ADDIS ABABA UNIVERSITY
SCHOOL OF GRAGUATE STUDIES
FACULTY OF INFORMATICS
DEPARTMENT OF INFORMATION SCIENCE

THE APPLICATION OF DATA MINING TO SUPPORT CUSTEMER RELATIONSHIP MANAGEMENT AT ETHIOPIAN AIRLINES

BY

DENEKEW ABERA JEMBERE

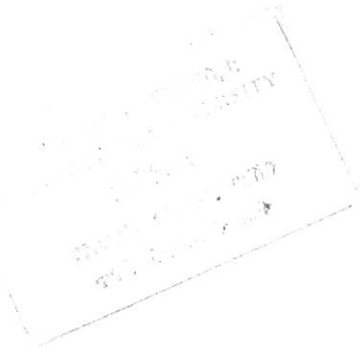
JUNE, 2003

THE APPLICATION OF DATA MINING TO SUPPORT CUSTEMER RELATIONSHIP MANAGEMENT AT ETHIOPIAN AIRLINES

BY

DENEKEW ABERA JEMBERE

A Thesis Submitted to the School of Graduate Studies, Addis Ababa University, Department of Information Science in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Science



June, 2003

ADDIS ABABA UNIVERSITY
SCHOOL OF GRAGUATE STUDIES

THE APPLICATION OF DATA MINING TO SUPPORT CUSTEMER RELATIONSHIP MANAGEMENT AT ETHIOPIAN AIRLINES

BY

DENEKEW ABERA JEMBERE

FACULTY OF INFORMATICS
DEPARTMENT OF INFORMATION SCIENCE

Approved by	Signature
Advisor	Signature
Advisor	Signature
Advisor	Signature
Examiner	Signature

ACKNOWLEDGEMENTS

Above all, I would like to glorify the almighty GOD for giving me the ability to be where I am. You have done so much for me, O Lord. No wonder I am glad! I sing for joy, Hallelujah!

I would like to thank my advisors Ato Mesfin Getachew, Ato Ermias Abebe and Ato Henok Wobishet for their constructive comments and overall guidance. But special thanks go to Ato Henok Wobishet, without whom this research would have not been a success. Henok, your helpful personality will always be a role in my heart.

I wish to thank Ato Ermias Alemu for being glad to provide his support. I would also like to thank Ato Mesfin Tassew, the Chief Information Officer of Ethiopian Airlines, for allowing me to carry out this research using the required data from the Airline.

I am very much grateful to my father Ato Abera Jembere, my mother W/ro Etaferahu Gichamo, my sisters W/ro Debrework Abera and W/rt Birhane Abera for their care and understanding during my study times. I am also grateful to Ato Engida Kasa (Gashe) and Ato Adamu Jember (Wondim Gashe) for they have been there to assist me morally and materially whenever I needed.

My heartfelt thanks also go to all my instructors and classmates at SISA for the lovely time and classes we have had together.

At last, but by no means the least, I would like to thank my brothers, and friends such as Nigusie Gezahegn, Mulugeta Jawoire, Admasu Engida, Habtamu Hailu and Misrak Genene for the constant assistance and encouragement they rendered to me since the time of my admission to the postgraduate program.

TABLE OF CONTENTS

Content **Page**

LIST OF TABLES

LIST OF ABBRIVATIONS

CLD- Customer Loyality Department

CRM- Customer Relation Management

CRISP-DM- CRoss Industry Standard Process

DBMS- Database Management System

OLAP- On-Line Analytical Processing

SQL- Strucvtured Query Language

EDDS- Electronic Data Distribution System

GLC- Golden Lion Club

ABSTRACT

Airlines are being pushed to understand and quickly respond to the individual needs and wants of their customers due to the dynamic and highly competitive nature of the industry. Most airlines use frequent flyer incentive programs and maintain a database of their frequent flyer customers to win the loyalty of their customers, by awarding points that entitle customers to various travel benefits.

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

This study is aimed at assessing the application of data mining techniques to support CRM activities at Ethiopian Airlines. The subject of this case study, the Ethiopian Airlines' frequent flyer program, has a database that contained individual flight activity and demographic information of over 35,000 program members.

Having the objective of filling the gap left by a related research, which was carried out by Henok (2002), this study has used the data mining database prepared by Henok (2002). In the course of using the database to attain the objective of this research, a data preparation tasks such as driving new attributes from the existing original attributes, defining new attributes and then preparing new data tables were done.

The data mining process in this research is divided into two major phases. During the first phase, since there has been an attempt to use three different data mining software, data was prepared and formatted into the appropriate format for the respective data mining software to be used.

The second phase, which is model building phase, was addressed in two sub-phases, the clustering sub-phase and the classification sub-phase, the major contribution of this study. In the clustering sub-phase the K-means clustering algorithm was used to segment individual customer records into clusters with similar behaviors. In the classification sub-phase, J4.8 and J4.8 PART algorithms were employed

to generate rules that were used to develop the predestined model that assigns new customer records into the corresponding segments.

As a final output of this research, a prototype of Customer Classification System is developed. The prototype enables to classify a new customer into one of the customer clusters, generate cluster results, search for a customer and find the cluster where the customer belongs, and also provides with the description of each customer clusters.

The results from this study were encouraging and confirmed the belief that applying data mining techniques could indeed support CRM activities at Ethiopian Airlines. In the future, more segmentation and classification studies by using a possible large amount of customer records with demographic information and employing other clustering and classification algorithms could yield better results

CHAPTER ONE

INTRODUCTION

1.1. Background

The focus of many service-providing industries has been towards a more evidently customer orientation. While competition is increasing, there is a need to better understand customers, and to quickly respond to their individual needs and wants. Thus, the objective of customer oriented industries became increasing the share of wallet for each individual customer and save costs by focusing on more targeted promotions.

The airline industry is one of the various service-providing industries. The international airline industry is complex, dynamic and subject to rapid change and innovation. The homogeneous nature of the airline product pushes airlines into making costly efforts to try to differentiate their product from that of their competitors. They do this by being first to introduce new aircraft types, by increasing their frequency of service, by spending more on in-flight catering and by advertising. Much of the promotions are aimed at trying to convince passengers that the product, which airlines offer, can be differentiated from that of their competitors.

In the airline industry, more targeted promotions are rather at a starting point. The prevalence of frequent flyer programs, for instance, makes targeted promotions easier and results in a wealth of data about the program's member customers. This in turn allows getting a better understanding of customer types and customer behavior. Being an incentive program, most airlines introduced the frequent flyer program to

results by making each marketing touch more appropriate and effective than is possible under any one-size-fits-all marketing scheme.

Customer Relationship Management, which is the overall process of exploiting customer-related information and using it to enhance the revenue flow from an existing customer, has become a strategic success factor for companies in all types of industries. Business organizations are continually faced with significant challenges in acquiring, engaging and retaining profitable customers. With rising customer expectations, these objectives are becoming more and more difficult to achieve.

In their interaction with customers, business organizations could gather transaction data. Customer transaction data alone is not enough to succeed in today's competitive environment. An organization must not only gather and store transaction data on individual customers, including purchase patterns, current status, contact history, demographic information, sales results and service trends; but that data must also be "actionable", so that managers and employees on the front lines can use it for decision support.

Consequently, customer focused organizations regard every record of transaction with a client or prospect as a learning opportunity. But learning requires more than simply gathering data. According to Berry and Linoff (1997), for learning to take place, data from many sources must first be gathered together and organized in a consistent and useful way or in short, the data has to be fed in to a *data warehouse.* Data warehousing allows the enterprise to remember what it has noticed about its customers. Next, the data must be analyzed, understood, and turned into actionable information. It is at this spot where data mining comes to play.

Ethiopian Airlines (*ETHIOPIAN*) currently has a FFP named "Sheba Miles" with over 35,000 members. Data pertaining to an individual member's flight activity details and personal details (like name, contact address, date-of-birth, etc) are stored in the program's database. Furthermore, the total member size increases as new members enroll and existing member's data is updated each time the member has a flight activity.

Based on the FFP, "ShebaMiles", a research on the possible application of data mining technique to support CRM activities at *ETHIOPIAN* has been done by Henok (2002). Henok has dealt on the application of data mining technology for customer segmentation on customers' data residing in the "ShebaMiles" database. According to Henok, the members of "ShebaMiles" can be clustered in to five clusters where customers in the same cluster share similar flight behavior than in a different cluster. Henok managed clustering the customers in to the five clusters by considering and selecting six (derived and original) attributes of the customers.

Henok's investigation was predestined only to get the possible customer segments based on the customers' data found in the "ShebaMiles" database (Henok, 2002). At the time of the investigation, Henok indicated that the size of the database members was about 22,000. It is observed that the number of the database members is increased by about 13,000 within a year, i.e., in 2003. However, for his investigation was predestined only to segment the customers, Henok did not proceed to deploy the data mining model that would enable to put new customers into one of the identified segments. The absence of such model could result in paying no attention to new customers that would fulfill the required behavior but not included in the already identified customer segments.

Thus, it seems sound and reasonable to do another investigation so as to fill the gaps left in the aforementioned research so that the deployment of the final data mining model would be effected. In this regard, the final clustering experiments that were done by Henok need to be repeated. Repeating these clustering experiments will enable to confirm the results found and be able to define meaningful variable values for the clusters. The values of the variables distinguishing the individual clusters were documented as if they were ordinal so that a new customer's record, which is entirely of quantitative values of variables, can not easily be put in to one of the clusters it might belong.

For instance, the centroids of the clusters identified were described using ordinal values (such as VH = Very High, L = Low, Lg = Long, R = Recent, etc.) of the variables selected for clustering. However, the corresponding numerical limits of these ordinal values of the variables were not documented so that implementing the final phase of the data mining process will not be easier. Therefore, repeating the final clustering experiments done by Henok and at the same time using his experience were believed to bring results that would facilitate successful completion of the data mining process.

It was, therefore, the aim of this investigation to build a data mining model that would enable to classify a new customer into one of the clusters identified in Henok's investigation. To come up with the data mining model, it was important to find the quantitative values of the cluster centroids of the clusters defined by qualitative values in Henok's investigation. To accomplish the clustering and then the classification tasks, the K-means algorithm, the See_5 algorithm (however not complete), and the J4.8 algorithm (which is the Weka implementation of See_4.5 or C4.5), were used.

1.2. Statement of the Problem and Justification

The airline industry can aptly be characterized as highly volatile and competitive, with unpredictable demand, variable pricing, and demanding customers. As the airline business becomes more competitive, gaining a competitive advantage becomes a common question of every airline industry to stay in the market. On the other hand, in such highly volatile and competitive environment, the homogeneous nature of the airlines business makes product differentiation very difficult and costly.

According to the IBM Corporation (2000), gaining competitive advantage is as clear as building close, long-term, and personalized relationships with high-value customers. This Corporation further pointed out that every customer likes to be treated as an individual.

As a result of their customers' behavior, airlines have, therefore, shifted their focus towards understanding their customers better so that they can quickly respond to the customers' individual needs and wants. In line with this, as a travel company, *ETHIOPIAN* needs to do the same to identify how important a customer is to its business, and treat that customer as an individual. This allows *ETHIOPIAN* to distinguish itself from its competitors, provide higher quality service, and thus retain today's customers for tomorrow's business.

As far as customer understanding is concerned, CRM is the key. Customer understanding involves *customer information* and *customer intelligence*. Customer information is related with collecting, organizing and cleaning all the data that the company has about its customers in a customer-centric view. Customer intelligence involves identifying trends and segments, assigning future values to customers,

and making use of this knowledge to develop and manage successful marketing and customer service programs (IBM Corporation, 2000).

Furthermore, CRM is also about using the customer information that a company holds to demonstrate its knowledge of the customer and treat them consistently at every contact point. In the airline industry, customer understanding (i.e., customer information and customer intelligence) can be gained through the introduction of a frequent flyer program (FFP), and use of the data of customers in the FFP database.

ETHIOPIAN, currently, has a FFP (called ShebaMiles) in order to increase and award the loyalty of its customers. The key program features are mileage accrua and mileage redemption, both on *ETHIOPIAN*. For this reason, the 'currency' of the FFP is miles, and this currency is used to identify customers with 'high value' and provide them with special benefits and services.

However, since the ShebaMiles mileage data does not include monitory measures, it could not be a good measure of customer profitability and may have less relevance for CRM. The *ETHIOPIAN* Customer Loyalty Department (CLD), which is managing the FFP, has switched to building a CRM environment. In relation to this, there has been an investigation contributed to CLD for its proper valuation of customers in response to CRM need of the department.

Carried out by Henok (2002), the investigation has identified possible customer segments based on the FFP members' data by applying data mining techniques. The resulting segments have been qualified using different original and derived attributes of the customers in the FFP database. The investigation has used the customers' data at the time of investigation, which amounts to about 22,000. Due to the

objective set for the investigation, further step to the next phase of the data mining process (the modeling phase) was not attempted, rather suggested.

However, from the survey made on the FFP (ShebaMiles) database, a year after the aforementioned investigation, the size of the member customers has increased by about 59%. This amounts the total customers in the database to about 35,000. As there is no model developed to classify the new customers in to the already identified clusters or segments, a timely marketing promotion is at stuck. This problem would drag the CLD back to the use of the 'old currency', mileage, which was reasonably disregarded as of no use to identify customers with their proper 'value'.

In this research an attempt has been made to develop a computer based customer-value revealing system or model, which enables the CLD classify and put a new customer into one of the pre-defined customer clusters. This model enables the CLD workers to use the massive historical customer data in the course of identifying the right customers for the right award at the time of selecting customers for various awards in response to targeted marketing promotion needs.

The primary objective of the customer-value system that has been attempted to develop in this research is not to predict customers' rejection from promotional rewards, rather to identify the right customers for the right reward.

In this investigation, the possible applications of data mining technique are used to predict customers' value based on different flight variables of the customers. Thus, the results of this data mining process could be used to differentiate customers with 'high value' from those with 'low value'. Moreover it could,

particularly, be used by the CLD of *ETIOPIAN* to distinguish whether valuable customers were being left unrewarded while less valuable ones enjoy privileges.

This investigation, being the application of data mining techniques on customer data at *ETHIOPIAN*, would make a contribution to strengthen *ETHIOPIAN*'s efforts to successfully implement CRM, and achieve a relatively sustainable competitive advantage.

Furthermore, the researcher believes that the classification and/or prediction model, which is the result of this study, could lead to a better understanding of airline customers' behavior, particularly in the African context. It is also the researcher's conviction that this study will, generally, contribute to the existing body of knowledge pertaining to the application of data mining techniques.

1.3. OBJECTIVES

1.3.1. General Objective

The general objective of this research is to support customer relationship management (CRM) activities at *ETHIOIAN* by employing appropriate data mining techniques on the frequent flyer customers' database in order to classify new customers into pre-defined clusters of customers.

1.3.2. Specific Objectives

The specific objectives are:

- Identify the type of customer data residing in the ShebaMiles database.
- Study and summarize the customer segments that have already been identified by Henok (2002).

- Collect data from the source(s) identified.
- Include additional important customer-data-attributes that were not considered by Henok(2002), if any.
- Prepare the data, for model building, by extracting and transforming the data into a format required for the data mining algorithm.
- Acquire appropriate data mining software that support clustering and classification.
- Apply clustering algorithm to check and meaningfully define the already identified customer segments.
- Apply classification algorithm to build and train a classification model, and test its performance.
- Build a customer-value informative system prototype using the model built
- Report the result and forward recommendations.

1.4. Research Methodology

In order to build good data mining models for a CRM system, there are a number of steps that one must follow (Edelstein, 2000). Accordingly, in the course of this investigation, the following steps were used.

1.4.1. Review of Related Literature

A review of relevant literature has been conducted to assess data mining technology, both concepts and techniques, and researches in this field. Various books, journals, magazines, articles, and papers from the Internet pertaining to the subject matter of data mining and Knowledge Discovery Process in

Databases (KDD) have been reviewed to understand the potential applicability of data mining in the practice of customer relationship management (CRM), particularly CRM in the airline industry.

1.4.2. Fact Finding Methods

To identify and analyze the business problems through making the business survey, fact-finding methods, i.e., observation and interviews have been made. As far as solving the business problems that were identified, the following steps (mentioned in 1.4.3 – 1.4.7) were considered to develop a model, employing data mining techniques.

1.4.3. Identifying Available Data Sources

The potential source of data used to undertake this research was mainly the *ETHIOPIAN*'s FFP database (ShebaMiles DB), which contains data pertaining to about 35000 members of ShebaMiles. ShebaMiles DB contains information on each of the flight segments the members flew, including origin and destination cities, booking class information, base mileage points awarded, and the members' frequent-flier number. In addition, each member's demographic data and his/her current status in the program are also available.

The other source of information is the revenue accounting database. Since ShebaMilesDB is basically used only for administrative purposes, unlike the availability of each member's flight activity, there was no corresponding revenue information. In order to assign revenue data to each flight activity in ShebaMiles DB, the corresponding revenue values were extracted from a revenue accounting database, where revenue information regarding individual flight activity is available.

1.4.4. Data Collection and Preparation

The primary sources of data for this research were the ShebaMiles and the revenue accounting databases that had been used by Henok (2002) for the customer segmentation experiments. Therefore, these databases were thoroughly studied and, as a result, four basic relations (tables) were found to be important but in different database formats.

However, for his segmentation experiments, Henok had to prepare a single table that had to be used by the Knowledge Studio software. In response to this, all the different tables had to be put together in a single database file so that preparing a single table would be possible. To this end, Henok had applied a number of procedures on the different database files and changed all the files to the same, MS Access, file format. Then he had put all the files together in a single MS Access database.

To avoid duplication of effort on what has already been done, the MS Access database that was put together by Henok was found useful and used for the current research.

In the data preparation phase, before directly getting in to the modeling process, two different file formats of the final data sets were prepared for two different data mining tools, Knowledge Studio and Weka.

The first format of the final data set was the MS Access table, to be fed into the Knowledge Studio Software, with the required field names. Since the Knowledge Studio had Open Database Connection (ODBC) facilities, the final data set table was directly imported from the MS Access database, without

changing it into other format. The second file was a *Comma Delimited* [2] text file in *ARFF* [3] format of the same final data set, to be fed into the Weka software.

1.4.5. Training and Building the Model

The model building phase in the data mining process of this investigation was carried out in two sub-phases, clustering and then classification rule generation.

Rules generated for classification purpose being the required final output of the model building phase, due to the reasons mentioned in section 5.6 of chapter five, the clustering sub-phase was inevitable step. Since this investigation was for the second time that segmentation at a member level was being done, there was an experience about how the segments could be found. So, the clustering run parameters for the investigation at hand were directly adapted from the previous investigation. To manage the clustering sub-phase, the Knowledge Studio and Weka software were used.

To carry out the classification sub-phase, two different software, Weka and See_5 (release 1.5) had been attempted to use. Since the See_5 was found to be a trial version, it couldn't process the total record used for the classification task. As a result, the classification model built using See_5 was not used in any further discussion. Therefore, further discussion in this investigation was done by making use of the models found from the selected Weka's algorithms runs.

---

[2] *Comma Delimited* applies for a list of records where the items are separated by commas.

[3] *ARFF* is an extension of a file format that the Weka software can read.

1.4.6. Performance Evaluation (Testing) the Model

Due to the reason mentioned in section 5.6.2 of chapter five, the type of classification model selected to be built was decision tree. In the classification sub-phase, a number of experiments with different setups were carried out and selections of a final decision tree with its corresponding extracted rule were made.

In the course of selecting a decision tree and the corresponding extracted rules, the different experiments were carried out using two different schemes (for attribute selection) and three different test modes (ways of feeding records to the algorithms), as briefly mentioned in section 5.6.2 of the experimental chapter.

To judge about the performance of the resulting classification model, the three test modes used were inputting:

- 100% of the records to the selected algorithm without any test,
- 100% of the records, but with a 10-fold cross-validation mode, and
- 75% of the records to train a model and then supply the unseen 25% of the record for testing the performance of the model.

1.4.7. Prototype Development

In this study an attempt was made to develop an operational application prototype named Customer Classification System that uses the selected classification rules generated from the decision tree learner in the classification sub-phase. The prototype classifies a customer into one of the customer clusters, provides with the description of each customer clusters, generates cluster reports, search for a customer and find the cluster where the customer belongs.

1.5. Scope and Limitation

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 classification model, and includes the deployment of only a prototype of the classification model.

Budget being a restrictive factor, the acquisition of appropriate data mining software for the study from the different data mining software available on market was not feasible. This limited the research to make use of classification algorithms supported by the Weka software, which was found through personal contact.

1.6. Research Contribution

The results from this research are meant to support the strategic decisions made by the Customer Loyalty Department (CLD) at *ETHIOPIAN*. In other words, it makes a contribution to strengthen *ETHIOPIAN*'s efforts to successfully implement CRM, and achieve a sustainable competitive advantage.

In general, the model, which is the result of this study, would enable CLD of *ETHIOPIAN* classify new customers in to the pre-defined segments so that different products and services, which are appealing to members of the specified group, can be produced and awarded accordingly. As a result, the marketing division can run specific and timely promotions that increase the average flight segments made by that customer group, thereby lowering marketing costs and increasing profitability through more effective and at-the-spot promotion and an improvement in customer satisfaction.

Apart from the specific problem area, the ShebaMiles FFP of *ETHIOPIAN,* this research may contribute a kind of experience that could be used to replicate the steps and develop a similar system in other customer oriented organizations.

1.7. Thesis Organization

This research report is organized into six chapters. The first Chapter briefly discusses background to the problem area, and states the problem, the general and specific objectives of the study, the research methodology, the scope and limitation, and application of the results of the research. Chapter two and three review the data mining technology and the customer relationship management respectively. The concepts pertaining to the data mining technology and its application in the problem are reviewed in chapter two. Chapter three is dedicated for the discussion of two basic issues, customer relationship management and survey of the frequent flyer program at *ETHIOPIAN* (the problem area). Chapter four explains the methods, the clustering techniques and decision trees, used in this research. Moreover the previous research, in which a researchable gap was left and become the concern of this research, is reviewed in this chapter, chapter four. Chapter five presents the experimentation phase of the study at hand. Results of the clustering and classification experiments were also discussed here. Finally, chapter six provides conclusion of the research, and also presents recommendation for future work.

CHAPTER TWO

DATA MINING

2.1. Introduction

Recently, capabilities of both generating and collecting data have been increasing rapidly. The ever increasing and wide spread of inexpensive computers has enabled organizations collect and store large volume of data. It is noted that organizational databases keep growing in number and size due to the availability of powerful and affordable database systems.

As a result of the explosive growth in data and databases, an urgent need for new techniques and tools has been generated. According to Fayyad, U. et al. (1996), these tools should assist humans to intelligently and automatically identify patterns; transform the processed data into useful information; and then extract knowledge from the rapidly growing volumes of digital data. Fayyad, U. et al. (ibid) further noted that these theories and tools are the subject of the new emerging field of knowledge discovery in databases (KDD).

The basic problem addressed by the KDD process is the mapping of low-level data, which are too voluminous to understand and digest easily, into other forms that might be more compact, more abstract, or more useful. At the core of the process is the application of specific data-mining methods for pattern discovery and extraction (Fayyad, U. et al., 1996).

2.2. Data Mining

Edelstein (1998) share the idea that most companies have accumulated large data while what they need is information. He then remarked that the newest and hottest technology to address these concerns is data mining. As described by Kurt Thearling (2000), data mining is the extraction of hidden predictive information from large databases; and is a powerful new technology with great potential to help companies focus on the most important information in their *data warehouses*[1].

Connolly, et al. (1999) viewed data mining as 'the process of extracting valid, previously unknown, comprehensible and actionable information from large databases and using it to make crucial business decisions.' The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. In this regard, Kurt Thearling (2000) notes that data mining tools can answer business questions that traditionally were too time-consuming to resolve. These tools scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

According to Chen, M.S. et al. (1997), data mining, which is also referred to as *knowledge discovery in databases (KDD),* is defined as a process of extracting nontrivial, implicit, previously unknown and potentially useful information (such as knowledge, rules, constraints, regularities) from data in databases. As observed in this definition, many authors use the phrases *data mining* and *knowledge discovery in databases* (KDD) interchangeably. The next section, therefore, gives a brief summary of the different views and usages of these phrases by different authors.

---

1. Data warehouse is a logical collection of information gathered from many different operational databases that supports business analysis activities and decision making tasks (Haag, S. et al., 1998).

2.3. Data mining and Knowledge Discovery in Databases (KDD)

Historically, the notion of finding useful patterns from data had been given a variety of names, including data mining, knowledge transaction, information discovery, information harvesting, data archeology, and data pattern processing. As Fayyad, U. et al. (1996) noted, the term data mining has mostly been used by statisticians, data analysts, and the management information system (MIS) communities. It has also gained popularity in the database field.

According to Piatetsky-Shapiro (1991), cited in (Fayyad, U. et al., 1996), the phrase Knowledge Discovery in Databases was coined at the first KDD workshop in 1989 to emphasize that knowledge is the end product of a data-driven discovery. It has been popularized in the Artificial Intelligence (AI) and machine learning fields.

Using the words of Fayyad, U. et al., cited in (Goebel and Le Gruenwald, 1999), a simple definition of KDD is; 'Knowledge discovery in databases is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.' Moreover, according to Frawley, W., et al., (1991), cited in (Carbone, 1999), Knowledge discovery is 'the non-trivial extraction of implicit, previously unknown, and potentially useful information from data'.

Many people treat data mining as a synonym for the phrase Knowledge Discovery in Databases or KDD. Alternatively others view data mining as simply an essential step in the process of knowledge discovery in databases. Han and Kamber (2001), Goebel and Le Gruenwald (1999), and Fayyad, U. et al. (1996) agree to the second view that KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. More specifically, according to

Brachman, R., and Anand quoted in (Goebel and Le Gruenwald, 1999), although at the core of the knowledge discovery process, the data mining step usually takes only a small part (estimated at 15% to 25 %) of the overall effort..

The data-mining component of the KDD process is the application of specific algorithms for extracting patterns from data and heavily relies on known techniques from machine learning, pattern recognition, and statistics. Although these fields provide some of the data-mining methods, KDD focuses on the overall process of knowledge discovery from data. These focuses of KDD include how the data are stored and accessed; how algorithms can be scaled to massive data sets and still run efficiently; how results can be interpreted and visualized; and how the overall man-machine interaction can usefully be modeled and supported (Fayyad, U. et al., 1996).

By reason of the popularity of the term data mining than the longer term Knowledge Discovery in Databases, Han and Kamber (2001) favored to adapt the broader view of data mining functionality, and defined data mining as ' the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repository.

Thus as to the usage of these two phrases, 'data mining' and 'knowledge discovery in databases', the broader view as it has been adapted and defined by Han and Kamper (2001) is adapted in this research. The reasons behind this adaption are, being consistent with major data mining projects, use the corresponding experiences, and avoid any confusion between the two phrases, 'data mining' and 'knowledge discovery in databases'.

Consequently, in this research, the terms data mining and knowledge discovery process will be used interchangeably. In line with this, the discussion of the data mining process is briefly summarized in the next section, section 2.3.1. The data mining process and the usage of the terms have been summarized from the steps followed in *the knowledge discovery process in databases* of Fayyad, U. et al. (1996) and that of the CRoss-Industry Standard Process for Data Mining (CRISP-DM) model of SPSS (Chapman, P. et al., 1999, 2000).

2.3.1. The Knowledge discovery process

The knowledge discovery process is iterative, and involves numerous steps with many decisions made by the user. This iterative process has been summarized by many researchers and the area professionals. Most of them agree that knowledge discovery process starts with a clear definition of the business problem or, equivalently, understanding of the application domain (Han and Kamber, 2001; Chapman, P. et al., 1999, 2000; Two Crows Corporation, 1999; Berry and Lineoff, 1997).

The process model for data mining (or .knowledge discovery process) provides an overview of the life cycle of a data mining project. It contains the phases of a project, their respective tasks and relationships between these tasks. At this description level, it is not possible to identify all relationships. Essentially, relationships could exist between any data mining tasks depending on the goals, the background and interest of the user and most importantly on the data (Two Crows Corporation, 1999).

The life cycle of a data mining project consists of six phases. Figure 2.1 shows the phases of a data mining process. The sequence of the phases is not rigid. Moving back and forth between different phases is always required. It depends on the outcome of each phase, for which phase or which particular task of

a phase, has to be performed next. The arrows indicate the most important and frequent dependencies between phases.

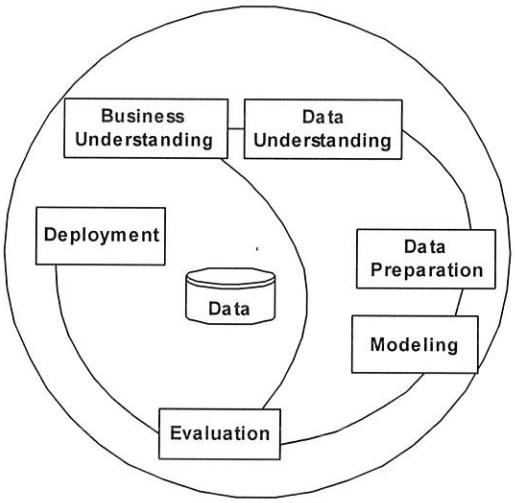


Figure 2.1: Phases of the CRISP-DM process cycle

The outer circle in Figure 2.1 symbolizes the cyclical nature of data mining itself. Data mining is not over once a solution is deployed. The lessons learned during the process and from the deployed solution can trigger new, often more focused business questions. Subsequent data mining processes will benefit from the experiences of previous ones.

Figure 2.1 shows the commonly agreed steps of knowledge discovery process. These steps are briefly summarized below.

Business understanding

This initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

Data understanding

The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

Data preparation

The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record and attribute selection as well as transformation and cleaning of data for modeling tools.

Modeling

In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often necessary.

Evaluation

At this stage of the project, one has to build a model (or models) that appear(s) to have high quality from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model and review the steps executed to construct the model to be certain it properly achieves the business objectives.

A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

Deployment

Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. It often involves applying "live" models within an organization's decision making processes. However, depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise. In many cases it is the user, not the data analyst, who carries out the deployment steps. However, even if the analyst will not carry out the deployment effort it is important for the user to understand up front what actions need to be carried out in order to actually make use of the created models.

2.4. Data mining and database management

Database management systems (DBMSs) provide a number of essential capabilities to data mining. These capabilities include persistent storage, a data model (e.g., relational or object-oriented), and a high-level query language, which allows users to request what data rather than how to access it (e.g., SQL). In addition, DBMS's provide transaction management and constraint enforcement to help preserve the integrity of the data. Database technology also provides efficient access to large quantities of data.

The KDD process implies that one is performing knowledge discovery against data which reside in one or more large databases. Typical properties of databases that complicate knowledge discovery include large volume; noise and redundancy; dynamicity of data; sparseness of data; multimedia data.

Recent advances in data warehousing – parallel databases – and on-line analytical processing tools have greatly increased the efficiency with which databases can support the large numbers of extremely complex queries that are typical of data mining applications. Finally, databases provide a metadata description that can be used to help understand the data, which is to be mined, and also aid in determining how the database should potentially change based on what has been learned.

2.5. Data mining and data warehousing

Since the introduction of computers into data processing centers in the 1960's, virtually every operational system in business has been automated. In companies, this automation has resulted in scads of data residing in dozens of disparate systems. In response to integrating the disparate systems, data

warehousing is the process of bringing diverse data together from throughout an organization for decision-support purposes (Berry and Lineoff, 1997).

Most of the time, the data to be mined is first extracted from an enterprise data warehouse into a data mining database or data mart. There is some real benefit if the data mining database is already part of a data warehouse. A data mining endeavor includes the effort to identify, acquire, and cleanse data. The problems of cleansing data for a data warehouse and for data mining are very similar. If the data has already been cleansed for a data warehouse, then it most likely will not need further cleaning in order to be mined (Two Crows Corporation, 1999; Berry and Lineoff, 1997).

Better yet, if the design of the data warehouse includes support for data mining applications, the warehouse facilitates and catalyzes the data mining efforts. In this case, the data mining database is a logical rather than a physical subset of the data warehouse. The two technologies work together to deliver value. According to Berry and Lineoff (1997), data mining fulfills much of the promise of data warehousing by converting an essentially inert source of data in to actionable information.

As useful as a data warehouse is, it is not the prerequisite for data mining and data analysis. Setting up a large data warehouse consolidates data from multiple sources, resolves data integrity problems, and loads the data into a query database. However, putting such a large database up can be an enormous task, sometimes taking years and costing millions of dollars. One could, however, mine data from one or more operational or transactional databases by simply extracting it into a read-only database. This new database functions as a type of data mart (or data mining database).

For the case of this investigation, an independent data mining database (a data mart) has been set up. This data mart is a database extracted form the different operational and transactional databases that are pertinent to the different services provided to the *ETHIOPIAN* customers, mainly of the ShebaMiles members.

2.6. Data mining and On-Line Analytical Processing (OLAP)

One of the most common questions of data processing professionals is about the difference between data mining and On-Line Analytical Processing (OLAP). Data mining and OLAP are very different tools that can complement each other (Graettinger, 1999; Two Crows Corporation, 1999; Edelstein, 1998)

OLAP was a term coined by E. F. Codd as quoted in (Fayyad, U. et al., 1996), and was defined by him as 'the dynamic synthesis, analysis and consolidation of large volumes of multidimensional data'. OLAP is a popular approach for analysis of data warehouses. It focuses on providing multidimensional data analysis, which is superior to Structured Query Language (SQL) in computing summaries and breakdowns along many dimensions.

Traditional query and report tools, or 'Information processing tools' to use the words of Han and Kamber (2001), describe what is in a database. OLAP goes further; it's used to answer why certain things are true. The user forms a hypothesis about a relationship and verifies it with a series of queries against the data. In other words, the OLAP analyst generates a series of hypothetical patterns and relationships and uses queries against the database to verify them or disprove them. Where as, data mining, instead of verifying hypothetical patterns, uses the data itself to uncover such patterns. In precise terms, when a user employs query tools such as OLAP to explore data, the user guides the exploration and when the

user employs data mining tools to explore data, the tools perform the exploration (Two Crows Corporation, 1999; Edelstein, 1998).

Han and Kamber (2001), and Fayyad, U. et al. (1996) also have the same position to the above fact that OLAP tools are targeted towards simplifying and supporting interactive data analysis, but the goal of data mining tools is to automate as much of the process as possible. Thus data mining is a step beyond what is currently supported by most standard database systems.

OLAP analysis is essentially a deductive process while data mining is an inductive process. When the number of variables being analyzed is in the dozens or even hundreds, it becomes much more difficult and time-consuming to find a good hypothesis and analyze the database with OLAP to verify or disprove it.

Since data mining involves more automated and deeper analysis than OLAP, data mining is expected to have broader applications. However, data mining does not replace but rather complements and interlocks with other decision support system capabilities such as OLAP. Before acting on a pattern, the analyst needs to know what the resulting implications would be of using the discovered pattern to govern the decision. The OLAP tool can allow the analyst to answer those kinds of questions. Furthermore, OLAP is also complementary in the early stages of the knowledge discovery (data mining) process because it can help to explore the data, for instance, by focusing attention on important variables, identifying exceptions, or finding interactions. This is important because the better one understands the data, the more effective the knowledge discovery (data mining) process will be (Two Crows Corporation, 1999; Edelstein, 1998; Fayyad, U. et al., 1996).

2.7. Data Mining, Artificial Intelligence (AI) and Statistics

Since the Renaissance, people have been looking at the world and gathering data to explain natural phenomena such as the movements of the sun, the moon, and the stars and created calendars to describe heavenly events. They managed to analyze data and look for patterns without the aid of computers even before recorded history began (Berry and Lineoff, 1997).

What started to change in the past few centuries have been the codification of the mathematics and the creation of machines to facilitate the taking of measurements, their storage, and their analysis. As Berry and Lineoff (1997) noted, traditional statistics has developed over the past two centuries to help scientists, engineers, and later business analysts to make sense of the data they have collected.

The history of data mining and data mining techniques is generally rather different, highlighting the influence of other disciplines. Data mining takes advantage of advances in the fields of artificial intelligence (AI) and statistics. Both disciplines have been working on problems of pattern recognition and classification, and have made great contributions to the understanding and application of neural nets and decision trees (Two Crows Corporation, 1999; Berry and Lineoff, 1997).

Statistics is very useful in providing a language and framework for quantifying the uncertainty, which results when one tries to infer general patterns from a particular sample of an overall population. However, it does not solve all data mining problems. Moreover, the computational complexity of statistical approaches does not grow well with larger data sets. On the other hand, data mining does not replace traditional statistical techniques. Rather, it is an extension of statistical methods that is in part the result of a major change in the statistics community.

The development of most statistical techniques was, until recently, based on elegant theory and analytical methods that worked quite well on the modest amounts of data being analyzed. The increased power of computers and their lower cost, coupled with the need to analyze enormous data sets with millions of rows, have allowed the development of new techniques with a powerful exploration of possible solutions (Two Crows Corporation, 1999; Edelstein, 1998; Berry and Lineoff, 1997; Fayyad, U. et al., 1996).

Recent techniques include relatively new algorithms like neural nets and decision trees, and new approaches to older algorithms. By virtue of bringing to bear the increased computer power on the huge volumes of available data, these techniques can approximate almost any functional form or interaction on their own. Traditional statistical techniques rely on the modeler to specify the functional form and interactions.

The key point is that data mining is the application of AI and statistical techniques to common business problems. Data mining is, therefore, a tool for increasing the productivity of people trying to build predictive models by making AI and statistical techniques available to the skilled knowledge workers as well as the trained professionals (Berry and Lineoff, 1997).

2.8. Data Mining Activities

According to Fayyad, U. et al. (1996), the two high-level primary goals of data mining in practice tend to be prediction and description. As per these authors, prediction involves using some variables or fields in the database to predict unknown or future values of other variables of interest, and description focuses on finding human-interpretable patterns describing the data. The goals of prediction and description can

be achieved using a variety of data mining methods. In response to this, more often than not, a data mining project involves a combination of different activities, which together solve a business problem.

Various problems of intellectual, economic, and business interest can be articulated in terms of the different data mining activities such as data description, clustering (or segmentation), concept (or class) description, classification, prediction (or regression), association analysis (or affinity grouping). Detailed and brief summaries of these commonly known and used data mining activities, and their corresponding applications can be read from the works of Han and Kamber, 2001; Chapman, P. et al., 1999, 2000; Two Crows Corporation, 1999; Berry and Lineoff, 1997; Fayyad, U. et al., 1996.

However, as the investigation at hand is directly related with clustering (segmentation), and is mainly concerned about classification, more emphasis is given to clustering and classification. Because of this a detailed discussion of this data mining activities is presented in section 4.2 and 4.3 of chapter four, respectively.

2.9. Application of Data Mining Technology

Data mining has been providing substantial contribution to the business environment, and becoming increasingly popular. It can be used to control costs as well as contribute to revenue increases (Two Crows Corporation, 1999).

Many organizations are using data mining to help manage all phases of the customer life cycle, including acquiring new customers, increasing revenue from existing customers, and retaining good customers. By determining characteristics of good customers (profiling), a company can target prospects with similar

characteristics. By profiling customers who have bought a particular product it can focus attention on similar customers who have not bought that product (cross-selling). By profiling customers who have left, a company can act to retain customers who are at risk for leaving (reducing churn or attrition), because it is usually far less expensive to retain a customer than acquire a new one (Two Crows Corporation, 1999).

As Two Crows Corporation (1999) noted, data mining offers value across a broad spectrum of industries. Telecommunications and credit card companies are two of the leaders in applying data mining to detect fraudulent use of their services(Fayyad, U. et al., 1996).

Insurance companies and stock exchanges are also interested in applying data mining technology to reduce fraud, to predict which customers will buy new policies, to identify behavior patterns of risky customers, identify fraudulent behavior (Two Crows Corporation, 1999). For instance, with the objective of developing a predictive model in support of insurance risk assessment, Tesfaye (2002) has applied the data mining technology and developed a prototype, named MIRS (Motor Insurance Renewal System).

Medical applications are another fruitful area: data mining can be used to predict the effectiveness of surgical procedures, medical tests or medications, to predict which medical procedures are claimed together, to characterize patient behavior so as to predict office visits, and to identify successful medical therapies for different illnesses. In this regard, Shegaw (2002), having the objective of developing a model that can support in preventing and controlling child mortality at the district of Butajira, Ethiopia, has applied the data mining technology and reported an achievement in the development of the model.

Companies active in the financial markets use data mining to determine market and industry characteristics as well as to predict individual company and stock performance (Two Crows Corporation, 1999). For instance, in the banking sector, data mining enables to detect patterns of fraudulent credit card use, identify 'loyal' customers, predict customers likely to change their credit card affiliation, determine credit card spending by customer groups, find hidden correlations between different financial indicators, identify stock trading rules from historical market data. With the objective of developing a model that can support the loan decision-making process at Dashen Bank S.C., Askale (2001), for instance has explored the potential applicability of data mining technology in the banking sector.

In marketing, the primary application of data mining is database marketing, which analyze customer database to identity different customer groups and forecast their behavior. Moreover, in this sector, retailers are making use of data mining to decide which products to stock in particular stores (and even how to place them within a store, i.e., to make market basket analysis), to assess the effectiveness of promotions and coupons, identify buying patterns from customers, to find associations among customer demographic characteristics (Two Crows Corporation, 1999; Fayyad, U. et al., 1996).

According to Two Crows Corporation (1999), pharmaceutical firms are also mining large databases of chemical compounds and of genetic material to discover substances that might be candidates for development as agents for the treatments of disease.

Information retrieval has typically been concerned with finding better techniques to query for and retrieve textual documents based on their content. Data mining is being applied to this area so that the vast

amounts of electronic publications currently available may be brought to users' attention in a more efficient manner (Carbone, 1997).

Data mining and information retrieval are being merged, according to Carbone, (1997), to provide a more intelligent "push" of information to a user. Information retrieval techniques have included the use of a user profile to help focus a search for pertinent documents. The addition of data mining techniques to the creation of a profile is currently being researched to improve the documents that are retrieved or brought to the user's attention.

In addition to the above areas, the data mining technology has many potential applications in other sectors such as transportation to determine the distribution schedules among outlets, and analyze loading patterns. As far as the transportation sector is concerned, the airline industry has drawn a considerable attention of a good number of data miners. Since the research at hand is exploring the potential application of data mining in the airline industry, the next sub-section reviews the possible applications of data mining in this industry.

2.9.1. Application of Data Mining in the Airline Industry

According to Feyen (n.d.), the most obvious applications of data mining in the airline business are related to frequent flyer programs. In a study conducted for a major European airlines alliance group (Qualiflier), Feyen et.al. (n.d.) noted 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 was to identify market segments containing customers with high profit potential.

The segments, according to Feyen (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 identifying core customers. The segments found can therefore be used for special marketing strategies.

According to Harris (n.d.), British Airways analyzed customer data to discover instances where a 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.

Electronic Data Distribution Service (EDDS) (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 alleviate the problem accordingly.

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.

CHAPTER THREE

CUSTOMER RELATIONSHIP MANAGEMENT

3.1. Loyalty and Customer Relationship Management

3.1.1 Overview

According to the Oxford English Dictionary (1997), Loyalty is defined as a true and faithful act or behavior. 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. In this regard, Edelstein (2000) notes, for almost every company, the cost of acquiring a new customer exceeds the cost of keeping good customers. However, recent studies indicate that despite heavy investments in customer satisfaction efforts and rewards programs, loyalty remains a vague 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 to whom the best value can be delivered by the firm over a sustained period of time (Reichheld, 1995). In a loyalty-based marketing, companies study their customers' database and segment it into those who are highly loyal and those who are less loyal and then 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. It is believed that the recent trend is about letting the

customers know 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).

There is a difference between loyalty/reward programs, and CRM. Loyalty/reward program is concerned with rewarding behavior that is assumed to be loyal, while CRM is concerned with managing behavior to create loyalty. In addition, loyalty/reward program deals with creating value for a customer, while CRM with developing value from a customer, and that the real value of CRM is when the company earns loyalty without reward.

3.1.2 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).

Under FFPs, 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. These reward programs were successful in creating a form of loyalty. But, from many of such programs have found that airlines are only as valuable to their customers as the last major awards.

One shortcoming of FFPs is that customers tend to fixate on the rewards. Consequently, product superiority becomes less of a priority. Moreover, with many of such programs, one reward is generally as good as another and creates cost for the company with no sustainable competitive advantage.

American Airlines Inc. were pioneers in launching AAdvantage, the first true FFP in the airline industry. Since the launching of AAdvantage, other airlines started to emulate 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.

Airlines often spend 3 to 6 percent of their revenue on frequent flyer programs (FFPs), which are concerned with rewarding behavior that is assumed to be loyal, compared to 3 percent on advertising (Chandler, 2001). However, frequency programs alone do not produce a very good return on investment if the airlines' aim is to retain their top customers.

Petersen (as quoted by Chandler, 2001) notes, 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 accumulated are not the measure of a good CRM program.

It is widely shared that loyalty programs could be an entrée into CRM, while frequency programs alone are not. Chandler (2001) believes that frequency programs are not loyalty programs; but legitimate loyalty programs often lead to CRM. Chandler further discusses, the primary focus of frequency programs is to build repeat business, while for loyalty programs, the focus is to build an emotional attachment to the brand.

More focused and more productive promotions are among the advantages 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. It will then be possible to run more targeted promotions geared towards the different customer segments. In addition, the new customer insight can be used to improve customer services.

According to Feyen (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 getting a better understanding of customer types and customer behavior.

3.2. Survey of the Frequent Flyer Program of *ETHIOPIAN*

This survey aims to assess the customer relationship management (CRM) process at *ETHIOPIAN*. The purpose is to conduct an analysis of the current CRM process, to identify the critical functions and activities involved, to review a research contribution that has been done in relation to CRM in *ETHIOPIAN*, and to identify and assess the available data sources that can support to derive customer classification model.

Ethiopian Airlines, *ETHIOPIAN*, launched its frequent flyer program (FFP) named "ShebaMiles" in January, 1999 with an objective to increase and award loyalty of customers. The name ShebaMiles is inspired by the legend of Makeda, the Queen of Sheba, who ruled Ethiopia around the 10th century B.C. in particular, by her famous trek across desert and sea to visit King Solomon in Jerusalem. (ShebaMiles Membership Application, n.d.).

Prior to launching ShebaMiles, *ETHIOPIAN* was running another program, called Golden Lion Club (GLC for short), that recognized its esteemed customers. Members of the GLC program (which no more exists)

included prominent personalities, both from government and the business community. GLC members were entitled to various benefits such as, separate check-in at the airport, access to special lounges, and priority in having their flight reservations confirmed as well as being upgraded to a higher class of service. Travel frequency was not a measure of this program.

By the time when ShebaMiles was launched, GLC members automatically became members of ShebaMiles, a frequent flyer program developed to reward individuals traveling frequently with *ETHIOPIAN.* This program has been 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.

To award different benefits and provide services to its members, the ShebaMiles program has terms and conditions that its members should follow. The three basic terms and conditions under which related rules have been declared are enrollment, termination of an individual membership, and also mileage accumulation rules. For instance, only individual persons aged 12 years or over, can be enrolled as a Member; and if a member accumulates less than 10,000 miles during the 12 months period following the enrollment date, the member is considered as he/she terminated the award program(ShebaMiles Membership Application, n.d.).

Currently, ShebaMiles has over 35,000 enrolled members. These members can earn two types of miles, *Base Miles* and *Bonus Mile*s and fall in one of the three Club levels of the program, which are Blue, Silver and Gold. Base Miles refer to the number of miles a passenger flies with *ETHIOPIAN*, and is awarded for the sector flown. Membership level is determined by the number of Base Miles flown annually (see table

3.1, which is extracted from ShebaMiles Membership Application (n.d.)). 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.

Members' club	Required Base Miles per year	*Awards Entitled*
Blue	Only enrollment as a member	*Bonus Miles and no enrollment fee to become a member of ShebaMiles and club upgrading when eligible*
Silver	> 25,000	*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.*
Gold	*> 50,000*	*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*

Table 3.1: The three Club levels of shebamiles, eligibility and awards

In *ETHIOPIAN,* the department responsible for running ShebaMiles is the Customer Loyalty Department (CLD). This department falls under the Market Development Division. Headed by a department manager, CLD engages in developing, coordinating and directing all activities pertaining to keeping the loyalty of customers. This department is responsible for the administration and all other related activities of the frequent flyer program, ShebaMiles.

3.2.1 Business Processes of the Frequent Flyer Program

Business processes are a set of activities that transform a set of inputs into a set of outputs (goods or services) for another person of process using people and tools. CLD being the central office running the frequent flyer program, it interacts with *frontline customer services* 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 Airlines lounges, etc. This survey has yielded Figure 3.1 (adapted from Henok, 2002) that depicts the overall flow of the program.

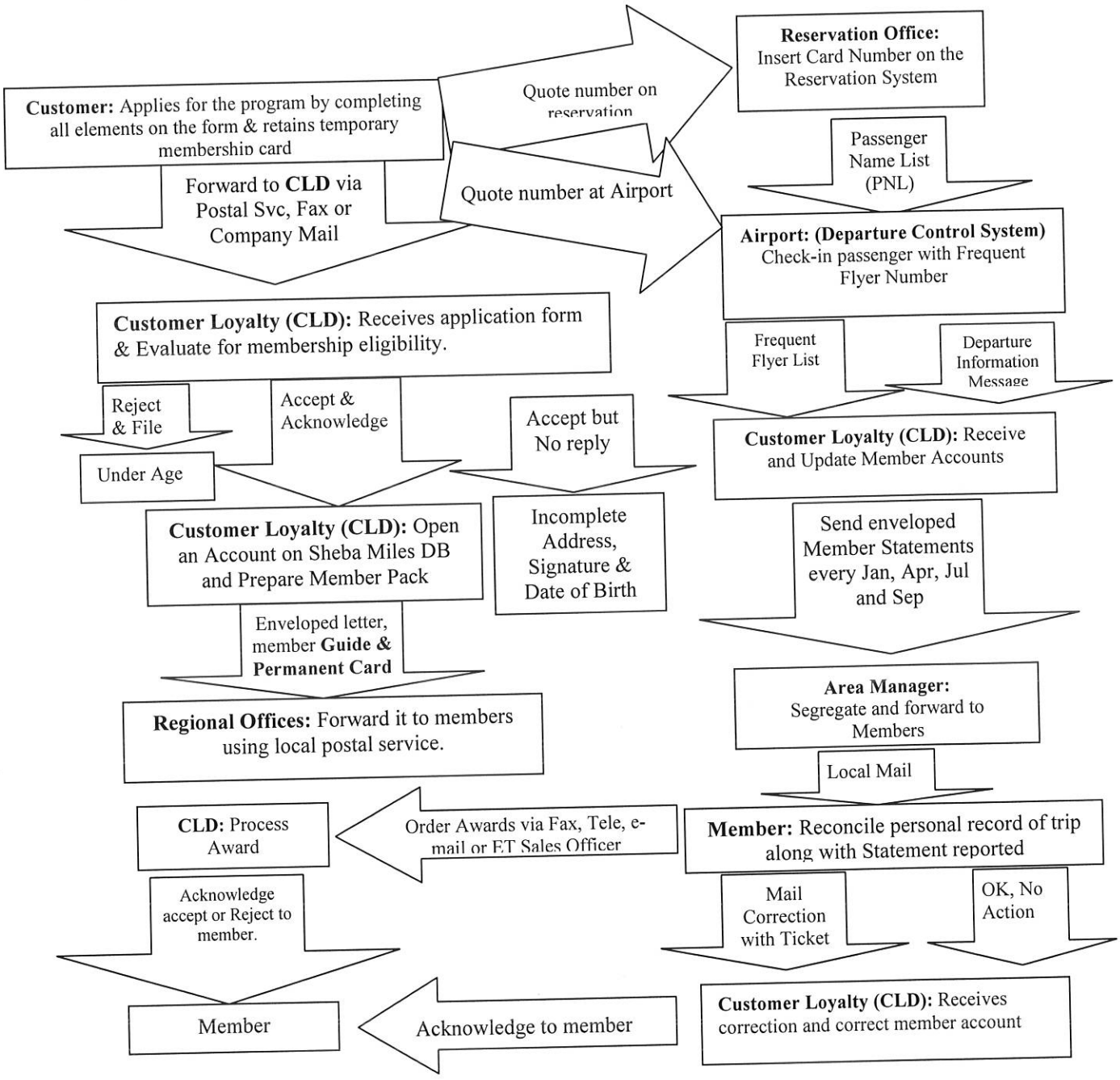


Figure 3.1: Business process of the ShebaMiles FFP Program at ETHIOPIAN

3.2.2 Overview of ShebaMiles' Database System

The CLD maintains a database in order to store information and manage the program. Data pertaining to a ShebaMiles member activity is obtained through a customized interface to the airline's as well as other automated data communication systems (DCS). 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). Figure 3.2, adapted from the Loyalty Management System (LMS) user guide, depicts the data flow of ShebaMiles' database system.

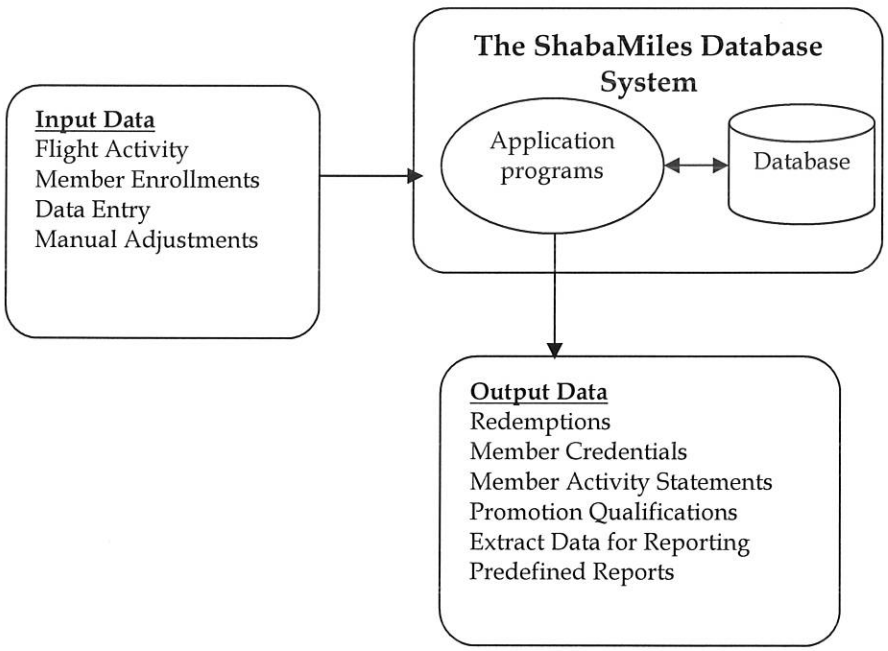


Figure 3.2: The data flow of ShebaMiles' database system

The main purposes of the ShebaMiles DB system are:

- 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 fulfil these purposes, 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 for viewing by each member.

3.2.3 Findings of the Survey

The CLD is in the process of building a CRM environment at *ETHIOPIAN*. In response to this, an investigation had been done and identified the possible customer segments of the ShebaMiles members. However, the resulting customer segments were applicable only for the customers' records that had been considered at the time of the investigation. The segments were not further used to build a model that enables the CLD classify new customer records. The department, therefore, wants to enrich its

knowledge of ShebaMiles members, latest or earliest, so that it could run more targeted promotions and improve customer services activities.

The ShebaMiles database, which runs on a Microsoft Visual Foxpro database management system, contains over 35,000 program members which have accumulated over 135,000 flight activities over the four years' lifetime of the program. However, the first investigation, by Henok (2002), has considered 22,000 of the program members and their 90,000 flight activities. The indicated differences in the number of program members and their corresponding flight activities are considered in one year.

Therefore, rather than going through all the segmentation steps whenever targeted promotion is needed, the CLD would like to have a model that handles the classification of each new customer in to one of the pre-defined segments. Moreover such a model would enable the CLD efficiently implement its CRM and also would get a timely response when there is a need to know who a customer is.

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, especially data mining tools can handle large amounts of data and can be used to extract inherent structures and patterns from data.

According to Feyen (n.d.), the most obvious applications of data mining in the airline business are related to the frequent flyer programs. Furthermore, Bounsaythip (2001) noted 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.

Customer information, if extracted from the data, would enable the CLD reduce its promotion costs and better recognize its core customers whenever needed. It is, therefore, the researcher's belief that the sheer volume of data and business requirements necessitates the use of data mining techniques for extracting information and generating rules and models.

The following chapter of this research project will provide summary of the important data mining methods used in this research in an attempt to derive the classification rules and then build the corresponding models respectively. The exploration of available databases by the use of data mining techniques and attempt to derive classification rules that would enable CLD classify each new customer's record in to its appropriate segments will be addressed in the experimental part of the project, chapter five.

CHAPTER FOUR

REVIEW OF APPLICABLE TECHNIQUES AND A RELATED RESEARCH

4.1 Introduction

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, and continued with improvements in data access such as more recently generating technologies that allow users to navigate through their data in real time.

Nowadays, different data mining methods are under use to explore data in databases and make a good use of the resulting patterns. Two of the commonly applicable data mining methods, clustering and classification, which will be used in the experimental part of this research, are briefly discussed in the following two sections, 4.2 and 4.3. At this junction, it should be noted that rather than the basic techniques, the related methods discussed in these sections are not used.

4.2 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. According to Berry et al., (2000), clustering 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)* [1] *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.

According to Two Crow Corporation (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.

Most commonly used clustering algorithms can be classified in two general categories: Hierarchical and Non-hierarchical.

Hierarchical procedures involve the construction of a hierarchy or treelike structure in a bottom-up and top-down approach.

In general, hierarchical procedures have the advantage of being fast and take less computing time. But they could be misleading and unreliable because undesirable early combinations may persist throughout the analysis and lead to artificial results. To reduce this possibility, an analyst should confirm the analyzed result with different clustering techniques.

[1] The training is entirely data-driven and no target results for the input are provided.

In contrast to hierarchical methods, non-hierarchical clustering procedures do not involve the tree-like construction process. Instead, the first step is to select a cluster center or seed, and all objects (data points) within a pre-specified threshold distance and included in the resulting cluster. The most well known non-hierarchical clustering algorithm is the K-Means algorithm.

In data mining, the most common methods used to perform clustering are K-means and Kohonen feature maps (or self-organizing maps or SOM) (Bounsaythip et al, 2001).

4.2.1 The K-Means Algorithm

Given a database of n objects or data tuples, a partitioning method constructs k partitions of the data, where each partition represents a cluster and k <= n. That is, it classifies the data into k groups, which together satisfy the requirements that, each group must contain at least one object; and each object must belong to exactly one group.

According to Berry et al. (2000), the *k*-means algorithm of cluster detection is the most widely used algorithm 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, "means" 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 $j = 1,..., K$, 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

$$J = \sum_{j=1}^{k} \sum_{n \varepsilon sj} [x^n - \mu_j]^2 ,$$

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

$$\mu_j = 1/N_j \sum_{n \varepsilon sj} x^n.$$

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

$$\Delta\mu_j = \eta(x^n - \mu_j),$$ 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 corresponds 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 4.1, which is adapted from Berry et.al. (2000).

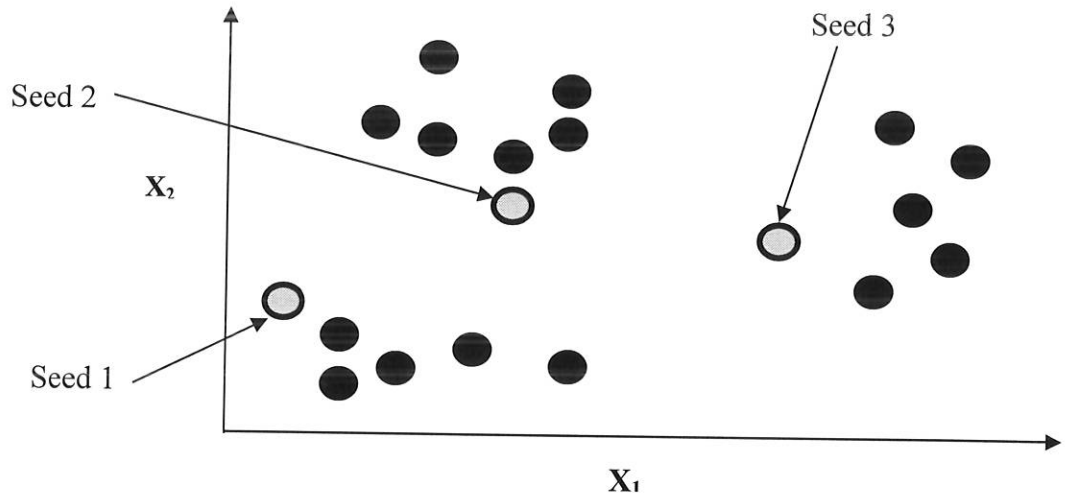


Figure 4.1: Initial Cluster Seeds

In Figure 4.2, which is adapted from Berry et al. (2000), 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.

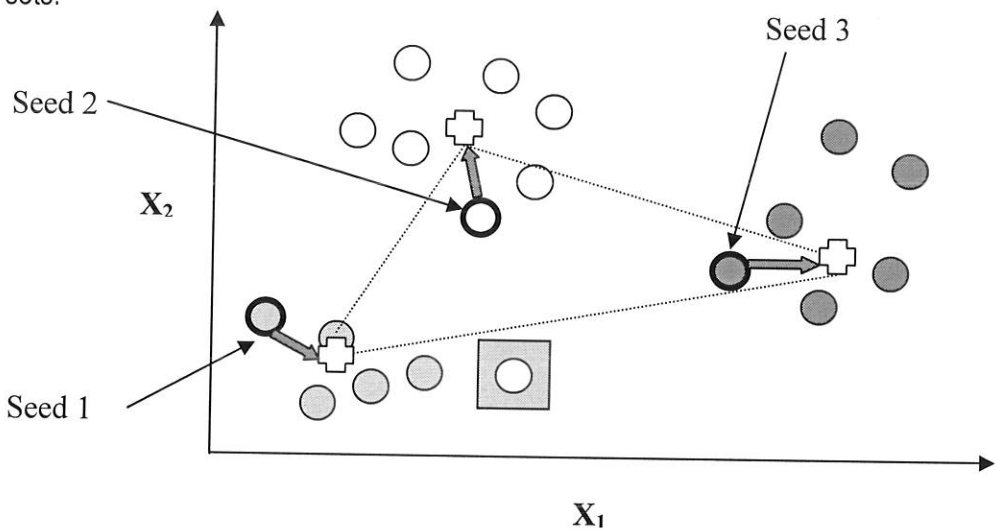


Figure 4.2: Cluster seeds after one iteration

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 states 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 may 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. (Ibid) 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. According to Berry et al. (2000), the three commonly used approaches to understand clusters are:

1. Using visualization to see how the clusters are affected by changes in the input variables.
2. Examining the differences in the distributions of variables from cluster to cluster, one variable at a time.

3. 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.

4.2.2 Self-Organizing Map (SOM)

The self-Organizing Map (SOM), introduced by Teuvo Kohonen in 1982, is one of the most popular unsupervised-learning neural network[1]. According to Kohonen (2001), SOM is a new, effective software tool for the visualization of high-dimensional data. Also known as the Kohonen's feature map, the SOM algorithm is quite a unique kind of neural network tool for visualization of high-dimensional data. In its basic form, SOM produces a similarity graph of input data (Kohonen, 2001). It consists of a finite set of models that approximate an open set of input data and the models are associated with nodes (neurons) that are arranged as a regular process that automatically orders them on the grid along with their mutual similarity.

SOM is believed to resemble processing that occur in the brain and is used for visualizing high-dimensional data in 2- or 3- dimensional space (Han and Kamber, 2001).

The map units, neurons, usually form a 2-D (two-dimensional) regular lattice where the location of a map unit carries semantic information. The SOM can thus serve as a clustering tool of high-dimensional data (Han and Kamber, 2001). Another important feature of the SOM is its capability to generalize (Kohonen, 2001). In other words, it can interpolate between previously encountered inputs.

---

1 Hardware or a computer program, which strives to simulate the information processing capabilities of its biological exemplar.

The basic SOM can be visualized like an array, the cells (or nodes) of which become specifically tuned to various input signal patterns of classes of patterns in an orderly fashion. Typically interconnected, the number of neurons may vary from a few dozen up to several thousand and are organized on a regular low-dimensional grid. Each neuron is represented by a d-dimensional weight vector $m= (m_1, m_2, \ldots, m_d)$, where d is equal to the dimension of the input vectors. The neurons are connected to adjacent neurons by a neighborhood relation, which dictates the topology of structure, of the map. The organized map avails itself ready to visualization and the properties of the data set can be illustrated in a meaningful manner.

4.3 Classification

The area of statistics is an important one to data mining. For many years, statistical methods have been the primary means for analyzing data.

There are three basic classes of statistical techniques with implementations employed in data mining today: linear, non-linear, and decision trees. Linear models describe samples of data or the relationship among the attributes or predictors, in terms of a plane so that the model becomes a global consensus of the pattern of the data. Linear modeling techniques include linear regression (for prediction) and linear discriminent analysis (for classification).

4.3.1 Decision Trees

A decision tree is a flow-chart- like tree structure, where each internal node denotes a test on an attribute, each brunch represents an outcome of the test, and leaf nodes represent classes or c lass

distributions. The top most node in a tree is the root node. A typical decision tree is shown in Figure 4.3 below.

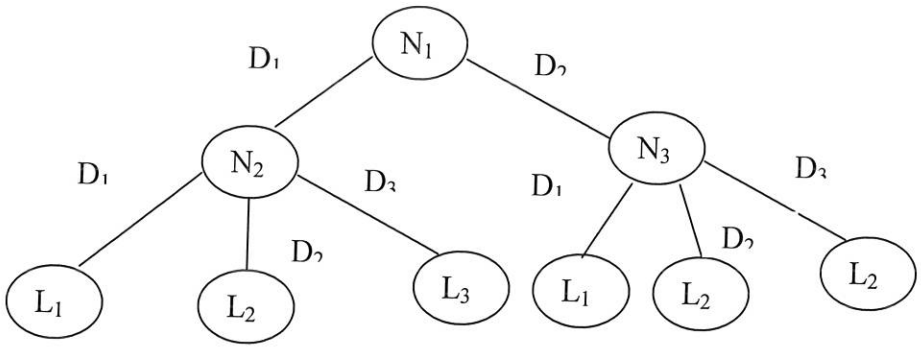


Figure 4.3: A Decision Tree with Decision (Ni) and Leaf (Li) nodes, and decisions (Di)

Depending on the algorithm, each node may have two or more branches. For example, CART (*Classification And Regression Trees*) generates trees with only two branches at each node. Such a tree is called a binary tree. When more than two branches are allowed it is called a multi-way tree.

Each branch will lead either to another decision node or to the bottom of the tree, called a leaf node. By navigating the decision tree you can assign a value or class to a case by deciding which branch to take, starting at the root node and moving to each subsequent node until a leaf node is reached. Each node uses the data from the case to choose the appropriate branch.

Decision trees models are commonly used in data mining to examine the data and induce the tree and its rules that will be used to make predictions. A number of different algorithms may be used for building decision trees including CHAID (Chi-squared Automatic Interaction Detection), CART (Classification And Regression Trees), Quest, and C5.0.

Decision trees are grown through an iterative splitting of data into discrete groups, where the goal is to maximize the "distance" between groups at each split. One of the distinctions between decision tree methods is how they measure this distance.

One can think of each split as separating the data into new groups which are as different from each other as possible. This is also sometimes called making the groups purer.

Decision trees which are used to predict categorical variables are called *classification trees* because they place instances in categories or classes. Decision trees used to predict continuous variables are called *regression trees.*

Trees can become very complicated. Imagine the complexity of a decision tree derived from a database of hundreds of attributes and a response variable with a dozen output classes. Such a tree would be extremely difficult to understand, although each path to a leaf is usually understandable. In that sense a decision tree can explain its predictions, which is an important advantage.

4.3.2 Decision Tree Induction

Early experiments implementing concept learning systems, conducted by Hunt, et. al. (1996), as quoted in (Carbone, P. L., 1997), provided a generic concept about how a decision tree is constructed. Until greedy algorithms were later created, searching for the simplest tree is a major time consuming task (Carbone, P. L., 1997).

According to Han and Kamber (2001), heuristic methods that explore a reduced search space are commonly used for attribute subset selection. These methods are typically *greedy* in that, while searching

through attribute space, they always make what looks to be the best choice at the time. Their strategy is to make a locally optimal choice in the hope that this will lead to a globally optimal solution. Such greedy methods are effective in practice and may come close to estimating an optimal solution.

According to Han and Kamber (2001), the basic algorithm for decision tree induction is a greedy algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner.

4.3.3 Decision Trees and Attribute Selection

Basic heuristic methods of attribute subset selection include the following techniques, where the stopping criteria for the techniques may vary:

Stepwise Forward Selection: The procedure starts with an empty set of attributes. The best of the original attributes is determined and added to the set. At each subsequent iteration or step, the best of the remaining original attribute is added to the set.

Stepwise Backward Elimination: The procedure stats with the full set of attributes. At each step, it removes the worst attribute remaining in the set.

Combination of Forward selection and Backward Elimination: The stepwise forward selection and backward elimination methods can be combined so that, at each step, the procedure selects the best attribute and removes the worst from among the remaining attributes.

The "best" (and "worst") attribute are typically determined using tests of statistical significance, which assume that the attributes are independent of one another. Many other attribute evaluation methods can

also be used. Here, we discuss one of the methods, *information gain measure,* used in building decision trees for classification.

The information gain measure is used to select the test attribute at each node in the tree. Such a measure is referred to as an *attribute selection measure* or a *measure of the goodness of split*. The attribute with the highest information gain (or greatest *entropy* reduction) is chosen as the test attribute for the current node. This attribute minimizes the information needed to classify the samples in the resulting partitions and reflects the least randomness or "impurity" in these partitions. Such an information-theoretic approach minimizes the expected number of tests needed to classify an object and guarantees that simple (but not necessarily the simplest) tree is found.

The expected information (or *the split information*) needed to classify a given sample S, which consists of s data samples residing in m distinct classes, {C1, C2, ..., Cm}, of the class label attribute, is given by

$$I(s_1, \ldots, s_m) = -\sum_{i=1}^{m} P_i \log_2(P_i) \qquad (1)$$

Where: $P_i = \frac{s_i}{s}$ is the probability that an arbitrary sample belongs to class C_i.

s_i – is the number of samples of S in class C_i.

To select the test attribute (i.e., the best attribute for splitting), the entropy and information gain need to be calculated for each attribute. Therefore, if an attribute A has v distinct values, $\{ a_1, a_2, \ldots, a_v \}$, then attribute A can be used to partition S in to v subsets, $\{ S_1, S_2, \ldots, S_v \}$, where Sj contains those samples in S that have value aj of A. If A were selected as the test attribute (i.e., the best attribute for splitting), then

$S_1, S_2, \ldots, S_v$ would correspond to the branch grown from the node containing the set S. The **entropy,** or *expected information* based on the partitioning into subsets by an attribute A, is given by:

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \ldots + s_{mj}}{s} . I(s_{ij}, \ldots, s_{mj}) \qquad (2)$$

Where: s_{ij} - is the number of samples of class Ci in a subset Sj

$\frac{s_{1j} + \ldots + s_{mj}}{s}$ - acts as the weight of the j^{th} subset and is the ratio of number of samples in the subset to total samples in S

The smaller the entropy value, the greater will be the purity of the subset partitions. It should be noted that, for a given subset S_j,

$$I(s_{1j}, s_{2j} \ldots, s_{mj}) = -\sum_{i=1}^{m} p_{ij} \log_2 (p_{ij}) \qquad (3)$$

Where: $p_{ij} = \frac{s_{ij}}{|S_j|}$ - is the probability that a sample in S_j belongs to class C_i

As a result, the encoding information that would be gained by branching on attribute A is:

$$Gain\ (A) = I(s_{1j}, s_{2j} \ldots, s_{mj}) - E\ (A) \qquad (4)$$

In other words, *Gain (A)* is the expected reduction in entropy caused by knowing the valued of the attribute A.

The attribute with the highest information gain is considered as the most discriminating attribute of the set under consideration. So, an attribute that yields maximum information gain will be chosen for data set partitioning. Then, a node is created and labeled with the chosen attribute, branches are formed for each value of the attribute, and the samples are partitioned accordingly. The same criteria will then be applied to each split sample. The iterative divide and conquer process executes until no further split is required.

This information gain inductive learning criterion is known as C4.5. Though the inductive learning algorithms offer an automatic decision trees/rules generation mechanism, the outcome is biased by

4.3.4 Controlling Tree Size

Trees left to grow without bound take longer to build and become unintelligible, but more importantly they over-fit the data. Tree size can be controlled by means of *stopping rules* that limit growth. One common stopping rule is simply to limit the maximum depth to which a tree may grow. Another stopping rule is to establish a lower limit on the number of records in a node and not do splits below this limit.

An alternative to stopping rules is to prune the tree. The tree is allowed to grow to its full size and then, using built-in heuristics or user intervention, the tree is pruned back to the smallest size that does not compromise accuracy. For example, a branch or sub-tree that the user feels is inconsequential because it has very few cases might be removed.

4.3.5 Advantages of Decision Trees

Decision trees make few passes through the data (no more than one pass for each level of the tree) and they work well with many predictor variables. As a consequence, models can be built very quickly, making them suitable for large data sets.

Decision trees handle non-numeric data very well. This ability to accept categorical data minimizes the amount of data transformations and the explosion of predictor variables inherent in neural nets.

Some classification trees were designed for non-numeric data and therefore work best when the predictor variables are also categorical. Continuous predictors can frequently be used even in these cases by converting the continuous variable to a set of ranges (binning). Some decision trees do not support continuous response variables (i.e., will not build regression trees), in which case the response variables in the training set must also be binned to output classes.

4.3.6 Limitations of Decision Trees

A common criticism of decision trees is that they choose a split using a "greedy" algorithm in which the decision on which variable to split doesn't take into account any effect the split might have on future splits. In other words, the split decision is made at the node "in the moment" and it is never revisited. In addition, all splits are made sequentially, so each split is dependent on its predecessor. Thus all future splits are dependent on the first split, which means the final solution could be very different if a different first split is made.

The benefit of looking ahead to make the best splits based on two or more levels at one time is unclear. Such attempts to look ahead are in the research stage, but are very computationally intensive and presently unavailable in commercial implementations.

Furthermore, algorithms used for splitting are generally univariate; that is, they consider only one predictor variable at a time. And while this approach is one of the reasons the model builds quickly —it limits the number of possible splitting rules to test — it also makes relationships between predictor variables harder to detect.

4.4 Summary of a related Research

As this investigation is predestined to fill the gap left by a related research done by Henok (2002), it is worth summarizing what had been accomplished by Henok. The summary, at this junction, is believed to provide the necessary background to the understanding and evaluation of the experiments that will be made in the current research. Therefore, the following few paragraphs are dedicated to summarize the achievements of Henok in his research so that the gap left can clearly be understood.

4.4.1 The Data Files Used

For his clustering experiments, Henok (ibid) had to prepare a single table that had to be used by the Knowledge Studio software. In the course of preparing a single MS Access table, all the data files from the different databases had to be changed to the same, MS Access, file format. After changing all the pertinent data files in to MS Access files by applying a number of procedures, Henok (ibid) put together

the different data files from different databases in a single MS Access database. Summary of the data files in Henok's MS Access database is given in **Table 4.1** below.

4.4.2 The Experiments Carried out

The general objective being identifying the possible customer segments and describing and summarizing the resulting customer clusters, Henok(ibid) has performed a total of 4 basic experiments.

Relations (Tables)	Number of records per relation	Type of attributes	Number of attributes	Total attributes	*Content description of each relation*
Member*[1]	22,022	Text	37	40	*Demographic data of each* ***member****[2]
		Date/Time	3		
Trips	90,833	Text	8	10	*Flight activity of each member*
		Date/Time	1		
		Number	1		
Revenue	20, 158	Text	4	6	*The total revenue of each member*
		Number	2		
Member Points*[1]	22,022	Text	1	5	*The total points posted to each member*
		Date/Time	1		
		Number	*3*		

Table 4.1 Summary of the data files in Henok's MS Access database

Under these 4 basic experiments, 9 different experimental setups (sub-experiments) were considered. All the experiments were done by using Knowledge Studio (version 3.0) of Angoss Software Corporation. The sub experiments, the basic tasks, the objectives, and the decision made at the end of each basic experiment are displayed in **Table 4.2**.

---

[1] *These tables were found directly form the original database and checked.

[2] **Member* in each relation's content description indicates the ShebeMiles frequent flier program member.

	Objective	Number of sub-experiments	Basic Tasks	Decision made on the result
Expt-	Selecting the variables to be included and number of clusters, K, to be used initially.	2- clustering runs	Inputting 9 customer variables	A selection of only 4 basic attributes with K = 4.
Expt-2	Building Clusters from a minimum set of variables selected in expt-1 with different number of iterations.	4- clustering runs	Inputting the selected 4 attributes with K = 5 and checking different number of iterations	Maximize the number of iterations with an addition of 1 variable.
Expt-3	See the effect of an additional variable with maximum iterations (10,000)	2- clustering runs	Inputting 5 attributes and checking with k = 4 and then K = 3	Maximize the number of variables with different K values
Expt-4	Keeping the number of iteration at 10,000 and observing the effect of different number of K values to make a final decision	3- clustering runs	Clustering runs with 6 variables, and K values of 6, 5, and then 4	Select the 6 variables, K = 5 and 10,000 iterations to segment the records.

Table 4.2 Summary of the basic experiments of Clustering by Henok (2002)

4.4.3 The Input Parameters Used and the Sub-Experiments

The input parameters (such as the input variables, the number of iterations and the number of clusters) used in each sub-experiment by Henok are summarized in **Table 4.3.**

In each of the experiments summarized in **Table 4.3**, out of the final sample size of 7,602 records, the number of records used was 4,000 (52.62% of the total records), leaving the rest 3,602 (47.83% of the total) records as a test set. However, there was no sound reason given for partitioning the final sample size in to these two 'training' and 'test' sets.

Experiment	Sub-Experiment	Variables	Iterations	*Number of Clusters*
Exp't-1	1-1	9	10,000	*6*
	1-2	9	10,000	*5*
Exp't-2	2-1	4	1,000	*5*
	2-2	4	5,000	*5*
	2-3	4	1,000	*4*
	2-4	4	5,000	*4*
Exp't-3	3-1	5	10,000	*4*
	3-2	5	10,000	*3*
Exp't-4	4-1	6	10,000	*6*
	4-2	**6**	**10,000**	***5***
	4-3	*6*	*10,000*	*4*

Table 4.3 Summary of the basic and sub- experiments by Henok (2002)

4.4.4 Output of the Sub-Experiment Selected

Among the three sub experiments in experiment 4, displayed in Table 4.3, Henok (2002) reported that the result from the ***sub-experiment 4-2*** was more meaningful as per the comment of the domain experts. The final summary of the clustering was, therefore, done by making use of the result of the clustering run from this sub-experiment and is presented in **Table 4.4** (Henok, 2002).

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

Table 4.4 Summary of the Clustering sub- experiment 4-2 by Henok (2002)

The corresponding meanings of letters in Table 4.3 are the following:

VH = Very High; **H** = High; **M** = Medium; **Lw** = Low;

VLw = Very Low; **R** = Recent; **Lg** = Long; **VLg** = Very Long.

And also, the corresponding meanings of numbered variables in the same table are the following:

Var 1 = TtlYearTrip **Var 2** = OneYearTrip **Var 3** = Tenur_Month

Var 4 = TotalRevenue **Var 5** = RevPerTenure **Var 6** = TripsPerTenure

4.4.5 The Class/Concept Description of each Cluster

In conclusion, the final clustering model, after being evaluated by the 3,602 records (the data set kept for 'test'), *a* concept/class description of each cluster in **Table 4.4** was done in collaboration with the domain experts (Henok, 2002). The corresponding class/concept description and remark of each cluster is summarized in **Table 4.5** below.

Cluster	Percent-age share	Concept/Class Description	Remark
1	52.6%	Members who are not worth giving immediate attention in terms of target promotion as compared with members in other groups	As recent they are, it is rather premature to judge potential value of customers in this group
2	2.9%	High tiered and most valuable group among the groups and is termed as the 'loyal' customer group	Addition of a fourth tier level (possibly platinum) might be good to keep the loyalty of this group.
3	10.5%	This group is worth giving immediate attention, with an enticing targeted promotion, as it has a potential to join the top tier group.	The likelihood of losing this segment to the competition is high due to its medium tenure in the program.
4	7.8%	This is the other top tier segment where customers belonging to this segment show characteristics of loyalty.	More targeted promotions could elevate the members' status from highly frequent travelers to that of very highly frequent travelers
5	26.3%	Customers with few trips but significant revenue contribution.	Targeted promotional campaigns might entice this group to make frequent trips.

Table 4.5 The class/concept description and remark of each cluster

4.5 Conclusion

As it has been attempted to briefly summarize the research done by Henok, the result of the research is a good experience for an investigation that can be done in the same problem area. Henok has identified the different databases and processed the required data for his investigation. However there were a number of tasks that had to be done so as to make the clustering results useful for customer classification tasks. Among the tasks that were not done by Henok; the cluster indexes of the individual customer record were not documented, the cluster centroids were not numerically expressed rather ordinal values were used to describe the customer clusters identified, the final data set was not used totally for the clustering experiment rather it was portioned in to training and testing data sets without any experience that supports training and testing a model in a clustering experiment.

Thus, in the next chapter, the exploration of available databases by the use of data mining techniques with the intention to complete the tasks that were not done by Henok and are mentioned above. To avoid duplication of efforts the experiences of Henok will be used, and the results that are important for the classification experiments will properly be documented.

Finally, unlike Henok's investigation, classification rules will be derived so that the gap would be filled and the data mining process will be complete. The derived classification rules will be used to develop a prototype that would enable the Customer Loyality Department (CLD) classify each new customer's record in to its appropriate segments. The classification rules are to be explainable to the CLD staff, and are also believed to add value to the implementation of efficient CRM processes.

CHAPTER FIVE

EXPERIMENTATION

5.1 Overview

In this chapter the researcher describes the source of data as well as the techniques that have been used in preprocessing and model building. Test results and corresponding discussions are also presented.

Thearling (2000) discusses, in order to enable successful CRM; the initial task is to identify market segments containing high profit potential. Since this initial phase has already been carried out by a previous investigation of Henok (2002), the use of the same data source will be made to carry out the final phase, customer classification. Accordingly, the main objective of this research was to provide a model that classifies customers with respect to the important dimensions of customers' travel behavior and corresponding value.

This research project incorporated the typical stages that characterize a data mining process. Accordingly, this section of the project is organized according to the CRoss-Industry Standard Process for Data Mining (CRISP-DM) process cycle, which is discussed in section 2.2.1 and depicted in Figure 2.1 of chapter two.

5.2 Data Mining Goals

The business survey undertaken by the researcher has revealed that current customer value is based on individual mileage, and that mileage is an arbitrary measure of customer profitability. Moreover, a customer segmentation, which disregards mileage as the only measure of customer profitability, had been carried out based on a reliable revenue value and flight activities of each customer. Thus, the first data mining goal was to identify the customer segments and describe the resulting clusters by making use of the same experience of the previous customer segmentation research.

Moreover, the variables that determine customer value will be used to derive the customer segments, and the subsequent classification rules. The most appropriate data mining techniques, which are clustering (or segmentation) and classification, will be used for this purpose.

In order to provide customer classification rules that can be explained to domain experts from the CLD, the main focus was on the inspection and use of the important attributes for the customer segmentation. This process will allow to get meaningful clusters where each member customer of a cluster will be identified by the cluster index of the group. This cluster index of each customer will in turn be used as input to the classification algorithms to generate the required classification rules.

The success criterion for this data mining research is the discovery of customer classification rules that would discover and differentiate customers with high profit potential and those with insignificant contribution. Provided that reasonable customer classification rules are discovered, the CLD could devise a means so as to reduce its promotion costs and better recognize its core customers whenever needed.

Moreover CLD could also design a special marketing strategy geared towards each customer that would enhance profitability as well as ensure the customer's loyalty.

5.3 Data Mining Tool Selection

The selection of an appropriate data mining tool for this project was done based on a certain criteria. Among the criteria used for the selection, the most important ones are:

- The data mining tasks that the tool is intended for
- The algorithms supported
- The computer architecture and operating system on which the software runs
- The possible formats for the data that is to be analyzed
- The maximum number of records the software can comfortably handle
- Visualization capabilities
- The availability of an evaluation version for download on the Internet

The tools that the researcher intended to get, the Knowledge Studio and Weka-3-2, were the ones that more or less fulfilled the above criteria. As a result Knowledge Studio version 3.0 of Angoss Software Corporation (www.angoss.com) and Weka-3-2, which were found from personal contact and used for this experiment.

5.4 Data Understanding

Having defined the data mining goals, the next step was the data understanding and the investigation of which data were available and useful for achieving the goals. In this regard, the primary source of data for this research is the ShebaMiles database that had been used by Henok (2002) for the customer segmentation experiments. Therefore, this database (ShebaMiles) was thoroughly studied and, as a result of the study, four basic relations (tables) were found to be important.

Among the four basic relations, two of the relations (*Member Table* and *Member-Point Table*) were found in MS Access database file format directly from the target database; one of the relations (*Revenue Table*) was found in ADABAS database file format from the revenue accounting database; and the fourth relation (*Trips Table*) was found in MS Visual FoxPro database file format.

For the segmentation experiments, Henok (ibid) prepare a single MS access database. Thus, all the above tables were put together in a single database file and preparing a single table that was used by the Knowledge Studio software was possible. To this end, Henok (ibid) had applied a number of procedures on the non–MS Access data file formats to change them to an MS Access file formats.

To avoid duplication of effort on what has already been done, the MS Access database that was put together by Henok (ibid) has been found appropriate to be used for the current research. The following section will, therefore, describe the data files (tables) that have been found in the database.

5.4.1 Description of the Data in Tables of the MS Access Database

1. Trips

This table contains a total of 90,833 records and eight fields of the aggregated data from the original tables of members' flight activities and the corresponding revenue table. 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[1] traveled
Revenue	Number	Revenue in US Dollars per each flight segment traveled

Table 5.1 Attributes of the Trips table

Since the fields (orig, dest, class) of the revenue table were identical in content the "Revenue" field was appended by a procedure, which checks the similarity of the three fields. Therefore, the revenue and trips relation became a single relation named by Trips as in table 5.1.

[1] A trip from one origin city to another destination city makes up one flight segment.

2. Member

This table contains 22,022 records within 40 fields describing demographic data pertaining to each member of the ShebaMiles program. However, only the 15 fields' names (avoiding repeated fields with very few and/or no corresponding customer input values) as well as their data types are 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 5.2 Attributes of the Member table (Partially)

3. Members Point

This table contains the number of points posted to each of the 22,022 members and 5 fields, which 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 5.3: Attributes of the Points table

5.4.2 Verification of Data Quality

Before using data files in Henok's MS Access database as they were, cross checking the data files with their corresponding original data files in ShebaMiles database were done. In the course of cross checking the data files, especially the non -MS Access files, it has been found that each pair of interrelated files contained approximately identical data regardless of their file formats. Moreover, the corresponding queries have been done to check the exact content of the data files for the cases where the original data files are not identical with that of the MS Access data files. As a result, it has been decided to use Henok's MS Access database for the investigation at hand.

5.5 Data Preparation

As per his report, Henok (2002) has used a final set of customer records having 9 input attributes selected for cluster modeling, as summarized in **Table 5.4**. The total number of customer-level records that were reported by him was 11,963. Out of this number of records, by using a time criteria, he selected a final sample of 7,602 customer-level records whose flight activities occurred during a 12 months period between April 01, 2001 and March 31, 2002.

Field	Data Type	Description	Remark
TtlYearTrips	Number	Total number of segments flown by member	Aggregated
OneYearTrips	Number	Total number of segments flown by member during the 12 months between April 2001 and March 2002	Aggregated
Ttl_Revenue	Number	Total revenue collected from member	Aggregated
Ttl_Points	Number	Total base mileage points awarded	Aggregated
Tenure_Months	Number	Number of months since member first enrolled in ShebaMiles	computed
RevPerPoints	Number	Ratio of Total Revenue to Total Mileage Points	Derived
RevPerTenure	Number	Ratio of Total Revenue to Member Tenure	Derived
RevPerTrips	Number	Ratio of Total Revenue to Total Number of Segments	Derived
TripsPerTenure	Number	Ratio of Total Number of Segments to Member's Tenure	Derived

Table 5.4 Selected attributes for Cluster modeling

However, the indicated final data set and the corresponding sample of customer records with the selected 9 input attributes were not found in the MS Access database. As a result, the cluster index, which is the important attribute for the classification experiments in the investigation at hand, was not found for each case of the sample customer record.

Therefore, since the MS Access database, to be used for the current research, has the aforementioned limitations, the data preparation process, which enables to get a suitable working data set, had to be repeated on the database.

5.5.1 Data Preprocessing

As discussed in section 2.2 of chapter two, although at the core of the knowledge discovery process, the data mining step usually takes only a small part (estimated at 15% to 25 %) of the overall effort. In other

words, the other stages of the process often require considerable effort. One of these stages is data preprocessing.

The purpose of the data preprocessing is to prepare the data as much as possible and put it into a form suitable for use in later stages. Starting from the data extracted from the MS Access database, a number of transformations had to be performed before a final data set, for model building, was found.

5.5.2 Preparing Data for Analysis

This process involves handling missing values, summarization, driving new fields, and finally preparing the data into a form that is acceptable to the appropriate data mining tools selected. As far as handling missing values is concerned, Henok (ibid) had used a commonly accepted way, as suggested in the Two Crows Corporation (1999), and developed a procedure that fills missing values such as revenue. However, the data summarization, and deriving new fields are the main tasks of this phase that need to be done, as the final refined data set is not there in the MS Access database.

The first task in this phase was the selection of 74,212 of the total 90833 records of the Trips table with complete revenue data. The remaining records, which had missing revenue data, were excluded as their inclusion would render the data mining goals to become useless.

The next task performed in this phase was the summarization of the member activity records in the Trips table. The summarization of the records was done by each member's records, thus reducing the total number of records from 74,212 to 11,970. This aggregation of records also necessitated the numeric representation of some attribute values. Moreover, while selecting the attributes to be used for this task of

aggregation, checking the attributes in the final data set's table and further consultation of domain experts was done. The new aggregated Trips table had the attributes shown in **Table 5.5** below.

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 5.5: Attributes of the Trips table aggregated at member level

As the attributes in **Table 5.5** were not the only ones used for the clustering experiments by Henok, the remaining attributes should be computed or derived from the existing attributes. In this regard, all of the derived attributes in **Table 5.4** could not be defined by using only the attributes in **Table 5.5.** At this junction, data integration was necessary as information pertaining to enrollment date, which is used to compute the ***Tenure Month*** of each member, was located in a different table that is the **Member** table described in **Table 5.2**.

Since the enrollment date is a date data type, the Tenure Month of each member, which is a numerical data type, should be computed for each member. This computation is necessitated to define two of the derived attributes (RevPerTenure and TripsPerTenure) in Table 5.4. The computation of Tenure Month for each member was, therefore, carried out by a procedure (refer to Annex I).

According to Saarenvirta (1998), data creation involves the creation of new variables by combining existing variables to form ratios, difference and so forth. Accordingly, the derived attributes, in **Table 4.4** are then defined using the existing attributes as follows:

$$\mathrm{Re}vPerPo\mathrm{int}s = \frac{Total\,\mathrm{Re}venue}{TotalPo\mathrm{int}s} \quad (1)$$

$$\mathrm{Re}vPerTenure = \frac{Total\,\mathrm{Re}venue}{Tenure\_Month} \quad (2)$$

$$\mathrm{Re}vPerTrips = \frac{Total\,\mathrm{Re}venue}{TotalYearTrips} \quad (3)$$

$$TripsPerTenure = \frac{TotalYearTrips}{Tenure\_Month} \quad (4)$$

Since the data set that Knowledge Studio accepts is a single table, all the above customer specific attributes were integrated into a single table by making use of the MS Access database query utilities. Consequently, for the next tasks of the investigation the final data set was then prepared. This data set contains 7,532 records selected from the total, 11,970 member level records by applying a time criteria on the customers' flight activity. The time criteria applied was, a 12 months period between April 18, 2001 and April 17, 2002 where April 17, 2002 is the last date on which members' flight activity is recorded.

Figure 5.1 shows the resulting data model of the ShebaMiles data mart, which contains the final data set that was input into the Knowledge Studio Software. The Trips and Member tables were joined to create a data set with records aggregated at a member level.

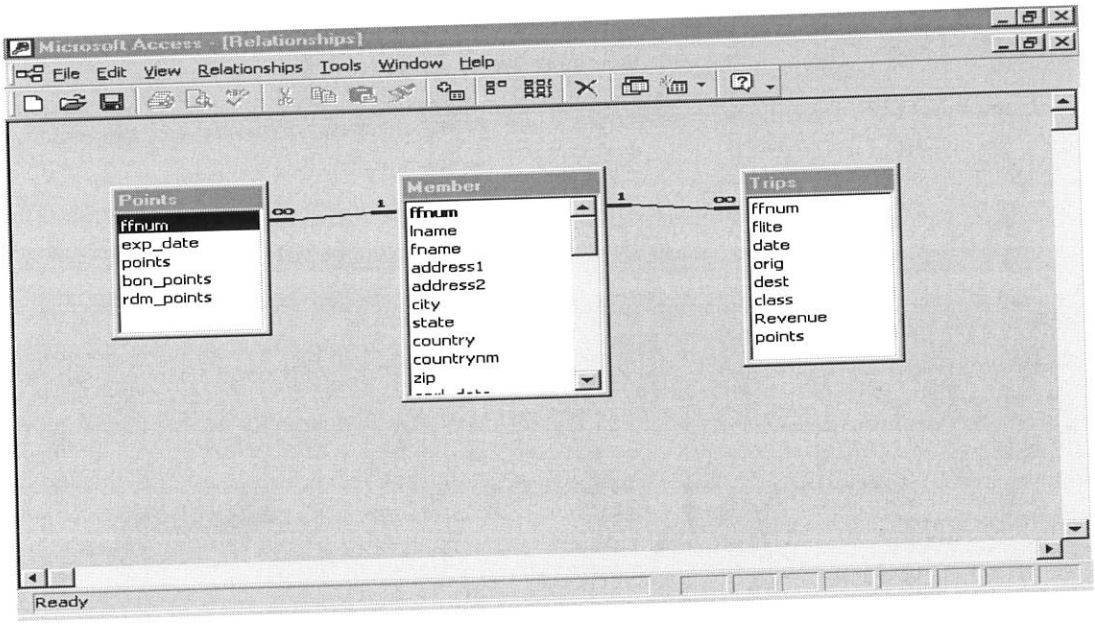


Figure 5.1: The ShebaMiles Data Mart data model

The next task at this phase was preparing the final data set in to file forms that are acceptable by the software that provide with the required outputs.

5.5.3 Data Formatting

Before directly getting in to the modeling process, two different file formats of the final data sets were prepared for two different data mining tools, Knowledge Studio and Weka.

The first format of the final data set was the MS Access table, to be fed into the Knowledge Studio Software, with the field names described in **Table 5.4**. The fact that Knowledge Studio had Open Database Connection (ODBC) facilities enabled the researcher to import the final data set directly from the MS Access database, without changing into other format.

The second file was a *Comma Delimited* [2] text file in *ARFF* [3] format of the same final data set, to be fed into the Weka software. Changing the final data set in the MS Access table format into this *ARFF* file format was done by exporting the data in to a file in comma-separate format. Then, loading the file into a text editor allows adding some basic lines and saving the file as *ARFF* file format.

5.6 Modeling

The modeling phase in the data mining process of this investigation was carried out in two sub-phases, clustering and then classification rule generation.

Rules generated for classification purpose being the required final output of this phase, the clustering sub-phase is inescapable step. For this, there are three reasons.

- The first reason was, as mentioned in section **5.5;** the cluster index for each customer record in the selected data set had to be found.
- The second reason was comparison between the clustering experimental result of this research and the previous investigation in relation to the customer records distribution in each cluster should be made.
- The third and the last reason was the ordinal values of each cluster, discussed in the sub-section **4.5.4**, had to properly be interpreted into their corresponding numerical cluster centroid values.

---

[2] *Comma Delimited* applies for a list of records where the items are separated by commas.

[3] *ARFF* is an extension of a file format that the Weka software can read.

5.6.1 The Clustering Sub-Phase

Since this investigation was for the second time that segmentation at a member level was being done, there were predefined segments. However, for the reasons mentioned above, employing a clustering algorithm was appropriate.

To carry out this sub-phase, 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 values in the final data set was almost none, K-means was used for this sub-phase of the study at hand, as it had been used in the previous investigation.

A critical task of using the K-means clustering algorithm was the choice of the right variables and the right scales. In this regard, since the variables to be fed into the clustering algorithm were convincingly selected in the previous investigation as summarized in section **4.4** (sub-sections **4.4.3** and **4.4.4**) of chapter four, no further trial had been done to reselect the input variables.

The K-means algorithm, in the course of segmenting the customer records, passes through each customer record, assigning each to the closest existing cluster center. To effectively handle this, the K-means algorithm requires that input variables should all be converted into numbers and the numbers be

further normalized to ensure that the influence of all variables is similar. The task of converting the values of the input variables to numbers has already been performed in the data preparation phase as described in **Table 5.4**. Furthermore, Knowledge Studio automatically normalizes the data.

The previous investigation partitioned the final sample size of 7,602 records into 4,000 records of training and 3,602 records of test sets based on no sound reason. Unlike the previous investigation, the final data set with 7,532 total records was used and imported into the Knowledge Studio environment for clustering in the investigation at hand.

Once the clusters were created, the comparison and interpretation of the resulting clusters of this research and that of the previous investigation were done by the domain experts and the researcher. The researcher used the following four approaches to compare and interpret the clusters.

1. Used the K-means clustering algorithm of the Weka software, which generated numerical values of the cluster centroids for the chosen number of clusters and type of variables used in the previous investigation.
2. Examined the proportional distribution of records in the clustering experimental results from the Knowledge Studio and Weka of this research with the previous one.
3. Associated the cluster indexes of the previous investigation and the investigation at hand (taking both the Knowledge Studio's and Weka's results in to account).
4. Finally, interpreted the ordinal cluster centroids of the clusters summarized in the previous investigation into their corresponding numerical cluster centroids, which were results of the Weka's clustering experiment.

Automatic Cluster Detection

As it had been mentioned earlier, the clustering run parameters for the investigation at hand were directly adapted from the previous investigation's selected sub-experiment, which was summarized in section 4.5.4 of chapter four in this investigation. As a result, the basic run parameters set in Knowledge Studio for K-means clustering were:

- Number of clusters (k = 5 in K-means) was selected
- Number of iterations (n = 10,000) was selected.

The Knowledge Studio Clustering Experiment

After importing the final data set from the single MS Access database's table to the Knowledge Studio environment and setting the run parameters as mentioned above, the following clustering results were observed. As it is displayed below, the clustering result in the Knowledge Studio environment can be seen using a tree structure for a selected variable. Figure 5.2 (a) shows the screen on which the final split is displayed. Figure 5.2 (b) shows the tree having 10 leaves corresponding to different total year trip values of the customers.

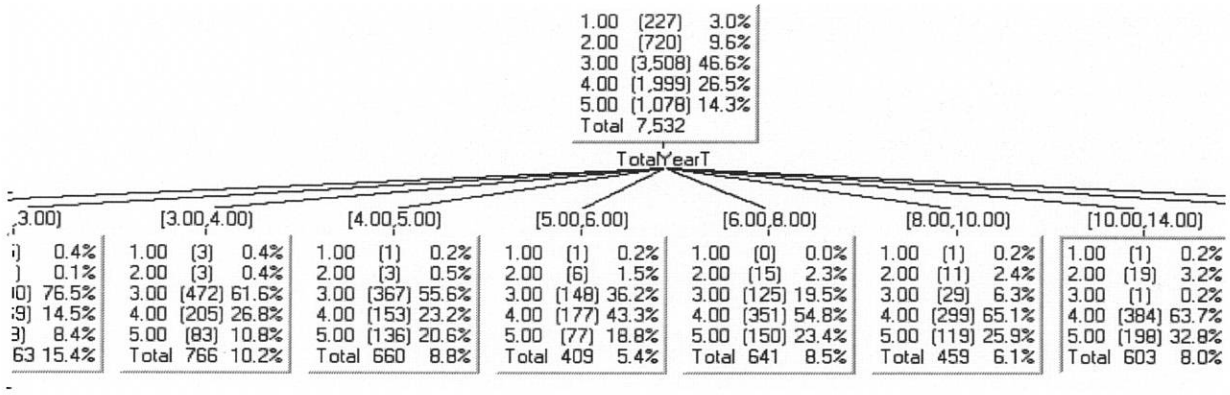


Figure 5.2: (a) TotalYearTrip Split

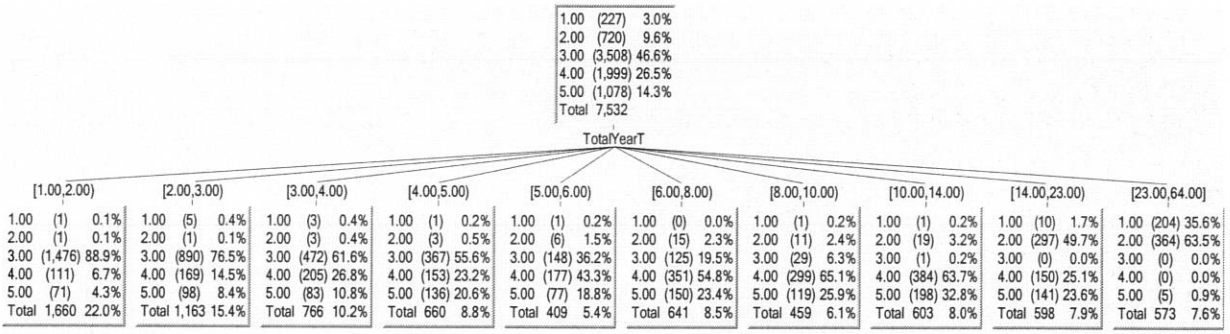


Figure 5.2: (b) Tree with the 10 leaves for different TotalYearTrip ranges

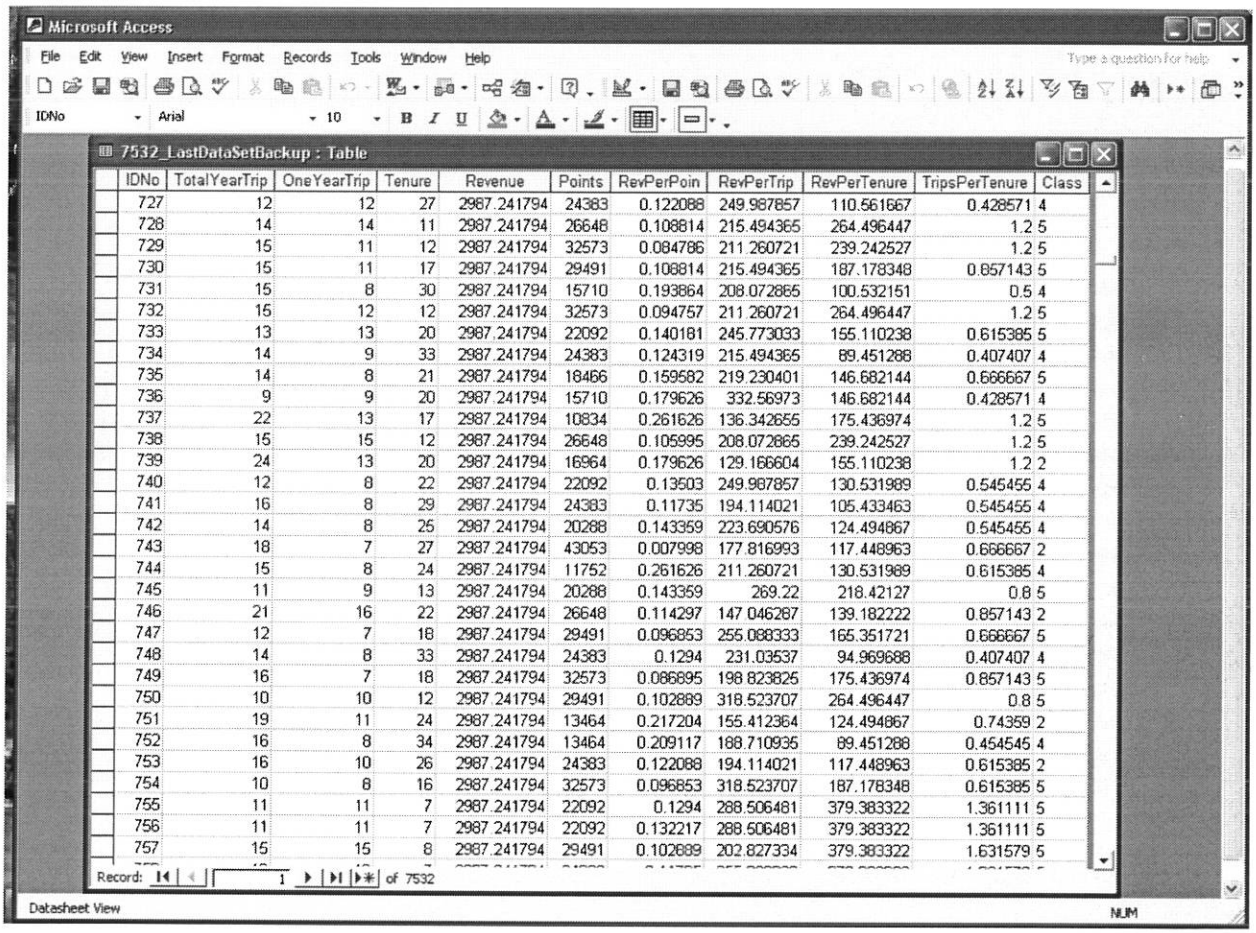
The following figure shows portion of the data set used and the corresponding cluster index of each customer record after it is being clustered by the K-Means algorithm of Knowledge Studio.

The final distribution and relative frequency of records in each of the 5 clusters (as shown in the root node of **Figure 5.2 (a)** and **(b)**) is displayed in Table 5.6.

Index	Number of Records	*Relative Frequency*
C1	227	*3.0%*
C2	720	*9.7%*
C3	3508	*46.5%*
C4	1999	*26.5%*
C5	*1078*	*14.3%*

Table 5.6: The distribution and relative frequency of records in the 5 clusters

(From Knowledge Studio)



IDNo	TotalYearTrip	OneYearTrip	Tenure	Revenue	Points	RevPerPoin	RevPerTrip	RevPerTenure	TripsPerTenure	Class
727	12	12	27	2987.241794	24383	0.122088	249.987857	110.561667	0.428571	4
728	14	14	11	2987.241794	26648	0.108814	215.494365	264.496447	1.2	5
729	15	11	12	2987.241794	32573	0.084786	211.260721	239.242527	1.2	5
730	15	11	17	2987.241794	29491	0.108814	215.494365	187.178348	0.857143	5
731	15	8	30	2987.241794	15710	0.193864	208.072865	100.532151	0.5	4
732	16	12	12	2987.241794	32573	0.094757	211.260721	264.496447	1.2	5
733	13	13	20	2987.241794	22092	0.140181	245.773033	155.110238	0.615385	5
734	14	9	33	2987.241794	24383	0.124319	215.494365	89.451288	0.407407	4
735	14	8	21	2987.241794	18466	0.159582	219.230401	146.682144	0.666667	5
736	9	9	20	2987.241794	15710	0.179626	332.56973	146.682144	0.428571	4
737	22	13	17	2987.241794	10834	0.261626	136.342655	175.436974	1.2	5
738	16	16	12	2987.241794	26648	0.105995	208.072865	239.242527	1.2	5
739	24	13	20	2987.241794	16964	0.179626	129.166604	155.110238	1.2	2
740	12	8	22	2987.241794	22092	0.13503	249.987857	130.531989	0.545455	4
741	16	8	29	2987.241794	24383	0.11735	194.114021	105.433463	0.545455	4
742	14	8	25	2987.241794	20288	0.143359	223.690576	124.494867	0.545455	4
743	18	7	27	2987.241794	43053	0.007998	177.816993	117.448963	0.666667	2
744	15	8	24	2987.241794	11752	0.261626	211.260721	130.531989	0.615385	4
745	11	9	13	2987.241794	20288	0.143359	269.22	218.42127	0.8	5
746	21	16	22	2987.241794	26648	0.114297	147.046267	139.182222	0.857143	2
747	12	7	18	2987.241794	29491	0.096853	255.088333	165.351721	0.666667	5
748	14	8	33	2987.241794	24383	0.1294	231.03537	94.969688	0.407407	4
749	16	7	18	2987.241794	32573	0.086895	198.823825	175.436974	0.857143	5
750	10	10	12	2987.241794	29491	0.102889	318.523707	264.496447	0.8	5
751	19	11	24	2987.241794	13464	0.217204	155.412364	124.494867	0.74359	2
752	16	8	34	2987.241794	13464	0.209117	188.710935	89.451288	0.454545	4
753	16	10	26	2987.241794	24383	0.122088	194.114021	117.448963	0.615385	2
754	10	8	16	2987.241794	32573	0.096853	318.523707	187.178348	0.615385	5
755	11	11	7	2987.241794	22092	0.1294	288.506481	379.383322	1.361111	5
756	11	11	7	2987.241794	22092	0.132217	288.506481	379.383322	1.361111	5
757	15	15	8	2987.241794	29491	0.102889	202.827334	379.383322	1.631579	5

Figure 5.3: Portion of the final data set after being clustered

As partly displayed in Figure 5.3 above, the cluster indexes were assigned to each of the customer records in the final data set. However, Knowledge Studio has no facility of computing and displaying the cluster centroids. As a result it was not possible to use the previous investigation's final summary of clusters (discussed in section 4.5.5 of chapter four) for interpretation. To manage this cluster centroid computation, the Weka-3-2 Software was used.

The Weka-3-2 Clustering Experiment

To associate the final summary of each cluster in the previous investigation with the cluster indexes of the research at hand, finding cluster centroids became an important step. To this end, the final data set was changed into *ARFF* file format (as mentioned in section 5.5 of same chapter), and was fed into Weka so that finding clusters and corresponding centroids happened to be possible. Following the run information (as was displayed in Weka's report window) shows attributes used, cluster centroids computed, Weka's cluster indexes, cluster-wise distribution of records, etc.

```
=== Run information ===
Scheme:     weka.clusterers.SimpleKMeans -N 5 -S 1
Relation:   7532_MemberLevelDataSet
Instances:  7532
Attributes: 10
            TotalYearTrips
            OneYearTrips
            Tenurel
            Revenue
            RevPerTenure
            TripsPerTenure
```

Ignored:

Points

RevPerPoints

RevPerTrip

Class

Test mode: evaluate on training data

===K-Means======

Cluster centroids:

Cluster Index	TtlYears Trips	OneYear Trips	Tenure	Revenue	RevPerTrip	TripPerTen
Cluster 0	2.524	2.186	9.077	637.028	82.457	0.292
Cluster 1	6.311	2.814	27.238	1626.858	62.699	0.231
Cluster 2	8.486	7.030	6.953	2110.234	586.900	1.167
Cluster 3	44.024	19.008	27.732	10337.444	536.830	1.510
Cluster 4	20.805	6.695	32.294	6461.808	241.208	0.629

Index	Number of Records	*Relative Frequency*
C0	3563	*47%*
C1	2093	*28%*
C2	875	*12%*
C3	246	*3%*
C4	*755*	*10%*

Table 5.7: The distribution and relative frequency of records in the 5 clusters

(From Weka)

Comparison of Results based on Distribution of Records in each Cluster

To properly use the summary made in the previous investigation, comparison of the clustering results in the previous research and the investigation at hand had to be made. In this regard, since the quantity of

records used in the two researches were different, the comparison was made based on the relative frequency of records in each cluster. The table below (**Table 5.8**) shows this comparison. The Knowledge Studio's cluster indexes for the research at hand being the reference, the relative frequency distribution of records observed in each cluster of the different clustering experiments were arranged accordingly.

Investigation at Hand						***Previous Investigation***		
Knowledge Studio			**Weka-3-2**			***Knowledge Studio***		
Cluster Index	**Number of Records**	**Relative Frequency**	**Cluster Index**	**Number of Records**	**Relative Frequency**	**Cluster Index**	**Number of Records**	**Relative Frequency**
C1	227	3.0%	**C3**	246	3.0%	**C2**	114	*2.9%*
C2	720	9.7%	**C4**	755	10.0%	**C4**	314	*7.8%*
C3	3508	46.5%	**C0**	3563	47.0%	**C1**	2103	*52.6%*
C4	1999	26.5%	**C1**	2093	28.0%	**C5**	1050	*26.3%*
C5	1078	14.3%	**C2**	875	12.0%	**C3**	419	*10.5%*
Total	*7,532*	*100%*		*7,532*	*100%*		*4,000*	*100%*

Table 5.8: Arrangement of Cluster Indexes based on Relative Distribution of Records

For the investigation at hand, to assign the cluster centroid (found from Weka) to each cluster (found from the Knowledge Studio) correctly, one additional checking mechanism was done. The K-Means algorithm deals with the assignment of records to a cluster based on a similarity measure, which is the Euclidian distance. For a new assignment, the mean of the objects in each cluster is calculated so that similarity of an object, which is to be assigned to a cluster, is computed with this mean. The final cluster centroid is, therefore, the mean of all the objects belonging to the same cluster.

Based on the above concept of final cluster centroid, an attempt was made to calculate mean for each of the final clusters found from the Knowledge Studio, using MS Excel. By taking the proportional records'

cluster indexes displayed in **Table 5.8**, the corresponding cluster centroids from Weka and the calculated mean from the MS Excel are displayed together in **Table 5.9** below.

Cluster Index[4]	Variables						*Records' Frequency*
	TtlYrTrip	OnYrTrip	Tenure	Revenue	RevPerTen	TripPerTen	
C1	41.444	21.440	23.264	9814.255	421.850	1.633	*227*
C3*	**44.024**	**19.008**	**27.732**	**10337.444**	**536.830**	**1.510**	***246***
C2	23.207	7.733	31.026	7172.047	282.922	0.740	*720*
C4*	**20.805**	**6.695**	**32.294**	**6461.808**	**241.208**	**0.629**	***755***
C3	2.257	1.926	9.516	565.318	68.147	0.264	*3508*
C0*	**2.524**	**2.186**	**9.077**	**637.028**	**82.457**	**0.292**	***3563***
C4	7.029	2.835	28.602	1853.185	67.686	0.248	*1999*
C1*	**6.311**	**2.814**	**27.238**	**1626.858**	**62.699**	**0.231**	***2093***
C5	7.494	6.218	7.851	1728.720	250.696	1.012	*1078*
C2*	***8.486***	***7.030***	***6.953***	***2110.234***	***586.900***	***1.167***	***875***

Table 5.9: Cluster centroids from Weka and the calculated mean from the MS Excel

As it can be seen from **Table 5.9**, the closeness of cluster centroids in each pair of clusters confirm that association of the cluster indexes (as shown in **Table 5.8**) from Weka and Knowledge Studio of the research at hand is reasonable.

In conclusion, to interpret the cluster indexes of the Knowledge Studio's clustering result of the current research into the corresponding concept/class description of the previous research (discussed in **section 4.5.5** of chapter four), the Weka's cluster centroids will be used. That is, the table below (**Table 5.10**) will be used.

[4] The C_i's where i = 1,2, 3, 4, 5, are the cluster indexes of the clusters formed by the Knowledge Studio. The C_j*'s where j = 0, 1, 2, 3, 4, are the cluster indexes of the clusters formed by Weka

Cluster Index	Variables						*Records' Frequency*
	TtlYrTrip	OnYrTrip	Tenure	Revenue	RevPerTen	TripPerTen	
C1	**44.024**	**19.008**	**27.732**	**10337.444**	**536.830**	**1.510**	*227*
C2	**20.805**	**6.695**	**32.294**	**6461.808**	**241.208**	**0.629**	*720*
C3	**2.524**	**2.186**	**9.077**	**637.028**	**82.457**	**0.292**	*3508*
C4	**6.311**	**2.814**	**27.238**	**1626.858**	**62.699**	**0.231**	*1999*
C5	***8.486***	***7.030***	***6.953***	***2110.234***	***586.900***	***1.167***	*1078*

Table 5.10[5]: Cluster Indexes and their corresponding Cluster Centroids

Interpretation of the Cluster Indexes

The interpretation of cluster indexes, displayed in **Table 5.10,** was done by using the class/concept description of clusters in the previous investigation. To facilitate this, the relative distributions of records in the clusters of both the previous and the current research, which are displayed in **Table 5.8,** were used. Moreover, to compare the ordinal values, displayed in **Table 4.4** of section 4.4.4, with the numerical values, the cluster centroids displayed in **Table 5.10** were used.

The combined information from the three tables (**Table 4.4, Table 5.8, and Table 5.10**) is displayed in **Table 5.11.** This table (**Table 5.11)** enabled to compare the possible ordinal values and the cluster centroids used to describe the resulting clusters in the previous and current researches respectively.

[5] The cluster indexes and total records are from the Knowledge Studio and the corresponding cluster centroids are from Weka

Cluster Index[6]	Variables						*Relative Frequency*
	TtlYrTrip	OnYrTrip	Tenure	Revenue	RevPerTen	TripPerTen	
C1	44.024	19.008	27.732	10337.444	536.830	1.510	*3.0%*
C2*	**VH**	**VH**	**Lg**	**VH**	**VH**	**H**	***2.9%***
C2	20.805	6.695	32.294	6461.808	241.208	0.629	*9.7%*
C4*	**H**	**VH**	**VLg**	**H**	**H**	**H**	***7.8%***
C3	2.524	2.186	9.077	637.028	82.457	0.292	*46.5%*
C1*	**VLw**	**VLw**	**R**	**VLw**	**VLw**	**M**	***52.6%***
C4	6.311	2.814	27.238	1626.858	62.699	0.231	*26.5%*
C5*	**Lw**	**M**	**Lg**	**Lw**	**VLw**	**Lw**	***26.3%***
C5	8.486	7.030	6.953	2110.234	586.900	1.167	*14.3%*
C3*	***H***	***H***	***M***	***M***	***H***	***H***	***10.5%***

Table 5.11: The Ordinal Values and Cluster Centroids of the Clusters in the Previous and Current researches respectively

As can be observed from **Table 5.11,** the attribute-wise association of the ordinal values with the numerical centroids is relatively consistent except in two of the six attributes, **OnYrTrip, RevPerTen.** The reason behind this might be the quantity of records in the clustering experiments. However, as it will be discussed in one of the classification experiments, out of the six attributes, the **OnYrTrip** attribute is not as important as the remaining five attributes. Therefore, the inconsistency of association in **OnYrTrip** attribute was found to have by far a minimum impact in the interpretation of cluster centroids.

Consequently, the cluster index and their corresponding class/concept description, for the investigation at hand were associated to the reasonable index. This association is displayed in the table below, **Table 5.12.**

---

[6] The C_i's where i = 1,2, 3, 4, 5, are the cluster indexes of the clusters for the investigation at hand.
The C_i*'s where i = 1, 2, 3, 4, 5, are the cluster indexes of the clusters in the previous investigation

Current Res Index	Previous Res Index	Relative Frequency	Concept/Class Description	Remark
C1	C2	3.0%	High tiered and most valuable group among the groups and is termed as the 'loyal' customer group	Addition of a fourth tier level (possibly platinum) might be good to keep the loyalty of this group.
C2	C4	9.7%	This is the other top tier segment where customers belonging to this segment show characteristics of loyalty.	More targeted promotions could elevate the members' status from highly frequent travelers to that of very highly frequent travelers
C3	C1	46.5%	Members who are not worth giving immediate attention in terms of target promotion as compared with members in other groups	As recent they are, it is rather premature to judge potential value of customers in this group
C4	C5	26.5%	Customers with few trips but significant revenue contribution.	Targeted promotional campaigns might entice this group to make frequent trips.
C5	*C3*	*14.3%*	*This group is worth giving immediate attention, with an enticing targeted promotion, as it has a potential to join the top tier group.*	The likelihood of losing this segment to the competition is high due to its medium tenure in the program.

Table 5.12: The cluster indices of both the current and previous researches and their corresponding class/concept description

At this point, it should be noted that, for reason of convenience, the cluster indexes that were used for classification are **i's** where **i = 1, 2, 3, 4, 5,** instead of **C1, C2, C3, C4,** and **C5** respectively.

5.6.2 The Classification Sub-Phase

Data mining tools might have to infer a model from databases. In the case of supervised learning inferring a model from databases requires one or more class labels. Such class labels could be given in advance, defined by the user or derived from clustering. Moreover, segmentation can either provide the class labels or restrict the dataset so that good classification models can be built (Chapman, P. et al., 1999, 2000). In this regard, the task that was done and reported in the clustering sub-phase allowed to get the class labels and then the classification process to be carried out. As a result, it was found that the

ShebaMiles database contained five important attribute values that denote the class-labels (cluster index) for a given customer record.

Classification is learning a function that maps (classifies) a data item in to one of several predefined classes. In relation to this, Chapman, P. et al. (1999, 2000) noted that when learning classification rules, the system had to find the rules that predict the class-label, which is the dependent or predicted attribute's value, from the independent or predicting attributes' value. Accordingly, it was the concern of this sub-phase to generate classification rules that would assign the correct class label to previously unseen and unlabeled customers.

In this sub-phase, the main concern was to build a classification model. Due to the fact that well-grown decision trees are as comparably useful as other classifiers, the type of classification model selected to be built was decision tree. The other reason for selecting decision tree model is, compared to other classification model types, decision tree has a significant advantage because it can be built manually - and so, is easily explained. Apart from this, decision tree operations are completely interactive and they benefit from powerful visualization features.

To carry out this sub-phase, two different software, Weka and See_5 (release 1.5) had been attempted to use. Since the See_5 was found to be a trial version, it couldn't process the total record used for the classification task. As a result, the classification model built using See_5 was not used in any further discussion. Therefore, further discussion in this investigation was done by making use of the models found from the selected Weka's algorithms runs.

Decision Tree Model Building

As decision tree is a classifier, any previously unseen record with the required degree of attributes can be fed into the tree. At each node, it will be sent either left or right to some test. Eventually, it will reach a leaf node and be given the label associated with that leaf. At this junction, we were interested to generating rules of assigning the *ETHIOPIAN* frequent flier program members to the class they belong.

With this objective, the total members' record considered in the clustering sub phase was used to construct a number of decision trees. The classification algorithms were implemented in Weka-3-2 to build a classification model in such a way that testing the model would be possible after training it. For most of the experiments carried out in this sub-phase, excepting the experiments where 10-fold cross-validation technique was used, the total record was partitioned in to two, the training and test, sub-data sets. These two sub-data sets were found from the final data set by using a stratified sampling technique where the different clusters, found in the clustering sub-phase, were considered as strata.

The reason behind applying stratified sampling technique was, partitioning the total record where the contribution of each cluster in the resulting sub-data sets could be proportional. To avoid the problem of over-fitting, 25% of the total record was selected as a test sub-data set and the remaining 75% as a training sub-data set. The proportional selection of records, for training and testing the different decision trees built, is shown in **Table 5.13** below.

Cluster	Total	Training Data	*Test Data*
1	227	170	*57*
2	720	540	*180*
3	3508	2631	*877*
4	1999	1499	*500*
5	1078	809	*269*
Total	*7532*	*5649*	*1883*

Table 5.13: Proportional selection of Training and Test sub-data sets for Weka-3-2

Experiments using Weka-3-2

Weka-3-2 supports many types of classification algorithms. Among the classification algorithms that Weka-3-2 supports, the J4.8 algorithm was used with different input parameters as well as different types of related classifiers. J4.8 algorithm is Weka's implementation of the C4.5 decision tree learner. The corresponding algorithm used to extract rules from the decision trees is J4.8 PART.

By making use of Weka-3-2 a total of 12 experiments were carried out, where 6 of the experiments were for constructing decision trees and the remaining 6 were for extracting the corresponding classification rules from the decision trees. In relation to this, J4.8 was the algorithm used to construct the decision trees in the 6 experiments. The extraction of the corresponding rules, in the remaining 6 experiments, from the decision trees was managed using J4.8 PART.

To display the run parameters and the outputs of the respective experiments, two tables (**Table 5.14** and **Table 5.15**) are used. As displayed in both tables, the different experiments were carried out by using two different schemes (for attribute selection) and three different test modes (ways of feeding records to the algorithms).

The two schemes used were:

1. Using all the 9 attributes of the records
2. Using an algorithm that selects the best attributes from the 9 attributes of the records

As a result, four different combinations of these two schemes and two algorithms, J4.8 and J4.8 PART, were used in the experiments. These combinations were:

a. S1: scheme one used with J4.8 decision learner algorithm
b. S2: scheme two used with J4.8 decision learner algorithm
c. S3: scheme one used with J4.8 PART decision rule extractor algorithm
d. S4: scheme two used with J4.8 PART decision rule extractor algorithm

The three test modes were:

1. T1: Inputting all the records without any test,
2. T2: Inputting all the records with a 10-fold cross-validation mode, and
3. T3: Inputting part of the records to train a model and then supply the unseen remaining part of the record for testing the performance of the model.

Experiment	Scheme	Number of Records	Number of attributes		Number of leaves	Size of Tree	Test Mode	Time Taken	Accuracy
			Inputted	Used					
1	S1	7532	9	9	87	173	T1	3.72	99.28%
2	S1	7532	9	9	87	173	T2	3.02	97.68%
3	S1	5649	9	9	82	163	T1	2.57	99.31%
		1883	9	9	82	163	T3	2.4	96.60%
4	S2	7532	9	5	58	115	T1	7.06	97.85%
5	S2	7532	9	5	58	115	T2	8.24	96.57%
6	S2	5649	9	5	68	135	T1	2.54	97.88%
		1883	*9*	*5*	*68*	*135*	*T3*	*5.54*	95.06%

Table 5.14: Input parameters and the resulting Decision Trees' output parameters

As can be observed from **Table 5.14**, the number of leaves and the corresponding sizes of the trees constructed from experiments 4 and 5 are less than those found form the other experiments. Moreover, these experiments (expt. 4 and 5) have used only 5 attributes to construct the decision tree unlike experiments 1 and 3, which used all the 9 attributes. However, the respective accuracies found from these experiments are not as good as from experiment 1 and 3.

To examine the corresponding rules extracted from each decision tree in order, the table below (**Table 5.15**) could be used.

Experiment	Scheme	Number of Records	Number of attributes		Number of Rules	Test Mode	Time Taken (sec.)	*Accuracy*
			Inputted	Used				
7	S3	7532	9	9	61	T1	5.83	*99.63%*
8	S3	7532	9	9	61	T2	5.34	*97.70%*
9	S3	5649	9	9	50	T1	3.71	*99.45%*
		1883	9	9	50	T3	3.78	*96.87%*
10	S4	7532	9	5	59	T1	17.1	*98.05%*
11	S4	7532	9	5	59	T2	9.08	*96.35%*
12	S4	5649	9	5	49	T1	8.57	*97.98%*
		1883	*9*	*5*	*49*	*T3*	*3.51*	*95.27%*

Table 5.15: Input parameters and the resulting extracted Decision Rules' output parameters

As can be observed from this table observed from this **Table 5.15**, experiments 10 and 11 have comparatively better than the other experiments with 59 extracted rules. This is because, unlike the other experiments, the number attributes used in these experiments is less than in experiments 7, 8, and 9. Moreover, regardless of the minimum number of attributes used in these experiments, 10 and 11, the accuracy found is not as such less than the experiments 7, 8, and 9.

Though experiments 4 and 10 were examined, respectively, with experiments 5 and 11, in the discussion above, they were simply done by feeding the algorithms with all the data having no test mode. Therefore, from **Table 5.14**, since experiment 5 was carried out to construct the required decision tree with a 10-fold cross-validation and had a reasonably good accuracy, it was selected. In addition to this, experiment 11, which is the corresponding rule extraction experiment from the decision tree constructed in experiment 5,

was selected. In this regard, generally, the reasons of selecting experiments 5 and 11 from all the experiments carried out could be mentioned follows:

- The number of records considered is relatively large.
- The number of attributes selected and used is minimum
- The number of leaves and size of the tree in experiment 5 are manageable; and the number of rules extracted in experiment 11 is reasonable.
- The test mode, which is the 10-fold cross-validation, used in both experiments is acceptable.
- The accuracy of the resulting model is comparatively good.

As a result, the full outputs of the selected working decision tree in experiment 5 and the corresponding rules extracted in experiment 11 are annexed for reference.

5.7 The Customer Classification System: A Prototype

A trained model in data mining can be used to process transactions and perform classification and prediction on data in the real time.

In this study an attempt was made to develop an operational application prototype named Customer Classification System that uses the classification rules generated from the decision tree learner in the classification sub-phase of this chapter. The prototype is used to classify a customer into one of the customer clusters, to generate cluster reports, search for a customer and find the cluster where the customer belongs, and also provides with the description of each customer clusters. The Customer Classification System contains

MS access database, the MS visual basic program hosting the classification rule. Some of the user interfaces and reports of this system are displayed in **Fig. 5.8** and **Fig 5.9**.

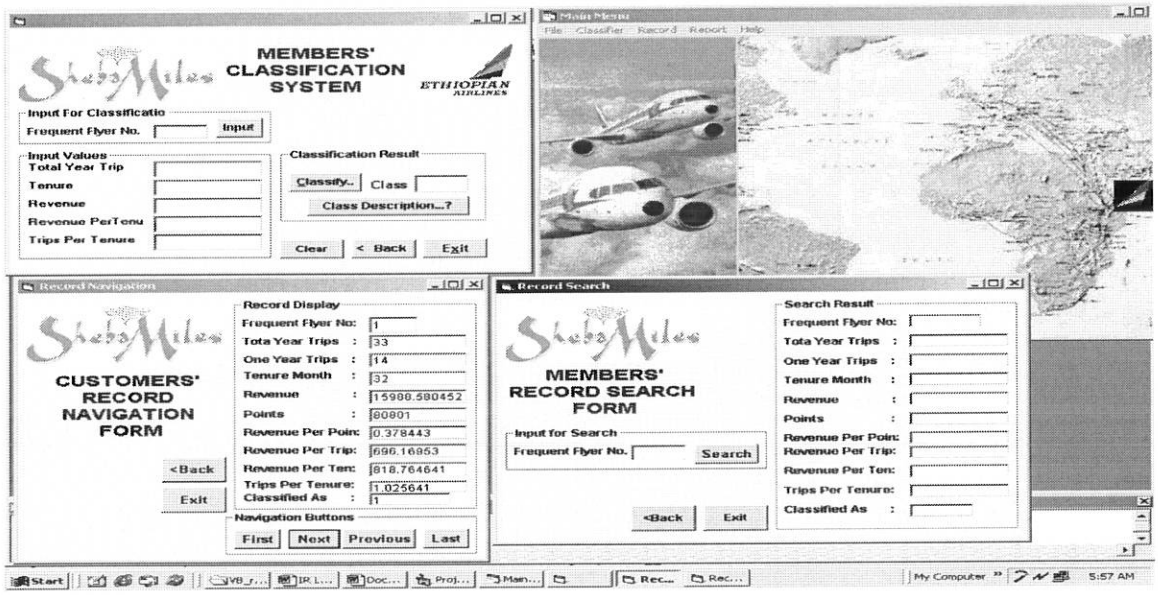


Figure 5.8: Portion of the Customer Classification System User Interface

Figure 5.8 shows the user interfaces that are developed using the MS Visual Basic programming language. In order of their appearance in the figure, the interfaces are; the members' classification System, the main menu, the customers' record navigation form and the members' record search form.

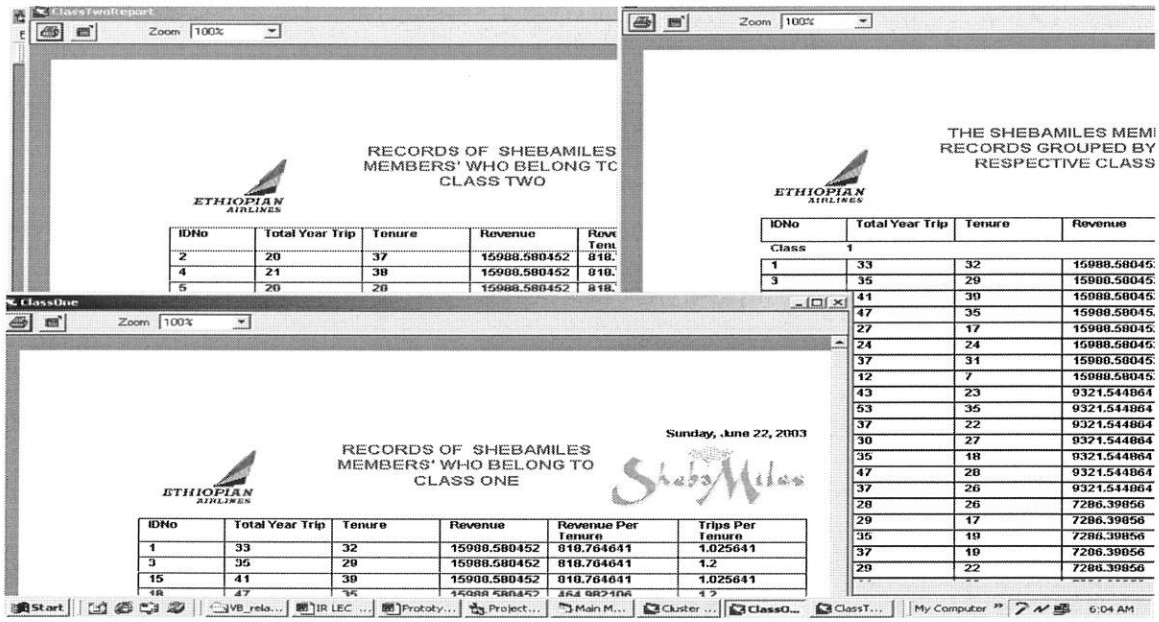


Figure 5.9: Portion of the Customer Classification System reports.

CHAPTER SIX

CONCLUSION AND RECOMMENDATIONS

6.1 Conclusion

Nowadays advances in communication technologies, on the one hand, and computer hardware and database technologies, on the other, have enabled organizations to collect, store and manipulate massive amounts of data. Having concentrated on the accumulation of data, the question is what to do next with this valuable resource? Indeed, the data contains and reflects activities and facts about the organization. But the data's hidden value, the potential to predict business trends and customer behavior, has largely gone unexploited. The increase in data volume causes great difficulties in extracting useful information and knowledge for decision support.

It is to bridge this gap of analyzing large volume of data and extracting useful information and knowledge for decision making that the new generation of computerized methods known as Data Mining or Knowledge Discovery in Databases (KDD) has emerged in recent years.

The application of data mining technology has increasingly become very popular and proved to be relevant for many sectors such as retail trade, airline industry, telecommunications, banking, and healthcare sectors Particularly in the airline industry, data mining technology has been applied for customer segmentation for effective and efficient customer relationship management (CRM), causes for delays at airports, airline pricing.

This research has tried to assess the application of data mining technology in customer relationship management (CRM) at Ethiopian Airlines, *ETHIOPIAN*, for developing a classification model. Such a classification model could enable the customer loyalty department (CLD) to implement CRM in *ETHIOPIAN* efficiently and effectively.

This investigation, conducted according to the CRISP-DM process model, was carried-out in five major parts namely: business understanding, data understanding, data preparation, model building, and evaluation. However, since a data mining task is an iterative process, these steps were not followed strictly.

A data set with 7,532 total frequent flyer program members' records was used to develop a classification model. Since this research was intended to fill a gap left by a related research, some valuable experiences of the previous research were used. However, though the previous research's objective was to identify the possible customer segments by using the same data, the cluster indexes were not documented along with each customer record. Because of this the model development phase of this research was done in two sub-phases, the clustering sub-phase and the classification sub-phase one after the other. In the clustering sub-phase, the K-means clustering algorithm; while in the classification sub-phase, the j4.8 algorithm, which is Weka's implementation of the C4.5 algorithm were used.

In order to select a classification model that can classify the frequent flyer program members, about 12 different models were built by employing the decision tree algorithms. In the decision tree selection process, more emphasis was given to important attributes to be used, the number of records considered and the size of the tree and the corresponding number of rules extracted from the tree. Though the

number of attributes selected and used for the customer segmentation, both in the previous and this investigations, were 9, only 5 of these attributes were found to be useful for the classification.

From the mini-experiments done using a trial version of See_5 (Release 1.5), it was observed that, for a given number of attributes, as the number of records used to develop a decision tree increases the corresponding number of rules generated will possibly increase. Due to this observation, not to get a minimum number of rules, among all the models developed for comparison, the models developed from the 7,532 records with a 10-fold cross-validation and attribute selection were given due attention. Accordingly, the better decision tree with the corresponding extracted rules was selected as a working model to classify members into their corresponding clusters. As a result, the classification accuracy of the selected decision tree seems convincing. That is, among the 7,532 data inputted to the model learner with a 10-fold cross-validation test mode, 96.575%, which is 7274 records were correctly classified.

The suggestions and opinions given by domain experts in the entire investigation were observed and found to be very important in the model development process, particularly, in the clustering sub-phase.

The results obtained in this research work have proved the potential applicability of data mining technology to classify airline customers into predefined clusters based solely on factors such as revenue contribution, travel frequency, number of months since the customer became a member of the program.

The overall model building process made by employing the clustering algorithm and then decision tree techniques demonstrated that data mining is a method that should be considered to support customer relationship management in organization and industries like airlines having a rich customer records.

6.2 Recommendations

This investigation has been conducted mainly for an academic purpose. However, it revealed the potential applicability of data mining technology to classify customers for an effective customer relationship management. Moreover, it is the researcher's belief that the contribution of this research work could be a good experience for a competitive study in customer-oriented businesses in the future. Apart from this, it is also the researcher's belief that the findings of the research would encourage business oriented organizations to work on the application of data mining technology to appreciate and employ customer relationship management, and as a result gain a competitive advantage.

It has been observed that important customer attributes were left unfilled with their required values and finding a full customer's profile has amounted to gathering data from different operational databases. The data collection and data preparation from different operational databases are very lengthy and tedious tasks that could hinder any customer-data driven marketing researches, like data mining. Therefore, the researcher strongly recommends the following:

- The airline should embark on developing an integrated customer data warehouse. This data warehouse should be an integration of customer oriented data sources, such as booking databases, departure control databases, sales information databases, frequent flyer databases, and customer services databases. Moreover, in each database, due attention should be given for each customer's record, and each customer attribute should be considered as valuable as the full customer record.

Since the airline business is highly data intensive, the airline needs to give appropriate credit for the application of data mining technology for gaining a competitive advantage and sustainable profits. So, the researcher's recommendation, in relation to this is:

- ✈ The airline should encourage data mining researches both on its customer and operational databases. This could be achieved by a data mining team formed from the business experts of the airline, the information technology professionals in the airline and external data mining professionals.

Due to the predestined multiple tasks to be carried-out and shortage of time, the model building phase of this research was not exhaustive in that it didn't consider all the frequent flier program's members. Though it has been done with the domain experts, it was a limited number of all the possible member attributes, which were available with their values that were used in the data mining task. However, it is known that a data mining task benefits from and provides an acceptable and valid result through the use of as much data and attributes as possible. In this regard, the researcher recommends the following for future related investigations:

- ✈ The inclusion of as much customer attributes as possible should be given due attention and more comprehensive models should be built by using large training and testing datasets taken from the relevant customer databases in the airline.

The model building process in this investigation was carried out in two sub-phases by making use of the K-means clustering algorithm and then the C4.5 decision tree learner algorithm, respectively. Though the results found were encouraging, refinement to the segmentation results using other clustering methods

and also a corresponding classification model development by using other classification methods could bring a good result. Based on this, the researcher recommends the following for further data mining projects:

- Further data mining projects should be undertaken by using neural clustering algorithms (SOM), where the clustering results found in this investigation can be confirmed or adjusted. Moreover the performance of the decision tree model for assigning new records to the appropriate clusters, which is currently equal to **96.575%,** need to be improved. Results could be improved through the use of other classification and prediction techniques such as neural networks, Belief Networks and making comparison between the accuracy of the corresponding classification models.

BIBILIOGRAPHY

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

Askale. (2001). The application of the Data Mining Technique in supporting loan disbursement activities at Dashen Bank S.C. *Unpublished Master's Thesis.*

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

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

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

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

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

Canaday, Henry. (1999). What does CRM mean for Airlines? Airlines International.

Carbone, P. L. (1997). Data Mining or "Knowledge Discovery in Databases": An Overview. The MITRE Corporation. USA

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

Chapman, P. et al. (1999, 2000). CRISP-DM 1.0 Step-by-step data mining guide SPSS Inc., U.S.A CRISPWP-0800

Chen, M.S., et al. (1997). Data Mining : An Overview from Database Perspective. National Taiwan University, Taipei, Taiwan.

Connolly, T., et al. (1999). Database Systems: A Practical Approach to Design, Implementation and Management. 2nd ed. Addison-Wesley, New York.

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

DSS Research. (2001) Understanding Market Segmentation.. Online. Internet.

http://www.dssresearch.com/library/segment/understanding.asp

Edelstein, Herb. (1998). Data Mining-Let's get Practical: How to identify strategic problem statement, prepare the Right data, and build and apply a robust model. Database programming & DesignMagazine. Online. Internet.
http://www.db2mag.com/98smEdel.htm

Edelstein, H. (2000). Building profitable customer relationships with data mining SPSS Inc. U.S.A CRMBPWP-0500

EDDS. (2001). Data Mining Case Studies.. Online. Internet. http://www.eds.ch

Ethiopian Airlines. (2002). Annual Report. Addis Ababa, Ethiopian Airlines.

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

Ethiopian Airlines. (n.d.). ShebaMiles Frequent Flyer Pogramme Membership Guide. Addis Ababa, Ethiopian Airlines.

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

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

Goebel, M. and Le Gruenwald. (1999). A Survey of Data Mining and Knowledge Discovery Software Tools

Gobena Mikael. (2000). Flight Revenue Information Support System for Ethiopian Airlines.
Addis Ababa University. *Unpublished Master's Thesis.*

Hagg, S., et al. (1998). Management Information Systems for the Information Age.
New York, McGraw-Hill.

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

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

Henok Wobishet. (2002). The application of Data Mining to support customer relationship

management at Ethiopian Airlines. Addis Ababa University. *Unpublished Master's Thesis.*

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

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

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

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

Kohonen, T. (2001). Self-Organizing Maps, 3rd ed. Springer, Berlin.

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

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

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

Tesfaye Hintsay. (2002). Predictive Modeling Using Data Mining Techniques in Support of Insurance Risk Assessment. Addis Ababa University, *Unpublished Masters Thesis*

Thearling, Kurt. (2000). Data Mining and Customer Relationships. Online. Internet.
http://www3.primuhost.com/~kht/text/whexcerpt/ whexcerpt.html

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

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

Witten, Ian H. and Eiber Frabk. (2000).Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. San Francisco: Morgan Kaufmann Publishers

APPENDICES

Appendix I

A procedure used to calculate Tenure Month for each member of the frequent flyer program

```
Dim cn As ADODB.Connection
Dim rs As ADODB.Recordset

Private Sub cmdTenureMonth_Click()

    Set cn = New ADODB.Connection
    Set rs = New ADODB.Recordset

    Dim d
    Dim ffN

    cn.Open "Provider = microsoft.jet.oledb.4.0;data source = D:\Data Cleaning\Sheba.mdb"


    Set rs = cn.Execute("FilteredMember")

    While Not rs.EOF

      ffN = rs.Fields("ffNum").Value

      d = DateDiff("m", rs.Fields("enrl_date").Value, #4/19/2002#)
      cn.Execute ("UPDATE FilteredMember SET FilteredMember.[TenureMonth] " & _
            "  = '" & d & " ' where FilteredMember.[ffNum]  = '" & ffN & "' ")
      rs.MoveNext

    Wend
    rs.Close
    Set rs = Nothing

End Sub
```

Appendix II

Decision Tree Constructed as an output of Experiment 5 using Weka-3-2

```
=== Run information ===

Scheme:       weka.classifiers.AttributeSelectedClassifier -B "weka.classifiers.j48.J48 -L -C 0.25 -M 2 -A" -
E "weka.attributeSelection.CfsSubsetEval " -S "weka.attributeSelection.BestFirst -D 1 -N 5"
Relation:     LastDataSet
Instances:    7532
Attributes:   10
              TotalYearTrips
              OneYearTrips
              Tenurel
              Revenue
              Points
              RevPerPoints
              RevPerTrip
              RevPerTenure
              TripsPerTenure
              Class
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

Selected attributes: 1,3,4,8,9 : 5
                     TotalYearTrips
                     Tenurel
                     Revenue
                     RevPerTenure
                     TripsPerTenure

Header of reduced data:
@relation 'LastDataSet-weka.filters.AttributeFilter-V-R1,3-4,8-10'

@attribute TotalYearTrips numeric
@attribute Tenurel numeric
@attribute Revenue numeric
@attribute RevPerTenure numeric
@attribute TripsPerTenure numeric
@attribute Class {1,2,3,4,5}
@data
```

```
Classifier Model
J48 pruned tree
------------------

Tenurel <= 19
|  TripsPerTenure <= 0.54
|  |  TotalYearTrips <= 5
|  |  |  RevPerTenure <= 218.42
|  |  |  |  Tenurel <= 18: 3 (3067.0)
|  |  |  |  Tenurel > 18
|  |  |  |  |  RevPerTenure <= 66.25: 3 (147.0/5.0)
|  |  |  |  |  RevPerTenure > 66.25
|  |  |  |  |  |  TotalYearTrips <= 3: 3 (4.0/1.0)
|  |  |  |  |  |  TotalYearTrips > 3: 4 (21.0/3.0)
|  |  |  RevPerTenure > 218.42
|  |  |  |  RevPerTenure <= 464.98: 3 (20.0)
|  |  |  |  RevPerTenure > 464.98
|  |  |  |  |  RevPerTenure <= 818.76: 2 (6.0)
|  |  |  |  |  RevPerTenure > 818.76: 1 (4.0)
|  |  TotalYearTrips > 5
|  |  |  Tenurel <= 17
|  |  |  |  TripsPerTenure <= 0.5
|  |  |  |  |  Revenue <= 2503.55: 3 (116.0/5.0)
|  |  |  |  |  Revenue > 2503.55
|  |  |  |  |  |  Tenurel <= 15: 3 (2.0)
|  |  |  |  |  |  Tenurel > 15: 4 (6.0/1.0)
|  |  |  |  TripsPerTenure > 0.5
|  |  |  |  |  Tenurel <= 16
|  |  |  |  |  |  TripsPerTenure <= 0.51: 3 (19.0/7.0)
|  |  |  |  |  |  TripsPerTenure > 0.51
|  |  |  |  |  |  |  RevPerTenure <= 100.53: 3 (6.0)
|  |  |  |  |  |  |  RevPerTenure > 100.53: 5 (14.0/3.0)
|  |  |  |  |  Tenurel > 16
|  |  |  |  |  |  Revenue <= 1421.85: 3 (2.0/1.0)
|  |  |  |  |  |  Revenue > 1421.85: 4 (3.0/1.0)
|  |  |  Tenurel > 17
|  |  |  |  Revenue <= 9321.54
|  |  |  |  |  RevPerTenure <= 52.99: 3 (10.0/3.0)
|  |  |  |  |  RevPerTenure > 52.99: 4 (127.0/11.0)
|  |  |  |  Revenue > 9321.54: 2 (4.0)
|  TripsPerTenure > 0.54
|  |  TotalYearTrips <= 19
```

```
|  |  |  Revenue <= 3606.24
|  |  |  |  TripsPerTenure <= 0.66
|  |  |  |  |  TotalYearTrips <= 3: 3 (79.0)
|  |  |  |  |  TotalYearTrips > 3
|  |  |  |  |  |  Revenue <= 543.82: 3 (6.0)
|  |  |  |  |  |  Revenue > 543.82: 5 (170.0/15.0)
|  |  |  |  TripsPerTenure > 0.66: 5 (825.0/3.0)
|  |  |  Revenue > 3606.24
|  |  |  |  Revenue <= 9321.54
|  |  |  |  |  Tenurel <= 16: 5 (46.0/2.0)
|  |  |  |  |  Tenurel > 16
|  |  |  |  |  |  Revenue <= 4534.37
|  |  |  |  |  |  |  Tenurel <= 18: 5 (5.0/1.0)
|  |  |  |  |  |  |  Tenurel > 18: 2 (5.0/1.0)
|  |  |  |  |  |  Revenue > 4534.37: 2 (10.0)
|  |  |  |  Revenue > 9321.54
|  |  |  |  |  Tenurel <= 7: 1 (9.0)
|  |  |  |  |  Tenurel > 7: 2 (5.0)
|  |  TotalYearTrips > 19
|  |  |  TripsPerTenure <= 1.36
|  |  |  |  Revenue <= 4028.96
|  |  |  |  |  TotalYearTrips <= 20: 5 (4.0)
|  |  |  |  |  TotalYearTrips > 20
|  |  |  |  |  |  Tenurel <= 17: 5 (3.0/1.0)
|  |  |  |  |  |  Tenurel > 17: 2 (5.0)
|  |  |  |  Revenue > 4028.96: 2 (29.0/6.0)
|  |  |  TripsPerTenure > 1.36: 1 (86.0/13.0)
Tenurel > 19
|  TotalYearTrips <= 15
|  |  Revenue <= 5169.55
|  |  |  Tenurel <= 21
|  |  |  |  RevPerTenure <= 33.32
|  |  |  |  |  Tenurel <= 20: 3 (20.0)
|  |  |  |  |  Tenurel > 20
|  |  |  |  |  |  TotalYearTrips <= 1: 3 (10.0)
|  |  |  |  |  |  TotalYearTrips > 1: 4 (18.0/1.0)
|  |  |  |  RevPerTenure > 33.32: 4 (77.0/7.0)
|  |  |  Tenurel > 21: 4 (1708.0/22.0)
|  |  Revenue > 5169.55
|  |  |  Revenue <= 6044.99
|  |  |  |  TotalYearTrips <= 12: 4 (7.0)
|  |  |  |  TotalYearTrips > 12: 2 (12.0/1.0)
   |  |  |  Revenue > 6044.99: 2 (43.0)
```

```
|  TotalYearTrips > 15
|  |  TripsPerTenure <= 1.02
|  |  |  RevPerTenure <= 130.53
|  |  |  |  TotalYearTrips <= 21
|  |  |  |  |  TripsPerTenure <= 0.67
|  |  |  |  |  |  Revenue <= 2987.24
|  |  |  |  |  |  |  RevPerTenure <= 105.43: 4 (27.0/3.0)
|  |  |  |  |  |  |  RevPerTenure > 105.43: 2 (3.0)
|  |  |  |  |  |  Revenue > 2987.24
|  |  |  |  |  |  |  TotalYearTrips <= 16: 4 (19.0/5.0)
|  |  |  |  |  |  |  TotalYearTrips > 16
|  |  |  |  |  |  |  |  Revenue <= 4028.96
|  |  |  |  |  |  |  |  |  TripsPerTenure <= 0.5
|  |  |  |  |  |  |  |  |  |  RevPerTenure <= 94.96: 2 (2.0)
|  |  |  |  |  |  |  |  |  |  RevPerTenure > 94.96: 4 (7.0/1.0)
|  |  |  |  |  |  |  |  |  TripsPerTenure > 0.5: 2 (33.0/6.0)
|  |  |  |  |  |  |  |  Revenue > 4028.96: 2 (7.0)
|  |  |  |  |  TripsPerTenure > 0.67: 5 (3.0/1.0)
|  |  |  |  TotalYearTrips > 21: 2 (39.0)
|  |  |  RevPerTenure > 130.53: 2 (477.0/19.0)
|  |  TripsPerTenure > 1.02
|  |  |  TotalYearTrips <= 35: 2 (27.0/5.0)
|  |  |  TotalYearTrips > 35
|  |  |  |  RevPerTenure <= 218.42
|  |  |  |  |  Revenue <= 4534.37: 1 (3.0)
|  |  |  |  |  Revenue > 4534.37
|  |  |  |  |  |  TripsPerTenure <= 1.2: 2 (5.0/1.0)
|  |  |  |  |  |  TripsPerTenure > 1.2
|  |  |  |  |  |  |  TotalYearTrips <= 43: 1 (2.0)
|  |  |  |  |  |  |  TotalYearTrips > 43
|  |  |  |  |  |  |  |  TotalYearTrips <= 47: 2 (2.0)
|  |  |  |  |  |  |  |  TotalYearTrips > 47: 1 (3.0/1.0)
|  |  |  |  RevPerTenure > 218.42: 1 (113.0/6.0)

Number of Leaves:          58
Size of the tree:          115
Time taken to build model:  8.24 seconds

=== Stratified cross-validation ===
=== Summary ===

Total Number of Instances               7532
Correctly Classified Instances          7274          96.5746 %
```

Incorrectly Classified Instances 258 3.4254 %

=== Detailed Accuracy by Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.855	0.005	0.855	0.855	0.855	1
0.896	0.009	0.911	0.896	0.903	2
0.988	0.008	0.991	0.988	0.989	3
0.971	0.015	0.959	0.971	0.965	4
0.953	0.007	0.956	0.953	0.954	5

=== Confusion Matrix ===

a	b	c	d	e	<-- classified as
194	28	0	0	5	| a = 1
17	645	0	46	12	| b = 2
0	0	3467	21	20	| c = 3
0	24	24	1941	10	| d = 4
16	11	9	15	1027	| e = 5

Rule set extracted as an output of Experiment 11 using Weka-3-2

=== Run information ===

Scheme: weka.classifiers.AttributeSelectedClassifier -B "weka.classifiers.j48.PART -C 0.25 -M 2" -E "weka.attributeSelection.CfsSubsetEval " -S "weka.attributeSelection.BestFirst -D 1 -N 5"
Relation: LastDataSet
Instances: 7532
Attributes: 10
TotalYearTrips
OneYearTrips
Tenurel
Revenue
Points
RevPerPoints
RevPerTrip
RevPerTenure
TripsPerTenure
Class
Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Selected attributes: 1,3,4,8,9 : 5
TotalYearTrips
Tenurel
Revenue
RevPerTenure
TripsPerTenure

Header of reduced data:
@relation 'LastDataSet-weka.filters.AttributeFilter-V-R1,3-4,8-10'

@attribute TotalYearTrips numeric
@attribute Tenurel numeric
@attribute Revenue numeric
@attribute RevPerTenure numeric
@attribute TripsPerTenure numeric
@attribute Class {1,2,3,4,5}

@data

Classifier Model

PART decision list

Tenurel <= 19 AND
TripsPerTenure <= 0.54 AND
TotalYearTrips <= 5 AND
RevPerTenure <= 218.42 AND
Tenurel <= 18: 3 (3067.0)

TripsPerTenure <= 0.51 AND
Tenurel > 21 AND
Revenue <= 4028.96 AND
TotalYearTrips <= 14: 4 (1577.0/2.0)

Revenue > 3606.24 AND
TripsPerTenure <= 1.02 AND
TotalYearTrips > 16 AND
TripsPerTenure <= 0.85 AND
RevPerTenure > 110.56: 2 (396.0/3.0)

TripsPerTenure > 0.66 AND
TotalYearTrips <= 18 AND
Revenue <= 3606.24 AND
Tenurel <= 18: 5 (799.0)

TotalYearTrips > 20 AND
TripsPerTenure > 1.36 AND
TotalYearTrips > 30: 1 (110.0/2.0)

Revenue <= 2503.55 AND
TripsPerTenure <= 0.5 AND
Tenurel <= 16: 3 (112.0)

Revenue <= 2987.24 AND
Tenurel <= 16 AND
Tenurel > 5 AND
TripsPerTenure > 0.54 AND
Revenue > 543.82: 5 (141.0/7.0)

Revenue <= 1934.56 AND
Revenue <= 1238.6 AND
Tenurel <= 20 AND

TotalYearTrips <= 4 AND
Tenurel <= 19: 3 (212.0)

TripsPerTenure > 0.61 AND
TotalYearTrips > 21 AND
TripsPerTenure <= 1.2 AND
Revenue <= 9321.54 AND
TotalYearTrips <= 43: 2 (122.0/13.0)

TripsPerTenure <= 0.61 AND
Revenue > 2987.24 AND
Revenue > 6044.99 AND
RevPerTenure <= 818.76: 2 (59.0)

TripsPerTenure <= 0.61 AND
TotalYearTrips > 12 AND
Revenue <= 2987.24 AND
TotalYearTrips > 14: 4 (46.0/5.0)

TripsPerTenure <= 0.61 AND
TotalYearTrips > 12 AND
Revenue <= 2987.24 AND
TripsPerTenure <= 0.58: 4 (5.0/1.0)

TripsPerTenure <= 0.61 AND
TotalYearTrips > 12 AND
Revenue > 2987.24 AND
Revenue > 5169.55: 2 (15.0/1.0)

TripsPerTenure <= 0.61 AND
TotalYearTrips > 12 AND
Revenue > 2987.24 AND
TotalYearTrips > 18 AND
TripsPerTenure > 0.51: 2 (11.0)

TripsPerTenure <= 0.61 AND
TotalYearTrips > 12 AND
Revenue > 2987.24 AND
TripsPerTenure > 0.36 AND
Revenue <= 4534.37 AND
TotalYearTrips <= 19: 4 (81.0/25.0)

TripsPerTenure <= 0.61 AND

Tenurel > 17 AND
Revenue > 1421.85 AND
TotalYearTrips <= 13 AND
TripsPerTenure <= 0.45 AND
Revenue > 1779.55: 4 (121.0)

TripsPerTenure <= 0.61 AND
Revenue <= 2987.24 AND
Tenurel <= 16 AND
Tenurel <= 15 AND
TotalYearTrips > 5 AND
Revenue > 1072.59 AND
TripsPerTenure > 0.45 AND
TripsPerTenure <= 0.51: 3 (14.0/6.0)

TripsPerTenure <= 0.61 AND
TotalYearTrips <= 8 AND
TripsPerTenure <= 0.45 AND
TotalYearTrips > 1 AND
Tenurel > 20: 4 (39.0/1.0)

TripsPerTenure > 0.61 AND
TotalYearTrips > 21 AND
TripsPerTenure <= 1.63 AND
TotalYearTrips > 32 AND
TripsPerTenure > 1.02 AND
Revenue <= 9321.54 AND
TripsPerTenure > 1.2: 1 (33.0/6.0)

TripsPerTenure <= 0.61 AND
Revenue <= 1072.59 AND
Revenue <= 722.55: 3 (39.0)

TripsPerTenure <= 0.58 AND
TotalYearTrips > 13 AND
TripsPerTenure > 0.36: 2 (18.0/4.0)

TripsPerTenure <= 0.58 AND
Tenurel > 16 AND
TotalYearTrips <= 8 AND
Tenurel > 18 AND
TotalYearTrips > 6: 4 (23.0)

TripsPerTenure <= 0.58 AND
Tenurel > 16 AND
TotalYearTrips > 8 AND
TotalYearTrips > 11: 4 (18.0/2.0)

TripsPerTenure <= 0.58 AND
Tenurel > 16 AND
TotalYearTrips > 8 AND
Tenurel > 18 AND
TotalYearTrips <= 10: 4 (14.0)

TripsPerTenure <= 0.58 AND
Tenurel > 16 AND
TotalYearTrips <= 8 AND
Tenurel > 17 AND
Revenue > 1238.6 AND
TotalYearTrips > 3 AND
Revenue <= 1656.76: 4 (31.0/2.0)

TripsPerTenure <= 0.58 AND
Tenurel > 16 AND
TotalYearTrips <= 8 AND
Tenurel > 19 AND
Revenue > 1007.11: 4 (6.0)

TripsPerTenure <= 0.58 AND
TotalYearTrips > 8 AND
Tenurel > 16 AND
Tenurel > 17: 4 (36.0/11.0)

TripsPerTenure <= 0.58 AND
TripsPerTenure <= 0.45 AND
TotalYearTrips > 2 AND
TotalYearTrips > 7: 4 (10.0/3.0)

TripsPerTenure <= 0.58 AND
TripsPerTenure <= 0.45 AND
TotalYearTrips > 2 AND
Revenue <= 2987.24 AND
TripsPerTenure > 0.33 AND
TotalYearTrips > 6 AND
RevPerTenure <= 105.43: 3 (8.0)

TripsPerTenure <= 0.58 AND
TripsPerTenure <= 0.45 AND
TotalYearTrips > 2 AND
Revenue <= 2987.24: 3 (52.0/20.0)

TotalYearTrips > 21 AND
TripsPerTenure > 1.36 AND
Revenue <= 6044.99: 1 (28.0/5.0)

Revenue > 6044.99 AND
RevPerTenure <= 818.76 AND
TotalYearTrips <= 41 AND
TotalYearTrips > 23: 1 (24.0/11.0)

Revenue > 6044.99 AND
Revenue > 9321.54 AND
TotalYearTrips > 30: 1 (17.0)

TripsPerTenure > 0.58 AND
Revenue > 6044.99 AND
Tenurel <= 7: 1 (9.0)

TripsPerTenure > 0.58 AND
TotalYearTrips <= 15 AND
RevPerTenure > 165.35 AND
Revenue <= 9321.54 AND
Tenurel <= 17: 5 (34.0)

Revenue <= 2704.85 AND
TotalYearTrips > 9 AND
Tenurel <= 23 AND
Tenurel <= 19 AND
RevPerTenure <= 146.68: 5 (24.0/1.0)

Revenue <= 2503.55 AND
Tenurel <= 16 AND
RevPerTenure <= 100.53: 3 (10.0/1.0)

TripsPerTenure > 0.58 AND
TotalYearTrips > 21 AND
Tenurel <= 37: 2 (22.0/3.0)

Revenue <= 2503.55 AND

TotalYearTrips > 5 AND
Tenurel <= 16 AND
TotalYearTrips > 7: 5 (6.0/1.0)

Revenue > 2704.85 AND
Tenurel > 21 AND
TripsPerTenure <= 1.02: 2 (22.0/2.0)

Revenue > 4028.96 AND
TripsPerTenure <= 1.02 AND
Tenurel > 7 AND
Tenurel > 17: 2 (13.0)

Revenue > 3262.72 AND
Revenue > 6044.99 AND
Revenue <= 9321.54: 1 (4.0)

RevPerTenure > 330 AND
RevPerTenure <= 464.98 AND
TotalYearTrips <= 19: 5 (9.0)

RevPerTenure > 330 AND
Revenue <= 9321.54: 1 (7.0/1.0)

TotalYearTrips <= 5 AND
RevPerTenure <= 818.76: 3 (4.0)

RevPerTenure > 464.98 AND
RevPerTenure <= 818.76: 2 (4.0)

TotalYearTrips > 15 AND
TripsPerTenure > 0.67 AND
Revenue <= 3262.72: 5 (9.0)

TotalYearTrips > 15 AND
Revenue > 2987.24 AND
Tenurel > 15 AND
TripsPerTenure > 0.8 AND
Revenue <= 4534.37 AND
TotalYearTrips > 17 AND
Tenurel > 16: 2 (6.0/2.0)

Tenurel > 8 AND

```
TotalYearTrips > 15 AND
TripsPerTenure > 0.8 AND
Revenue <= 4534.37: 5 (11.0)

TotalYearTrips > 15: 2 (9.0/2.0)

TotalYearTrips > 4 AND
Tenurel <= 13 AND
TripsPerTenure <= 0.54 AND
Revenue <= 1656.76: 5 (5.0)

TotalYearTrips > 4 AND
Revenue > 2302.6 AND
TripsPerTenure > 0.45 AND
RevPerTenure > 165.35: 5 (12.0)

Revenue <= 9321.54 AND
TotalYearTrips > 6 AND
RevPerTenure > 94.96 AND
RevPerTenure <= 130.53: 4 (5.0)

RevPerTenure <= 818.76 AND
TotalYearTrips > 6 AND
Revenue > 2503.55 AND
RevPerTenure <= 175.43 AND
Revenue > 2704.85 AND
TotalYearTrips > 13: 4 (5.0/2.0)

RevPerTenure <= 818.76 AND
Revenue > 2503.55 AND
TotalYearTrips <= 11: 4 (4.0)

TripsPerTenure > 0.5 AND
TripsPerTenure > 0.54: 5 (5.0/1.0)


Tenurel <= 8: 1 (4.0)

TotalYearTrips > 7: 5 (3.0/1.0)

: 3 (2.0)

Number of Rules:  59
```

Time taken to build model: 9.08 seconds

=== Stratified cross-validation ===
=== Summary ===

Total Number of Instances	7532	
Correctly Classified Instances	7257	96.3489 %
Incorrectly Classified Instances	275	3.6511 %

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.819	0.004	0.877	0.819	0.847	1
0.903	0.011	0.897	0.903	0.9	2
0.985	0.008	0.99	0.985	0.988	3
0.97	0.016	0.956	0.97	0.963	4
0.951	0.008	0.953	0.951	0.952	5

=== Confusion Matrix ===

a	b	c	d	e	<-- classified as
186	36	0	0	5	| a = 1
18	650	0	45	7	| b = 2
0	0	3457	27	24	| c = 3
0	26	20	1939	14	| d = 4
8	13	14	18	1025	| e = 5