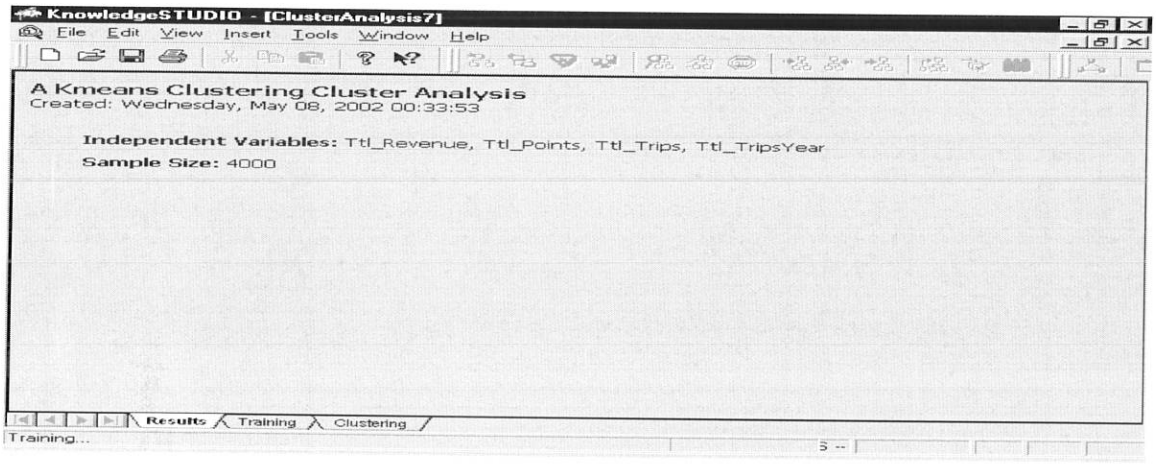


Figure 4.16 Summary of the third cluster run

The clusters were analyzed by visualizing the input variables and splits one at a time. Consequently, the following pattern was discovered in the five cluster groups with respect to each of the input variables. The data overview report from Knowledge Studio was used to set the threshold values to classify the records as follows:

1. Ttl_Trips: Very Frequent; Moderately Frequent; Frequent; Not Frequent
2. Ttl_TripsYear: Very Frequent; Moderately Frequent; Frequent; Not Frequent
3. Ttl_Revenue: High Revenue; Moderately High Revenue; Average Revenue; Low Revenue
4. Ttl_Points: High Points; Average Points; Low Points

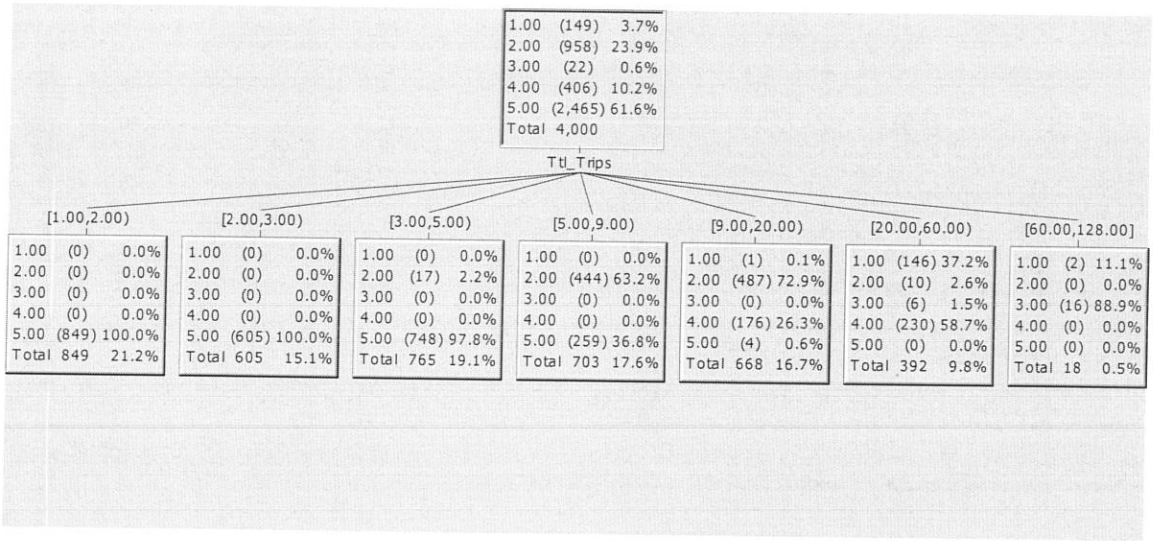


Figure 4.17 Output of the third cluster run

Cluster	Frequency of Records	Ttl_Trips (Frequency)	Ttl_Trips Year	Ttl_Revenue	Ttl_Points
1	149 (3.7%)	Very High	Very High	Very High	Very High
2	958 (23.9%)	Medium	Medium	Medium	Medium
3	22 (0.6%)	Very High	Very High	Very High	Very High
4	406 (10.2%)	High	High	High	Very High
5	2465 (61.6%)	Low	Low	Low	Low

Table 4.9 Summary of results from the third cluster run

Further analysis of the clusters revealed that members who belonged to clusters 1 and 3 made very frequent trips, generated very high revenue and had accumulated the highest total points awarded. Those belonging to cluster 2 were moderate members with medium values with respect to the variables under study. Cluster 4 was an interesting group with very high total points and the rest high. Cluster 5 consisted of members with all the variables low.

A second cluster run was conducted to check the effect of the number of iterations on the clustering outcome. The same data set was input into the algorithm, where the number of clusters was set at 5 and the maximum number of iteration to 5000.

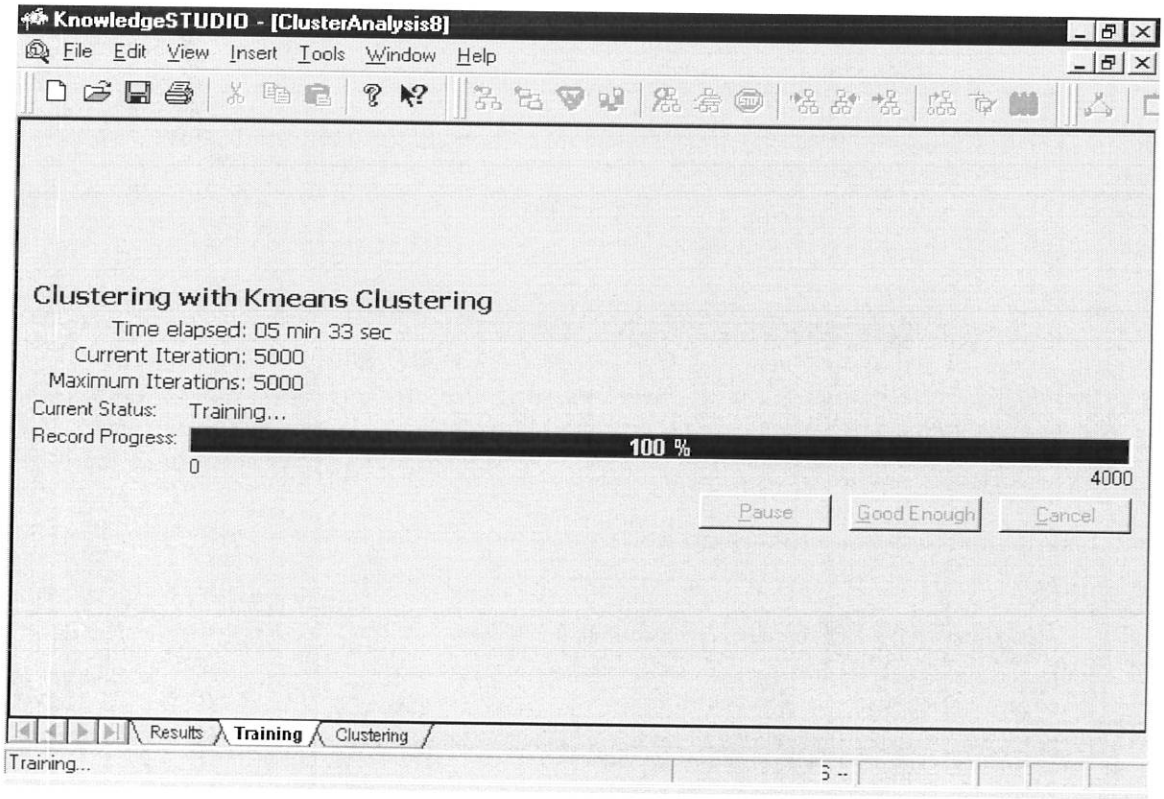


Figure 4.18 Fourth cluster run

The model took twice the time it took to train the first model. The clustering result shown in Figure 4.19 revealed that the five clusters from the second run exhibit similar characteristics to the cluster outputs of the first run. A summary of the second clustering results is shown in Table 4.10.

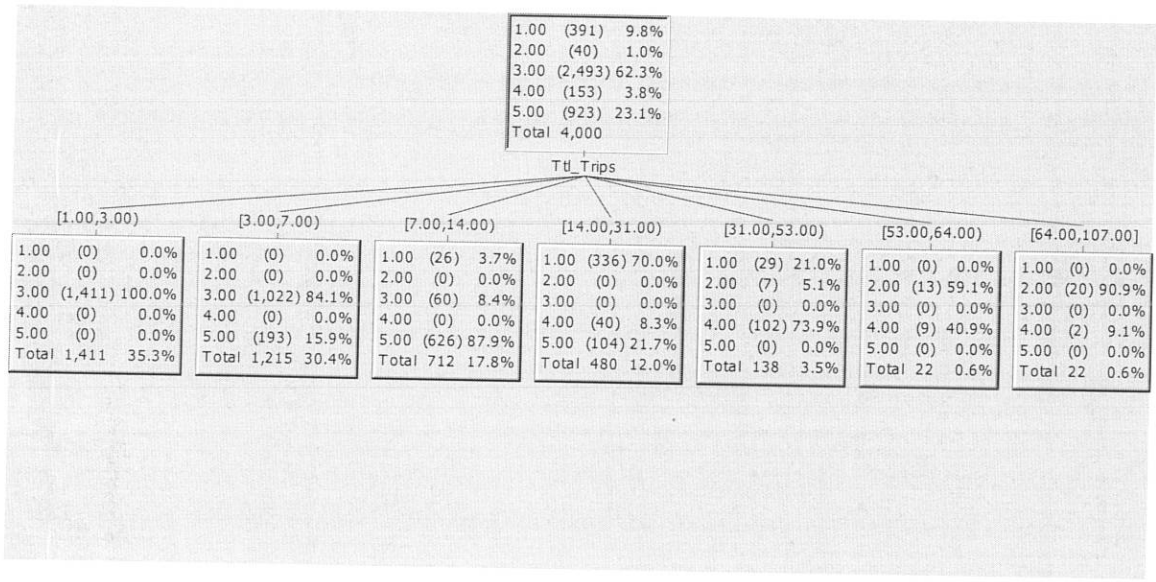


Figure 4.19 Output of the fourth cluster run

Cluster	Frequency of Records	Ttl_Trips (Frequency)	Ttl_Trips Year	Ttl_Revenue	Ttl_Points
1	391 (9.8%)	High	High	High	Very High
2	40 (1.0%)	Very High	Very High	Very High	Very High
3	2493 (62.3%)	Low	Low	Low	Low
4	153 (3.8%)	Very High	Very High	Very High	Very High
5	923 (23.0%)	Medium	Medium	Medium	Medium

Table 4.10 Summary of results from the fourth cluster run

As could be seen from tables 4.9 and 4.10, though the results do not exactly match, similar patterns were identified in the results of both the clustering runs. The minimal difference in the number of records belonging to clusters from the two models exhibiting similar behaviors could be attributed to differences in the maximum number of iteration set for the two runs.

The cluster groups showing similar characteristics are shown on the same row in Table 4.11. The comparison elaborates the similarity of the results from the two runs.

Cluster	Frequency of Records	Cluster	Frequency of Records
1	149 (3.7%)	4	153 (3.8%)
2	958 (23.9%)	5	923 (23.0%)
3	22 (0.6%)	2	40 (1.0%)
4	406 (10.2%)	1	391 (9.8%)
5	2465 (61.6%)	3	2493 (62.3%)

Table 4.11 Comparison of results from the third and fourth cluster runs

Since Clusters 1 and 3 had the same 'very high' values, and assuming that reducing the cluster numbers would produce better segments, a second run was conducted with the number of clusters set to 4. The same training data set, which was used in the first experiment, was again used to train the third clustering model. The maximum number of iteration was set to 1,000. The results are shown in Figure 4.20.

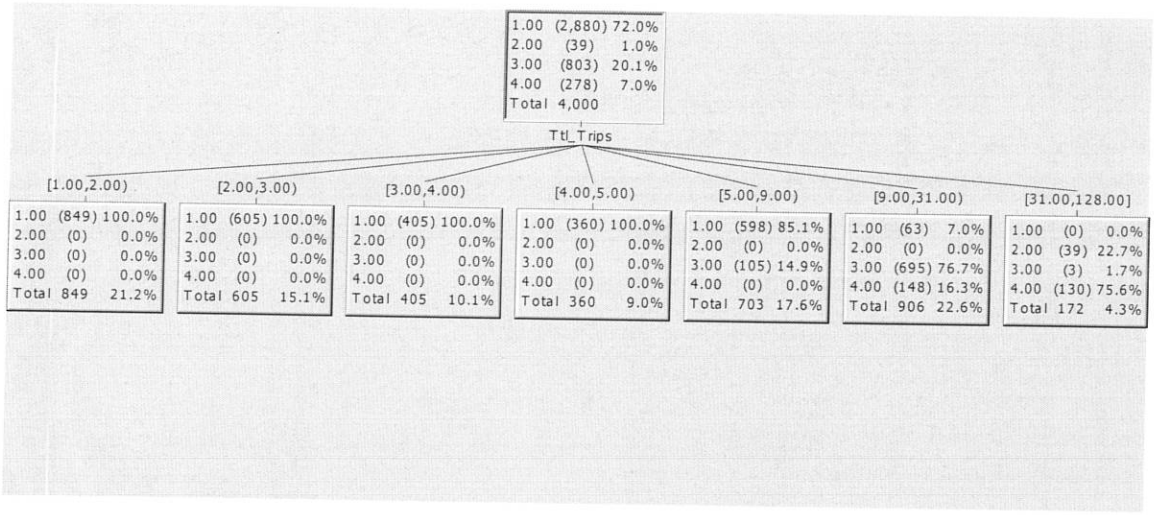


Figure 4.20 Output of the fifth cluster run.

Again, the resulting clusters were analyzed making use of Knowledge Studio's decision tree. The four cluster groupings in the third run revealed a slightly different pattern than the first run. Since the total number of clusters in this case is less by one than in the first run, no two clusters represented very similar pattern. The output and interpretation of the model is summarized in Table 4.12.

Cluster	Frequency of Records	Ttl_Trips (Frequency)	Ttl_TripsYear	Ttl_Revenue	Ttl_Points
1	2880 (72.0%)	Low	Low	Low	Low
2	39 (1.0%)	Very High	Very High	Very High	Very High
3	803 (20.1%)	Medium	High	High	High
4	278 (7.0%)	Very High	High	Very High	Very High

Table 4.12 Summary of results from the fifth cluster run

Once again, another clustering run was conducted to check the effect of the number of iterations on the third clustering outcome. The same data set was input into the algorithm, where the number of clusters was set at 4 and the maximum number of iteration to 5000.

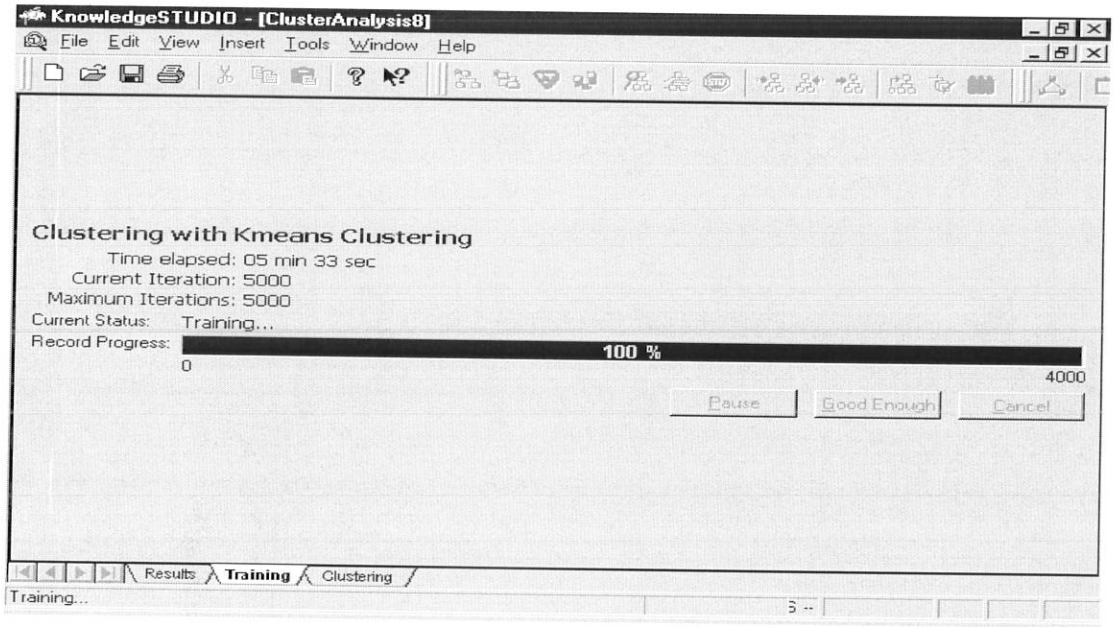


Figure 4.21 Sixth cluster run

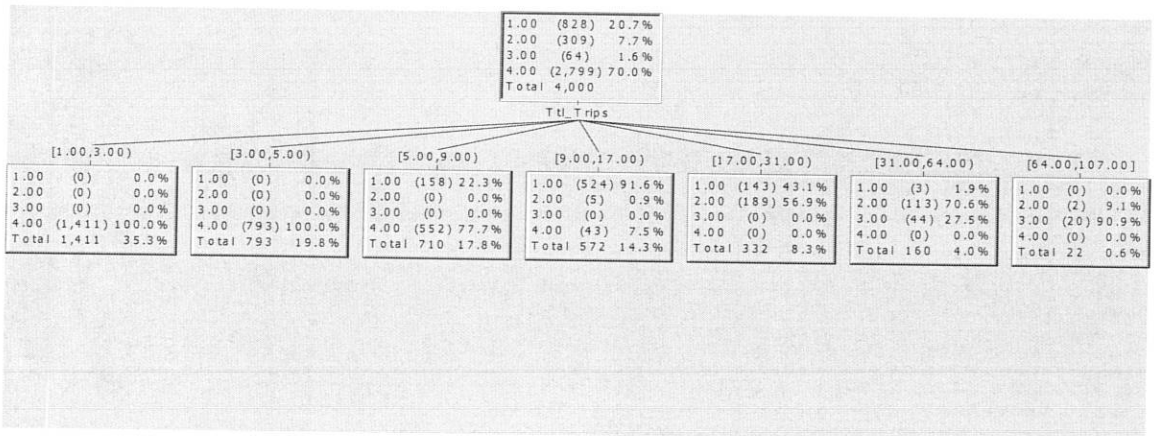


Figure 4.22 Output of sixth cluster run

Again, the resulting clusters were analyzed making use of Knowledge Studio's decision tree. The four cluster groupings in the fourth run revealed a slightly different pattern than the third run. Since the total number of clusters in this case is less by one than in the first run, no two clusters represented exactly similar patterns with respect to the input variables. The results of the model is summarized in Table 4.13.

Cluster	Frequency of Records	Ttl_Trips (Frequency)	Ttl_TripsYear	Ttl_Revenue	Ttl_Points
1	828 (20.7%)	Medium	High	High	High
2	309 (7.7%)	Very High	High	Very High	Very High
3	64 (1.6%)	Very High	Very High	Very High	Very High
4	2799 (70.0%)	Low	Low	Low	Low

Table 4.13 Summary of results from the sixth cluster run

Similar to earlier cluster outputs, though the cluster numbers of the two results from Tables 4.12 and 4.13 do not exactly match, similar patterns have been identified from the data. This minimal difference between the two models was once again attributed to the maximum number of iteration.

Cluster	Frequency of Records	Cluster	Frequency of Records
1	2880 (72.0%)	4	2799 (70.0%)
2	39 (1.0%)	3	64 (1.6%)
3	803 (20.1%)	1	828 (20.7%)
4	278 (7.0%)	2	309 (7.7%)

Table 4.14 Comparison of results from the fifth and sixth cluster runs

Similar to Experiment 1, the cluster groups showing similar characteristics are shown on the same row in Table 4.14. The comparison elaborates the similarity of the results from the two cluster runs.

According to Berry et.al. (2000), the best set of clusters may be simply the ones that show some expected pattern in the data. The results obtained from the above cluster models were encouraging to continue the experimentation adding more of the input variables as well as making use of differing parameter settings.

Experiment 3

In this phase of the experimentation, one more variable, Tenure_Months, was added in addition to the four variables that were used in Experiment 2. The intention was to see the effect of an additional variable on the distribution of the cluster outputs. The parameters used for the first run during this phase were:

- Number of clusters = 4
- Maximum number of iteration = 10,000

The process of clustering the data set, the analysis, and the cluster outputs are shown in figures 4.23, 4.24 and 4.25.

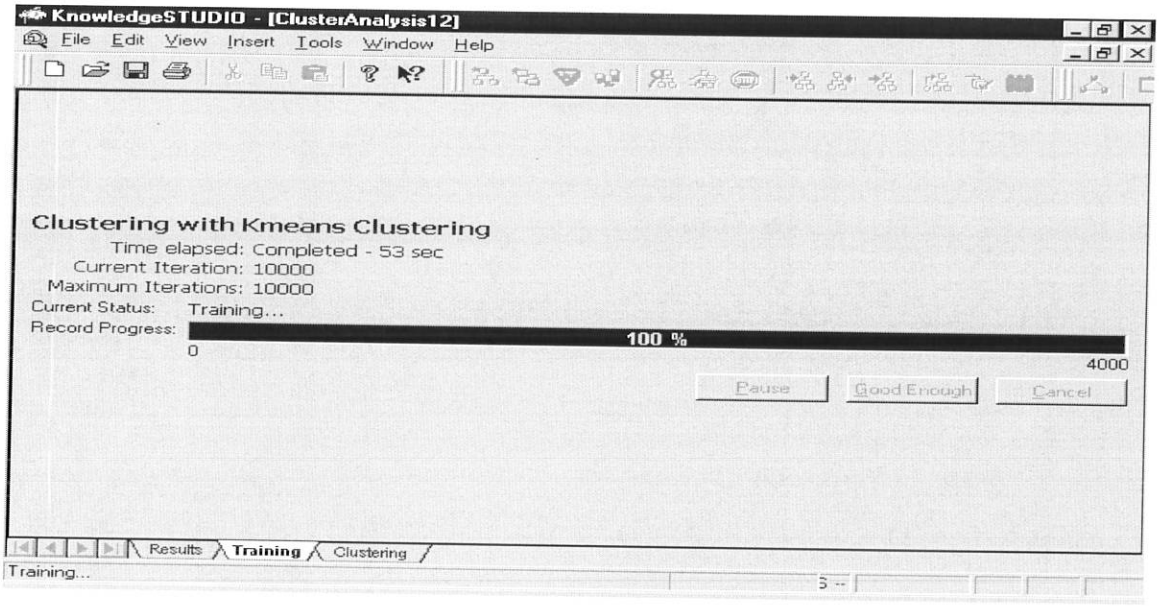


Figure 4.23 Seventh cluster run

The introduction of the additional variable brought about a change in the distribution of the records among the 4 clusters. According to their tenure in the program, member records were identified as long, moderate (average), and recent. The clustering output has been summarized in Table 4.15.

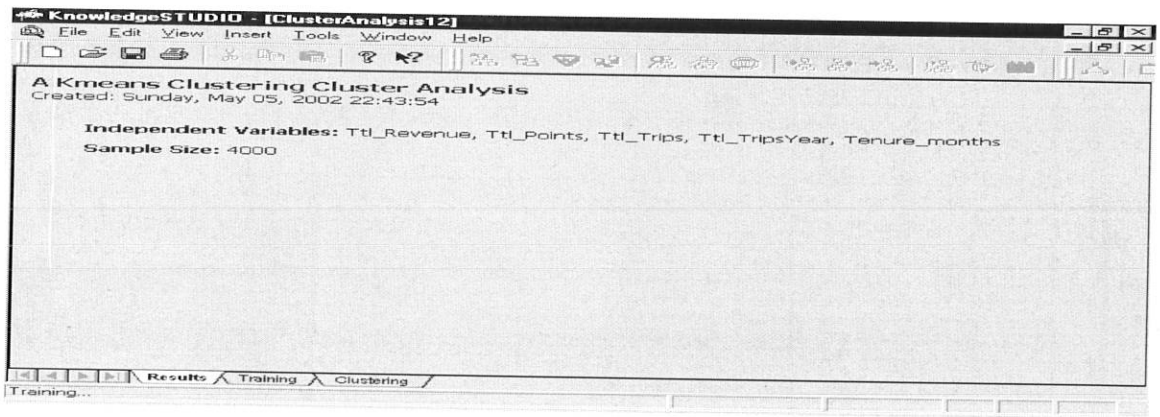


Figure 4.24 Summary of the seventh cluster run

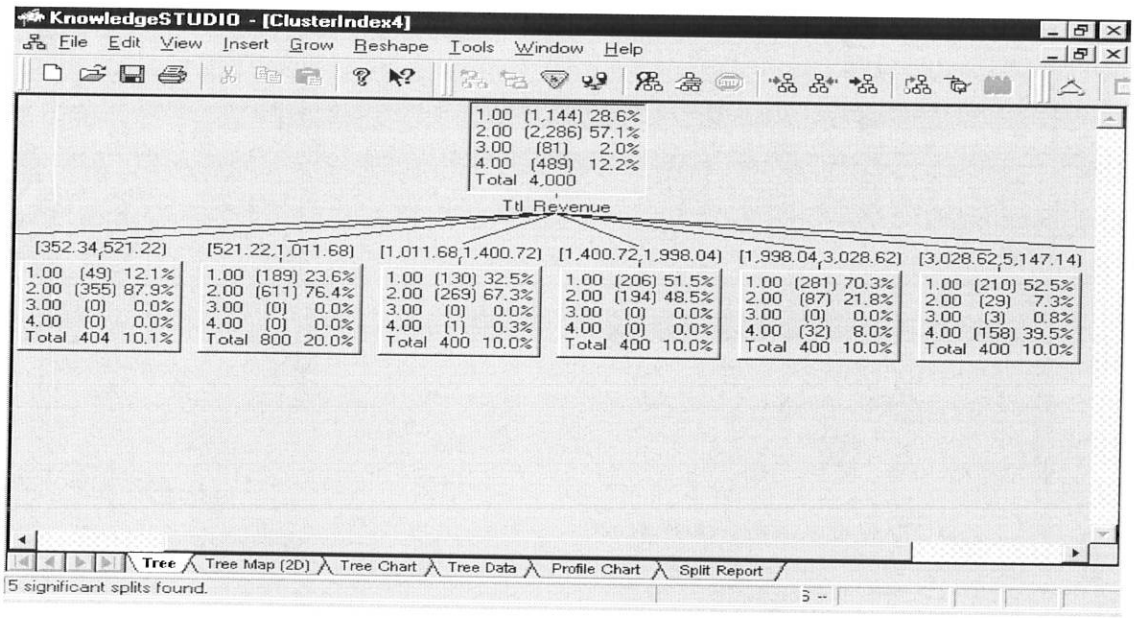


Figure 4.25 Output of the seventh cluster run

Cluster	Frequency of Records	Ttl_Trips	Ttl_TripsYear	Ttl_Rev	Ttl_Pts	Tenure
1	1144 (28.6%)	Low	Low	Low	Low	Long
2	2286 (57.1%)	Low	Low	Low	Low	Recent
3	81 (2.0%)	High	High	High	High	Medium
4	489 (12.2%)	Medium	Medium	Medium	Medium	Medium

Table 4.15 Summary of results from the seventh cluster run

The above summary of results show that the patterns discovered in the four clusters were different from the ones discovered during earlier experiments, in that Ttl_Revenue was the most determining variable than the others. Furthermore, the patterns discovered from this cluster grouping were more meaningful. This result established the fact that the inclusion of the variable Tenure_Months enhanced the robustness of the clustering model. Another cluster run was conducted with the number of clusters set to 3, keeping the variables and the maximum number of iterations the same.

Cluster	Frequency of Records	Ttl_Trips	Ttl_Trips Year	Ttl_Rev	Ttl_Pts	Tenure_Months
1	299 (7.5%)	High	High	High	High	Medium
2	2386 (59.2%)	Low	Low	Low	Low	Recent
3	1333 (33.3%)	Medium	Medium	Medium	Medium	Long

Table 4.16 Summary of results from the eighth cluster run

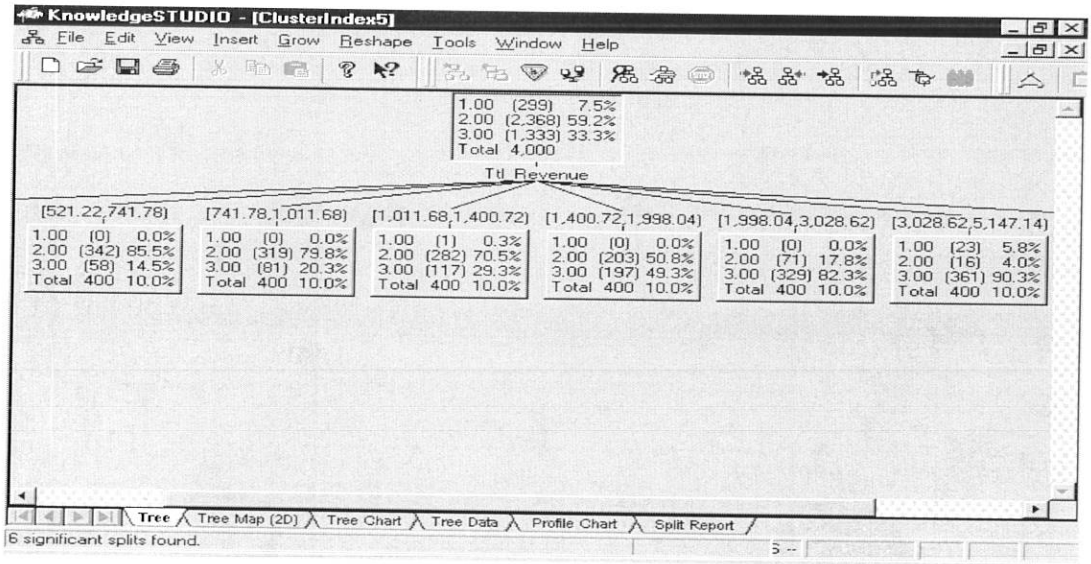


Figure 4.26 Output of the eighth cluster run

Results from the seventh cluster run give consistent results in that Ttl_Revenue was used as the first splitting variable, indicating that total revenue is an important variable that could measure a customers' value.

Following these results, the domain experts were keen on running more experiments in order to arrive at clusters that were more descriptive. In the next experiment, the variable Ttl_Points was not considered as results from the experiment indicated that it was not important in terms of determining customers' value. This was also shared by the domain experts, based on their business knowledge. Furthermore, Tenure_Months was replaced by RevPerTenure and TripsPerTenure, the latter of which were believed to be more descriptive

Experiment 4

Based on results from the previous experiment and domain expert's comments, six variables were used as input to the clustering run during this phase of the experimentation.

- Total number of segments flown by member (Ttl_Trips) (Var 1)
- Total number of segments flown by member during the 12 months between April 2001 and March 2002 (Ttl_TripsYear) (Var 2)
- Total revenue collected from member (Ttl_Revenue) (Var 3)
- Number of months since member first enrolled in ShebaMiles (Tenure_Months)
- (Var 4)
- Ratio of Total Revenue to Member's Tenure (RevPerTenure) (Var 5)
- Ratio of Total Number of Segments to Member Tenure (TripsPerTenure) (Var 6)

Cluster run	No. of variables	No. of records	No. of clusters	No. of iteration
1	6	4,000	6	10,000
2	6	4,000	5	10,000
3	6	4,000	4	10,000

Table 4.17 Summary input parameters for the cluster runs in Experiment 4

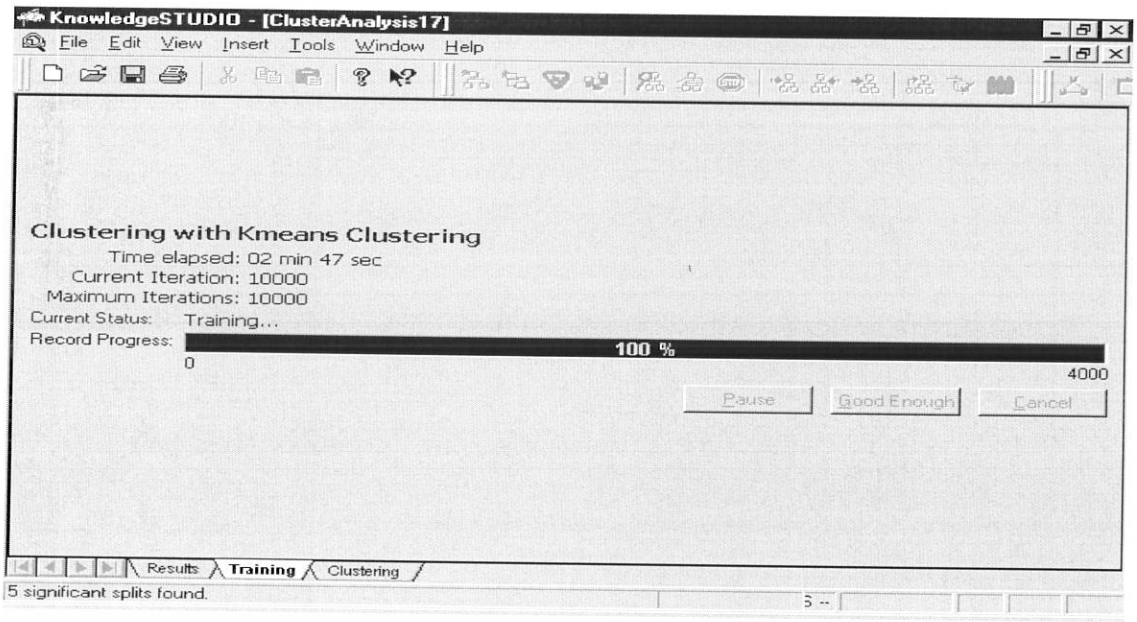


Figure 4.27 ninth, tenth and eleventh cluster runs

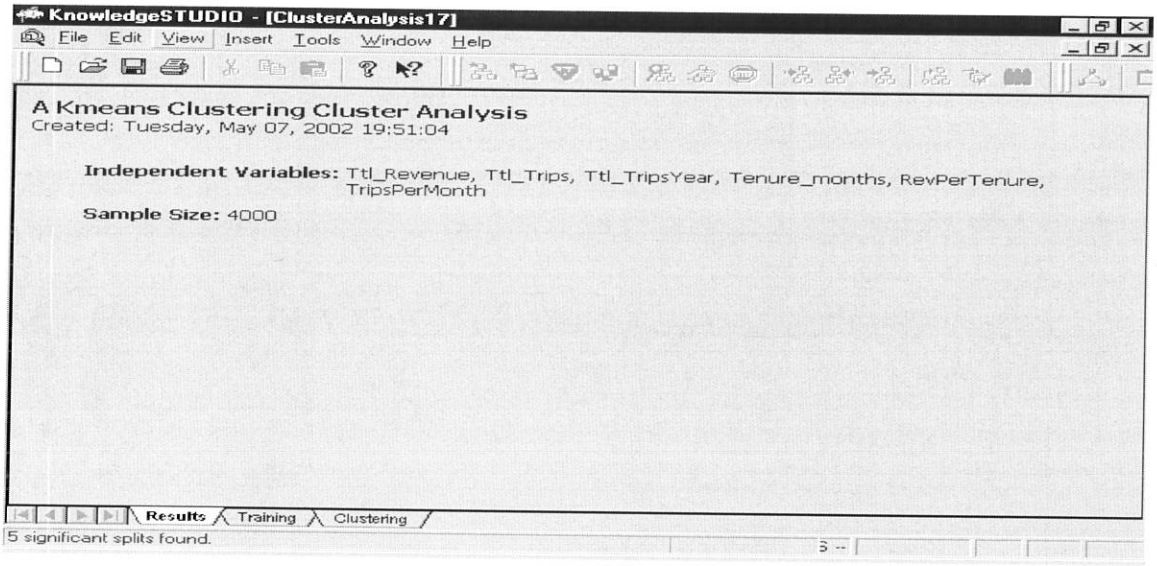


Figure 4.28 Summary of the ninth, tenth and eleventh cluster runs

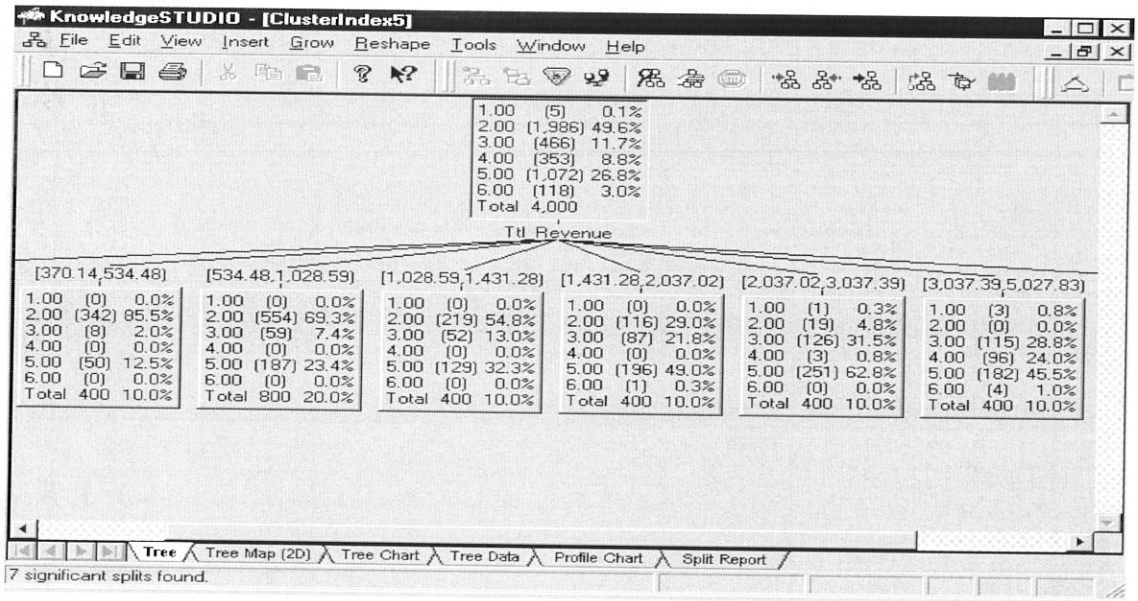


Figure 4.29 Output of the ninth cluster run

Results from the ninth cluster run showed that Ttl_Revenue was the initial splitting variable, thus indicating that it is a determining variable. The first Cluster form this cluster run contained only five records (0.1%) of the total, thus indicating that further experimentation by reducing K was valid.

Additional runs were conducted by setting K to 5 and 4 respectively. The outputs, which are shown in Figures 4.30 and 4.31, again show that Ttl_Revenue was the first splitting variable. This pattern seemed to suggest that total revenue contribution was an important variable in the segmentation of customers.

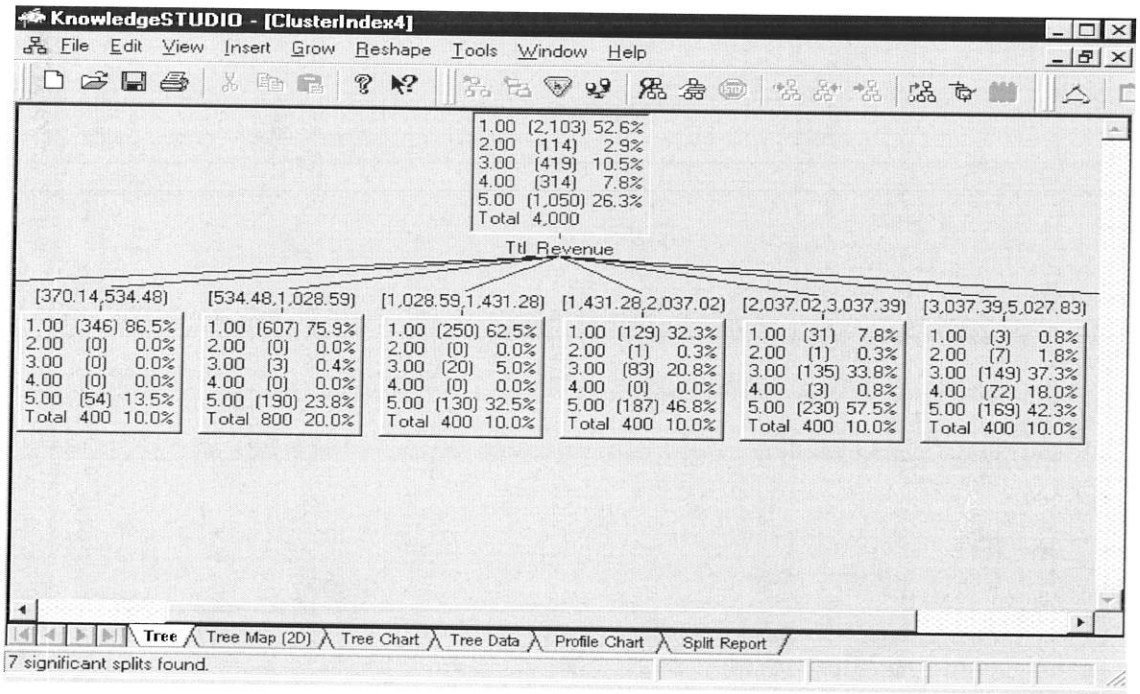


Figure 4.30 Output of the tenth cluster run

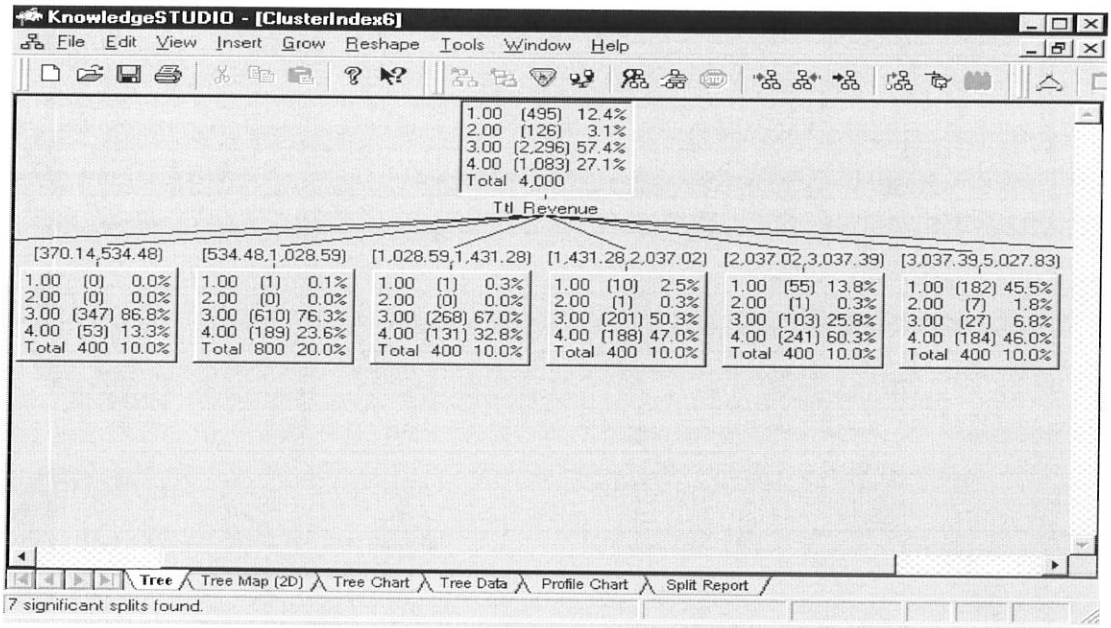


Figure 4.31 Output of the eleventh cluster run

Metrics were developed to describe the characteristics of each variable, based on the distribution of records in the segments. Representations and the corresponding meaning of the metrics used were the following:

VH = Very High; H = High; M = Medium; Lw = Low; VLw = Very Low; R = Recent; Lg = Long; VLg = Very Long.

Cluster	Frequency of Records	Var 1	Var 2	Var 3	Var 4	Var 5	Var 6
1	2103 (52.6%)	VLw	VLw	VLw	R	VLw	M
2	114 (2.9%)	VH	VH	VH	Lg	VH	H
3	419 (10.5%)	M	H	H	M	H	H
4	314 (7.8%)	H	H	VH	VLg	H	H
5	1050 (26.3%)	Lw	Lw	M	Lg	VLw	Lw

Table 4.18 Summary of the tenth cluster run, with k = 5

Cluster	Frequency of Records	Var 1	Var 2	Var 3	Var 4	Var 5	Var 6
1	495 (12.4%)	H	H	H	M	H	H
2	126 (3.1%)	VH	VH	VH	Lg	H	H
3	2296 (57.4%)	Lw	Lw	Lw	R	M	M
4	1083 (27.1%)	M	M	M	Lg	Lw	Lw

Table 4.19 Summary of the eleventh cluster run, with k = 4

4.5.4 Summary of the Cluster Results

An important requirement of segmentation is that the derived segments can be explained to domain experts. The discovery of patterns requires that there is close interaction with domain experts, which allows them to interact with the output. The evaluation of the final clustering results heavily depended on the domain experts.

In order to confirm the discovery of meaningful segments, plausibility checks were performed based on additional important variables. Several runs of the K-means algorithm were conducted with number of clusters set between 4 and 6. The entire data set output from the different cluster runs was available with the cluster index appended to the end of each record. This enabled the domain experts to compare resulting customer segments from the different cluster runs.

Except the cluster outputs in Experiment 1, where the results were difficult to decipher, the results of the cluster runs in Experiment 2, 3 and 4 revealed important knowledge as to the characteristics of each discovered segment. According to domain experts, the results of the

models were relevant to the business problem, since each segment differed between high and low potential customer groups. Pritscher (n.d.) tell us that there is no actual quantitative definition of a good segmentation, thus assessing the groups by investigating their revenue distribution (customer value) is worthwhile.

The clustering results from Experiment 4 were more meaningful, where the input variables were: Ttl_Trips, Ttl_TripsYear; Ttl_Revenue; Tenure_Months; RevPerTenure
; TripsPerTenure. The model with 5 clusters, created from the 4,000 records' training data set, led to customer segments with distinct travel behavior. In addition to the meaningful patterns identified in these 5 segments by domain experts, the resulting decision tree found Ttl_Revenue to be the best splitting variable. The following is a brief interpretation, with domain experts, of the fifth cluster run results with K = 5 from Experiment 4. Furthermore, the results are summarized in Table 4.20.

1. Cluster 1 consists of customers with very low frequency of trips traveled, very low revenue generating and quite recent in their tenure as members. The majority of the records (52.6%) belong to this segment.

2. Cluster 2 contains very highly frequent, very high revenue generating as well as long tenured customers. This segment makes up the smallest number of records (2.9%) out of the total. This cluster represents a segment with high tiered and a valuable group of customers.
3. Cluster 3 contains medium frequency, high revenue and medium tenured customers. The total revenue generated per the members' tenure as well as the number of trips per the members' are both high for this group. This segment makes up 10.5% of the total records.

4. Cluster 4, making up 7.8% of the total contains high frequency, very high revenue generating and very long tenured customers. This is the other cluster where another top tier segment was identified.

5. Cluster 5 comprises of low frequency, medium revenue generating, and long tenured members. Making up 26.3% of the total records, this group's activity is quite insignificant compared to its long tenure in the program. On the other hand, the revenue contribution from the few trips made by this group has been significant.

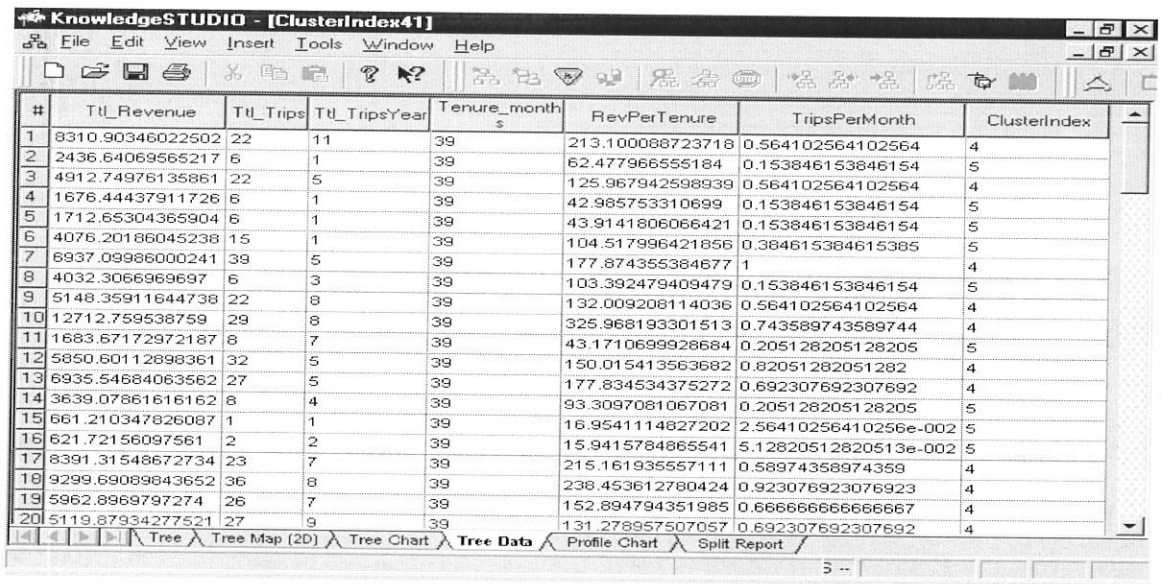
Cluster	Description	Percentage
2	Very high revenue, very frequent trips, long tenure	2.9%
4	Very high revenue, frequent trips, very long tenure	7.8%
3	High revenue, medium trips, medium tenure	10.5%
5	Medium revenue, few trips, long tenure	26.3%
1	Very low revenue, very few trips, recent tenure	52.6%

Table 4.20 Summary of identified clusters based on travel behavior

The result of the clustering model was further evaluated by applying it to a data set comprising of 3,602 new records. The model clustered the new records into homogenous groups, the results of which are shown in Table 4.21.

Training Cluster	Frequency of Records (4,000 records)	New Cluster	Frequency of Records (3,602 records)
1	2103 (52.6%)	1	1890 (52.5%)
2	114 (2.9%)	2	97 (2.7%)
3	419 (10.5%)	3	339 (9.4%)
4	314 (7.8%)	4	302 (8.4%)
5	1050 (26.3%)	5	974 (27.0%)

Table 4.21 Cluster comparison of the training and newly clustered data sets



#	Ttl_Revenue	Ttl_Trips	Ttl_TripsYear	Tenure_months	RevPerTenure	TripsPerMonth	ClusterIndex
1	8310.90346022502	22	11	39	213.100088723718	0.564102564102564	4
2	2436.64069565217	6	1	39	62.477966555184	0.153846153846154	5
3	4912.74976135861	22	5	39	125.967942598939	0.564102564102564	4
4	1676.44437911726	6	1	39	42.985753310699	0.153846153846154	5
5	1712.65304365904	6	1	39	43.9141806066421	0.153846153846154	5
6	4076.20186045238	15	1	39	104.517996421856	0.384615384615385	5
7	6937.09986000241	39	5	39	177.874355384677	1	4
8	4032.3066969697	6	3	39	103.392479409479	0.153846153846154	5
9	5148.35911644738	22	8	39	132.009208114036	0.564102564102564	4
10	12712.759538759	29	8	39	325.968193301513	0.743589743589744	4
11	1683.67172972187	8	7	39	43.1710699928684	0.20512820512820 5	5
12	5850.60112898361	32	5	39	150.015413563682	0.820512820512820	4
13	6935.54684063562	27	5	39	177.834534375272	0.692307692307692	4
14	3639.07861616162	8	4	39	93.3097081067081	0.20512820512820 5	5
15	661.210347826087	1	1	39	16.9541114827202	2.56410256410256e-002	5
16	621.72156097561	2	2	39	15.9415784865541	5.12820512820513e-002	5
17	8391.31548672734	23	7	39	215.161935557111	0.589743589743590	4
18	9299.69089843652	36	8	39	238.453612780424	0.923076923076923	4
19	5962.8969797274	26	7	39	152.894794351985	0.666666666666667	4
20	5119.87934277521	27	9	39	131.278957507057	0.692307692307692	4

Figure 4.32 Training data view of the cluster model from Experiment 4

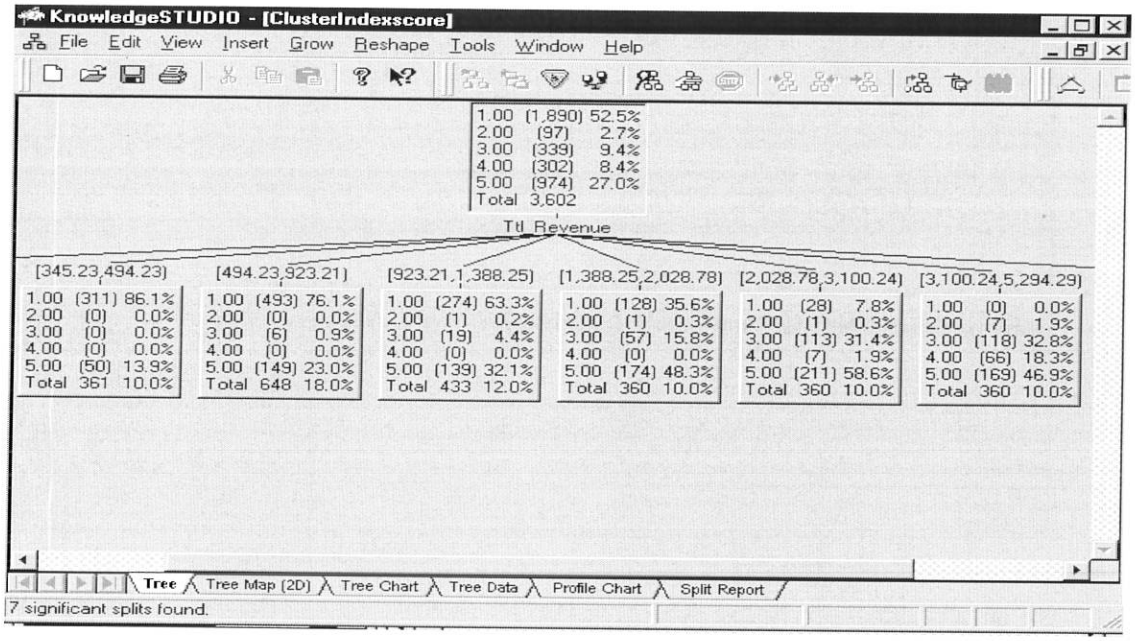


Figure 4.33 Output of the application of the clustering model on a new data set

According to Saarenvirta (1998), to be confident in data mining results, one should observe current business knowledge in results. Observing current business knowledge provides confidence that data selection and data preparation efforts have been valid. Saarenvirta continues that if results are observed that were previously unknown, one can have confidence in them.

The cluster results confirm the current business knowledge that frequent travelers generally generate high revenue. The clusters also revealed that the revenue contribution of customers with the same high travel frequencies differed, with one segment generating more revenue that the other. This view of customer segments was new and also made business sense to the domain experts.

4.5.5 Building Decision Tree Model

During this phase, the clustering model that segmented the records into meaningful groups was used to build a decision tree using Knowledge Studio. The cluster index was used as the dependent variable and was used to derive rules explaining how to assign new records to the correct cluster. The six variables used in deriving the cluster model were used as independent variables.

A total of 40 decision tree models were grown using Knowledge Studio's decision tree algorithm. The models differed according to the variables used to make the first split. Once the models were trained, they were used to assign (score) new records to the appropriate cluster. Six models which scored high training and prediction have been reported.

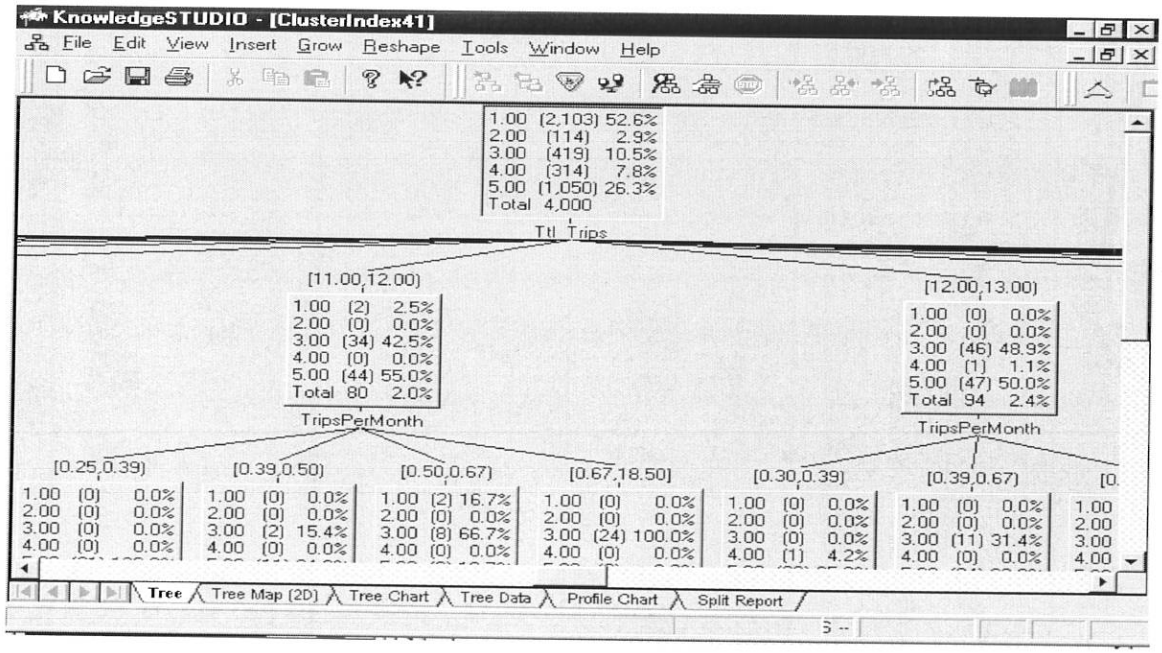


Figure 4.34 A fully-grown decision tree clustering model

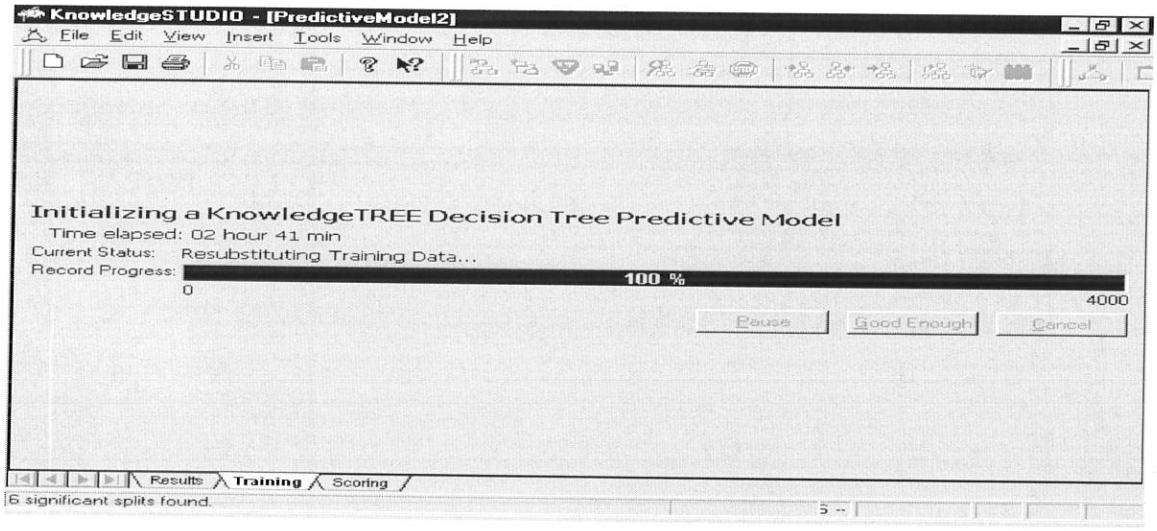


Figure 4.35 Training the decision tree predictive model

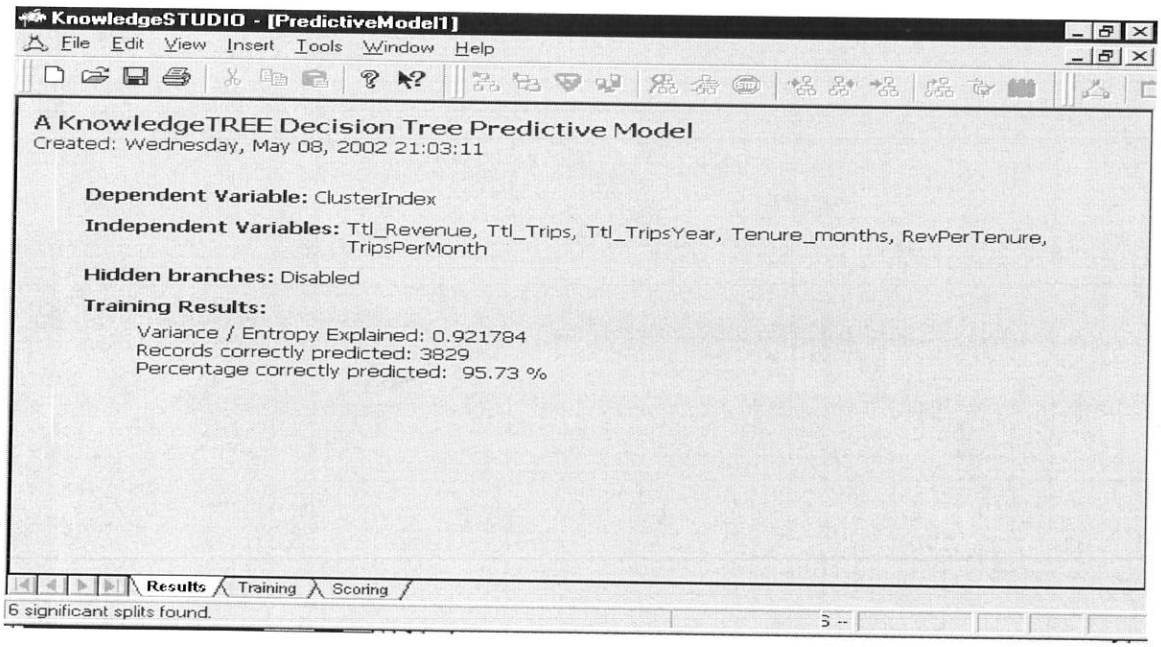


Figure 4.36 Training and clustering results of Model 1

Model	Splitting Variable	Training Result (%) (4000 Records)	Scoring (Prediction) Result (%) (3602 Records)
1	Ttl_Revenue	95.73	92.18
2	Ttl_Trips	96.60	93.09
3	Ttl_TripsYear	96.30	92.90
4	Tenure_Months	96.55	93.27
5	RevPerTenure	96.00	92.10
6	TripsPerTenure	95.97	92.53

Table 4.22 Summary of the clustering models' training and scoring results

In the training results, shown in Table 4.22, the models with Ttl_Trips and Tenure_Months as the first splitting variables had the highest training results respectively. Ttl_Revenue had the least result compared to the other variables. The difference between the highest and lowest training results was 0.87%, which was negligible. Furthermore, prediction results were also in the range of 92% – 93%.

These decision tree models were used to assign 3,602 new records to one of the five clusters. The decision tree models with Tenure_Months and Ttl_Trips predicted 93.27% and 93.09% of the new records respectively. Model 1, with Ttl_Revenue as the splitting variable, predicted 92.18% of the records correctly.

4.6 Evaluation

During this phase, the degree to which the model meets the business objectives was assessed. Furthermore, the whole process has been reviewed and the next steps were determined.

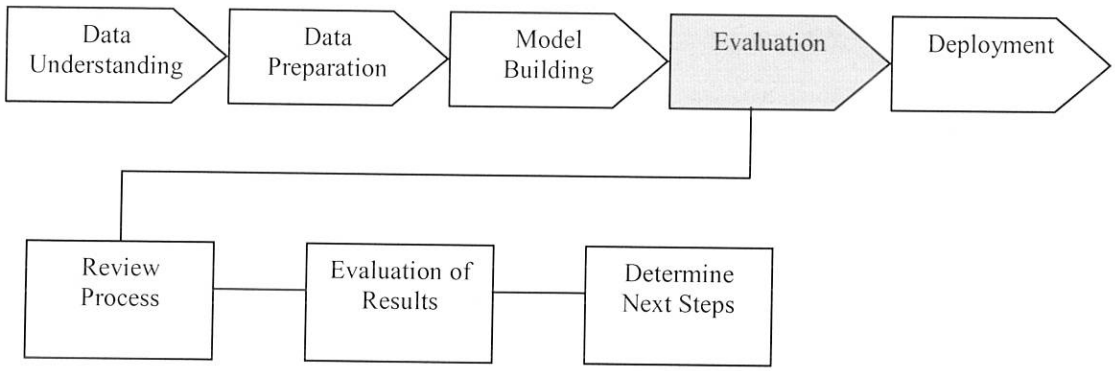


Figure 4.37 The evaluation phase

The data pre-processing phase culminated with the building of models. The model-building phase was divided into two. The first phase comprised of building a clustering model using the k-means algorithm. This exercise, which was iteratively conducted, yielded various cluster models that segmented the customer data.

The variables used for the cluster runs were those variables, which described a customer's travel behavior. Each of the cluster outputs were analyzed in order to establish their business worth as meaningful segments that revealed information pertaining to customers' value.

The analysis, which was closely undertaken with domain experts revealed that the segments indeed revealed patterns that grouped customers that exhibited similar characteristics with regards to the variables that were used to segment them. Further experiments were conducted to arrive at a better clustering model, which involved testing the effects that the variables had on the

resulting clusters. Decision tree models were used to identify which of the variables determined the clustering results.

The clustering results from Experiment # 4, which had 5 clusters was the 'best' from the ones in the other runs in terms of meaningful segmentation results they produced. 'Best' by no means implies that better results cannot be found. It is the researchers belief that more experimentation could yield more meaningful clusters. Time constraint was the major reason for not having built more models. Furthermore, Ttl_Revenue was found to be an important variable in clustering customer's records.

According to the evaluation made on the results of the 'best' cluster results, the majority of the records (52.6%) belonged to Cluster 1. This cluster contains customers with very low frequency of trips traveled, very low revenue generated, and relatively very recent in their tenure as members. According to the domain experts, this segment contains members who are not worth giving immediate attention in terms of targeted promotions as compared to members in the other segments. Furthermore, since the tenure of members in this segment is rather recent, it is rather premature to judge on the potential value of members of this group. In the near future, targeted promotions could be designed that would entice this group to be more active.

Cluster 2 contains very highly frequent, very highly revenue contributing as well as long tenured customers. This segments makes up the smallest number of records (2.9%) out of the total training records. This represents the high tiered and most valuable group among the members. Special care should be given to retain this group in order to avoid customer attrition to the competition. Further analysis revealed that customers belonging to this group are what are termed

as 'loyal'. Marketing strategies such as the addition of a fourth tier level (possibly 'Platinum') might be a good way to maintain the loyalty of this group.

Cluster 3 contains medium frequency, high revenue generating and medium tenured customers. The total revenue generated per the members' tenure as well as the number of trips per the members' are both high for this group. This segment, which makes up 10.5% of the total training records, has a potential to join the top tier group provided targeted promotions are run, which are enticing. This customer segment is worth giving immediate attention, as the likelihood of losing this segment to the competition is high due to its medium tenure in the program.

Cluster 4, making up 7.8% of the records, contains high frequency, very high revenue generating and very long tenured customers. This is the other top tier segment where more targeted promotions could elevate the members' status from highly frequent travelers to that of very highly frequent travelers, thereby benefiting from the very high revenue this group generates. Customers belonging to this segment show characteristics of loyalty.

Cluster 5 comprises of low frequency, medium revenue generating, and long tenured members. Making up 26.3% of the total records, this group's activity is quite insignificant compared to its rather long tenure in the program. Furthermore, the revenue contribution from the few trips made by this group has been significant. Targeted promotional campaigns might be prepared to entice this group to make frequent trips.

The results seem to confirm the 80/20 Pareto Principle, which states that 80% of the revenue is generated by 20% of the customers. Accordingly, clusters 2, 3, and 4, which are clusters

containing the most valuable customers in terms of high revenue generation, make up 21% of the total records. Clusters 1 and 5 make up 79% of the records.

Once the clustering model that segmented the customers into meaningful groups was selected, the next activity was to build a decision tree with the cluster index as the dependent variable and derive rules explaining how to assign new records to the correct cluster.

Training results showed that the final six models with the six variables as the first splitting variables, Ttl_Trips and Tenure_Months had the highest training results respectively. Ttl_Revenue had low training and prediction results compared to the other variables, but the percentage difference between the highest and lowest scores, which ranged between 0.87% - 1.17%, was quite negligible

Earlier experiments have shown that Ttl_Revenue is a more determining variable, and Model 1 has been chosen as a working model as a result of this study. The researcher believes that the training and predictive results of Model 1 could be further improved by employing pruning techniques, which was not fully exhausted during this study, mainly due to time constraints.

4.7 Model Deployment

The segmentation results were encouraging, and this knowledge could be used for more targeted marketing. Since the rules behind the segments are rather opaque, it is necessary to derive rules, which are as simple as possible and which can be applied to the entire member database. This has been included in list of further works.

Chapter 5

Conclusion and Recommendations

5.1 Conclusion

This research attempted to study the possible application of data mining techniques, and especially clustering techniques, to support CRM at ETHIOPIAN. The study was conducted in five major phases, namely business understanding, data understanding, data preparation, model building, and evaluation.

The data collection and preparation were major tasks due to the dispersed nature of the required data. Next, K-means clustering algorithms were applied to segment the customer data into meaningful groups. The parameters used by K-means algorithm were user defined, which made the experiment lengthy. The lack of predefined customer segments to validate the results against made the experiment to depend highly on domain experts' opinions in discovering new ones.

According to initial clustering results, 'total trips' was the variable that best described the records, though resulting segments did not make business sense. In subsequent clusters that made more business sense, the most determining variable was 'total revenue'. Although it is difficult to generalize based on these results, findings of the study seem to validate the business norm that 'customer value' is based on the total revenue contribution.

The cluster model, which according to the domain experts made business sense, segmented the records into five clusters. Three of the clusters contained 21% of customer records that generated

the highest revenue, and differed in the total frequency of trips and tenure of customers. The cluster containing medium and low revenue generating customers contained 27% and 52% of the customer records used in the study respectively.

In addition to confirming current business knowledge, the clusters provided a new view of customer segments with different travel behavior. The cluster model was later validated, where it clustered new records into five homogenous segments.

The decision tree model that was generated from the cluster results correctly assigned 92.18% of new records to the five clusters, with 'Total Revenue' as the splitting variable. This result was found to be lower than the results obtained with 'Total Trips' and 'Tenure in Months' as splitting variables, but the difference was quite negligible, and the decision tree model with 'Total Revenue' making the initial split was chosen as a working model.

In general, the results from this study were encouraging. It was possible to segment customer data using data mining techniques that made business sense. It is the researcher's belief that a more thorough study using data mining techniques can increase business leverage from customers and support CRM activities at ETHIOPIAN. Furthermore, knowledge of data mining techniques, marketing strategies and airline business processes should be integrated to successfully implement CRM.

5.2 Recommendations

The researcher makes the following recommendations based on the findings of this study.

Further data mining studies

Even though results from this study were encouraging, refinement to the segmentation results could prove valuable. Based on the results, further data mining projects should be undertaken, including further work on segmentation using more detailed customer demographic data, several predictive models for target (direct mail) marketing, and opportunity identification using association rule algorithms within the segments discovered. In addition, other neural clustering algorithms (SOM) can also be used to test the validity of the cluster results obtained using the K-means algorithm.

The researcher believes that the performance of the decision tree models for assigning new records to the appropriate clusters, which are currently in the range of 92-93%, could be improved through techniques such as pruning. Furthermore, more tests should be conducted to find variables that are more descriptive.

Even though the deployment of the model was beyond the scope of this study, the profiling of clusters, by generating SQL codes to the main customer database, would enable the assessment of the potential value of each cluster.

Members of the frequent flyer program are a subset of the airline's customer base. A further task is to find ways and means to integrate information regarding flight activities of all of the airline's customers. This information can be used to identify future potential customers.

Develop a customer data warehouse

This study has confirmed that data collection and data preparation from different operational databases is a very lengthy and tedious task, especially when there are time and resource constraints. The researcher strongly recommends that the airline embark on developing a customer data warehouse. This data warehouse should be fed by customer oriented data sources – booking databases, departure control databases, sales information databases, frequent flyer databases, and customer services databases.

Building up a core data mining competency internally

The airline business, which is highly data intensive, cannot ignore the potential of data mining. Not limited to customer databases, data mining can be applied on operational data generated in other lines of the airline's activities. The first step would be to form a data mining team with members from both the IT and user departments after assessing the business areas where data mining could add value.

Provide online frequent flyer information via the Internet

Extending ShebaMiles information over the Internet will be a leap forward, which enables members to view their account and get all the necessary information at their convenience at any time and any place. However, first the current mileage tracking system has to be improved, and up to date and accurate real-time information has to be communicated to the member. Enabling the passenger to view the following information is recommended:

- Year to date balance of points along with the respective expiry dates
- An update of last flight activity
- Online request for eligibility of awards (such as Redemption, upgrade, excess baggage and other services)
- Tier status
- Extra miles required to earn awards

References

Angoss Software Corporation. Knowledge Studio Data Mining Software User Guide. 2001. On request. Internet. http://www.angoss.com.

Berry, Michael J.A. and Gordon Linoff. Data Mining Techniques For Marketing Sales and Customer Support. New York: John Wiley & Sons. 1997.

---. Mastering Data Mining: The Art and Science of Customer Relationship Management. New York: John Wiley & Sons. 2000.

Bigus, Joseph P. Data Mining with Neural Networks: Solving Business Problems from Application Development to Decision Support. New York: McGraw-Hill. 1996.

Bishop, Christopher M. Neutral Networks for Pattern Recognition. Oxford: Clarendon Press. 1995.

Bounsaythip, Catherine and Esa Rinta-Runsala. Overview of Data Mining for Customer Behavior Modeling. Version 1. *VTT Information Technology*. 2001. Online. Internet. http://www.vtt.fi/tte/.

Canaday, Henry. What does CRM mean for Airlines? *Airlines International*. 5 (1999): 1-3.

Chandler, Jerry C. Guide to Travel Loyalty & CRM. Rockville, MD: Garrett Communications. 2001.

---. Defining the Value Proposition at Six Continents' Priority Club. *Travel Loyalty & CRM.* September 2001: 3-5.

CRISP-DM. CRISP-DM 1.0: Step-by-step data mining guide. 2000. Online. Iternet. http://www.crisp-dm.org.

DCI I.T. Airline Pricing and Data Mining. Data Warehouse Report. 1998. Online. Internet. http://www.datawarehouse.dci.com/articles/1998/11/3reno.htm

Doganis, Rigas. Flying Off Course: The Economics of International Airlines. London: Routledge. 1991.

Donoghue, J.A. Getting It Wired. *Air Transport World.* April 2002. 24-26.

DSS Research. Understanding Market Segmentation. 2001. Online. Internet. http://www.dssresearch.com/library/segment/understanding.asp

EDS. Data Mining Case Studies. 2001. Online. Internet. http://www.eds.ch

Elder, John F. and Dean Abbott. A Comparison of Leading Data Mining Tools. KDD-98. August 28, 1998. Online. Internet. http://www.kdnuggets.com

Ethiopian Airlines. Annual Report 2001. Addis Ababa, Ethiopian Airlines.

Ethiopian Airlines. Bringing Africa Together: The Story of an Airline. Nairobi: Camerapix Publishers International. 1988.

Ethiopian Airlines. ShebaMiles Frequent Flyer Pogramme Membership Guide. Addis Ababa, Ethiopian Airlines.

Ethiopian Airlines. Worldwide Timetable: Summer 2002. Dubai: Printing Services.

Fayyad, Usama, Gregory Piatetsky-Shapiro and Padhraic Smyth. From Data Mining to Knowledge Discovery in Databases. *AAAI*. (1996). Online. Internet. http://www.kdnuggets.com.

Fickel, Louise. Window Seat. *CIO Magazine*. July 15, 2000. Online. Internet. http://www.cio.com/archive/071500/window_content.html.

Gobena, Mikael. Flight Revenue Information Support System for Ethiopian Airlines. 2000. A Thesis Submitted in Partial Fulfillment of the requirement for the Degree of M.Sc. I.S. Addis Ababa University: Addis Ababa.

Goebel, Michael and Le Gruenwald. A Survey of Data Mining and Knowledge Discovery Software Tools. SIGKDD Explorations. June, 1999. Online. Internet. http://www.kdnuggets.com

Han, Jiawei and Micheline Kamber. Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers. 2001.

Harris, Jeanne G. Finding the Customer in Transaction Data. n.d. Online. Internet. http://www.crmproject.com

Holtz, Herman. Databased Marketing. New York: John Wiley & Sons. 1992.

IBM. Customer Relationship Management. 2000. Online. Internet. http://www.ibm.com/solutions/travel

Kotler, Philip. Marketing Management: Analysis, Planning, Implementation and Control. 9th edition. New Delhi, Prentice Hall of India. 1998.

Levin, Nissan and Jacob Zahavi. Data Mining. 1999. On request. Internet. http://www.kdnuggets.com.

McDaniel, Carl D. and Roger Gates. Marketing Research Essentials. Minneapolis/St. Paul: West Publishing Company. 1995.

McDonald, Michele. AAdvantage Marks 20th birthday. *Travel Weekly*. April 30, 2001. Online. Internet. httlp://www.findarticles.com

McKinsey&Company. The New Era of Customer Loyalty Management. 2001. Online. Internet. htttp://www.marketing.mckinsey.com

Oxford English Dictionary. Oxford: Oxford University Press. 1997.

Piatetsky-Shapiro, Gregory. Knowledge Discovery in Databases: 10 Years After. SIGKDD Explorations. 1 (2000). Online. Internet. http://www.kdnuggets.com/gpspubs/sigkdd-explorations-kdd-10-years.html

Pritscher, Lisa and Hans Feyen. Data Mining and Strategic Marketing in the Airline Industry. n.d. Online. Internet. http://www.luc.ac.be/iteo/articles/pritscher1.pdf

Quee, Wong Toon. Marketing Research. 3rd edition. Oxford: Butterworth-Heinemann. 1999.

Reichheld, Frederick F. Loyalty and the Renaissance of Marketing. *Relationship Management for Competitive Advantage.* Ed. Adrian Payne et.al. Oxford: Butterworth-Heinemann. 1995. 232-253.

Saarenvirta, Gary. Mining Customer Data. 1998. Online. Internet. http://www.db2mag.com/db_area/archives/1998/q3/98fsaar.html

Thearling, Kurt. An Introduction to Data Mining. n.d. Online. Internet. http://www3.primuhost.com/~kht/text/dmwhite/dmwhite.html

---. Increasing Customer Value by Integrating Data Mining and Campaign Management Software. 1999. Online. Internet. http://www3.primuhost.com/~kht/text/integration/integration.html

Trybula, Walter J. Data Mining and Knowledge Discovery. ed. Williams, Martha E. Annual Review Information Science and Technology., Vol. 32, Information Today, Inc. Medford, New Jersey, USA. 1997.

Two Crows Corporation. Introduction to Data Mining and Knowledge Discovery. 3rd ed. 1999. On request. Internet. http://www.twocrows.com.

Glossary of terms

Loyalty Programs: It is an integrated system of network by which service giving organizations identify individual market segments that use their services with exceptional frequency and make plan as to what can be done to recognize their loyalty.

Frequent Flyer Program: One form of loyalty program whereby individuals and groups who frequently travel on ETHIOPIAN earn recognition of their patronage to ETHIOPIAN.

Frontline Customer Services: These are offices that have direct contact with the passengers of an airline. It shall include Ticket Offices, Travel Agencies, General Sales Agents, Consolidators, Sales Offices, Airport, Reservations Offices, In-flight Services, Commissary Unit, ETHIOPIAN Lounges, etc.

ShebaMiles DB: ShebaMiles' frequent flyer program administration database system.

Departure Information Form (DIF): is a carbonated paper form to be filled out and signed by ETHIOPIAN ShebaMiles member before flight departure, which must also be verified and signed by an airport staff.

Credential: Program materials that are supplied by the Customer Loyalty Department (CLD) to the Member.

Member: is defined as Ethiopian's Frequent Flyer Program Member whose name is printed on the card, who has been assigned a membership number, the account number, and who has signed the card.

Departure Information Message (DIM): The Departure Information Message is an ETHIOPIAN standard Frequent Flyer Program Message. It is a list containing the details of Departure Information Forms (DIFs) for a given flight. It is to be composed and sent as soon as a flight has departed.

Frequent Traveler List (FTL): an automatically generated list, containing ShebaMiles members' flight activity information, which is input to the ShebaMiles DB.

Elite levels: these are two-step membership levels in the ShebaMiles program namely, Silver Club and Gold Club.

Silver Club: An elite club level with a requirement of 25,000 base miles.

Gold Club: An elite club level with a requirement of 50,000 base miles.

Annex 1

Annex 1: Procedures Used for Data Collection

This annex contains SQL procedures (developed using Microsoft Visual Basic 6.0's ADO interface), which were used to collect data from different sources.

a) The following procedure was used to read through 306 ShebaMiles DB input data files, and merge the individual records into a single table named **Trips Load,** which contained 101,189 records.

```
Sub OpenFile()
    NFile = FreeFile
    With FrmImportingData
    .CommonDialog1.Filter = "*.Txt|*.Txt"
    .CommonDialog1.ShowOpen
    .CommonDialog1.CancelError = True
     Open .CommonDialog1.FileName For Input As #NFile
End With
End Sub
Public Sub CollectTripData()
CountFile = Dir$("C:\Reward Data\Cleaning Data\*.trp")
  counting = 0

While CountFile <> ""
  counting = counting + 1
   FrmImportingData.TxtCount.Text = counting
   Open "C:\Reward Data\Cleaning Data\" & CountFile For Input As #1

    While Not EOF(1)

    Line Input #1, Val
      MemNum =  Mid(Val, 52, 11)
      FltNum   =  Mid(Val, 27, 4)
      Class     =  Mid(Val, 51, 1)
      Origin    =  Mid(Val, 21, 3)
      Destn     =  Mid(Val, 24, 3)
      TktNum =  Mid(Val, 64, 13)
      Dte        =  Mid(Val, 17, 2)
      Mn = Mid(Val, 15, 2)
```

```
            Yr   = Mid(Val, 13, 2)
            FltDate = Mn & "/" & Dte & "/" & Yr
        If (FltDate <> "  /  /  ") And (FltDate <> "//") Then
        On Error Resume Next
        With DEShebaMiles
           .ComInsertTripData MemNum, FltNum, FltDate, Class, Origin, Destn, TktNum
        End With
        End If
           Wend
           Close #1
           CountFile = Dir$
        Wend
        End Sub
```

b) Since there were records with temporary member numbers in the **Trips Load** table, the following procedure was used, which read through another matching table in the ShebaMiles DB (named **Alias**), and retrieved the corresponding permanent member numbers:

```
Public Sub CollectAlias()
  Call OpenFile
    While Not EOF(NFile)
        Line Input #NFile, Val
        PerNum    = Mid(Val, 1, 11)
        TempNum = Mid(Val, 12, 8)
        With DEShebaMiles
           .ComINsertAliasInfo PerNum, TempNum
        End With
     Wend
End Sub
```

c) The following procedure was used to match unique flight activity records in the **Trips** table with those in the **Trips Load** table, extract the required *'ticket number'* information from the latter, and insert it as new field in the **Trips** table:

```
Public Sub TktNumCollector()

  Dim  Cn   As New ADODB.Connection
  Dim  Rs1  As New ADODB.Recordset
  Dim  Rs2  As New ADODB.Recordset

  Cn.Open  "Provider=Microsoft.Jet.OLEDB.4.0;Data Source=C:\Reward Data\Reward-
           Extracts\ShebaMiles DB.mdb;Persist Security Info=False"

  Set Rs1 = Cn.Execute("Select * from Trips order By ffnum")

  While Not Rs1.EOF

      MemNum = Rs1.Fields("ffnum")
      FltDate = Rs1.Fields("date")
      FltNum = Rs1.Fields("Flite")

      Set Rs2 = Cn.Execute("SELECT [Ticket Number] From [Trip Load] WHERE
                ((([Trip Load].[Member Number])= '" & MemNum & "') AND (([Trip
                Load].[Flight Number])= '" & FltNum & "') AND (([Trip Load].[Flight
                Date])= #" & FltDate & "#))")

   If  Not Rs2.EOF  Then

    TktNumber = Rs2.Fields("Ticket Number")

    Cn.Execute ("UPDATE Trips SET Trips.TktNum = '" & TktNumber & "' WHERE
               (((Trips.ffnum)= '" & MemNum & "') AND ((Trips.flite)='" & FltNum &
               "') AND ((Trips.date)=#" & FltDate & "#))")

   End If

   Rs1.MoveNext

  Wend

End Sub
```

d) The following procedure was used to extract revenue data from a revenue accounting database and insert it into the **Trips** table:

```
Public Sub RevDataCollector()
  Dim  Cn   As New ADODB.Connection
  Dim  Rs1  As New ADODB.Recordset
  Dim  Rs2  As New ADODB.Recordset

  Cn.Open  "Provider=Microsoft.Jet.OLEDB.4.0;Data Source=C:\Reward Data\Reward-
           Extracts\ShebaMiles DB.mdb;Persist Security Info=False"

  Set Rs1 = Cn.Execute("Select * from RevData")

  While Not Rs1.EOF

    TktNmbr = Rs1.Fields("TktNumber")
    Org = Rs1.Fields("Org")
    Destn = Rs1.Fields("Destn")
    RevValue = Rs1.Fields("Revenue")

    Set Rs2 =  Cn.Execute("SELECT * From TripsFinal WHERE (((TripsFinal.[TktNum])=
           "' & TktNmbr & "') AND ((TripsFinal.[Orig])= "' & Org & "') AND
           ((TripsFinal.[Dest])= "' & Destn & "'))")

    If Not Rs2.EOF Then

'     Rs2.Fields("Revenue").Value = RevValue
'     Rs2.Update

      Cn.Execute ("UPDATE TripsFinal SET TripsFinal.Revenue = "' & RevValue & "'
                 WHERE (((TripsFinal.TktNum)= "' & TktNmbr & "') AND
                 ((TripsFinal.Orig)="' & Org & "') AND ((TripsFinal.Dest)="' & Destn
                 & "'))")

    End If

    Set Rs2 = Nothing

    Rs1.MoveNext

  Wend

End Sub

Public Sub RevenueToTrips()

  Dim  Cn  As New ADODB.Connection
  Dim  Rs  As New ADODB.Recordset
```

```
    Cn.Open "Provider=Microsoft.Jet.OLEDB.4.0;Data Source=C:\Reward Data\Reward-
             Extracts\ShebaMiles DB Refined.mdb;Persist Security Info=False"

    Set Rs = Cn.Execute("Select * from TripsWithRevData ")

    While Not Rs.EOF

      MemNum = Rs.Fields("ffnum")
      FltDate = Rs.Fields("date")
      FltNum = Rs.Fields("Flite")
      Rvnu = Rs.Fields("Revenue")

      Cn.Execute ("UPDATE TripsOriginal SET TripsOriginal.revenue = '" & Rvnu & "'
                  WHERE (((TripsOriginal.[ffnum])= '" & MemNum & "') AND
                  ((TripsOriginal.[flite])= '" & FltNum & "') AND ((TripsOriginal.[date])= #"
                  & FltDate & "#))")

       Rs.MoveNext

    Wend

    Set Rs = Nothing
    Set Cn = Nothing

End Sub
```

Annex 2

Annex 2: Procedures Used for Data Preparation

This annex contains the procedure (developed using Microsoft Visual Basic 6.0's ADO interface) that was used to fill records with missing values in the data preparation phase.

The following procedure was used to compute and insert missing revenue values in the **Trips** table:

```
Public Sub RevenueFiller()
    Dim   Cn    As New ADODB.Connection
    Dim   Rs    As New ADODB.Recordset
    Dim   Rs2   As New ADODB.Recordset
    Dim   Org, Destn, Clss, MemNum As String
    Dim   AVRevenue     As Double
    Dim   FltDate       As Date

   Cn.Open "Provider=Microsoft.Jet.OLEDB.4.0;Data Source=C:\Reward Data\Reward-
            Extracts\ShebaMiles DB Refined.mdb;Persist Security Info=False"

   Set    Rs = Cn.Execute("Select * from TripsWithRevFillUp")

   While  Not Rs.EOF
       On Error Resume Next
       Org = Rs.Fields("Orig")
       Destn = Rs.Fields("Dest")
       Clss = Rs.Fields("class")
       FltDate = Rs.Fields("Date")
       MemNum = Rs.Fields("ffnum")
       On Error Resume Next
       If Clss = "C" Or Clss = "D" Then
       On Error Resume Next

    Set   Rs2   =   Cn.Execute("SELECT   Sum(TripsWithRevData.Revenue)   AS
          SumOfRevenue,   TripsWithRevData.orig,TripsWithRevData.dest,   Count
          (TripsWithRevData.orig)   AS   CountOforig   From   TripsWithRevData
          Where(((TripsWithRevData.orig)='"&Org&"')And ((TripsWithRevData.Dest)
          = '" & Destn & "') And ((TripsWithRevData.Class) ='"& Clss & "')) GROUP
          BY TripsWithRevData.orig,TripsWithRevData.dest,TripsWithRevData.class ")
```

```
If  Not Rs2.EOF  Then

        AVRevenue = Rs2.Fields("SumOfRevenue") / Rs2.Fields("CountOfOrig")
        Cn.Execute ("UPDATE TripsWithRevFillUp SET TripsWithRevFillUp.Revenue = '"
                    & AVRevenue & "' WHERE (((TripsWithRevFillUp.orig)='" & Org &
                    "') AND ((TripsWithRevFillUp.dest)='" & Destn & "') AND
                    ((TripsWithRevFillUp.date)=#" & FltDate & "#) AND
                    ((TripsWithRevFillUp.ffnum)='" & MemNum & "'))")

        Set Rs2 = Nothing
 Else

        On Error Resume Next

        Set  Rs2  =  Cn.Execute("SELECT  Sum(TripsWithRevData.Revenue)  AS
                     SumOfRevenue, TripsWithRevData.orig, TripsWithRevData.dest,
                     Count(TripsWithRevData.orig)  AS  CountOforig  From
                     TripsWithRevData Where (((TripsWithRevData.orig) = '" & Org
                     & "') And ((TripsWithRevData.Dest) = '" & Destn & "')) And
                     ((TripsWithRevData.class)='"  &  C  &  "'))  OR
                     (((TripsWithRevData.class)='"  &  D  &  "'))GROUP  BY
                     TripsWithRevData.orig, TripsWithRevData.dest")

If  Not Rs2.EOF  Then

    AVRevenue = Rs2.Fields("SumOfRevenue") / Rs2.Fields("CountOfOrig")
    Cn.Execute ("UPDATE TripsWithRevFillUp SET TripsWithRevFillUp.Revenue = '" &
                AVRevenue & "' WHERE (((TripsWithRevFillUp.orig)='" & Org & "')
                AND  ((TripsWithRevFillUp.dest)='"  &  Destn  &  "')  AND
                ((TripsWithRevFillUp.date)=#"  &  FltDate  &  "#)  AND
                ((TripsWithRevFillUp.ffnum)='" & MemNum & "'))")

    Set Rs2 = Nothing

    End If

End If

Else
        'If Class is Different From C & D

            '***************

   Set Rs2 = Cn.Execute("SELECT Sum(TripsWithRevData.Revenue) AS SumOfRevenue,
             TripsWithRevData.orig,                    TripsWithRevData.dest,
             Count(TripsWithRevData.orig) AS CountOforig From TripsWithRevData
             Where  (((TripsWithRevData.orig)  =  '"  &  Org  &  "')  And
             ((TripsWithRevData.Dest) = '" & Destn & "') And ((TripsWithRevData.Class)
             =  '"  &  Clss  &  "'))  GROUP  BY  TripsWithRevData.orig,
             TripsWithRevData.dest, TripsWithRevData.class ")
```

```
On Error Resume Next

If  Not Rs2.EOF  Then

    AVRevenue = Rs2.Fields("SumOfRevenue") / Rs2.Fields("CountOfOrig")
    Cn.Execute ("UPDATE TripsWithRevFillUp SET TripsWithRevFillUp.Revenue = '"
                & AVRevenue & "' WHERE (((TripsWithRevFillUp.orig)='" & Org &
                "') AND ((TripsWithRevFillUp.dest)='" & Destn & "') AND
                ((TripsWithRevFillUp.date)=#" & FltDate & "#) AND
                ((TripsWithRevFillUp.ffnum)='" & MemNum & "'))")

    Set Rs2 = Nothing
Else
    On Error Resume Next

    Set  Rs2  =Cn.Execute("SELECT  Sum(TripsWithRevData.Revenue)  AS
                SumOfRevenue, TripsWithRevData.orig, TripsWithRevData.dest,
                Count(TripsWithRevData.orig)  AS  CountOforig  From
                TripsWithRevData Where (((TripsWithRevData.orig) = '" & Org &
                "') And ((TripsWithRevData.Dest) = '" & Destn & "')) And
                ((TripsWithRevData.class)<>'" & C & "')) OR
                (((TripsWithRevData.class)<>'" & D & "'))GROUP BY
                TripsWithRevData.orig, TripsWithRevData.dest")

If  Not Rs2.EOF  Then
    AVRevenue = Rs2.Fields("SumOfRevenue") / Rs2.Fields("CountOfOrig")
    Cn.Execute  ("UPDATE TripsWithRevFillUp SET TripsWithRevFillUp.Revenue
                = '" & AVRevenue & "' WHERE (((TripsWithRevFillUp.orig)='" &
                Org & "') AND ((TripsWithRevFillUp.dest)='" & Destn & "') AND
                ((TripsWithRevFillUp.date)=#" & FltDate & "#) AND
                ((TripsWithRevFillUp.ffnum)='" & MemNum & "'))")

    Set Rs2 = Nothing

      End If

    End If

  End If

  Rs.MoveNext

 Wend

End Sub
```

Annex 3

Annex 3: Training and Clustering Results of Decision Tree Models

The following figures are the clustering results of the decision tree models that have been used to assign new records to the appropriate clusters. The variables used to make the initial splits of the decision trees were Ttl_Trips, Ttl_TripsYear, Tenure_Months, RevPer Tenure, and TripsPerTenure respectively. Furthermore, the decision tree model with Ttl_Revenue making the initial split has been chosen as a working model (refer to Model 1 in Section 4.5.5).

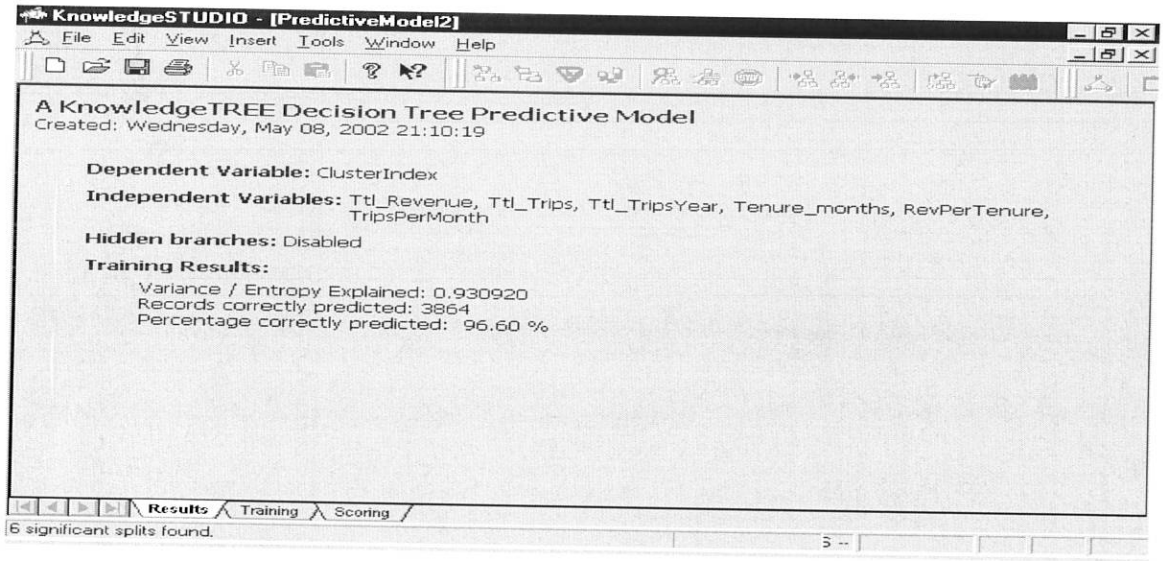


Figure 1. Training and clustering results of Model 2

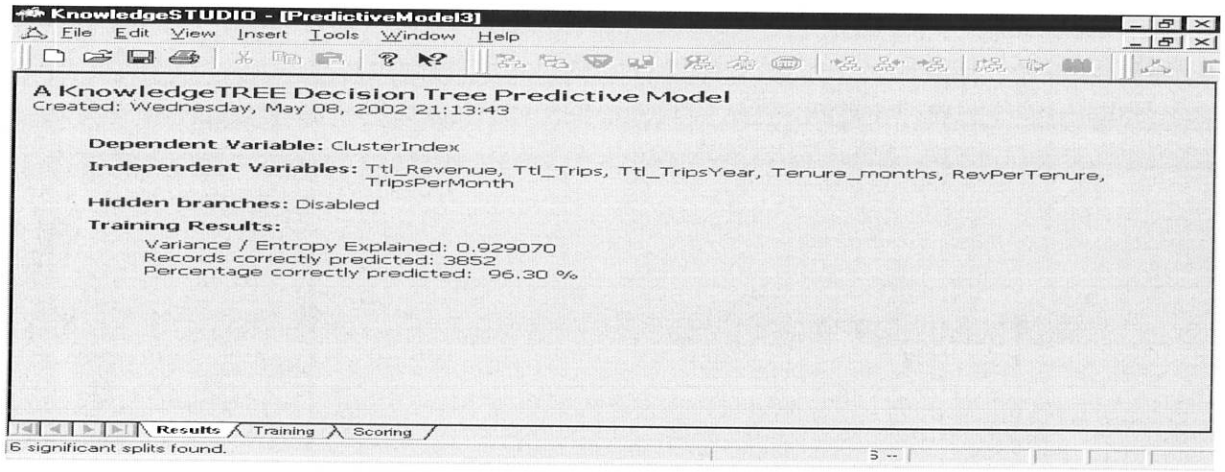


Figure 2. Training and clustering results of Model 3

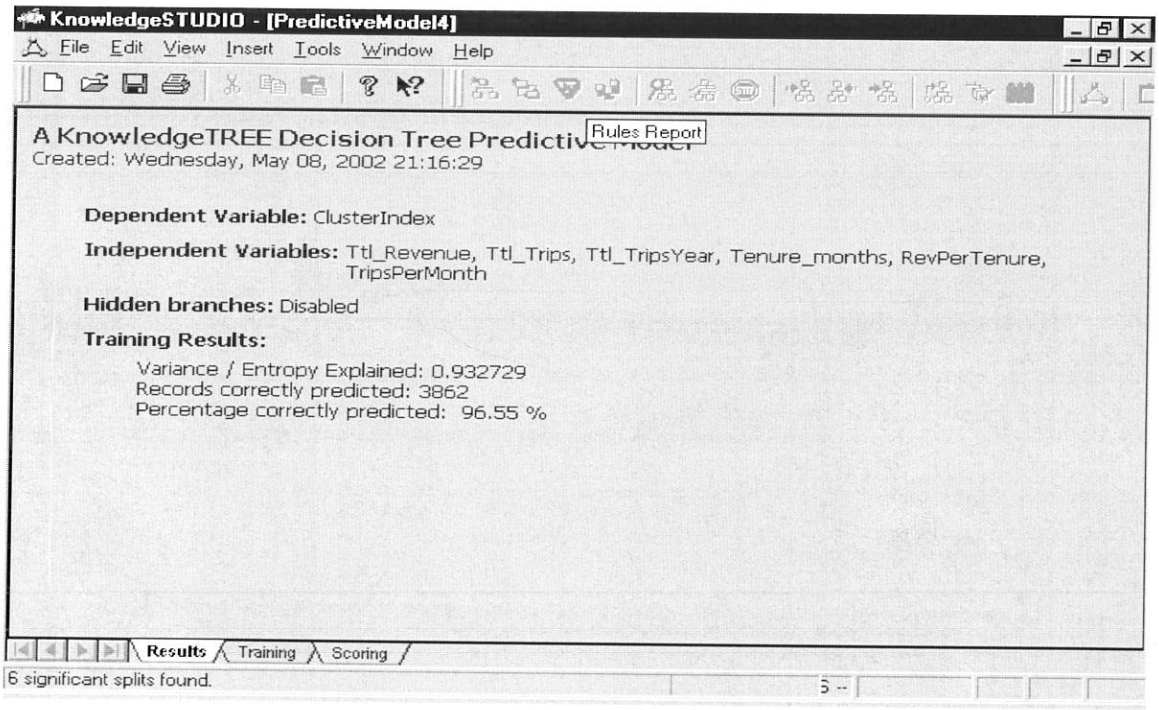


Figure 3. Training and clustering results of Model 4

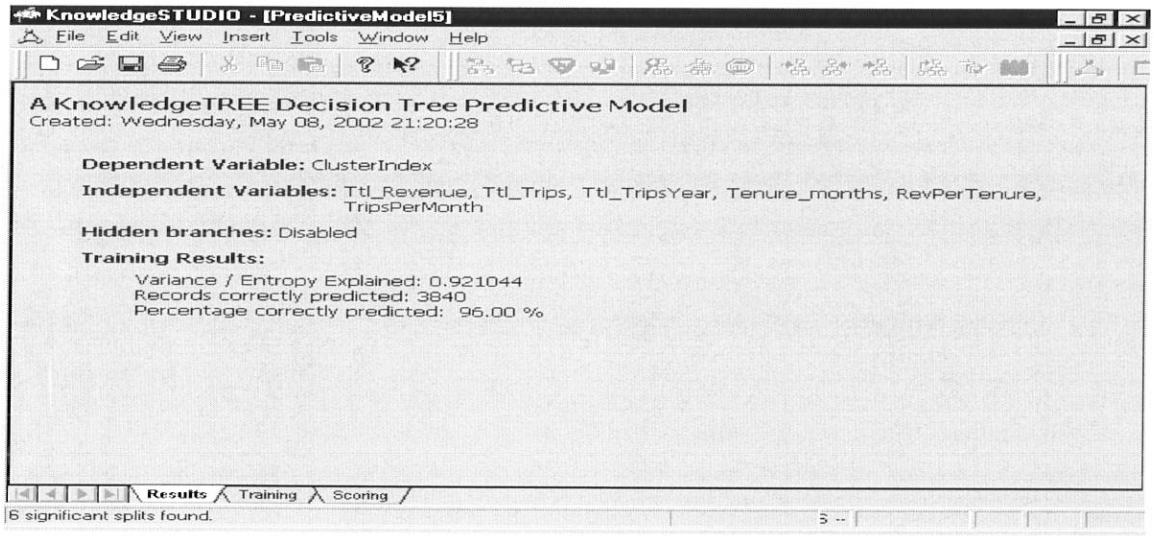


Figure 4. Training and clustering results of Model 5

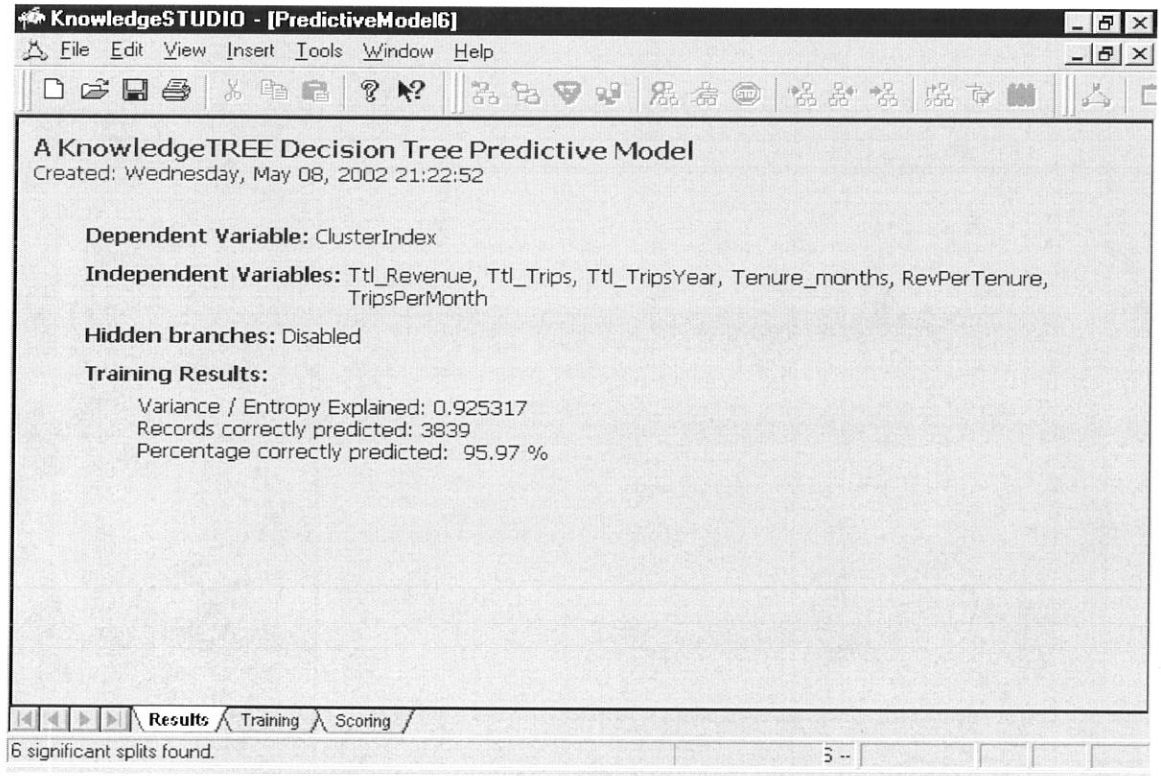


Figure 5. Training and clustering results of Model 6

Annex 4

Annex 4: Format of the Departure Information Message

This annex contains the format of the departure information message (DIM), which is one of the messages from ETHIOPIAN's departure control system (DCS) regarding the flight activities of members of ShebaMiles. Data from this message is used as input to the ShebaMiles DB, in order to track member flight activity.

i. A sample of DIM

```
>ZZ CLP MSR
DIF        CCYYMMDD        CC     FLHT  ORI
DES  NAME          CL ACCOUNTNUMBR  TICKETNUMBERS
_____END _____
```

ii. Description of Fields on the Departure Information Message

1) ZZ = ETHIOPIAN's host message type and office address codes (the addressed office codes are placed on this line).

2) DIF= Message identifier - Departure Information Form

3) CCYYMMDD= Flight date format.

   CC= Century

   YY= Year

   MM= Month

   DD= Date

   Example: June 08, 2002 should be written as "20020608"

4) CC = Flight Operating Airline 2 letter Code (Example, "ET" for ETHIOPIAN)

5) FLHT = Flight Number of the Operating Airline

6) ORI = 3 letters IATA Departure Airport/City code

7) DES = 3 letters IATA Arrival Airport/City code

8) NAME = Name of the Member not exceeding 20 Characters

9) CL = Ticketed class (class of service on the paid ticket)

10) Account Number = The membership number on the membership card of the member as it appears on the card. Example 00010001000.

NOTE: the Account Number should not be more than 11 digits long.

11) Ticket Numbers = the ticket number of used coupon. It shall only be 13 characters long.

Example - 0712404235236
071 = Airline code
2404 = Form
235236 = Serial Number

iii. Sample of a completed DIM

```
>ZZ CLP MSR
DIF  19991102 ET 0829 ADD
DES     NAMECL     ACCOUNTNUMBR  TICKETNUMBERS
EBB     A. Soofie      C      00010017534  0712404238291
EBB     ADALANO OYVIND         S  00010001084    0712404451382
EBB     ABADINADIRI MEBI B  00010027920     0714404000001
```

Annex 5

Annex 5: Generation of a Frequent Traveler List

In order to permit departure control systems (DCS) a means to exclusively report ShebaMiles members' flight activities to the ShebaMiles DB as may be necessary, airlines or airport handling agents automatically generate a Frequent Traveler List (FTL) message.

The following procedure is used to check-in a ShebaMiles member in ETHIOPIAN's DCS and generate a FTL:

">PA: 1, 2/20,FQTV ET/0000000". Where,

i. "PA" is the command to check-in from the list,
ii. "1" is the order of the member on a check-in list,
iii. "2/20" is the piece and weight for the luggage,
iv. "FQTV" is a request to initiate FTL message for the flight in general and specifically to the member,
v. ET is the membership card issuing carrier,
vi. "/" Is the separator
vii. "0000000" is the membership number

Annex 5

Annex 5: Generation of a Frequent Traveler List

In order to permit departure control systems (DCS) a means to exclusively report ShebaMiles members' flight activities to the ShebaMiles DB as may be necessary, airlines or airport handling agents automatically generate a Frequent Traveler List (FTL) message.

The following procedure is used to check-in a ShebaMiles member in ETHIOPIAN's DCS and generate a FTL:

">PA: 1, 2/20,FQTV ET/0000000". Where,

i. "PA" is the command to check-in from the list,
ii. "1" is the order of the member on a check-in list,
iii. "2/20" is the piece and weight for the luggage,
iv. "FQTV" is a request to initiate FTL message for the flight in general and specifically to the member,
v. ET is the membership card issuing carrier,
vi. "/" Is the separator
vii. "0000000" is the membership number

DECLARATION

The thesis is my original, has not been presented for a degree in any other university and that all sources of material used for the thesis have been duly acknowledged.

Henock Woubishet Tefera

July 2002

The thesis has been submitted for examination with our approval as university advisors.

Ermias Abebe

Million Meshesha

Nigussie Tadesse

July 2002