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

APPLICATION OF DATA MINING TECHNIQUES TO SUPPORT CUSTOMER RELATIONSHIP MANAGEMENT (CRM) FOR ETHIOPIAN SHIPPING LINES (ESL)

A THESIS SUBMITTED IN PARTIAL FULFILMETN OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATION SCIENCE

BY
KUMNEGER FIKRE

July 2006

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
Faculty of Informatics
Department of Information Science

APLICATION OF DATA MINING TECHNIQUES TO SUPPORT CUSTOMER RELATIONSHIP MANAGEMENT (CRM) FOR ETHIOPIAN SHIPPING LINES (ESL)

BY
KUMNEGER FEKRIE

Name and Signature of Member of the Examining Board

Ato Dereje Teferi , Chairman, Examining Board

Prof. B.R.K. Rao, Advisor

Dr. Kumudha Raimond, External Examiner

Chairman, Faculty _____ Signature _____ Date 01/08/06

Chairman, Graduate Council _____ Signature _____ Date _____

Acknowledgement

I would like to thank my advisor, Prof. Bandaru Rama Krishna Rao for his heart felt advise, expert guidance and material support. His guidance was, of course the corner stone of this research work.

My appreciation and deepest gratitude goes to the staff members of the Marketing Department, Ethiopian Shipping Lines. Especially Ato Tewodros and W/t Bethlehem Who were very cooperative to extend their unreserved support whenever I asked for it.

I am also indebted to my beloved husband, Ato Samuel Fisseha and my lovely daughter, Ruth for their patience, love and support.

List of Abbreviations

AI- Artificial Intelligence

CRISP- Cross Industry Standard Process

CRM- Customer Relationship Management

ESL- Ethiopian Shipping Lines

KDD- Knowledge Discovery in Databases

OLAP- On Line Analytical Processing

SOM- Self organizing Map

List of Figures and Tables

List of Tables

List of figures

Table of Content

Abstract

Nowadays, the global marketing strategy is making business extremely competitive, dynamic and subject to rapid change. Hence, businesses should be highly concerned to needs and wants of their customers in order to respond accordingly. Customer Relationship Management is the overall process of exploiting customer – related information and using it to enhance the revenue flow from an existing customer. Data mining techniques are used to extract important customer information from available data bases.

The major objective of this study is testing the application of data mining techniques to support CRM activities for Ethiopian shipping Lines. The customer profile file of ESL contains individual shipment activities of more than 20,000 records, out of which about 4,000 are unique customers. After the data is collected, the necessary preprocessing steps are conducted on it in order to make it applicable for the modeling process.

K – Means clustering algorithm was used to segment individual customer records in to clusters with similar behaviors. Different parameters were used to run the clustering algorithm before arriving at customer segments that made business sense to domain experts. After the clustering is made, decision tree classification techniques were employed to generate rules that could be used to assign new customer record to the segments.

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

CHAPTER ONE

Introduction

1.1 Background

"Customer is the King", says the old adage. So, customers are vital components for the survival of any business. They have a high value to the business. As a result, businesses have found it essential to acquire new customers as well as retain the existing ones. Nowadays, Customer Relationship Management (CRM) is one of the hot issues towards the success of businesses.

CRM is an integration of technologies and businesses used to satisfy the needs of a customer during any given interaction move. Specifically it involves acquisition, analysis, and use of knowledge about customer in order to sell more goods or services and to do it more efficiently. CRM has been increasingly recognized as a business strategy to effectively understand, manage and sustain customer relationship with advanced information and communication technologies. It principally revolves around marketing and begins with a deep analysis of consumer behavior. It uses information technology to gather data, which can then be used to develop information required to create a more personal interaction with the customer. (Bose, 2002, Fekadu 2004, Wu Tie, 2003)

Marketing is the process of planning and executing the conception, pricing, promotion and distribution of ideas, goods and services to create exchanges that satisfy individual and organizational objective. Nowadays, customers that have real value to a company are the center of marketing strategies. One of the major focuses in marketing is *customer segmentation.* (Schiffaman & Kanuk,1991).

Customer segmentation refers to the process of dividing customers in to homogenous groups on the basis of common attributes and is at the heart of CRM. Segmentation describes the segments or clusters, within the data. By determining similar classes of

customers, more targeted communication is possible and marketing return on investment can be enhanced since marketing messages are accurately reaching those customers most likely to respond. Furthermore, different marketing strategies can be developed that are more appealing to members of the specified group. Segmentation requires the collection, organization and analysis of customer data (Bounsaythip, 2001, Schiffaman etal, 1991).

Database marketing techniques have been used to identify customer groups with high revenue potential, select criteria for mailing list and improve customer retention rate. The knowledge derived from these segments will enable one to focus on more targeted promotion. Furthermore, knowing customer needs better and treating them accordingly can increase their lifetime value. One of the major component /input for market analysis is the need for customer data such data/database should contain relevant attributes (fields) about virtually each customer. Nowadays, Business databases are growing dramatically but not by same rate with technologies to analyze and extract information from these huge amount of data.

Recent break through, especially in the area of data mining and artificial intelligence, are very promising in handling and extracting valuable information for effective planning and decision-making purpose (Witten et al 2001).

Data mining is a new kind of business information analysis technique. It aims to find out 'hidden' correlations among data by extracting, converting, analyzing and modeling from huge amount of transaction data in business database. It is the process of extracting information in order to discover hidden facts contained in the database using a combination of machine learning, statistical analysis, modeling techniques and database technology in the areas such as decision support, prediction, forecasting and estimating. The goal of data mining is to create models for decision making that predict future behavior based on analysis of past activity. In recent years, much importance has been attached to data mining from business perspectives mainly because of extensive use of enterprise database, data warehouse and urgent need of acquiring valuable information. These key data can be applied into a variety of areas such as: business administration,

1.2 Statement of the Problem and Justification of the study

A company that provides high quality customer service will have a competitive advantage that will ensure its supremacy over less capable competitors and guarantee its survival. Many experts believe that consistently providing a superior service is a main avenue to differentiating a company. There are no magic formulas, however; each company must strive for quality by taking into consideration its own special features and the demands of its customers.

Data mining technologies are extremely important to extract information from customer data /database for the purpose of supporting decision making. Especially in the presence of good collection of customers' data, businesses will undoubtedly obtain competitive advantage over their competitors by using these technologies and are very useful in creating value to the business in the long run.

Companies, which involve in providing service like shipping lines, air lines, banks, insurance, telecommunication companies and super markets, are potential users of data mining techniques for their overall customer relationship management. As a result a number of researches have been and are being done on the application of data mining techniques for such type of businesses world wide. CRM is one of the potential areas where data mining can be applied. In the context of Ethiopia, Henock (2002) and Fekadu (2003) have conducted researches in CRM using data mining techniques.

This research will focus on the application of data mining techniques for *Ethiopian Shipping Lines (ESL)* in supporting its efforts in customer relationship management. So as to enhance its responsiveness to the current and prospective customers needs and create strong market bases.

ESL is a multimillion capital share company that operates a regular liner service. On one hand the organization has rich customer related data on which the application of data mining techniques, especially classification and clustering methods could result in

valuable information for marketing decision making. On the other hand, there is no such an integrated system or model being applied to segment customers and hence no clear cut group/class based marketing orientation.

This research is therefore, to avoid this gap by applying data mining techniques on the huge data with the objective of segmenting the customers, which can be used for a better customer relationship, and to improve the profitability of the company and efficiency in dealing with customers needs. Segmentation helps to identify customers that generate high revenue to the company. This in turn may help the company to treat this group of customers according to their needs in order to satisfy them and attain their loyalty. To accomplish this task the company needs to replace the traditional way of customer relationship management with IT based CRM.

The company can plan and implement various marketing strategies to gain competitive advantage over the rivals and create long-term value to the investment. This in turn proves better results. Besides, the environment is being rapidly changing because of information and communication technologies (ICTs). Hence, adoption of this recent trend is becoming mandatory.

1.3 Objectives of the Study

The general objective of this research is to help ESLSC maintain an appropriate Customer Relationship Management through the application of data mining techniques for the purposes of developing a good customer relations, improved efficiency, service and profitability.

The specific objectives of the study include:

- To collect the necessary data from the company
- To prepare the data for analysis, this involves extracting the data and transforming into the format required for the data mining algorithm.

- To select the data mining tools and algorithms for clustering.
- .To design and develop the clustering model.
- To evaluate (test) the model.

1.4 Methods

The general approach of the study is basically quantitative as the core process is to collect and organize customer data. However, as the resulting clusters should be meaningful for marketing decision making, a close relationship with the domain experts and understanding of the business operation is mandatory. In this regard it could have qualitative aspect as well.

- Literature Review: Relevant literatures (books, journals, magazines and the Internet) pertaining to the subject matter of data mining, customer relationship management and customer segmentation are reviewed.

- Business Understanding: Interviews, observations, and document review are made to assess the needs of users, analyze the business problems, and have good background knowledge in interpreting results of the data mining process. In addition questionnaires will be distributed to frequent customers to seek suggestions for better service.

- Data mining methodology: To solve the business problems and meet the identified objectives, the researcher follows the following steps in order to develop a data mining model and employ data mining techniques.

- Identifying Available data sources: The data source of this research is the data containing customer records of Ethiopian Shipping Lines. The data contains information pertaining to customers who use the shipping service provided by the organization. The data set contains customer types and their respective attributes.

- Data collection and preparation for analysis: Before subjecting the raw customer data to analysis, it has been converted into a format suitable for analysis. Data understanding and preparation activities, such as analysis, editing, coding, cleaning, integration and transformation are conducted.

- Build and train the data mining model: After the data is cleaned, formatted and transformed, it has been used to build clustering and classification models. A training, testing, and validation data sets are used to generate clusters and an explanation (rules) of the dependent in terms of the independent (input) variables.

- Evaluating (testing) the model: In order to check the output of the model's performance, the process of the modeling and the results are counterchecked with users and domain experts of the organizations in order to provide useful information for making optimal customer related decisions.

1.5 Application of Results

The results of this research will support the routine and strategic decisions made by the Ethiopian Shipping Lines. By planning and implementing an appropriate marketing strategy and by providing an attractive offer through the right channel and at the right time, each customer contact is more likely to achieve its goal.

As a result, customers will get improved services and their profitability will increase. The research is considered to contribute a lot for the company in addressing inefficiencies of the existing customer relationship.

The research is also believed to initiate further researches in the area, as it is an initial attempt for exploiting the potentials of data mining techniques in the shipping lines industry in the area of customer relationship management in Ethiopia.

In general, the results may support marketing to develop /improve new/existing products and services based on the needs of customers.

1.6 Thesis Organization

This thesis consists of five chapters. The first chapter deals with a general background, motivation, statement of the problem and justification of the study, objectives, research methods, scope and limitation of the study, and the possible applications of the research work. The second chapter deals with general CRM and its current status at ESL. The third chapter covers the different aspects of data mining and the data mining techniques and algorithms used for the segmentation process. Chapter four, the most relevant and important chapter, discusses the different stages of experimentation towards building the data mining model and interpretation of results. The final chapter, chapter five, deals with results, conclusions and recommendations based on investigation of the study.

CHAPTER TWO

Customer Relationship Management and Data Mining

2.1 Loyalty and Customer Relationship Management

2.1.1. Overview

Loyalty refers to a true and faithful act or behavior. Customer loyalty is commonly defined based on a customer's purchase behaviors. A customer is classified as a loyal customer of a company, as long as this customer maintains an active account with the company. In some industries a loyal customer is defined by his/her purchase details: recency, frequency and monetary value. Creating and maintaining customer loyalty is a great concern for businesses. Businesses believe that it costs more to find a new customer than to keep and grow an existing one. Loyalty is an elusive goal in almost every industry.

Loyalty based marketing effort enables the firm to find and retain the right customers. According to Richheld (1995) the right customers are those to whom the best value can be delivered by the firm over a sustained period of time. Companies study their customers' base and segment it into those who are highly loyal and those who are less loyal. In response to these findings, companies focus all their marketing activities on the loyal customer segment.

The recent trend in loyalty management is changing from a reward- based relationship to one that is defined through sharing information with customers. As quoted by Henock (2003) Petersen (n.d.) believes it is about letting the customers decide that the company understands "who they are rather than what they are". It is the understanding of 'who' the customer is that underlies what is known as CRM.

CRM stands for Customer Relationship Management. It is defined as an all-embracing approach integrating sales, customer service, marketing and other functions that touch

customers so that by integrating strategy, people, process and technologies, relationships with customers, distributors, and suppliers are maximized. Basically CRM is a notion regarding how well an organization can keep the most profitable customers at the same time reduces costs, increase value of interaction and hence maximize profit. (Bose, 2002).

It is a strategy used to learn more about customers' needs and behaviors in order to develop stronger relationships with them. After all, good customer relationships are at the heart of business success. There are many technological components to CRM, but thinking about CRM in primarily technological terms is a mistake. The more useful way to think about CRM is as a process that will help bring together lots of pieces of information about customers, sales, marketing effectiveness, responsiveness and market trends.

The idea of CRM is that it helps businesses use technology and human resources to gain insight into the behavior of customers and the value of those customers. If it works as hoped, a business can:

- Provide better customer service
- Make call centers more efficient
- Cross sell products more effectively
- Help sales staff close deals faster
- Simplify marketing and sales processes
- Discover new customers
- Increase customer revenues

The purpose of CRM is to improve marketing productivity. Productivity is measured in terms of efficiency, effectiveness and economy. This can be achieved through creating cooperative and collaborative processes that help reduce transaction costs, increase revenue and finally create value to businesses during the lifetime of a customer. It is an integrated effort to identify, maintain and build up a network for the mutual benefit of both sides through interactive, individualized and value added contacts over a long period

of time. In some cases CRM is regarded as a dominant/ core paradigm of marketing. It is regarded as a shift of marketing role from manipulating the customer to genuine involvement with customer through appropriate communication and sharing of knowledge (Parvatlyar et al., 2002).

CRM can also be seen as business strategy aimed at gaining a long-term competitive advantage of optionally delivering customer value and extracting business value simultaneously. For the vast majority of businesses, the ability to acquire, retain and enhance customer relationship is the last place left to find an advantage. (Bull, 2003).

Recently many companies are adopting customer centric strategies programs, tools and technology for efficient and effective CRM. This is because of the understanding that for making optimal decisions an integrated and detail information, which is reliable and relevant is mandatory.

CRM has become number one focus in today's competitive market. More than ever, the ability to understand and manage close relationship with customers has become a prerequisite to achieve business goals. Past and present trends imply the relevance of customer information and knowledge to build strong relationships with their customers over longer period of time through providing customer satisfaction, at the same time, earn business value (Kim, Shu, & Hwang, 2003).

According to Parvatiyar & Sheth (2002) CRM has attracted the expanded attention of practitioners and scholars. More and more companies are adopting customer-centric strategies, programs, tools, and technology for efficient and effective customer relationship management. They are realizing the need for in-depth and integrated customer knowledge in order to build close cooperative and partnering relationships with their customers. The emergence of new channels and technologies is significantly altering how companies interface with their customers, a development bringing about a greater degree of integration between marketing, sales, and customer service functions in organizations.

The nature and scope of CRM is not yet clear and many researchers in the area of marketing are undergoing for the development of conceptual foundations of managing relationships with customers. Researchers in computer and information science disciplines are also searching for methodologies, techniques and software tools to assist decisions in the management of relationships with customers.

Many argue that customers' centric orientation (one-one relationship) is a subset/extension of the marketing orientation (group based relationship). Others disagree to this argument and describe the trend as a fundamental shift from managing market to managing a specific customer. In the first case (marketing orientation) the firm has control over the marketing mix, however, in the second case, the firm is directed by customer tastes and preferences. CRM principally revolves around marketing and begins with the deep analysis of customer behavior. CRM is based on the customer centric orientation to deal with different behavior of individual customer to obtain and maintain a share of each customer rather than a share of the entire market with the help of appropriate ICT. (Xu, Yen, Lin, & Chou, 2002).

2.1.2 Principles and Tasks of CRM

2.1.2.1 CRM Principles

There are three basic principles of CRM implementation, namely, *personalization, loyalty and lifetime value*. The first one deals about treating customers individually so that products and services are designed and offered based on the preferences and behavior of the customer. Loyalty refers to the company's retention capacity through continuous contacts/ relationships so that the customer gets satisfied and less likely to swift to other companies. The last one focuses on the selection of good customers from 'bad' through analysis of their respective behaviors. Decision would be taken to drop bad ones and keep the good ones as the motive of the organization is to maximize its profits in the long –run and has limited resources to spend for customer care. From economical perspective, it is less costly to retain a customer than to find a new one. The major goals of CRM are increasing revenue growth though customer satisfaction and reduce cost of

sales, distribution and minimize customer support costs (Parvatlyar & Sheth, 2001, Gray & Byun, 2001).

Key CRM Tasks

The basic question CRM tries to answer are basically two: to know the status of customers to find out their profitability level and to focus on strategies and ways the customers can grow to derive maximize profit to the business and at the same time keep customer satisfaction and remain loyal.

The basic tasks to address these questions can be categorized into four tasks/ processes: customer identification, customer differentiation, customer interaction and customization (Gray & Bynum., 2001).

- **Customer Identification**

It is the first step and refers to the attraction and or knowing of customers through marketing channels, transactions and interactions over time for the purpose of growing to a profitable one.

- **Customer Differentiation**

This refers to segmenting of customers into different perspective from the company's point of view as each customer has their own lifetime value.

- **Customer Interaction**

As the customer is to be exploited for maximum long-term value, analysis of his or her behavior over time is mandatory to know and offer the right goods and services at the right time.

- **Customization (Personalization)**

The final goal of CRM is to treat each customer uniquely so that each customer long-term value and/or loyalty increases.
The above tasks are greatly facilitated with the use of IT.

2.1.3 CRM and IT

Companies that effectively use IT will be the ones that best improve customer service, whether those customers are external (e.g. clients) or internal (e.g. employees, stockholders).These companies will make decisions quickly, act efficiently, and directly touch their customers in positive ways.

The combination of good customer information, data mining, and relevant technology enables companies to better understand their customer base and communicate with them more effectively. Once a firm is actively using customer information to make decisions about how, when and what to market to customers, they often increase the volume of targeted customer contacts. This increase leads many firms to look for new ways to automate mining and marketing processes to make the most of their newfound learnings about customers. (Ro King).

CRM technology facilitates communication and management of customers through automating the information channels like face-to-face, mail, phone, fax, web, and e-mail. Moreover the technology enables the companies improve performance almost in every functional areas and business processes including sections that have closer contact with the customer like marketing, sales, field service, contact center etc. Data mining and data warehousing tools and other related technologies help to analyze and extract hidden knowledge from customer data and have become the backbone of CRM systems (Bose, 2002).

Though potentially all firms can benefit from CRM technology, for certain companies, it is best suited. These companies include those which accumulate a lot of data on each customer on the course of their business like financial and service providing companies (IBM, 2001).

2.1.4 The Right Customers

The best customer is one a company already has and who can spend the most money at the property the company produces. This person will spend that money when the products and services the company provides are the best fit for the problems she or he needs solved. Consequently, the best customers for the property are also those who receive the greatest benefit from once services. They are willing to pay more than others (and more than they pay now), buy more frequently and remain more loyal because the company is solving their problems.

The problem with this optimistic outlook is that many executives are not sure how to determine which are their best customers even though the answer may already be at hand. Customer records, service and product checks, cash register receipts, and the records of strategically allied business partners contain most of the information any property's managers will need to determine the ideal customer base. With that information in hand, revenue enhancement becomes a matter of creatively packaging products and services for those customers.

2.1.4 Nature of Customers

As customers are not equally profitable to the business, appropriate segments should be made to select those who are legible for a given marketing program or to determine what kind of marketing effort is necessary for each segment of customers to finally maximize the profitability of the firm.

There are three distinct types of customer relationships: the top, middle and lower groups. The tops are those with higher productivity and loyalty. These customer groups need a good treatment to keep them for longer period through offering the possible best offer, products and services. The middle ones are those who have the potential to be grown to the top group. Here a lot to be done to make this group profitable and loyal. The lower group is marginally profitable. The task of targeting this group should be made in care, as the cost of such activity may be greater than the benefit derived. Treating this group of customer would affect the treatment of the other better groups, as the resources in any organization are limited to perform CRM projects. The identification of these three groups is one of the major tasks of CRM so that appropriate CRM strategies are designed and implemented for each group as CRM primarily involved in targeting the right customer at the right time for the right price. (Bull, 2003).

2.1.5. Revenue Enhancement

2.1.5.1. Overview

Revenue enhancement comprises a suite of strategies for increasing the amount of money a business makes. The key concept of revenue enhancement is that a business can make more money from its existing customer base. The lag time between taking steps to enhance revenue and seeing revenue improvements is health non existent. While that claim may seen unlikely, the truth is that most managers are so busy putting out daily fires that they overlook opportunities to increase profits. In the daily press of business, managers overlook customers who are actually begging to buy more. These clients have problems (needs) that are not being completely resolved or wishes that are not being fulfilled and those situations represent opportunities for better profits and, at the same time, happier & more satisfied customers.

2.1.5.2. Revenue Enhancement Myths

Quain et.al (1998) explode ten common misconceptions about revenue enhancement. Overcoming the fears created by these myths is the first step in creating great profits. These myths are discussed below.

Myth 1. *Increasing revenue will make customers unhappy.* No one makes money by making their customers unhappy. In its most effective form, revenue enhancement is a long term strategy for winning and keeping more customers by making them more happy, not less happy. Long term, it is impossible to increase revenue with out making your customers happy.

Myth 2. *Revenue enhancement is too time consuming.* Many busy executives miss revenue opportunities because they are concerned about the time commitment. Certainly, if you decide on a complete over haul of all management systems to take advantage of every revenue opportunity you will never start to increase revenues. On the other hand, just taking one revenue enhancing step, no matter how small, will make more money than doing nothing.

Myth 3. *Revenue enhancement is just yield management.* Yield management is a technique that can enhance revenues. True revenue enhancement is much more than any single technique, however. It is a service philosophy.

Myth 4. Revenue enhancement can be achieves by cutting costs. While careful cost control is essential in any enterprise, cutting costs generally decreases the value of the service to the customer. While customers may not notice the loss of some services, a better approach is to control costs while offering customers great benefits.

Myth 5. *Developing additional revenue sources requires hiring special managers.* The truth is that any employee can enhance revenues. Existing Company executives are completely capable of initiating changes and new programs that will increase revenue and customers satisfaction. The only change necessary is a revision in Philosophy.

Myth 6. *Revenue enhancement can be initiated only by using computers and hard – to – work formulas.* Computers can be an asset to many forms of revenue enhancement, and the more technical versions, such as yield managements, require complex computer algorithms. However, simple techniques abound for restructuring profitability.

Myth 7. *When a property is already producing as much money as it was designed to make, no more is possible.* Even if a company is 100% functioning, one can still find ways to increase revenue. One may be exceeding revenue goals but, for sure, could make even more.

Myth 8. *One can make more money simply by working harder.* Most company managers are already working hard- and so are most employees. Rather than put in more hours (with a guaranteed diminishing return), managers should instead work smarter. Revenue enhancement can create two precious resources – time and money.

Myth 9. *To increase revenues, one needs to serve more customers.* Serving more customers may not be a bad idea, but difficult to approach long term revenue potential unless one is serving the right customers. Instead of increasing customer counts, finding the right customers may mean serving fewer, lucrative customers in place of a large batch of customers who spend little. The sad fact is, sometimes one must discourage the "wrong" customers (i.e. those who cost the company more than they make it) from using the services. This is done by placing restrictions through pricing, promotion, or product use.

Myth 10. *Making more money is easy: Just tell people to produce more.* Ordering people to make more money is rarely a solution to any problem. If anything, one will engender hard feelings and passive aggression. The approach proposed here represents a long- term commitment to increasing revenues by identifying the right customers and serving them repeatedly. In fact, for true, sustained revenue enhancement, one may want to induce his/her employees to work "smart" by offering them rewards that are specifically tied to the company's profitability.

2.2 Customer Life Cycle

Companies that are truly customer centric use every customer touch pint to stimulate interest, close business, satisfy a need and pre-demonstrate commitment to the relationship. And that is exactly what customer life cycle care does. It ensures that every touch point fulfills potential regardless of whether that touch happens in marketing, sales or customer service.

CLC does this by capturing and delivering high value information at every touch point across the customer relationship life cycle. This 360-degree perspective is important, because of the cross departmental nature of customer relationship in the real world.

All of this business processes are supported by business analytics. Business analytics tracks all life cycle processes in order to optimize performance, quality and efficiency. At every- touch point through out the life cycle, companies use existing information abut the customer and capture new information about the customer. The goal of a customer from a business analytics view point is to:

- Effectively capture every and all useful information wherever it becomes available in the life cycle and
- Make any and all useful information available to any employee who needs it anywhere in the life cycle

If executed properly, this results in a true 360-degree view of the customer.

CLC fulfills this goal by integrating the capture and delivery of information across all departments. By doing so, CLC allows marketers, salespeople and service staff to treat customers as individuals-something customers really like! With account. CLC also eliminates duplication of effort by making sure no one has to capture the same data twice. And it ensures the overall quality of the information that everyone is using by replicating best data management practices across the company.

Effective CLC also optimizes use of all communication channels, giving customers the freedom to use phone, email, online self-service, chat, fax and or mail as they see fit at various points throughout the relationship lifecycle. In addition to catering to the customer's needs, this approach benefits the vendor by driving customer integrations to more efficient channels.

Simply put, CLC results in more revenue, greater profitability and happier customers. And perhaps best of all, it enables companies to realize the benefits of CRM with less cost and risk than traditional CRM implementations- which typically requires significant process re-engineering and major enterprise software implementations.

CLC accomplishes this by leveraging a powerful resource that CRM initiatives have historically underutilized: the customer service organization.

o Customer Segmentation

2.3.1Overview

It is known that consumer wants, needs and preferences differ according to peoples' backgrounds such as occupation, education, age, income, religion, etc., resulting in different interests for a variety of products and services. User segmentation concerns finding homogeneous groups of customers representing similar needs/wants, preferences and profiling these clusters in terms of other characteristics such as age, income, religion, etc. Consequently, marketing activities such as pricing, promotion and product assortment can specifically be tailored towards certain segments of customers.

Customer Segmentation is the subdivision of a market into discrete customer groups that share similar characteristics. Customer Segmentation can be a powerful means to identify unmet customer needs. Companies that identify underserved segments can then outperform the competition by developing uniquely appealing products and services.

Customer Segmentation is most effective when a company tailors offerings to segments that are the most profitable and serves them with distinct competitive advantages. This prioritization can help companies develop marketing campaigns and pricing strategies to extract maximum value from both high- and low-profit customers. A company can use Customer Segmentation as the principal basis for allocating resources to product development, marketing, service and delivery programs.

Customer Segmentation requires:

- Divide the market into meaningful and measurable segments according to customers' needs, their past behaviors or their demographic profiles;
- Determine the profit potential of each segment by analyzing the revenue and cost impacts of serving each segment;
- Target segments according to their profit potential and the company's ability to serve them in a proprietary way;
- Invest resources to tailor product, service, marketing and distribution programs to match the needs of each target segment;
- Measure performance of each segment and adjust the segmentation approach over time as market conditions change decision making throughout the organization.

Companies can use Customer Segmentation to:

I. Prioritize new product development efforts;
II. Develop customized marketing programs;
III. Choose specific product features;
IV. Establish appropriate service options;
V. Design an optimal distribution strategy;
VI. Determine appropriate product

2.3.2 Bases and Variables of Segmentation

One of the challenging tasks during segmentation is to chose and derive appropriate segmentation bases and variables in which the segmentation process is performed and resulting segments are interpreted. This greatly depends on the type of segmentation problem. Variables can be categorized as geographical, demographical, psychological and behavioral (Basgoze & Gkturk, 2003).

Behavioral segmentation clusters end users based on the requirements of the decision makers or in respect to the problem to be addressed. In many instances, behavioral based segmentation is recommended if the prime motive of the firm is to maximize benefit (profit) by identification of marketing opportunities. This is s because of the fact that customer behavior analysis enable to know the firm what are the needs and preferences of customers so that an appropriate marketing programs are designed and implemented accordingly that result in increased long term profitability through maximizing revenue and cutting costs. Moreover, this kind f segmentation creates relatively homogenous solutions according to selected characteristics. As the resources, skills and values are directed to maximize return, such kind of segmentation has got popularity these days (Bunsaytthip et al., 2001).

However, problem of accurately and usefully describing segments is cumbersome. There is also a possibility of using combined variables taken from different categories but does not guarantee better segmentation results.

2.3.3 Criteria for Successful Segmentation

There are six basic criteria for a given segmentation to be relevant and lead to profitable strategies and programs namely identifiablity, substantiality, accessibility, stability, responsiveness and actionability. Identifiablity refers to how well the segments are separated into clusters that are dissimilar to each other but contain similar members within them. Substantiality refers to the size of a given segment whether it is legible for a

design of a particular marketing strategy/ program. Accessibility refers to the extent at which customers are reached through different channels including direct while stability is about the continuity of segments with their attributes in the future. Responsiveness refers to the level of initiation of customers in a particular segment to a given message over a specified period uniquely from their segments. The final criteria, actionability, is abut the extent segments react and show marketing efforts in line with the organizations expectations and competencies (Vriens, 2001).

2.4 Ethiopian Shipping Lines Share Company (ESLSC)

2.4.1 Overview

ESLSC was established before 40 years in 1964 and became operational in 1966. Its share was jointly owned by a U.S company (49%) and Ethiopian Government (51%) until 1970; since then the Ethiopian Government fully owns it.

During this period, the government has given full support to strengthen its capability through provision of finance and credit guarantee to build new vessels compatible to the then trading requirements. This is a useful practice exercised by both developed and developing countries for shipping companies operating with their flag.

Latter in 1992 following the deregulation and the move to market oriented economy policy, ESLSC was forced to work under free market economic policy conditions. Consequently, the company has been exposed to unfair competition from global carriers and illegitimate practices; owing to the fact that it owns traditional service in line with old, cost ineffective and unsuitable vessels for the current and prospective market requirements. Ethio-Ship(2005).

2.4.2 CRM in ESLSC

In order to improve customer satisfaction, ESLSC provides frequent sailing schedules as per their requirements either by using space or slot chartered agreement or connecting carrier arrangement. Complaint handling system has also been established. The complaints of a customer are normally handled by the respective department and by the

Managing Director as necessary. Customer complaints and subsequent measures are collected, compiled and reported every quarter to PESA. Generally, the company relies on the staff members especially on the sales team to keep track of customer satisfaction.

Even though efforts are being made to renovate service delivery of ESLSC, there is a lag in implementation due to several factors. In order to make the current strategic thinking more transparent and practical ESLSC should implement modern CRM techniques. So that it can strengthen its relation with its customers.

CHAPTER THREE

Data Mining

3.1 Overview

Witten & Frank (2000) defines data mining as the process of discovering patterns in data. The process must be automatic or (more usually) semi-automatic. The patterns discovered must be meaningful in that they lead to some advantage, usually an economic advantage. The data is invariably present in substantial quantities.

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge driven decisions. The automated, prospective analysis offered by data mining move beyond the analyses of past events provided by retrospective tools typical decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

3.2 Data Mining as KDD Process

As explained by Han & Kamber (2001), many people treat data mining as a synonym for another popularly used term - Knowledge Discovery in Databases (KDD). However, data mining is a step in the KDD process consisting of particular data mining algorithms.

KDD process is the process of using data mining methods (algorithms) to extract (identify) what is deemed knowledge according to the specifications of measures and

thresholds, using the database along with any required preprocessing, sub sampling and transformation.

It is an iterative and interactive process involving numerous steps with many decisions made by the user. Some of the basic steps in the KDD process are discussed below:

i. **Data cleaning and preprocessing**-in this step basic operations such as the removal of noise or outliers if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, accounting for time sequence information and known changes.
ii. **Data reduction and projection**- finding useful features to represent the data depending on the goal of the task. Using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data.
iii. **Choosing the data mining task**- deciding whether the goal of the KDD process is classification, regression, clustering, etc.
iv. **Choosing the data mining algorithm(s)**-selecting method(s) to be used for searching for patterns in the data. This includes deciding which models and parameters may be appropriate.
v. **Data mining**-is an essential process where intelligent methods are applied in order to extract data patterns. The user can significantly aid the data mining method by correctly performing the preceding steps.
vi. **Pattern evaluation**-to identify the truly interesting patterns representing knowledge based on some interestingness measures
vii. **Knowledge presentation**-a step where visualization and knowledge representation techniques are used to present the mined knowledge to the user

3.3 Data Mining and Related Fields

Data mining has a strong relationship and considerable overlap with other fields as it is a recently emerged but fast growing discipline. It is of interest to researchers in machine

learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization.

3.3.1 Data Mining and Data Warehousing

When beginning work on data mining problem, it is first necessary to bring all the data together into a set of instances. In a real business application, it will be necessary to bring data together from different departments. For example, in a marketing study data will be needed from sales department, the customer-billing department, and the customer service department.

Practically, integrating data from different sources usually presents many challenges. Different departments will use different styles of record keeping, different conventions, different time periods, different degree of data aggregation, different primary keys, and will have different kinds of error. The data must be assembled, integrated, and cleaned up. The idea of enterprise wide database integration is known as **data warehousing**. Data warehouses provide a single consistent point of access to corporate or organizational data, transcending departmental divisions. They are the place where old data is published in a way that can be used to inform business decisions. The movement toward data warehousing is a recognition of the fact that the fragmented information that an organization uses to support day to day operations at a departmental level can have immense strategic value when brought together. Clearly the presence of data warehouse is a very useful precursor to data mining, and if it is not available, many of the steps involved in data warehousing will have to be undertaken to prepare the data for mining.

3.3.2 Data Mining and Database Management

Database Management System (DBMS) also provide good ground to data mining as the broader purpose of data mining is to discover knowledge from large databases. DBMS offer essential capabilities to data mining as it contains a consistent data model and advanced high level query language that help users to get the required information for

their need. The data about data (meta data) found in the databases is also valuable for data mining as the nature of the attributes can be easily understood. (Connoly & Begg, 2000).

These greatly facilitates the data mining process as data collected, cleaned, and integrated for modeling from different databases. In such environment, the data-mining task could be done with less cost and time but with relevant output.

3.3.3 Data Mining and OLAP

There exist a kind of confusion with about the distinction of the two terms. OLAP (On Line Analytical Processing) greatly utilizes the application of complex queries to analyze multidimensional data mostly from the data warehouse query languages. The user exploits OLAP capabilities to check his/her generalization and relationship in the data by forming the necessary query. That is, the user should first hypothesize and use OLAP and use the data for support. However in data mining, generalizations and relationships are generated from the data itself. It requires in depth analysis to induce new information from the data using relatively complex techniques (Connoly & Begg., 2000).

OLAP and data mining are integral parts of any decision support process, which has increasingly become a focus of the database industry. OLAP provides higher-level querying idioms based on the multidimensional data model, while data mining provides the most abstract analysis operations. To date, however, most OLAP activities work separately from data mining activities. For example, OLAP systems focus on providing access to multidimensional data; while data mining systems deal with influence analysis of data along a single dimension.

3.3.4 Data Mining, Artificial Intelligence (AI) and Statistics

Data mining was first introduced when AI, and Statistical techniques were applied to common business problems. Well, the scope of data mining now widens from business problems. Its scientific application, for example, has dramatically increased for variety of

purposes. The basic capability of data mining provided by these fields is the capacity of pattern recognition from data using complex and powerful techniques and tools like neural nets and decision trees. The advancement of information and communication technology (ICT) has also accelerated the growth of data mining both in terms of techniques and tools as well as scope of application (Mitchell, 1997).

3.4 Data Mining Functionalities

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. In general data mining tasks can be classified into two categories: *descriptive and predictive*. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions.

In some cases, users may have no idea, which kinds of patterns in their data may be interesting, and hence may like to search for several different kinds of patterns in parallel. Thus it is important to have a data mining system that can mine multiple kinds of patterns to accommodate different user expectations or applications. Furthermore, data mining systems should be able to discover patterns at various granularities. Data mining systems should also allow users to specify hints to guide or focus the search for interesting patterns. Sine some patterns may not hold for all of the data in the database, a measure of certainty or "trustworthiness" is usually associated with each discovered pattern.

Data mining functionalities and the kinds of patterns they can discover are discussed below.

3.4.1 Concept/ Class Description: Characterization and Discrimination

It can be useful to describe individual classes or concepts in summarized, concise, and yet precise terms. Such descriptions of a class or a concept are called class /concept descriptions. These descriptions can be derived via:

- Data characterization, by summarizing the data of the class under study (often called the target class) or data discrimination, by comparison of the target class with one or a set of comparative classes (often called the contrasting classes) or both data characterization and discrimination. Data characterization is a summarization of the general characteristics or features of a target class of data. The data corresponding to the user-specified class are typically collected by a database query.

- Data discrimination is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes. The target and contrasting classes can be specified by the user, and the corresponding data objects retrieved through database queries.

3.4.2 Association Analysis

Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. Association analysis is widely used for market basket transaction data analysis

3.4.3 Classification and Prediction

Classification is the process of finding a set of models (or functions) that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data.

Classification can be used for predicting the class label of data objects. However in many applications, users may predict some missing or unavailable data values rather than class labels. This is usually the case when the predicted values are numerical data and is often specifically referred to as prediction. Although prediction may refer to both data value prediction and class label prediction, it is usually confined to data value prediction and

thus is distinct from classification. Prediction also encompasses the identification of distribution trends based on the available data.

Classification and prediction may need to be preceded by relevance analysis, which attempts to identify attributes that do not contribute to the classification or prediction process. These attributes can then be excluded.

3.4.4 Cluster Analysis

Unlike classification and prediction, which analyze class label data objects, clustering analyzes data objects without consulting a known class label. In general, the class labels are not present in the training data simply because they are not known to begin with. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity. That is clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Each cluster that is formed can be viewed as a class of objects, from which rules can be derived. Clustering can also facilitate taxonomy formation, that is, the organization of observations into a hierarchy of classes that group similar events together.

3.4.5 Outlier Analysis

A database may contain data objects that don't comply with the general behavior or model of the data. These data objects are *outliers*. Most data mining methods discard outliers as noise or exceptions. Outliers may be detected using statistical tests that assume a distribution or probability model for the data, or using distance measures where objects that are a substantial distance from any other cluster are considered outliers. Rather than using statistical or distance measures, deviation based methods identify outliers by examining differences in the main characteristics of objects in a group.

3.4.6 Evolution Analysis

Data evolution analysis describes and models regularities or trends for objects whose behavior changes overtime. Although this may include characterization, discrimination, association, classification, or clustering of time - related data, distinct feature of such analysis include time-series data analysis, sequence or periodicity pattern matching, and similarity-based data analysis.

- **Data Mining Methods for Customer Segmentation**

3.5.1 Clustering Techniques and Algorithms

Clustering techniques apply when there is no class to be predicted but rather when the instances are to be divided into natural groups. These clusters presumably reflect some mechanism at work in the domain from which instances are drawn, a mechanism that causes some instances to bear a stronger resemblance to one anther than they do to the remaining instances. Clustering naturally requires different techniques to the classification and association learning methods that we have considered so far.

There are different ways in which the results of a cluster can be expressed. The groups that are identified may be exclusive, so that any instance belongs in only one group. Or they may be overlapping, so that an instance may fail into several groups. Or they may be probabilistic, whereby an instance belong to each group with a certain probability. Or they may be hierarchical, such that there is a crude division of instances into groups at the top level, and each of these groups are refined further. The choice between these possibilities should be dictated by the nature of the mechanisms that are thought to underlie the particular clustering phenomena. However because these mechanisms are rarely known- the very existence of clusters is, afterall, something that we are trying to discover- and for pragmatic reasons to, the choice is usually dictated by the clustering tools that are available.

Basically, there are three different clustering methods. The first is the classic *k-means* algorithm, which forms clusters in numeric domains, partitioning instances into disjoint clusters. It is a simple and straightforward technique that has been used for several decades. The second is an *incremental clustering* technique that was developed in the late 1980s and is embodied in a pair of systems called COBWEB (for nominal attributes) and CLASSIT (for numeric attributes). These methods come up with a hierarchical grouping of instances. The third is a *statistical clustering* method based on a mixture model of different probability distributions, one for each cluster, which -unlike the other two methods- assign instances to classes probabilistically, not deterministically.

Requirements of Clustering Algorithm

Ability to deal with different types of attributes: Many algorithms are designed to cluster interval-based (numerical) data. However, applications may require clustering other types of data, such as binary, categorical (nominal), and ordinal data, or mixtures of these data types.

Discovery of cluster with arbitrary shape: Many clustering algorithms determine clusters based on Euclidean or Manhattan distance measures. Algorithms based on such distance measures tend to find spherical clusters with similar size and density. However, a cluster could be of any shape. It is important to developments that can detect clusters of arbitrary shape.

Minimal requirements for domain knowledge to determine input parameters: Many clustering algorithms require users to input certain parameters in cluster analysis (such as the number of desired clusters). The clustering results can be quite sensitive to input parameters. Parameters are often hard to determine, especially for data sets containing high-dimensional objects. This not only burdens users, but also makes the quality of clustering difficult to control.

Ability to deal with noisy data: Some clustering algorithms are sensitive to the order of input data; for example, the same set of data, when presented with different orderings to such an algorithm, may generated dramatically different clusters, it is important to develop algorithms that are insensitive to the order of input.

High dimensionality: A database or a data warehouse can contain several dimensions or attributes. Many clustering algorithms are good at handling low-dimensional data, involving only tow to three dimensions. Human eyes are good at judging the quality of clustering for up to three dimensions. It is challenging to cluster data objects in high-dimensional space, especially considering that such data can be very sparse and highly skewed.

Constraint-based clustering: Real-world applications may need to perform clustering under various kinds of constraints. Suppose that our job is to choose the locations for a given number of new automatic cash-dispensing machines (i.e., ATMs) in a city. To decide upon this, we cluster households while considering constraints such as the city's rivers and highway networks, and customer requirements per region. A challenging task is to find groups of data with good clustering behavior that satisfy specified constraints.

3.5.1.1 The K-Means Method

The *K-means* is an iterative distance based method. First, you specify in advance how many clusters are being sought: this is the parameter ***k*** then *k* points are chosen at random as cluster centers. Instances are assigned to their closet cluster center according to the ordinary Euclidean distance function. Next, the centroid or mean of all instances in each cluster is calculated- this is the ***'means'*** part. These centroids are taken to be new center values for their respective clusters. Finally the whole process is repeated with the new cluster centers. Iteration continues until the same points are assigned to each cluster in consecutive rounds, at which point the cluster centers have stabilized and will remain the same thereafter.

This clustering method is simple and reasonably effective. As with all practical clustering techniques, the final cluster centers do not represent a global minimum but only a local one, and completely different final clusters can arise from differences in the initial randomly chosen cluster centers.

A large number of variants of the basic *k*-means procedure have been developed. Some produce a hierarchical clustering by applying the algorithm with $k = 2$ to the overall dataset and then repeating, recursively, within each cluster. Others concentrate instead on speeding up clustering. The basic algorithm can be rather time consuming because a substantial number of iterations may be necessary, each involving finding the distance of the k cluster centers from every instance to determine the closet. There are simple approximations that will speed it up considerably, for example by leading with projections of the dataset and making cuts along selected axes instead of the arbitrary hyper plane divisions implied by choosing the nearest cluster center, but they inevitably compromise the quality of the resulting clusters.

3.5.1.2 Cluster Interpretation

Once the clusters have been created using clustering algorithms, they need to be interpreted. Though there are several approaches of understanding clusters, according to Berry & Linoff (2000), the three that are commonly used are:

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

3.5.2 Decision Tree

Decision trees which recursively divide the space into different regions have sharp breaks in their estimation surfaces, allowing great local responsivity. In decision trees, the variables selected for splits may be different in each adaptively-partitioned region of the space. The flexibility of the method seems often, in practice, to make up the crude basis function (a constant). Trees are natural for classification, but are also useful in difficult estimation problems where their simple piecewise-constant response surface and lack of smoothness constraints make them highly robust to outliers in either the predictors or the response variable. They automatically select variables, and construct models quite rapidly for an adaptive method. Importantly, trees are also probably the easiest model form to interpret (so long as they are shallow) which greatly improves the model's chance of actually being used. The main problem with trees is that they devour data at a rate exponential with depth; so to uncover complex structure, extensive data is required. (Fayyad, et.al, 1996).

3.5.2.1 Decision Tree Building

The basic algorithm for decision tree induction is a "greedy" algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner. The algorithm selects an attribute from the rest of attributes with a strategy of searching a local optimum solution (at each node) that leads to a global optimum solution. However this is not always true. For searching a local optimum solution, various methods can be employed to calculate the attribute selection measure. The iterative divide and conquer process executes until no further split is required. (Witten & Frank, 2001).

The problem of constructing a decision tree can be expressed recursively. First, select an attribute to place at the root node and make one branch for each possible value. This splits up the example set into subsets, one for every value of the attribute. Now the process can be repeated recursively for each, using only those instances that actually reach the branch. If at any time all instances at a node have the same classification, stop developing that part of the tree.

3.5.2.2 Decision Tree Pruning

While constructing the decision tree, a stopping criterion is usually used to limit the depth of the tree and the minimum number of members required in a given node for further splitting. Alternatives for using a stopping criterion is to let the tree finish growing but use pruning methods to reduce the tree to a smaller size possible from its full size but without compromising accuracy. (Han & Kamber, 2001).

CHAPTER FOUR

Experimentation

4.1. Overview

This chapter comprises the core component of the research. It deals with the description of the date mining process undertaken based on the Cross- Industry Standard Process for Data Mining (CRISP-DM) model. This model has six phases: business understanding, data understanding, data preparation, model building, evaluation and deployment.

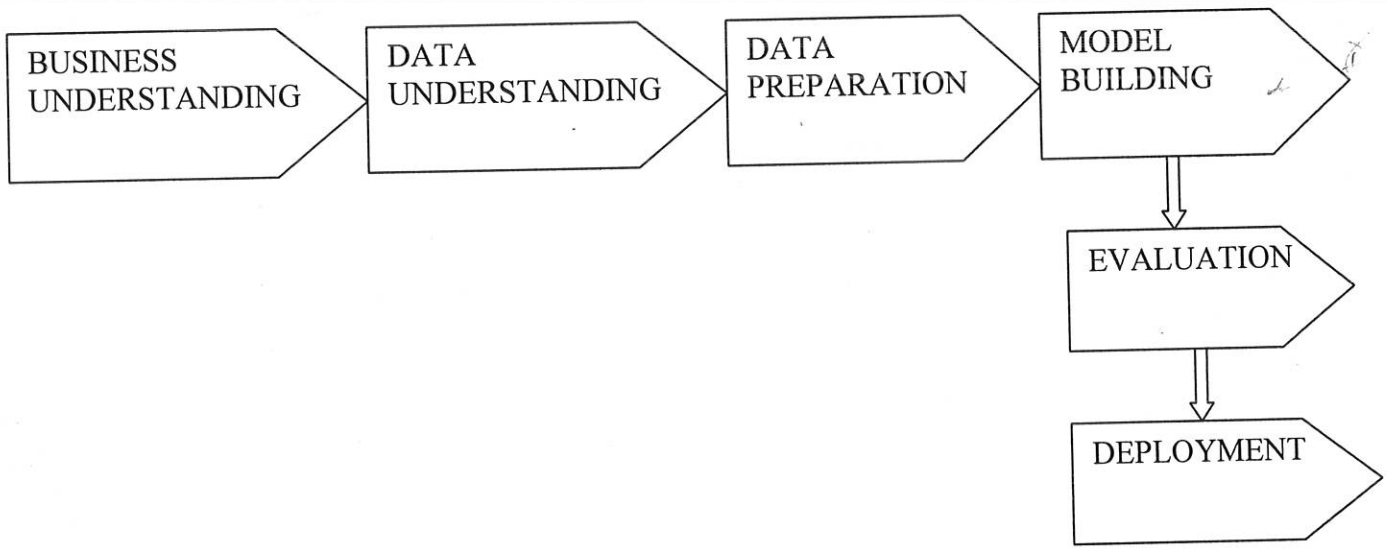


Fig. 2.1 CRISP-DM Model

As noted before, the general objective of the research was to segment customers based on their behavior, so as to identify the most profitable ones and to treat them according to their needs.

4.2. Business Understanding

Based on the evaluation made to understand the general background, processes employed and the business problem to be addressed, profitability of a customer can be evaluated based on type of customer type, amount of money, trade route, country and date.
Hence, high value customers can be defined as those with higher revenue. Consequently, low value customers are those with lower revenue. There are intermediary ones (variants) between these kinds of customers.

4.2.1. Date Mining Tool Selection

Finding free data mining software with strong capabilities that performs the required tasks of data mining i.e. clustering and classification, was one of the challenging tasks the researcher faced. This was so because of scarcity of funds for such purpose and the high cost of data mining tools even for research purposes.
The researcher had to first set criteria for tool selection and search for the one that fulfills the criteria. The criteria used to select one tool from the other were the following;

- The performance of the tool in terms of speed and quality.
- The time allowed to use the tool.
- The application area the tool proved its performance (Shipping lines, Customer Segmentation).
- The clustering and classification algorithms supported.(MS-Excel, MS-Access, and MS- word)
- The number of records (rows) and attributes (column) the tool can handle.
- User friendliness.

After considerable search and contact with tool providers, and review of tools available, three data mining packages were obtained freely which were *WEKA*

version 3.2.3, Knowledge Studio Version 4.1.1 and Ghost Miner version 2.0. After detail evaluation of the tools as per the criteria, knowledge Studio Version 4.1.1 was selected. The researcher was satisfied by the capabilities of the selected software as its overall performance was very good during the course of the data mining process.

4.3. Data Understanding

Next to understanding of the problem to be addressed clearly and the selection of an appropriate tool, the succeeding task is to analyze and understand the content and structure of the data available. To fulfill this requirement, raw data was initially collected regarding customer behavior from the ESL Invoice Office and careful analysis of the data and its structure is done together with the domain experts by evaluating the relationships of the data with the problem at hand and the particular data mining tasks to be performed.

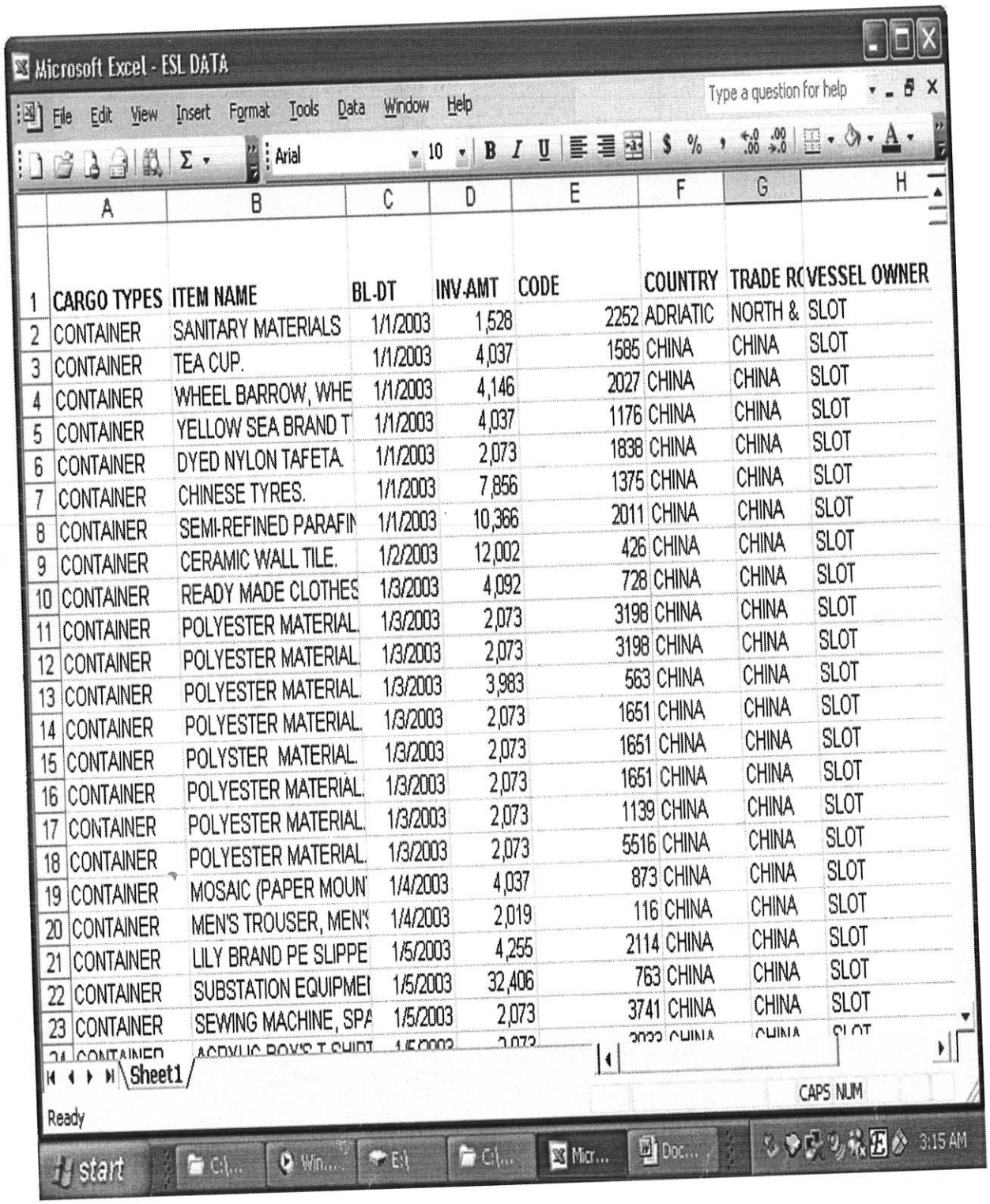
4.3.1 Initial Data Collection

As indicated above, the major data source was the Invoice and Booking Office of the company. The Center is a rich data source since registration of customers is made in this section and it keeps track of every detail data about the customers. As a result the item type, the amount of money paid, country, trade route, etc of every customer is maintained in the section. The researcher took a 27 months data. From this data most important (or determinant) attributes are selected by discussing with the domain experts.

To make it manageable for sampling, the data needed was divided into smaller size, further, this text data required transformation into an excel format to allow some processing and be fed to another software that could further process it to a format ready for data mining. Generation of additional derived attributes was also required.

4.3.2. Description of the Data Collected

The data collected for analysis, as indicated above, was of 27 months transaction data (from January 2003 to March 2005). It had a size of about 27,000 records containing customer detail. It has more than seven attributes, and a sample sheet is shown in Table 4.1.



	A	B	C	D	E	F	G	H
1	CARGO TYPES	ITEM NAME	BL-DT	INV-AMT	CODE	COUNTRY	TRADE RO	VESSEL OWNER
2	CONTAINER	SANITARY MATERIALS	1/1/2003	1,528	2252	ADRIATIC	NORTH &	SLOT
3	CONTAINER	TEA CUP.	1/1/2003	4,037	1585	CHINA	CHINA	SLOT
4	CONTAINER	WHEEL BARROW, WHE	1/1/2003	4,146	2027	CHINA	CHINA	SLOT
5	CONTAINER	YELLOW SEA BRAND T	1/1/2003	4,037	1176	CHINA	CHINA	SLOT
6	CONTAINER	DYED NYLON TAFETA.	1/1/2003	2,073	1838	CHINA	CHINA	SLOT
7	CONTAINER	CHINESE TYRES.	1/1/2003	7,856	1375	CHINA	CHINA	SLOT
8	CONTAINER	SEMI-REFINED PARAFI	1/1/2003	10,366	2011	CHINA	CHINA	SLOT
9	CONTAINER	CERAMIC WALL TILE.	1/2/2003	12,002	426	CHINA	CHINA	SLOT
10	CONTAINER	READY MADE CLOTHES	1/3/2003	4,092	728	CHINA	CHINA	SLOT
11	CONTAINER	POLYESTER MATERIAL.	1/3/2003	2,073	3198	CHINA	CHINA	SLOT
12	CONTAINER	POLYESTER MATERIAL.	1/3/2003	2,073	3198	CHINA	CHINA	SLOT
13	CONTAINER	POLYESTER MATERIAL.	1/3/2003	3,983	563	CHINA	CHINA	SLOT
14	CONTAINER	POLYESTER MATERIAL.	1/3/2003	2,073	1651	CHINA	CHINA	SLOT
15	CONTAINER	POLYSTER MATERIAL.	1/3/2003	2,073	1651	CHINA	CHINA	SLOT
16	CONTAINER	POLYESTER MATERIAL.	1/3/2003	2,073	1651	CHINA	CHINA	SLOT
17	CONTAINER	POLYESTER MATERIAL.	1/3/2003	2,073	1139	CHINA	CHINA	SLOT
18	CONTAINER	POLYESTER MATERIAL.	1/3/2003	2,073	5516	CHINA	CHINA	SLOT
19	CONTAINER	MOSAIC (PAPER MOUN	1/4/2003	4,037	873	CHINA	CHINA	SLOT
20	CONTAINER	MEN'S TROUSER, MEN'S	1/4/2003	2,019	116	CHINA	CHINA	SLOT
21	CONTAINER	LILY BRAND PE SLIPPE	1/5/2003	4,255	2114	CHINA	CHINA	SLOT
22	CONTAINER	SUBSTATION EQUIPMEI	1/5/2003	32,406	763	CHINA	CHINA	SLOT
23	CONTAINER	SEWING MACHINE, SPA	1/5/2003	2,073	3741	CHINA	CHINA	SLOT

Table 4.1 Sample data sheet

4.3.3. Data Exploration

The collected data provide information indirectly. The frequency of customers and the total revenue they generate to the company are derived from the code and amount columns. The date at which the customers come for the first time is derived from the code and date columns.

4.3.4. Data Quality Verification

The collected data contains missing, incomplete and irrelevant data. In many of these cases, data in the customer detail record column is missing or incomplete. The reliability of data and completeness of records are relatively good as the data is produced electronically. For some of the records the amount attribute was missed. Since it is imperative that this value be assigned to each shipment in order to provide reliable customer information for the complete individual history, it is recalculated based on the other variables by the experts.

4.4. Data Preparation

Data preparation involves a series of steps to provide the final data set for modeling. It includes data selection, cleaning, construction, integration, and formatting.

4.4.1 Data Selection

The list of attributes selected for the data preparation process, as noted earlier, includes customer code, date and amount columns.

Since the data mining task to be performed (clustering) needs relatively higher number of records, all of the 27 months data (about 26,000 records) were taken. As the purpose of the research is to explore and generalize about the population,. It is beneficiary to take the entire population.

4.4.2. Data Cleaning

The data was cleaned by removing the records that had incomplete (invalid) data and/or missing values under each column. Removing of such records was done as the records with this nature are few and their removal does not affect the entire data set. The researcher makes use of MS-Excel for cleaning the data. MS-Excel is selected since the data was first in Excel format and since it is familiar to the researcher.

4.4.3. Data Construction

Almost all attributes were derived from the ESL Customer's Profile File. This step of data preparation took the researcher considerable time and effort. This kind of construction had to be done since the customer file contains raw data, which is not in the way appropriate for the business goal to be addressed, and the corresponding data mining tasks to be performed. The attributes were derived from the file for each customer using MS-Excel.

4.4.4. Data Integration

Since the required data was available in a single file, this step was not necessary for this particular research.

4.4.5 Data formatting

This step involves changing of the data into a format suitable for the data mining tool (algorithm). The tool (algorithm) selected doesn't require preliminary formatting except for the outliers in the data. Outliers, which may mislead the k-means algorithm, are replaced with a maximum value that is set together with domain expert for each attribute.

The final processed data set including the attributes and their description after the data preparation phase looks as follows.

Field name	Data type	Description
Code	Number	Customer identification number
TotalShipment	Number	Total number of shipments made by customer
TotalRevenue	Number	Total revenue collected from customer
Ntenure	Number	Number of months since customer comes for the first time
Rev_Tenure	Number	Ratio of total revenue to customer tenure
Rev_Shipment	Number	Ratio of total revenue to total number of shipment
Shipment_Tenure	Number	Ratio of total number of shipment to customer tenure

Table 4.2 List of all attributes with their type and description

4.5 MODELING

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

4.5.1 Selection of Modeling Technique

The major objective of this research being the generation of strategic customer segments among the ESL customers, a clustering algorithm was applied in order to identify groups which were different from each other according to their product mix as well as to their

value, but whose members were very similar to each other. Since no predefined segmentation existed, employing clustering algorithms was appropriate.

The data mining techniques selected for customer segmentation are automatic cluster detection and decision tress. The selection was made because of the fact that these techniques are widely applied for segmentation problems. Moreover, the techniques are implemented well in the selected data mining tool and hence there was an opportunity to choose among related algorithms.

For clustering purpose, the available algorithms in the tool are *K-Means and Expectation Maximization (EM).* The *K-Means* algorithm is selected as it is a very good general-purpose clustering algorithm and is recommended for most situations. More over it is good in handling discrete and numeric attributes. However, the EM algorithm is recommended when there is a large amount of missing data (Angoss.2002).

Regarding *decision trees*, the algorithms found in the tool are *Knowledge SEEKER and Heat SEEKER.* The first one was selected as it is a powerful, flexible algorithm that is especially good for exploration purposes. It can handle a large amount of variables with either a continuous or discrete dependent variable. However, the later works for few attributes and does not have the strengths the former offers. (Angoss, 2002)

The *k-means* clustering algorithm passes through each customer record, assigning each to the closest existing cluster center. According to Pritcher (n.d), a critical task of using the *k-means* clustering algorithm is the choice of the right variables and the right scales.
Once the clusters were created, the interpretation of the clusters rested with the domain experts and the researcher. The following three approaches were employed to understand the clusters using KnowledgeStudio:

1) Visually analyzed how the clusters were affected by changes in the input variables.
2) Examined the differences in the distributions of variables from cluster to cluster, one variable at a time.
3) Finally, automatically grow a decision tree with 'cluster index' as the dependent variable, and used it to derive rules explaining how to assign new records to the correct cluster.

4.5.2. Test Design

Before starting modeling, a plan should be first set to guide the training, testing and evaluating process and put a test for the quality and validity of the model.
The entire population (100%) was used to train the clustering model whereas for the decision tree model, 60% was used for training the model, 30% served as a test data and the remaining 10% were employed as a validation set.

The process of segmentation and the interpretation of resulting segments were examined together with domain experts so that the final output provides good basis for possible design and implementation of an appropriate CRM strategies and program. The variables used in this phase are:

1. Total number of segments shipped by customer (TotalShipment)
2. Total revenue collected from the customer (TotalRevenue)
3. Number of months since the customer first enrolled in ESL (NTenure)
4. Ratio of Total Revenue to total customer's tenure (Rev_Tenure)
5. Ratio of Total Revenue to total number of segments (Rev_Shipment)
6. Ratio of total number of segments to customer tenure (Shipment_Tenure)

4.5.3. Model Building

The model building phase is divided into three subsections namely *attribute selection models, automatic cluster detection model and decision tree models*. The first section

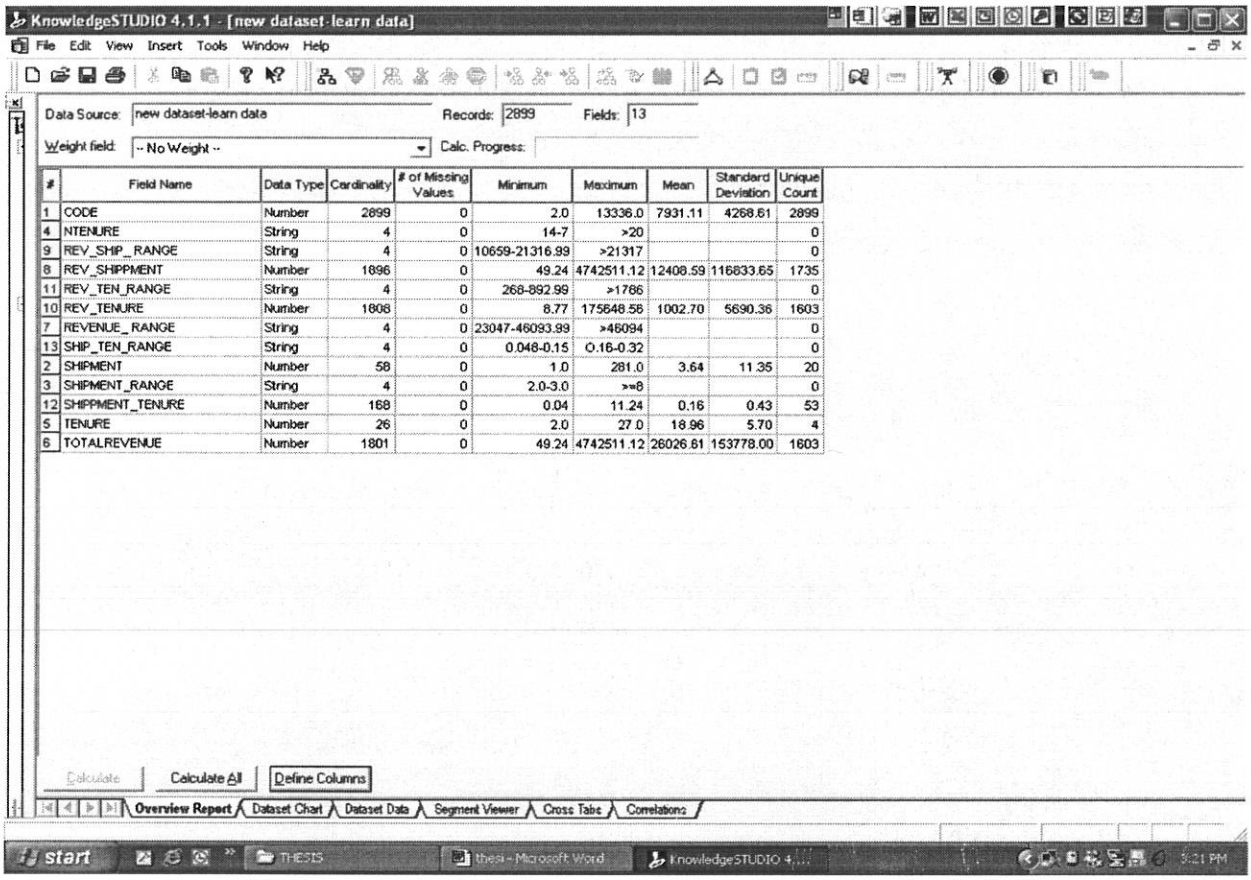
involves the identification of best attributes to be used for the next modeling, automatic cluster detection. The later involves automatic formation of clusters using selected attributes. Brief analysis and interpretation of clusters is also made in this section. This section ends by choosing the best classifiers for the decision tree model. The last subsection deals with building of decision tree and develops rules by taking the final clusters as dependent variables.

4.5.3.1 Attribute Selection

The construction of both clustering and classification models are very important so that best attributes are selected form the resulting decision tree built.

Six different models (three automatic cluster detection and three decision tree models) were built at different values of *"k"* to distinguish those attributes which have higher information content so that the next clustering will be made based on these attributes that would result a better model to understand and easier segments to interpret. The attributes used for this clustering includes all those, except code, listed in the data set after completion of the data preparation phase.

The basic parameters that can take different values in the automatic cluster detection models are the number of clusters to be created (k) and the number of iterations required (i). The number of customer segments (k), which can range from 2 to 20 in most of the time, is also dependent on the capacity of the firm to properly manage from four to six clusters. The other parameter, the number of iteration (i) refers to the maximum number of times the algorithm reads the data to form clusters. The number runs may be between 10-10,000, and as it increases, the algorithm runs longer and the results are accurate. The algorithm will stop running if it reaches this limit. Since the purpose of this experiment involves in identifying best attributes, all attributes are used in the initial model building at the same level for maximum value of iterations (i=10,000), but different values of k (3, 4 and 5). The attributes together with their details displayed by the Knowledge Studio are presented herein Fig. 4.2.



#	Field Name	Data Type	Cardinality	# of Missing Values	Minimum	Maximum	Mean	Standard Deviation	Unique Count
1	CODE	Number	2899	0	2.0	13336.0	7931.11	4268.61	2899
4	NTENURE	String	4	0	14-7	>20			0
9	REV_SHIP_RANGE	String	4	0	10659-21316.99	>21317			0
8	REV_SHIPMENT	Number	1896	0	49.24	4742511.12	12408.59	116833.65	1735
11	REV_TEN_RANGE	String	4	0	268-892.99	>1786			0
10	REV_TENURE	Number	1808	0	8.77	175648.56	1002.70	5690.36	1603
7	REVENUE_RANGE	String	4	0	23047-46093.99	>46094			0
13	SHIP_TEN_RANGE	String	4	0	0.048-0.15	0.16-0.32			0
2	SHIPMENT	Number	58	0	1.0	281.0	3.64	11.35	20
3	SHIPMENT_RANGE	String	4	0	2.0-3.0	>=8			0
12	SHIPMENT_TENURE	Number	168	0	0.04	11.24	0.16	0.43	53
5	TENURE	Number	26	0	2.0	27.0	18.96	5.70	4
6	TOTALREVENUE	Number	1801	0	49.24	4742511.12	26026.81	153778.00	1603

Fig. 4.2 Overview Report of the Training Dataset

Experiment 1

When k=3 and i=1000, 5000, 10000

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

Cluster run	No. of variables	No. of records	No. of clusters	No. of iterations
1	7	2899	3	1000
2	7	2899	3	5000
3	7	2899	3	10000

Table 4.3 Summary of Input Parameters for the first three runs in Experiment 1

Experiment 1a

When i=1000

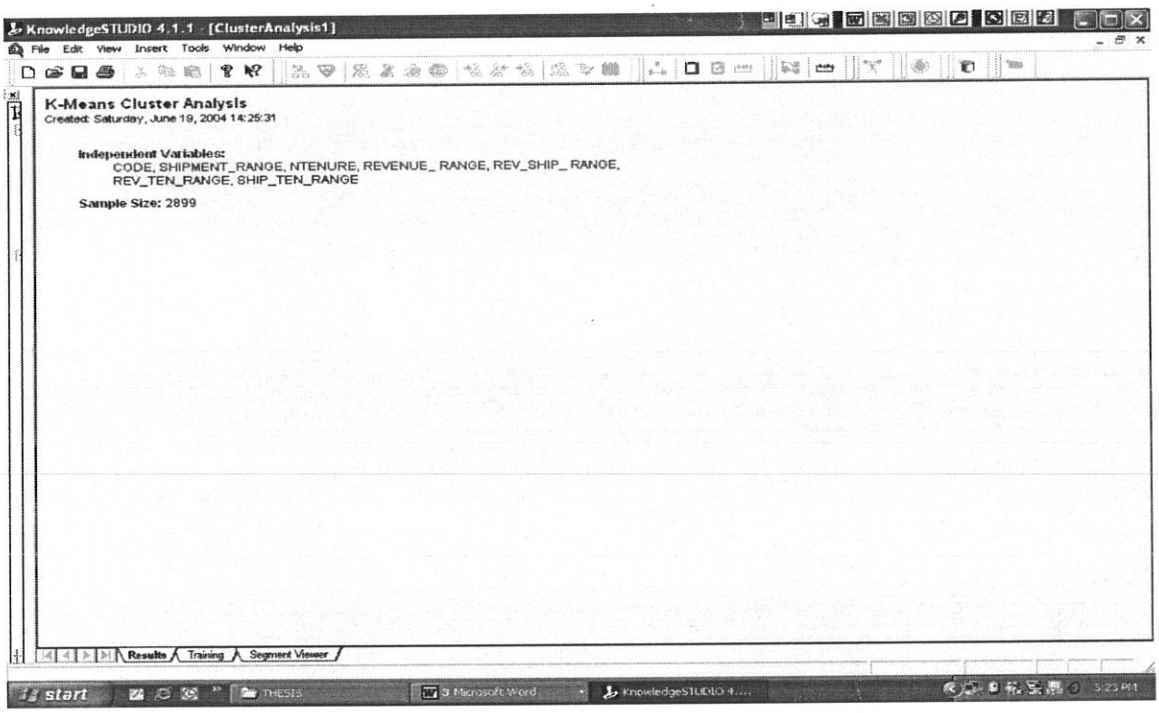


Fig. 4.3 Summary of the cluster analysis for experiment 1a

As depicted in the figure above, all the variables are taken as independent variables. It also shows the size of the training data set i. e. 2899 records (60% of the total dataset) is taken in the cluster analysis.

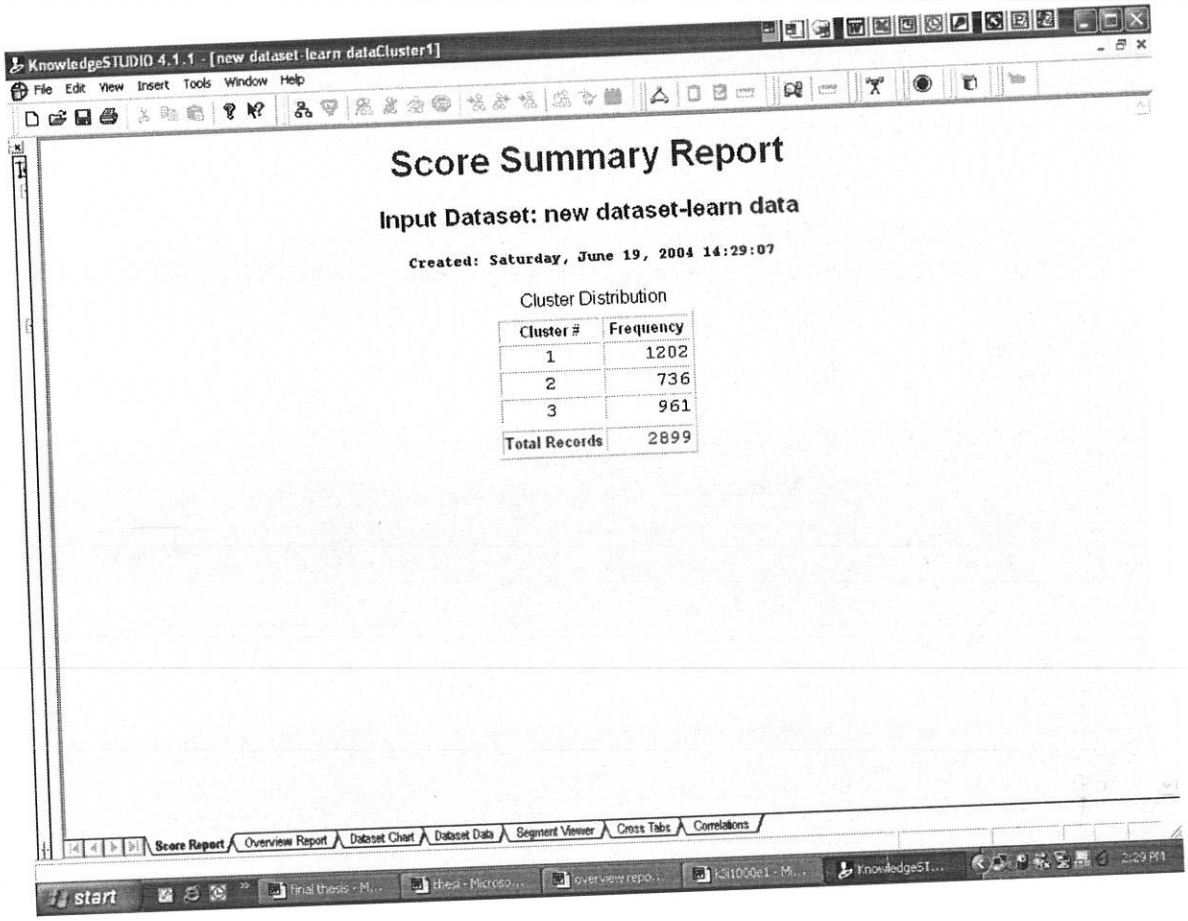


Fig. 4.4 Score report for experiment 1a

The above score report shows the three clusters with their corresponding frequency. As the figure reveals the first cluster consists majority of the customers (1202 customers out of 2899) belong to this cluster. Cluster 3 accounts for the second major group (961) customers and cluster 2 consists the least number of customers relative to the others. However; the number of customers in a given cluster alone does not reflect the importance of that cluster; it is the revenue they generate that matters most.

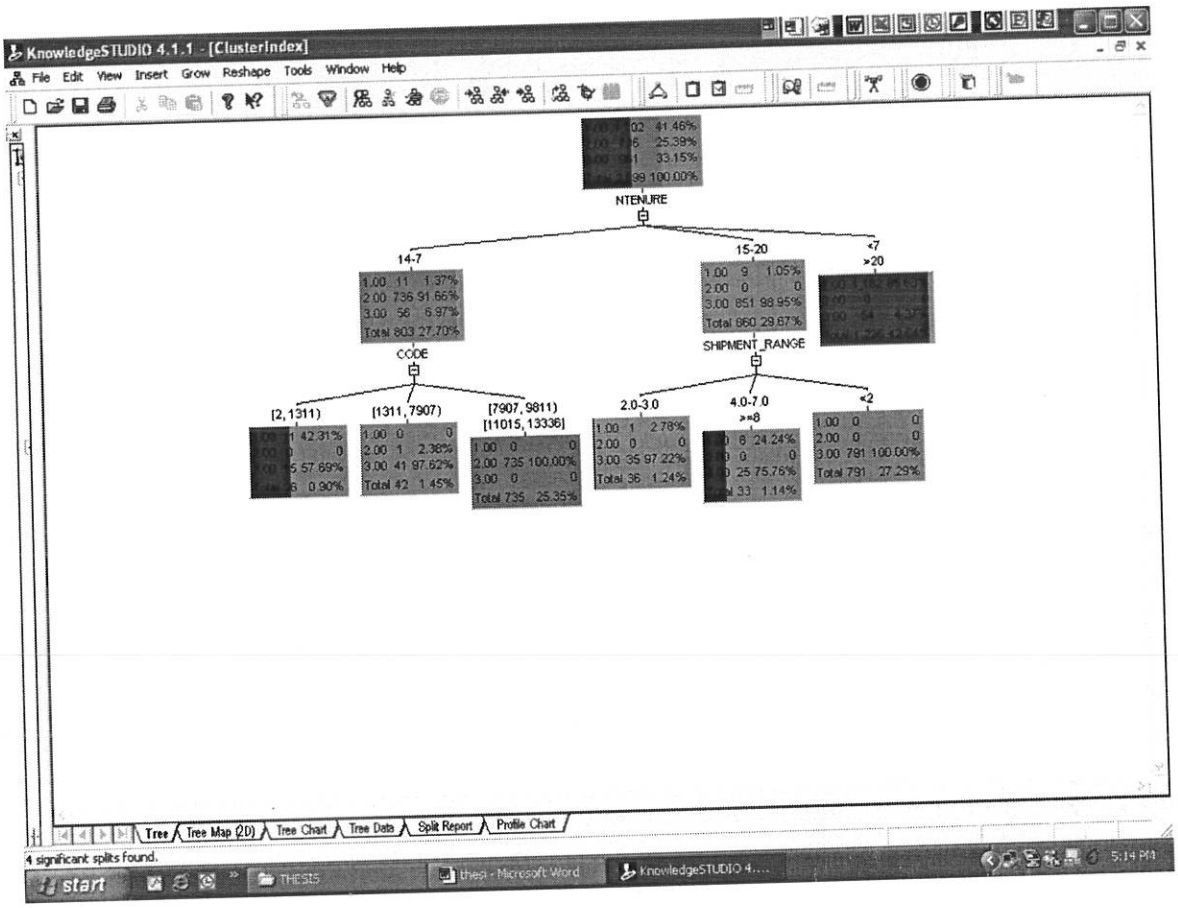


Fig 4.5 Decision tree for experiment 1a

The output of the three cluster runs (K=3 & i=1000, 5000, 10000) was a decision tree with the cluster index as the dependent variable. The decision tree provided a descriptive classification model of the clusters, thus enabling exploration and detection of the characteristics of each cluster.

Analysis of the output of the clusters revealed that it was quite difficult to detect patterns that identify the characteristics of each cluster.

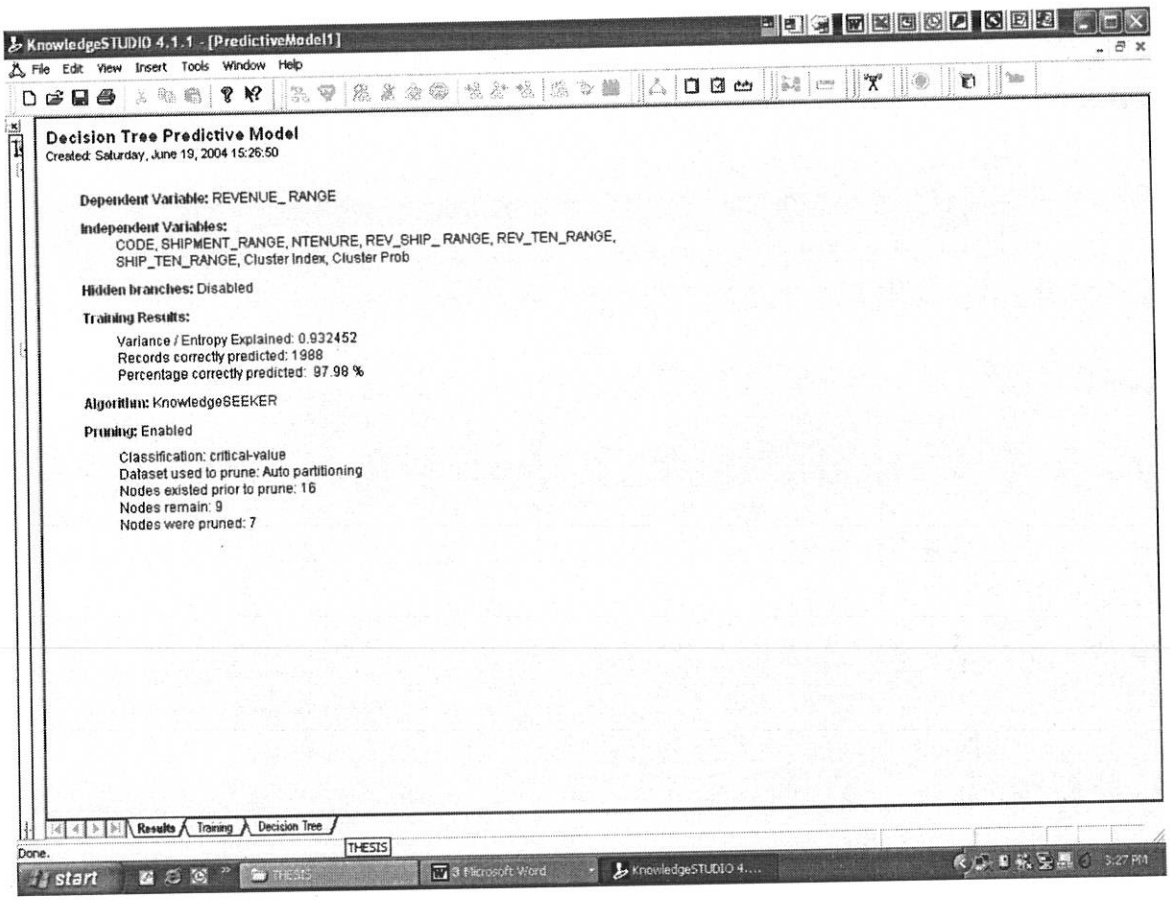


Fig 4.6 Predictive Model for Experiment 1a

The above predictive model is developed by taking cluster index as the dependent variable and all the others as independent variables. As the figure reveals 1988(97.98%) of the records were correctly predicted.

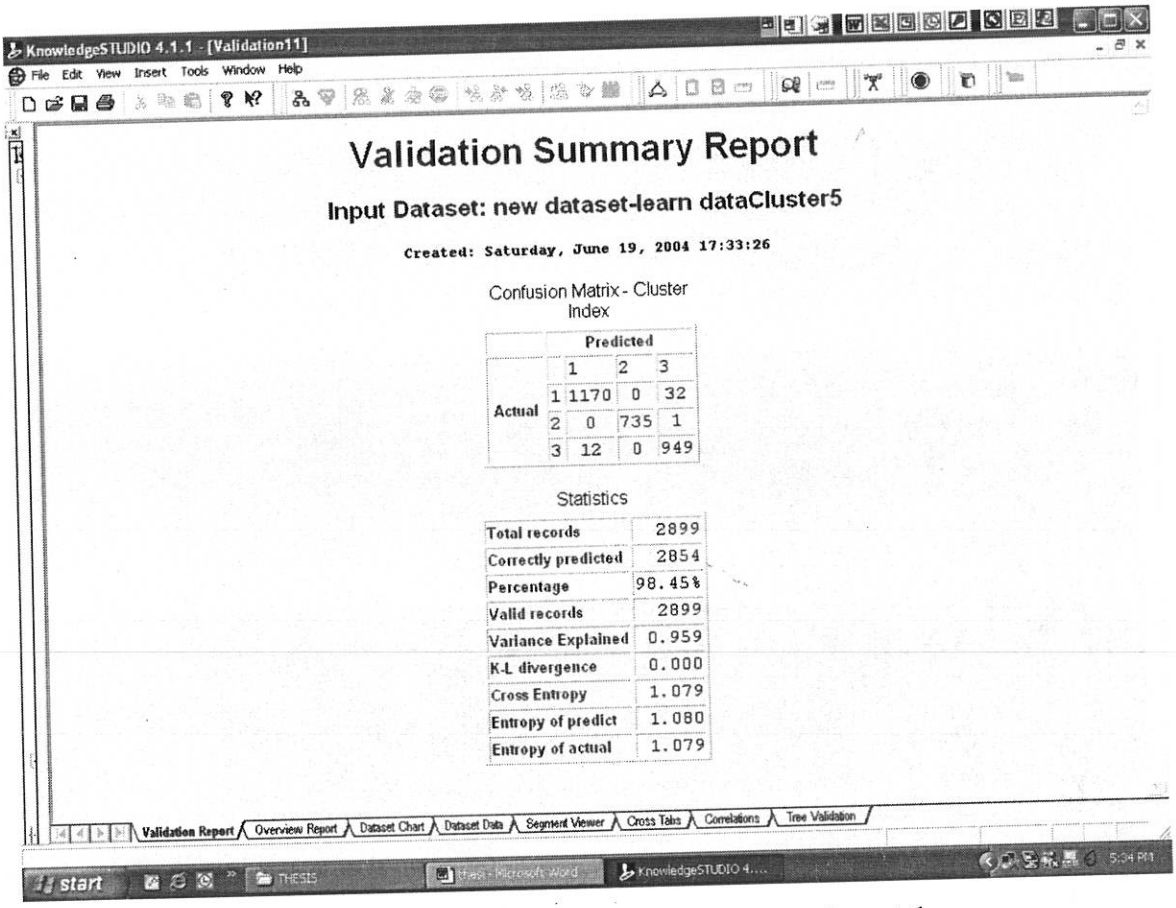


Fig 4.7 Validation Summary of the Decision Tree for experiment 1a

The validation summary shows that out of the 1202 customers of **Cluster 1**, 1170 were correctly predicted. The rest 32 were predicted as if they belong to Cluster 3. Out of the 736 records in Cluster 2, only one was wrongly predicted as a Cluster 3 record. 949 of the records in Cluster 3 were correctly predicted whereas 12 records were wrongly predicted as Cluster 1. Generally, 2854 out of 2899 records (98.45%) were correctly predicted

The same process was repeated When k= 3 but by making the value of i = 5000 and 10,000 but there was no change in the performance of the model so the next experiment shows the changes in the model when the value of k changes.

The Data Overview report from Knowledge Studio helps to evaluate the experiment based on the variables that are supposed to determine the behavior of customers. The overview report is used to set thresholds values in order to classify records as follows:

- TotalShipment: Very Frequent; Moderately Frequent; Frequent; Not Frequent
- TotalRevenue: High Revenue; Moderately High Revenue; Average Revenue; Low Revenue.
- Tenure: Long tenured, Moderately tenured, Recent tenure, Quite recent tenure

Variable	Cardinality	Minimum	Maximum	Mean
TotalRevenue	1771	49.25	2288259.43	21694.49
TotalShipment	53	1	249	3.31
Tenure	25	2	27	18.93
Rev_Tenure	1778	8.59	84750.35	842.19
Rev_Shipment	1869	49.25	2288259.43	9543.35
Shipment_Tenure	165	0.04	9.58	0.15

Table 4.4 Summary of the data overview report

Based on the values of the report the threshold values are set as follows:

Variable	Very frequent	Moderately frequent	Frequent	Not Frequent
ShipmentInterval	>=8	4-7	2-3	<2
Shipment_Tenure_Int	>0.32	0.16-0.32	0.048-0.15	<0.048

Table 4.5 Threshold for TotalShipment and Shipment_Tenure

Variable	High	Moderately high	Average	Low
RevenueInterval	>46094	23047-46093.99	6914-23046.99	<6913.99
Rev_Tenure_Int	>1786	893-1785.99	268-892.99	<=267.99
Rev_Shipment_Int	>21317	10659-21316.99	3197-10658.99	<3197

Table 4.6 Threshold for TotalRevenue, Rev_Tenure and Rev_Shipment

Variable	Long tenured	Moderately tenured	Recent tenure	Quite recent tenure
TenureInterval	>20	15-20	7-14	<7

Table 4.7 Threshold for Tenure

On the basis of the above threshold values and by splitting the decision tree with each variable the following summary is developed for **Experiment 1**.

When *K*=3 and the value of *i*=1000, 5000, 10000

- Cluster 1 consists both long tenured and recent customers, high revenue_shipment generators, high, moderately high and average revenue_tenure generators, high, moderately high and average revenue generators, very frequent, moderately frequent and frequent shipment_tenure, very frequent, moderately frequent and frequent shipment.
- Cluster 2 consists of recent, low revenue_shipment, low revenue_tenure, low revenue generators and not frequent customers.
- Cluster 3 consists moderately tenured, average and low revenue_shipment, low revenue generators and not frequent customers.

The details of each cluster enumerated above can be generalized in a table as follows:

CLUSTER #	REVENUE	TENURE	FREQUENCY	REMARK
1	HIGH, MODERATE, AVERAGE	LONG & RECENT	VERY-FREQUENT, MODERATE,FREQUENT	IMPORTANT CLUSTER
2	LOW	RECENT	NOT FREQUENT	
3	AVERAGE & LOW	MODERATE	NOT FREQUENT	

Table 4.8 Summary of Clusters when K=3

As the summary shows customers in cluster 1 are quite important. But there is no clear distinction of customers in this cluster. It consists all types of revenue generators except the Low category; all types of shipments made except the Not Frequent category. This is not a criterion for a good clustering technique. So, by taking this as an input, the next experiment segments the customers into four clusters.

Experiment 2

When the value of *k* is 4 and *i* is 10000

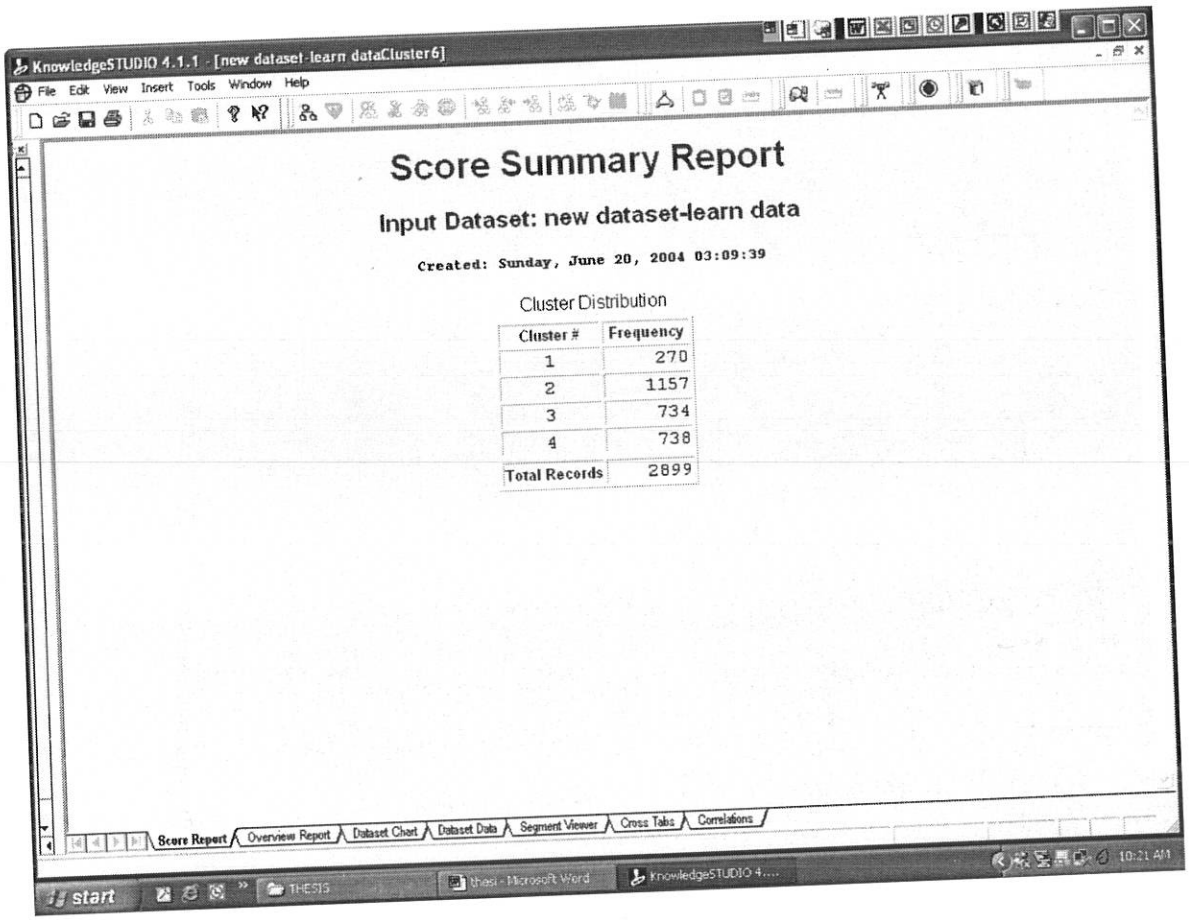


Fig 4.8 Cluster analysis when *k*=4 and *i*=10000

Majority of the customers fall under Cluster 2 (1157 out of 2899). Cluster 4 consists the second larger group. Cluster 3 is the third one & cluster 1 consists the least number of customers (270 out of 2899).

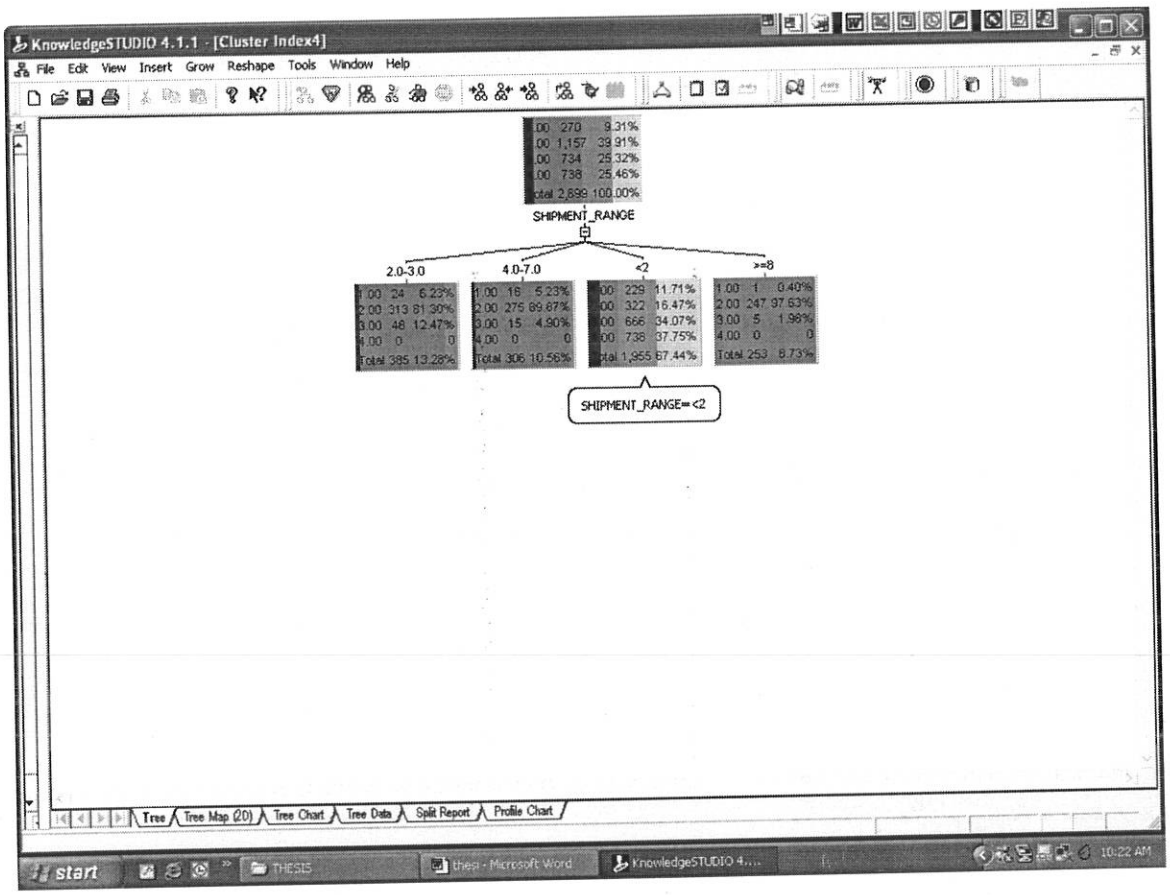


Fig. 4.9 Decision tree when k=4 and i=10000

The decision tree provided a descriptive classification model of the clusters, thus enabling exploration and detection of the characteristics of each cluster.

Using the decision tree it was possible to detect interesting patterns about the customers of ESL.

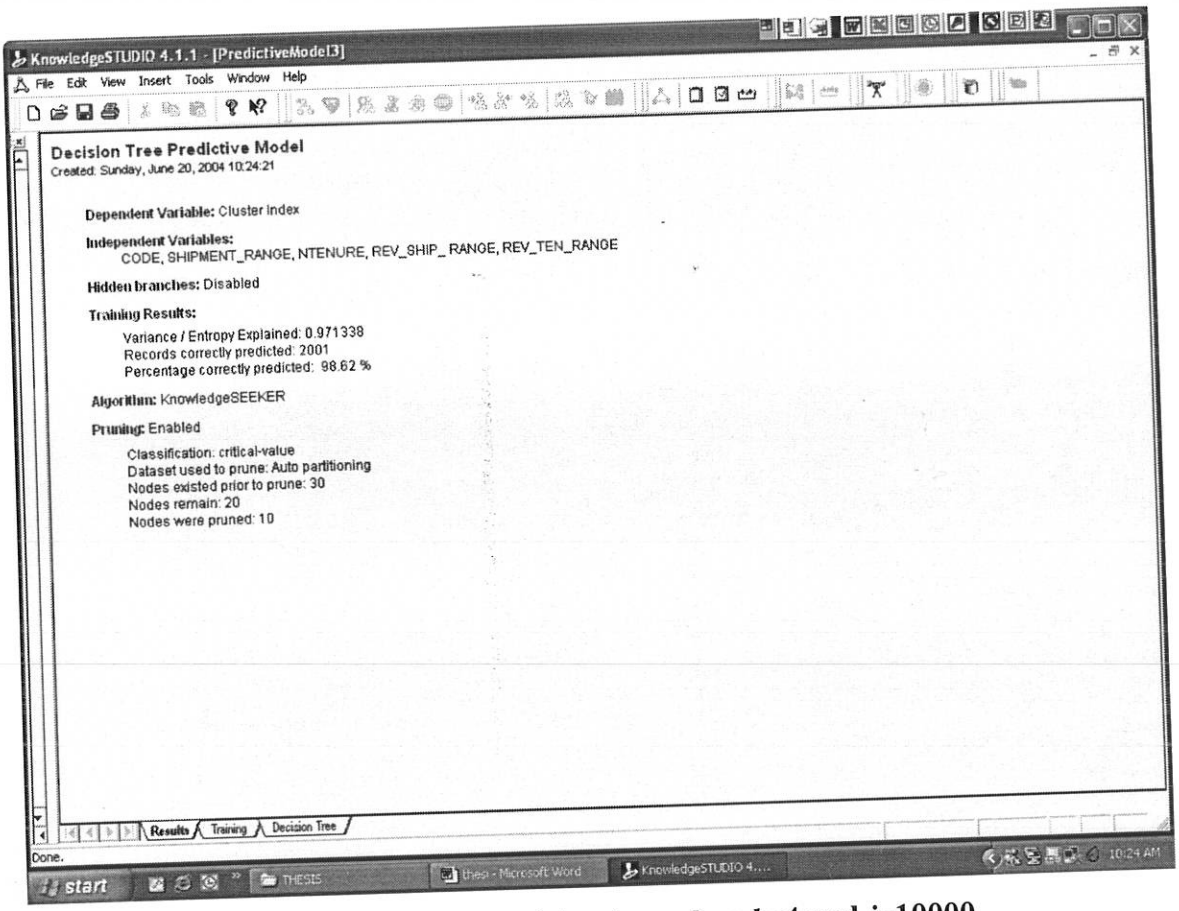


Fig. 4.10 Predictive model for the decision tree when k=4 and i=10000

The above predictive model is developed by taking cluster index as the dependent variable and all the other as independent variables. As the figure reveals 2001i.e 98.62% of the records were correctly predicted.

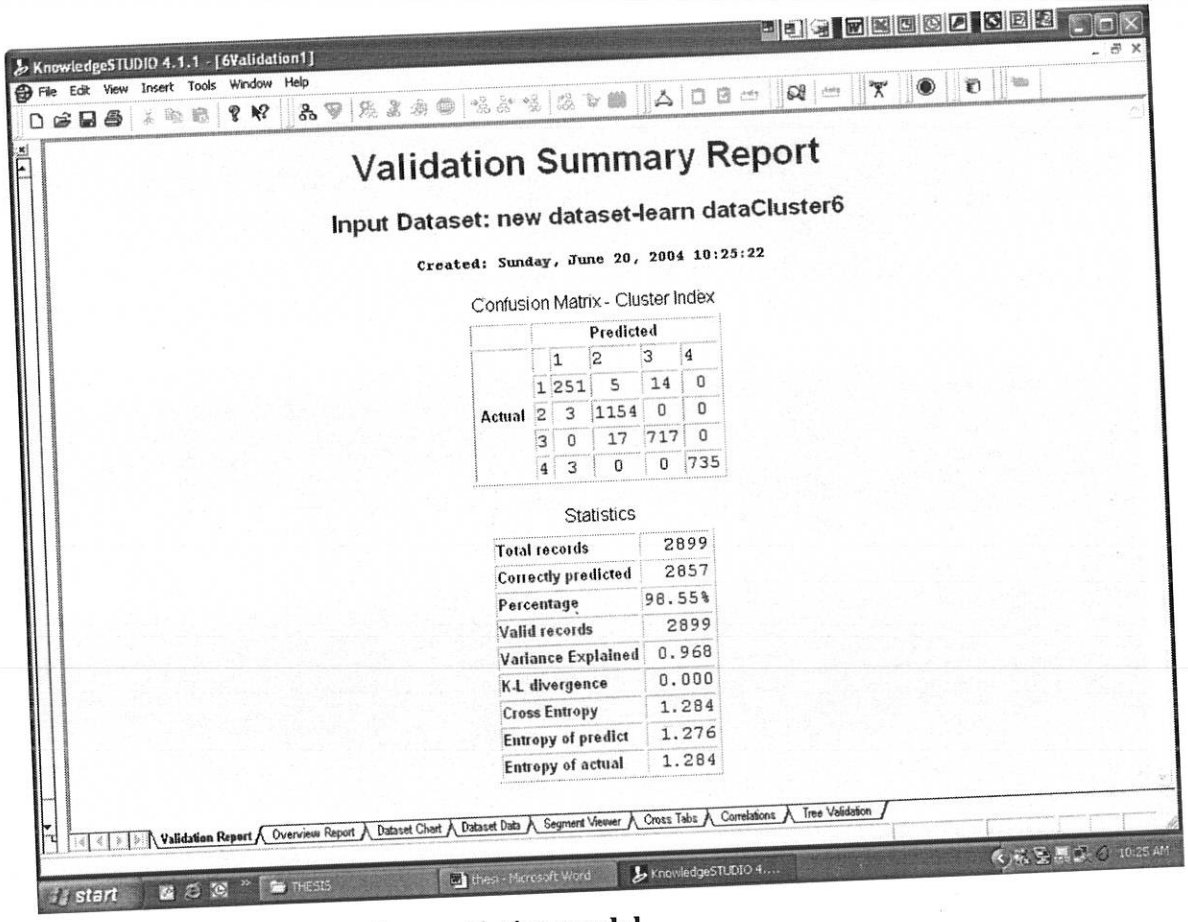


Fig. 4.11 Validation for the predictive model

The validation summary shows that out of the 270 customers of **Cluster 1**, 251 were correctly predicted. Out of the 1157 records in Cluster 2, only 3 were wrongly predicted as Cluster 1 records. 717 of the records in Cluster 3 were correctly predicted whereas 17 records were wrongly predicted as Cluster 2. Out of the 738 records in Cluster 4, only 3 were wrongly predicted as Cluster 1 customers. Generally, 2857 out of 2899 records (98.55%) were correctly predicted

From the above Experiment, the following behaviors are detected from each cluster:

CLUSTER 1

This cluster consists:

- 4.24% of long tenured customers, 16.63% of moderately tenured customers, 8.09% of recent tenure customers and all quite recent customers.
- No high revenue_shipment generators, no moderately high revenue_shipment generators, 11.39% of average revenue_shipment generators and 10.43% of low revenue_shipment generators.
- No high revenue_tenure generators, no moderately high revenue_tenure generators, no average revenue_tenure generators, 14.97% of low revenue_tenure generators.
- No high, moderately high and average revenue generators, 14.18% of low revenue generators.
- 3.31% very frequent shipment_tenure customers, 6.29% of moderately frequent shipment_tenure, 10.68% of frequent shipment_tenure and 8.26% of Not Frequent customers.
- 0.40% of Very Frequent, 5.23% of Moderately Frequent,6.23% of Frequent and 11.71% of Not Frequent customers.

CLUSTER 2

This cluster consists:

- 94.37% of Long Tenured customers, no Moderately Tenured, Recent Tenure and Quite Recent Tenure customers.
- 100% of high revenue_shipment, 100% of moderately high revenue_shipment, 67.31% of Average revenue_shipment and 18.53% of low revenue_shipment generators.
- 100% of high revenue_tenure, 100% of moderately high revenue_tenure,100% of Average revenue_tenure and 3.44% of low revenue_tenure generators.
- 100% of high revenue, 100% of moderately high revenue, 100% of average revenue and 8.51% of low revenue generators.

- 93.80% of very frequent shipment_tenure, 89.4%of Moderately frequent shipment_tenure, 16.87% of Frequent shipment_tenure and 91.74% of Not Frequent shipment_tenure customers.
- 97.63% of Very Frequent,89.87% of Moderately Frequent, 81.3% of frequent customers and 16.47% of Not Frequent customers..

CLUSTER 3

This cluster consists:

- 1.39% of Long Tenured customers, 83.37% of Moderately Tenured, no Recent Tenure and Quite Recent Tenure customers.
- No high revenue_shipment, 31.60% of moderately high revenue_shipment customers, 21.30% of Average revenue_shipment and no low revenue_shipment generators.
- No high revenue_tenure, 38.55% of Moderately high revenue_tenure, no Average revenue_tenure customers, no low revenue_tenure generators.
- No high revenue, 18.51% moderately high revenue and no average revenue and low revenue generators.
- 2.89% of very frequent shipment_tenure, 4.3%of Moderately frequent shipment_tenure, 35.63% of Frequent shipment_tenure and no customer belongs to the Not Frequent shipment_tenure customers.
- 1.98% of Very Frequent,4.9% of Moderately Frequent, 12.47% of Frequent customers and 34.07% of Not Frequent customers.

CLUSTER 4

This cluster consists:

- No Long Tenured customers and Moderately Tenured, 91.91% of Recent Tenure and no Quite Recent Tenure customers.
- No high revenue_shipment, moderately high revenue_shipment, 39.53% of Average revenue_shipment customers, no low revenue_shipment generators.

- No high revenue_tenure, Moderately high revenue_tenure, 40.91% of Average revenue_tenure customers and no low revenue_tenure generators.
- No high revenue, moderately high revenue and average revenue generators, 38.76% of low revenue generators.
- No very frequent shipment_tenure and Moderately frequent shipment_tenure, 36.83% of Frequent shipment_tenure and no customer belongs to the Not Frequent shipment_tenure customers.
- No Very Frequent, Moderately Frequent and Frequent, 37.75% of Not Frequent customers.

Summary of customers in each cluster when k=4

- Cluster 1 consists of customers that are Not Frequent, Low revenue generators and quite recent in their tenure as customers. This cluster accounts for the least number of records, (9.3%) belong to this segment. It is also the least important of all the clusters.
- Cluster 2 contains Very frequent, high revenue generators and long tenured customers. This segment makes up the majority of the customers (39.91%) out of the total. This cluster represents a valuable group of customers.
- Cluster 3 contains frequent, moderate revenue generators and that are moderately tenured in the company. This segment makes up 25.31% of the total records.
- Cluster 4 consists of customers that are Not frequent, average revenue generators and recent in their tenure as customers. 25.45% of customers belong to this cluster.

The Table below shows the behavior of the clusters in a summarized way:

CLUSTER #	REVENUE	TENURE	FREQUENCY	REMARK
1	LOW	QUITE RECENT	NOT FREQUENT	
2	HIGH	LONG	VERY FREQUENT	IMPORTANT CLUSTER
3	MODERATE	MODERATE	FREQUENT	
4	AVERAGE	RECENT	NOT FREQUENT	

Table 4.9. Summary of Clusters when K=4

This experiment results in a very interesting pattern of customer clustering. It relatively fulfills the criteria of good clustering technique. There is a high intraclass similarity and a low interclass similarity. Though customers are well clustered in this experiment, another experiment by changing the value of k to 5 is made in order to compare the different experiments and to choose the best among them.

Experiment 3

When the value of k=5 and i=10000

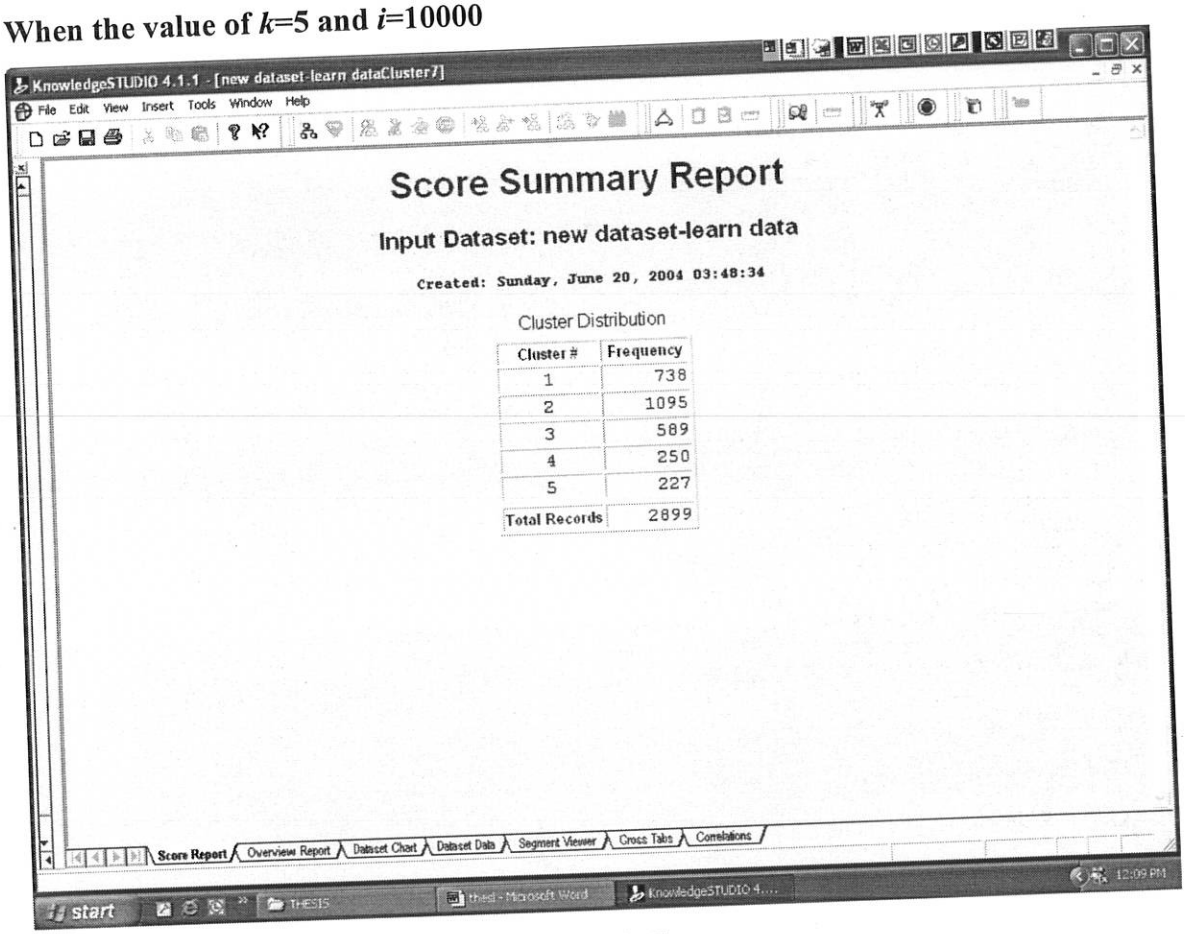


Fig 4.12 cluster analysis when the value of k=5

Cluster 2 comprises the highest number of customers (1095 out of 2899). Cluster 3 accounts for the second largest group. Cluster 1 is the third. Cluster 4 and 5 accounts for the 4^{th} & 5^{th} largest classes respectively.

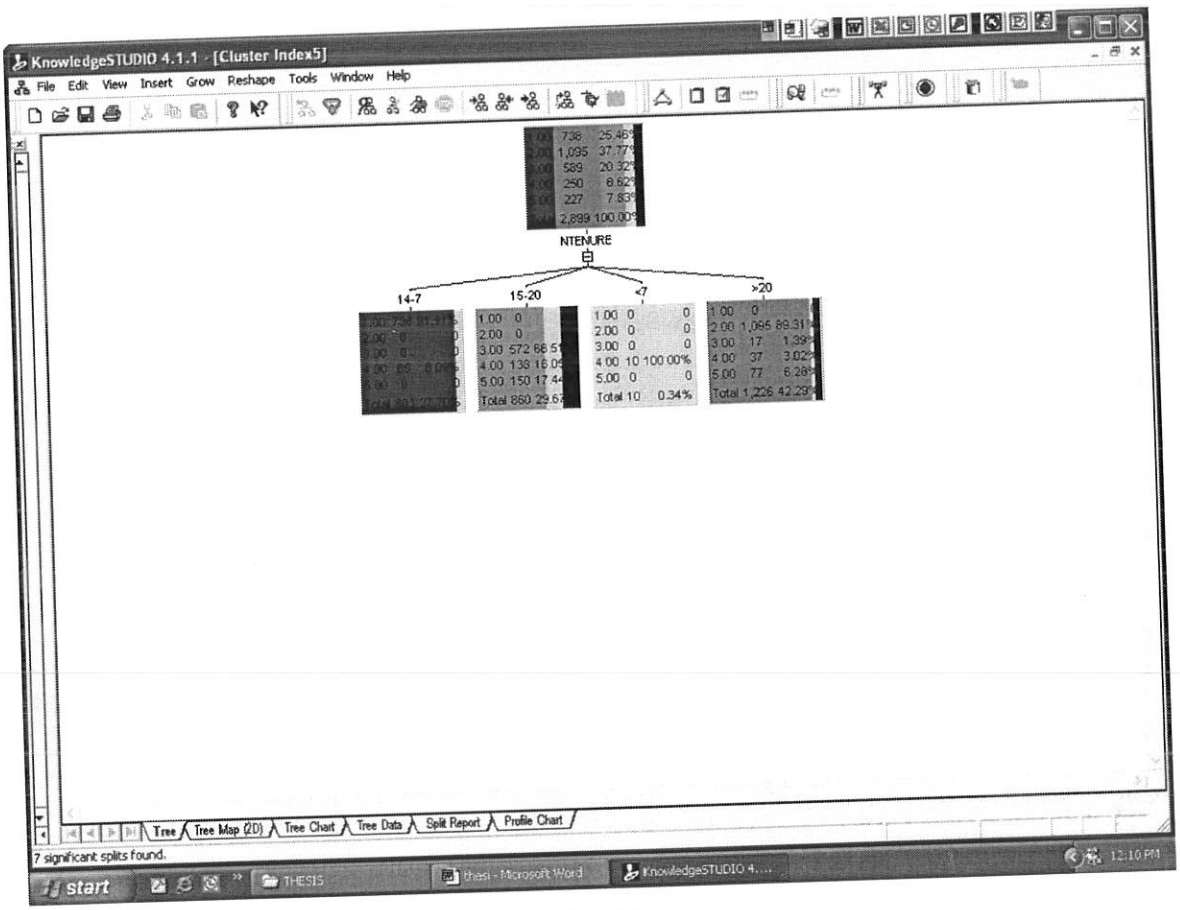


Fig 4.13 Decision tree when the value of k=5

The output of the cluster run when K=5 & i= 10000 was a decision tree with the cluster index as the dependent variable. The decision tree provided a descriptive classification model of the clusters so as to enable exploration and detection of the characteristics of each cluster. However; analyzing the output of the clusters in this experiment was quite difficult.

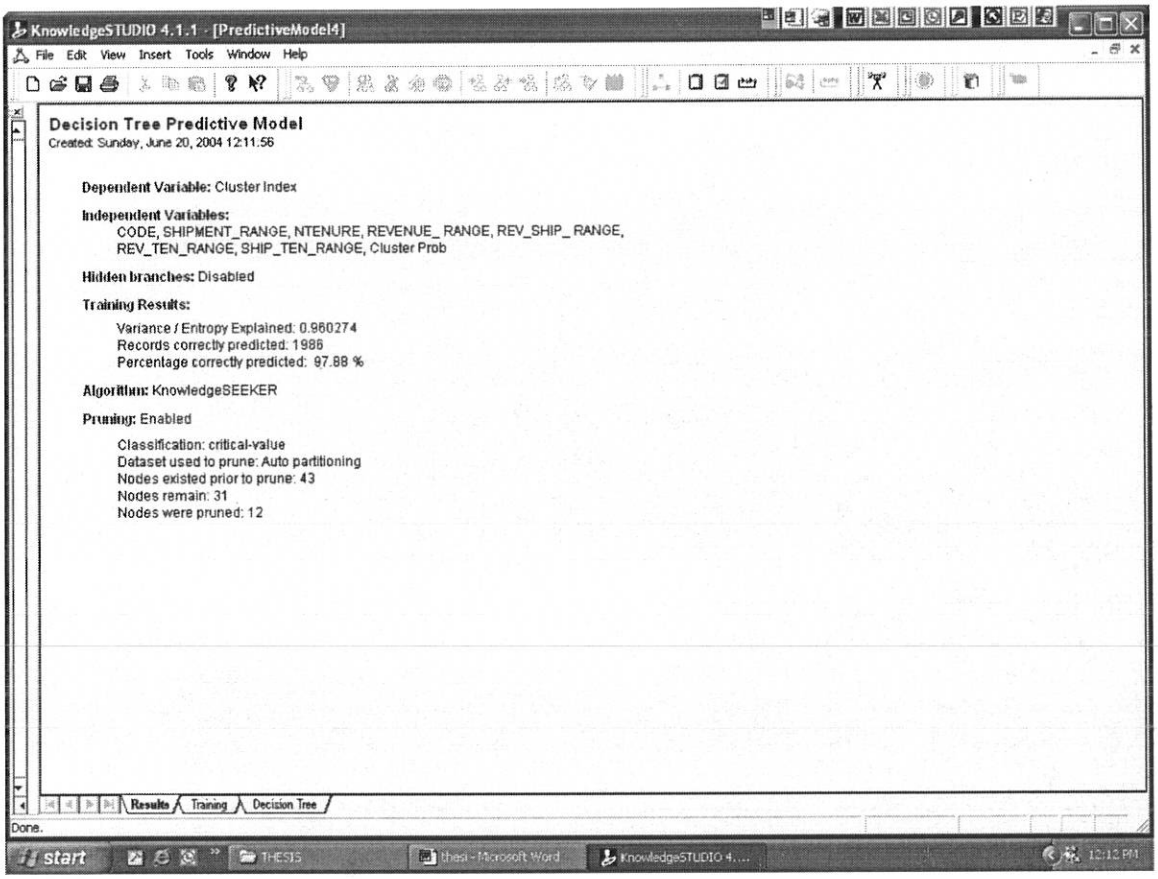


Fig. 4.14 Predictive model for the decision tree when k=5

The above predictive model is developed by taking cluster index as the dependent variable and all the other as independent variables. As the figure reveals 1986i.e 97.88% of the records were correctly predicted.

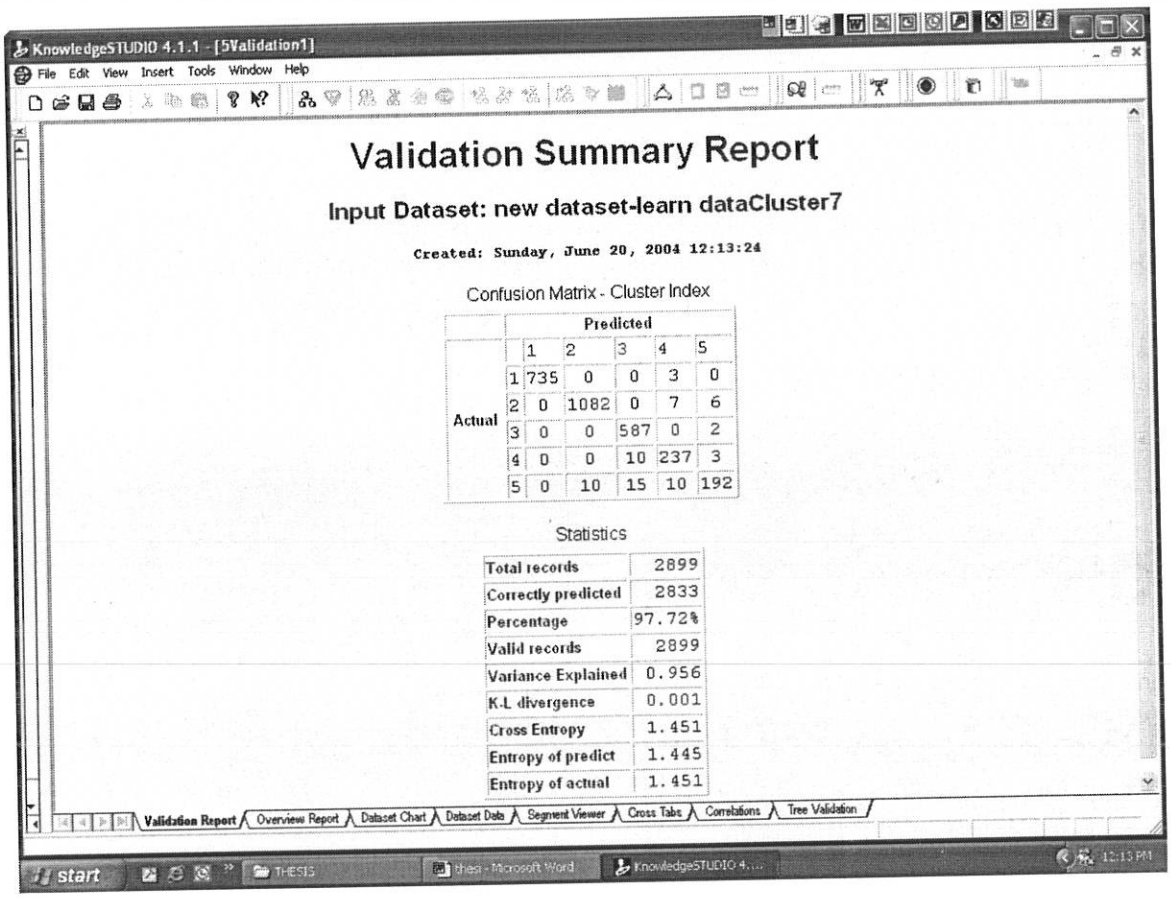


Fig. 4.14 Validation of the predictive model when k=5

The validation summary shows that out of the 738 customers of **Cluster 1**, 735 were correctly predicted. Out of the 1095 records in Cluster 2, only 13 were wrongly predicted. 587 of the records in Cluster 3 were correctly predicted whereas only 2 records were wrongly predicted as Cluster 5. Out of the 250 records in Cluster 4, 13 were wrongly predicted. Out of the 227 records in Cluster 5, 35 were wrongly predicted. Generally, 2833 out of 2899 records (97.72%) were correctly predicted

CLUSTER 1

This cluster consists:

- No long tenured & moderately tenured customers, 91.91% of recent tenure customers and no quite recent customers.
- No high revenue_shipment generators, no moderately high revenue_shipment generators, no average revenue_shipment generators and 39.53% of low revenue_shipment generators.
- No high revenue_tenure generators, no moderately high revenue_tenure generators, no average revenue_tenure generators, 40.91% of low revenue_tenure generators.
- No high, moderately high and average revenue generators, 38.76% of low revenue generators.
- 3.31% very frequent shipment_tenure customers, no moderately frequent shipment_tenure, 36.83% of frequent shipment_tenure and no Not Frequent shipment_tenure customers.
- No Very Frequent, no Moderately Frequent, no Frequent and 37.75% of Not Frequent customers.

CLUSTER 2

This cluster consists:

- 89.31% of long tenured, no moderately tenured customers, no recent tenure and no quite recent customers.
- 100% of high revenue_shipment generators, 100% of moderately high revenue_shipment generators, 57.54% of average revenue_shipment generators and 18.75% of low revenue_shipment generators.
- 100% of high revenue_tenure generators, 100% of moderately high revenue_tenure generators, 95.91% of average revenue_tenure generators, 1.44% of low revenue_tenure generators.
- 100% of high, moderately high and average revenue generators, 5.25% of low revenue generators.

- 94.21% very frequent shipment_tenure customers, 90.40% moderately frequent shipment_tenure, 16.87% of frequent shipment_tenure and 72.93% of Not Frequent shipment_tenure customers.
- 97.63% of Very Frequent, 91.18% of Moderately Frequent, 81.30% of Frequent and 13.09% of Not Frequent customers.

CLUSTER 3

This cluster consists:

- 1.39% of long tenured, 66.51% of moderately tenured customers, no recent and no quite recent tenure customers.
- No high revenue_shipment generators, no moderately high revenue_shipment generators, no average revenue_shipment generators and 31.55% of low revenue_shipment generators.
- No high revenue_tenure generators, no moderately high revenue_tenure generators, no average revenue_tenure generators, 32.65% of low revenue_tenure generators.
- No high, moderately high and average revenue generators, 30.93% of low revenue generators.
- 2.89% of very frequent shipment_tenure customers, 4.30% of moderately frequent shipment_tenure, 28.39% of frequent shipment_tenure and no Not Frequent shipment_tenure customers.
- 1.98% of Very Frequent, 4.90% of Moderately Frequent, 12.47% of Frequent and 26.55% of Not Frequent customers.

CLUSTER 4

This cluster consists:

- 3.02% of long tenured, 16.05% of moderately tenured customers, 8.09% of recent and 100% of quite recent tenure customers.
- No high revenue_shipment generators, no moderately high revenue_shipment generators, 8.88% of average revenue_shipment generators and 10.18% of low revenue_shipment generators.

- No high revenue_tenure generators, no moderately high revenue_tenure generators, no average revenue_tenure generators, 13.86% of low revenue_tenure generators.
- No high, moderately high and average revenue generators, 13.13% of low revenue generators.
- 2.89% of very frequent shipment_tenure customers, 5.30% of moderately frequent shipment_tenure, 10.43% of frequent shipment_tenure and 5.13% of Not Frequent shipment_tenure customers.
- 0.40% of Very Frequent, 3.92% of Moderately Frequent, 6.23% of Frequent and 10.9% of Not Frequent customers.

CLUSTER 5

This cluster consists:

- 6.28% of long tenured, 17.44% of moderately tenured customers, no recent and quite recent tenure customers.
- No high revenue_shipment generators, no moderately high revenue_shipment generators, 33.58% of average revenue_shipment generators and no low revenue_shipment generators.
- No high revenue_tenure generators, no moderately high revenue_tenure generators, 4.09% of average revenue_tenure generators, 11.14% of low revenue_tenure generators.
- No high, moderately high and average revenue generators, 11.92% of low revenue generators.
- No very frequent shipment_tenure customers, no moderately frequent shipment_tenure, 7.49% of frequent shipment_tenure and 21.94% of Not Frequent shipment_tenure customers.
- No Very Frequent, Moderately Frequent and Frequent and 11.61% of Not Frequent customers.

Summary of customers in each cluster when k=5

- Cluster 1 consists Recent, Low Revenue Generators and Not Frequent, Frequent shipment_tenure customers. 25.46% of the customers belong to this cluster.
- Cluster 2 consists Long Tenured (Loyal), High Revenue generators and Very Frequent. Customers they are also characterized by being high revenue_shipment and high revenue_tenure generators. 37.77% of the total customers belong to this cluster.
- Cluster 3 consists Moderately Tenured, Low Revenue generators , Not Frequent customers. 20.32% of the total customers belong to this cluster.
- Cluster 4 consists Quite Recent Tenure, Low Revenue generators and Not Frequent customers. 8.62% of the total customers are in this cluster.
- Cluster 5 consists Moderately tenured, Not Frequent, Low Revenue generators. 7.83% of the total customer.

The behavior of each cluster can be summarized in a Table as follows:-

CLUSTER	REVENUE	TENURE	FREQUENCY	REMARK
1	LOW	RECENT	NOT FREQUENT	
2	HIGH	LONG	VERY FREQUENT	IMPORTANT CLUSTER
3	LOW	MODERATE	NOT FREQUENT	
4	LOW	QUITE RECENT	NOT FREQUENT	
5	LOW	MODERATE	NOT FREQUENT	

As the summary indicates customers in cluster 3 and 5 are very similar. According to the criteria of a good clustering technique the customers should be in one cluster in order to guarantee high intra class similarity and low interclass similarity.

The results of the above three experiments can be generalized as follows:

Value of k	Prediction (%)	Validation (%)
3	98.37	98.45
4	98.62	98.55
5	97.88	97.72

Table4.18 summary of the three experiments

From the summary of the behaviors resulted from the three experiments and from the above generalization it is very meaningful to segment the customers into four clusters. Hence, the researcher is interested to focus on the second experiment and say few points about each cluster in that experiment. For the sake of discussion let's bring the Table that generalizes Experiment 2 once again.

The behavior of each cluster can be summarized in a Table as follows:-

CLUSTER #	REVENUE	TENURE	FREQUENCY	REMARK
1	LOW	QUITE RECENT	NOT FREQUENT	
2	HIGH	LONG	VERY FREQUENT	IMPORTANT CLUSTER
3	MODERATE	MODERATE	FREQUENT	
4	AVERAGE	RECENT	NOT FREQUENT	

Table 4.9. Summary of Clusters when K=4

Based on the above summary, the clusters can be listed on the basis of their importance to ESL as follows:-

Cluster #	Description	Percentage
2	High Revenue, Very Frequent, Long Tenure	39.91
3	Moderately High Revenue, Frequent, Moderately Long Tenure	25.31
4	Average Revenue, Not Frequent, Recent Tenure	25.45
1	Low Revenue, Not Frequent, Quite Recent	10.33

Table 4.19 Summary of the clusters when K=4 in order of their importance

Hence, ESL should give particular emphasis to the customers who belong to Cluster 2 & 3. Customers in Cluster 2 are extremely important since they are high revenue generators, long tenured or loyal and very frequent ones. These individuals are the backbones of the company. So ESL should give more individual attention, be more responsive and develop customization in order to maintain these customers and maximize its profit. Focusing on its profitable customers enable the company to identify their needs and act accordingly to satisfy them. This in turn makes the customers to be loyal to the company.

Rule generation

The rules applicable for the selected decision tree as generated by the tool were very detail and many. Sample of these is annexed as appendix 1.

4.6 Evaluation

Once an optimum model is built, critical assessment of the model against the business goal to be achieved (business problem to be addressed) is very important.

The business goal was to segment customers into meaning full groups so that an appropriate CRM strategies and programs can be designed and implemented. The basis of segmentation was according to customers' value to the business. And, customer value was defined based on revenue and its derivative variables. The data

mining process having a task of clustering was performed in order to achieve the business objectives set.

The results of the data mining process were encouraging and at least provide a way of reaching data mining solutions for market segmentation for the organization, if not yield a possible solution. Customers who generate similar revenue were grouped in the same group whereas the groups formed were different from each other. This is the underlying criterion of segmentation (clustering). Moreover, the decision tree model provided a description of the segments and rules for assigning new records to segments.

The researcher believes that with further analysis of the results by marketing experts and IT specialists, the data mining process could be revisited to produce an optimal segmentation scheme so that relevant customer related information for informed decision making is acquired that help the organization manage its scarce resources efficiently, effectively and economically.

4.7. Deployment of results

Application of a segmentation scheme involves the consumption of considerable resources since people, business processes and technology together are integrated and directed based on the information obtained form the segmentation plan. Therefore, the results of this study can be deployed for marketing decision making after a thorough evaluation and integrating the necessary adjustments by group of domain experts.

CHAPTER FIVE

Conclusion and Recommendations

5.1 Conclusion

The major objective of this thesis was to segment customers into similar groups based on their revenue generating behavior. Hence, the number of shipments made, the revenue collected, the number of months since the customer first enrolled in the company, the ratio of total revenue to total customer's tenure, ratio of total revenue to total number of shipments and ratio of total number of shipments to customer tenure are used as a criteria to segment customers.

As the result reveals it was possible to segment customers based on their profitability and hence as per their long term potential to generate revenue.
The results of the research were encouraging as the domain experts in the ESL accepted it. It can derive benefit through an appropriate utilization of the research to improve its customer relationship management that is one of the hottest issues to be addressed in today's customer oriented, dynamic, and competitive market environment.

In general, the study focused in the application of data mining in the area of CRM and more specifically for the purpose of customer centric market segmentation at Ethiopian Shipping Lines. To this effect, related literature on data mining techniques, CRM and market segmentation was reviewed. In the experimentation part, the CRISP data mining process model was followed to complete the data mining task.

5.2 Recommendations

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

- Need to build a data warehouse: This study was a victim of a very lengthy data preparation process that consumes considerable time, effort and other resources. It is very important to build a data warehouse not only for data mining purposes but also other important data analysis tasks as optimal decision making is only feasible if it is based on reliable and relevant information. Customers' data at various contact points should be collected and integrated as such data in these days becomes much valuable.

- Undertaking further data mining researches: There is always a room for improvement in data mining as it is an interactive and iterative process. It needs refinement, update and enhancement as business problems always become diverse and complex. In these context, this research can be further refined through changing the techniques and algorithms, the data, and the modeling parameters used in the study. More specifically,

    - ❖ The employment of the neural network for clustering and classification, which is very popular data mining technique, may yield better results.
    - ❖ The type and number of models built can also be enhanced so that an optimal model may be selected after analysis of relatively high number of potential models.

- Enhancing CRM and data mining understanding at all levels: In order to compete in the today's global marketing, ESLSC should give much emphasis to create value to the business at each and every contact with the customer. Hence, the

entire employee should act towards this common goal so as to ensure its feasibility. On the other hand, the Management should take and accomplish the responsibility of proper employee training and continuous upgrading. This is extremely important to properly serve the cluster and benefit from the ever-advancing Information and Communication Technologies.

- Provide support for researches: Researchers would not lose their considerable time and effort if there were a concerned body in the organization that facilitates such work. If information and consultancy services together with material support are forwarded to the researchers. Research problems related to research work will be greatly reduced and researches may end in success.

- Need for strong commitment for research and change: Application of customer relationship management and related technologies like data mining require continuous and flexible approach, and analysis of the dynamic nature of customers. Such researches should be initiated and conducted in house, and resulting outputs should be utilized to design, implement, and continually improve CRM strategies and programs at all level of the organization.

References

1. Angoss software corporation, knowledge studio user manual, 2003.
   www.angoss.com
1. Basgoze, A,& Gokturk, M, Building customer profiles using data mining techniques, Turkish symposium on artificial intelligence and neural networks, 2003,
2. Berry J.& Linoff,S., Mastering data mining, john wiley & sons Inc, 2000.
3. Bishop, M., Neural Networks for pattern recognition, oxford University press., 1998.
4. Bose, R., Customer relationship Management: key components for IT success, Industrial Management and Data System 102/2, 89-97 2002.
   a. www. Emeraldinsight, com
5. Bounsaythip, C. & Rinta-Runsala, E., Overview of Data Mining for customer behavior modeling, LOUHI, 2001,
   http://www. Vtt.fi/tte/
6. Bull, C., Strategic Issues in Customer Relationship Management implementation, Christopher, Business process Management journal, volume 9,No 5,2003.
7. Connely M,T. & Begg E.C, Data Base Systems. A practical approach to design, implement and management, 3rd Ed, 2000.
8. CRISP-DM., CRISP-DM 1.0: Step-by-step data mining guide. 2000.
   http://www.crisp-dm.org
9. DSS Research. Understanding Market Segmentation, 2001.
   http://www.dssresearch. Com/liabrary/segment/ understanding.asp
10. Ethio-Ship, Ethiopian Shipping Lines, Vol.11, No.11, Jan.18, 2005
11. Fekadu Mekonnen, Application of Data Mining Techniques to Support Customer Relationship Management at Ethiopisn Telecommunications Corporation, Department of Information Science, Faculty of Informatics, Addis Ababa University, 2004
12. Forcht, K.A & Cochran, K., Using Data Mining and Data werhousing techniques, Industrial Management and data Systems, pp 189-1999.
    www.emeraldinsight.com

13. Gray,P.and Byun,J., Customer Relationship Management, March 2001.
14. Han, F,and Kamber, M., Data Mining, Concepts and Techniques, Morgan Kaufmann publishers, academic press, 2001.
15. Henock Woubishet, Application of Data Mining Technology to Support Customer Relationship Management at Ethiopian Airlines, Department of Information Science, Faculty of Informatics, Addis Ababa University,2002.
16. IBM Corporation, enhance Your Business Applications: Simple Integration of Advanced Data Mining Functions Advanced data Mining Functions, 2002. http://www:redbook.ibm.com/redbooks/pdfs/sg246879.pdf
17. IBM Corporation, Mining Your Own Business in Telecoms Using DBA Intelligent Miner for data,2001. http://www.redbooks.ibm.com/redboods/pdfs/sg24627.pdf
18. Kellen, V., Customer relationship Management Measurement Frameworks, 2002.
19. Kim,J., suh, E.& Hwang, H., A Model for Evaluating the Effectiveness of Customer relationship Management Using the Balanced Scorecard, Journal of Interacting Marketing, Volume 17,No 2,2003.
20. Kotler, P., Marketing Management: Analysis. Planning, Implementation and Control, 9th Edition, New Delhi, Prentice Hall of India, 1998.
21. Mikinsey Marketing Solutions (MKMS), Tactical CRM: Three Steps to Mining profits, Not Data. http://www.mckinsey. Com
22. Mckinsey Marketing Solutions, The New Era of Customer Loyalty Management. http://www. Mckinsey.com
23. Mckinsey Marketing Soultions, Unlock Hidden Potential your customer Relationship Management Investments. http://www.mckinsey.com
24. Mitchell, M., Machine Learning. The MCGraew-Hill Companies, 1997.
25. Parvatlyar, A and Seth, N.J, Customer Relationship Management emerging practice, process and discipline,
26. Pritscher L.& Feyen, H., Data Mining and Strategic Marketing in the Airline Industry.

http://www.iuc.ac.be/iteo/articles/pritscher1.pdf

27. Quain, William J., Sansbury, Michael & Quinn, Dennis. Revenue Enhancement: A simple Approach to Yield Management. Cornell Quarterly, April 1999, Vol. 40, No. 2. Cornell University.
28. Saarenvirta,G., Mining Customer data, A step by step look at a powerful clustering and segmentation methodology. 1998.
    http://www.db2mag.com/db area/archives/1998/q3/98fsaar.html
29. Schiffaman, G.L and Kanuk, L.L, Consumer Behavior, Prentice Hall, Inc, 4th Edition. 1991.
30. Two Crows Corporation, Introduction to Data Mining and Knowledge Discovery, 3rd Ed,1999.
    http://www.twocrows.com
31. Ulwich, T.& Jan A.Elsenhauer, A,J, The Natural Order of Segmentation: Aligning Company culture with its customers,
    http://www.diwings.ch/e/druck/Market Segmetation.pdf
32. Vriens, M., Market segmentation, Analytical developments and applications Guidelines, Technical Review series, March 2001.
33. Witten, H.I and Frank, E, Data Mining, Practical machine Learning Tools and Techniques with JAVA implementation,2001.
34. Xu, Y., Yen, C.D., Lin, B., & Chou, C.D. Adopting Customer Relationship Management Technology, Industrial Management and Data systems 102/8, PP 442-552,2002.
    www.emeraldinsight.com/
35. Yang Y.& Padmanabhan, B., Data Mining for customer Segmentation, A behavioral pattern-based Approach, Operations and information Management department, University of Pennsylvania, Jan 2004.
36. Ziberna, A. & Zabkar, V. Application of End user Segmentation Methods, 2003.
    http://mrvar.fdv.uni-lj.si/pub/mz/mz19/zabkar.pdf

English Language Rule # 1:
There is a 3.13901 percent chance that Cluster Index will be 1, a 36.2539 percent chance that Cluster Index will be 2, a 8.31321 percent chance that Cluster Index will be 3 and a 52.2939 percent chance that Cluster Index will be 4.

English Language Rule # 2:
If NTENURE is equal to 14-7 then there is a 0 percent chance that Cluster Index will be 1, a 0 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 100 percent chance that Cluster Index will be 4.

English Language Rule # 3:
If NTENURE is equal to 15-20 then there is a 3.6036 percent chance that Cluster Index will be 1, a 0 percent chance that Cluster Index will be 2, a 20.1577 percent chance that Cluster Index will be 3 and a 76.2387 percent chance that Cluster Index will be 4.

English Language Rule # 4:
If NTENURE is equal to <7 then there is a 60 percent chance that Cluster Index will be 1, a 0 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 40 percent chance that Cluster Index will be 4.

English Language Rule # 5:
If NTENURE is equal to >20 then there is a 4.6473 percent chance that Cluster Index will be 1, a 87.2199 percent chance that Cluster Index will be 2, a 5.14523 percent chance that Cluster Index will be 3 and a 2.98755 percent chance that Cluster Index will be 4.

English Language Rule # 6:
If NTENURE is equal to 14-7 and REV_SHIP_ RANGE is equal to <3197 then there is a 0 percent chance that Cluster Index will be 1, a 0 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 100 percent chance that Cluster Index will be 4.

English Language Rule # 7:
If NTENURE is equal to 14-7 and REV_SHIP_ RANGE is equal to <3197 and REV_TEN_RANGE is equal to <=267 then there is a 0 percent chance that Cluster Index will be 1, a 0 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 100 percent chance that Cluster Index will be 4.

English Language Rule # 8:
If NTENURE is equal to 14-7 and REV_SHIP_ RANGE is equal to <3197 and REV_TEN_RANGE is equal to <=267 and REVENUE_ RANGE is equal to <6913.99 then there is a 0 percent chance that Cluster Index will be 1, a 0 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 100 percent chance that Cluster Index will be 4.

English Language Rule # 9:
If NTENURE is equal to 14-7 and REV_SHIP_ RANGE is equal to <3197 and REV_TEN_RANGE is equal to <=267 and REVENUE_ RANGE is equal to <6913.99 and SHIP_TEN_RANGE is equal to 0.048-0.15 then there is a 0 percent chance that Cluster Index will be 1, a 0 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 100 percent chance that Cluster Index will be 4.

English Language Rule # 10:
If NTENURE is equal to 14-7 and REV_SHIP_ RANGE is equal to <3197 and REV_TEN_RANGE is equal to <=267 and REVENUE_ RANGE is equal to <6913.99 and SHIP_TEN_RANGE is equal to 0.048-0.15 and SHIPMENT_RANGE is equal to <2 then there is a 0 percent chance that Cluster Index will be 1, a 0 percent chance that Cluster Index will be 2, a

0 percent chance that Cluster Index will be 3 and a 100 percent chance that Cluster Index will be 4.

English Language Rule # 11:
If NTENURE is equal to 15-20 and REV_SHIP_ RANGE is equal to 3197-10658.99 then there is a 0 percent chance that Cluster Index will be 1, a 0 percent chance that Cluster Index will be 2, a 100 percent chance that Cluster Index will be 3 and a 0 percent chance that Cluster Index will be 4.

English Language Rule # 12:
If NTENURE is equal to 15-20 and REV_SHIP_ RANGE is equal to <3197 then there is a 4.5134 percent chance that Cluster Index will be 1, a 0 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 95.4866 percent chance that Cluster Index will be 4.

English Language Rule # 13:
If NTENURE is equal to 15-20 and REV_SHIP_ RANGE is equal to 3197-10658.99 and REV_TEN_RANGE is equal to <=267 then there is a 0 percent chance that Cluster Index will be 1, a 0 percent chance that Cluster Index will be 2, a 100 percent chance that Cluster Index will be 3 and a 0 percent chance that Cluster Index will be 4.

English Language Rule # 14:
If NTENURE is equal to 15-20 and REV_SHIP_ RANGE is equal to 3197-10658.99 and REV_TEN_RANGE is equal to <=267 and REVENUE_ RANGE is equal to <6913.99 then there is a 0 percent chance that Cluster Index will be 1, a 0 percent chance that Cluster Index will be 2, a 100 percent chance that Cluster Index will be 3 and a 0 percent chance that Cluster Index will be 4.

English Language Rule # 15:
If NTENURE is equal to 15-20 and REV_SHIP_ RANGE is equal to 3197-10658.99 and REV_TEN_RANGE is equal to <=267 and REVENUE_ RANGE is equal to <6913.99 and SHIP_TEN_RANGE is equal to 0.048-0.15 then there is a 0 percent chance that Cluster Index will be 1, a 0 percent chance that Cluster Index will be 2, a 100 percent chance that Cluster Index will be 3 and a 0 percent chance that Cluster Index will be 4.

English Language Rule # 16:
If NTENURE is equal to 15-20 and REV_SHIP_ RANGE is equal to 3197-10658.99 and REV_TEN_RANGE is equal to <=267 and REVENUE_ RANGE is equal to <6913.99 and SHIP_TEN_RANGE is equal to 0.048-0.15 and SHIPMENT_RANGE is equal to <2 then there is a 0 percent chance that Cluster Index will be 1, a 0 percent chance that Cluster Index will be 2, a 100 percent chance that Cluster Index will be 3 and a 0 percent chance that Cluster Index will be 4.

English Language Rule # 17:
If NTENURE is equal to 15-20 and REV_SHIP_ RANGE is equal to <3197 and REV_TEN_RANGE is equal to <=267 then there is a 4.5134 percent chance that Cluster Index will be 1, a 0 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 95.4866 percent chance that Cluster Index will be 4.

English Language Rule # 18:
If NTENURE is equal to 15-20 and REV_SHIP_ RANGE is equal to <3197 and REV_TEN_RANGE is equal to <=267 and REVENUE_ RANGE is equal to <6913.99 then there is a 4.5134 percent chance that Cluster Index will be 1, a 0 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 95.4866 percent chance that Cluster Index will be 4.

English Language Rule # 19:

If NTENURE is equal to 15-20 and REV_SHIP_ RANGE is equal to <3197 and REV_TEN_RANGE is equal to <=267 and REVENUE_ RANGE is equal to <6913.99 and SHIP_TEN_RANGE is equal to 0.048-0.15 then there is a 0 percent chance that Cluster Index will be 1, a 0 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 100 percent chance that Cluster Index will be 4.

English Language Rule # 20:

If NTENURE is equal to 15-20 and REV_SHIP_ RANGE is equal to <3197 and REV_TEN_RANGE is equal to <=267 and REVENUE_ RANGE is equal to <6913.99 and SHIP_TEN_RANGE is equal to >0.32 then there is a 100 percent chance that Cluster Index will be 1, a 0 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 0 percent chance that Cluster Index will be 4.

English Language Rule # 21:

If NTENURE is equal to 15-20 and REV_SHIP_ RANGE is equal to <3197 and REV_TEN_RANGE is equal to <=267 and REVENUE_ RANGE is equal to <6913.99 and SHIP_TEN_RANGE is equal to O.16-0.32 then there is a 100 percent chance that Cluster Index will be 1, a 0 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 0 percent chance that Cluster Index will be 4.

English Language Rule # 22:

If NTENURE is equal to 15-20 and REV_SHIP_ RANGE is equal to <3197 and REV_TEN_RANGE is equal to <=267 and REVENUE_ RANGE is equal to <6913.99 and SHIP_TEN_RANGE is equal to 0.048-0.15 and SHIPMENT_RANGE is equal to 2.0-3.0 then there is a 0 percent chance that Cluster Index will be 1, a 0 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 100 percent chance that Cluster Index will be 4.

English Language Rule # 23:

If NTENURE is equal to 15-20 and REV_SHIP_ RANGE is equal to <3197 and REV_TEN_RANGE is equal to <=267 and REVENUE_ RANGE is equal to <6913.99 and SHIP_TEN_RANGE is equal to 0.048-0.15 and SHIPMENT_RANGE is equal to <2 then there is a 0 percent chance that Cluster Index will be 1, a 0 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 100 percent chance that Cluster Index will be 4.

English Language Rule # 24:

If NTENURE is equal to 15-20 and REV_SHIP_ RANGE is equal to <3197 and REV_TEN_RANGE is equal to <=267 and REVENUE_ RANGE is equal to <6913.99 and SHIP_TEN_RANGE is equal to >0.32 and SHIPMENT_RANGE is equal to 4.0-7.0 then there is a 100 percent chance that Cluster Index will be 1, a 0 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 0 percent chance that Cluster Index will be 4.

English Language Rule # 25:

If NTENURE is equal to 15-20 and REV_SHIP_ RANGE is equal to <3197 and REV_TEN_RANGE is equal to <=267 and REVENUE_ RANGE is equal to <6913.99 and SHIP_TEN_RANGE is equal to >0.32 and SHIPMENT_RANGE is equal to >=8 then there is a 100 percent chance that Cluster Index will be 1, a 0 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 0 percent chance that Cluster Index will be 4.

English Language Rule # 26:

If NTENURE is equal to 15-20 and REV_SHIP_ RANGE is equal to <3197 and REV_TEN_RANGE is equal to <=267 and REVENUE_ RANGE is equal to <6913.99 and SHIP_TEN_RANGE is equal to O.16-0.32 and SHIPMENT_RANGE is equal to 4.0-7.0 then there is a 100 percent chance that Cluster Index will be 1, a 0 percent chance that Cluster Index will be

2, a 0 percent chance that Cluster Index will be 3 and a 0 percent chance that Cluster Index will be 4.

English Language Rule # 27:
If NTENURE is equal to <7 and REV_TEN_RANGE is equal to <=267 then there is a 60 percent chance that Cluster Index will be 1, a 0 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 40 percent chance that Cluster Index will be 4.

English Language Rule # 28:
If NTENURE is equal to <7 and REV_TEN_RANGE is equal to <=267 and REVENUE_ RANGE is equal to <6913.99 then there is a 60 percent chance that Cluster Index will be 1, a 0 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 40 percent chance that Cluster Index will be 4.

English Language Rule # 29:
If NTENURE is equal to <7 and REV_TEN_RANGE is equal to <=267 and REVENUE_ RANGE is equal to <6913.99 and SHIP_TEN_RANGE is equal to >0.32 then there is a 0 percent chance that Cluster Index will be 1, a 0 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 100 percent chance that Cluster Index will be 4.

English Language Rule # 30:
If NTENURE is equal to <7 and REV_TEN_RANGE is equal to <=267 and REVENUE_ RANGE is equal to <6913.99 and SHIP_TEN_RANGE is equal to O.16-0.32 then there is a 100 percent chance that Cluster Index will be 1, a 0 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 0 percent chance that Cluster Index will be 4.

English Language Rule # 31:
If NTENURE is equal to <7 and REV_TEN_RANGE is equal to <=267 and REVENUE_ RANGE is equal to <6913.99 and SHIP_TEN_RANGE is equal to >0.32 and SHIPMENT_RANGE is equal to <2 then there is a 0 percent chance that Cluster Index will be 1, a 0 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 100 percent chance that Cluster Index will be 4.

English Language Rule # 32:
If NTENURE is equal to <7 and REV_TEN_RANGE is equal to <=267 and REVENUE_ RANGE is equal to <6913.99 and SHIP_TEN_RANGE is equal to O.16-0.32 and SHIPMENT_RANGE is equal to <2 then there is a 100 percent chance that Cluster Index will be 1, a 0 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 0 percent chance that Cluster Index will be 4.

English Language Rule # 33:
If NTENURE is equal to >20 and REV_TEN_RANGE is equal to 268-892.99 then there is a 3.40909 percent chance that Cluster Index will be 1, a 96.5909 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 0 percent chance that Cluster Index will be 4.

English Language Rule # 34:
If NTENURE is equal to >20 and REV_TEN_RANGE is equal to 893-1785.99 then there is a 0 percent chance that Cluster Index will be 1, a 100 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 0 percent chance that Cluster Index will be 4.

English Language Rule # 35:
If NTENURE is equal to >20 and REV_TEN_RANGE is equal to <=267 then there is a 26.3158 percent chance that Cluster Index will be 1, a 0 percent chance that Cluster Index will be

2, a 46.6165 percent chance that Cluster Index will be 3 and a 27.0677 percent chance that Cluster Index will be 4.

English Language Rule # 36:
If NTENURE is equal to >20 and REV_TEN_RANGE is equal to >1786 then there is a 0 percent chance that Cluster Index will be 1, a 100 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 0 percent chance that Cluster Index will be 4.

English Language Rule # 37:
If NTENURE is equal to >20 and REV_TEN_RANGE is equal to 268-892.99 and REVENUE_ RANGE is equal to 23047-46093.99 then there is a 0 percent chance that Cluster Index will be 1, a 100 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 0 percent chance that Cluster Index will be 4.

English Language Rule # 38:
If NTENURE is equal to >20 and REV_TEN_RANGE is equal to 268-892.99 and REVENUE_ RANGE is equal to 6914-23046.99 then there is a 0 percent chance that Cluster Index will be 1, a 100 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 0 percent chance that Cluster Index will be 4.

English Language Rule # 39:
If NTENURE is equal to >20 and REV_TEN_RANGE is equal to 268-892.99 and REVENUE_ RANGE is equal to <6913.99 then there is a 21.875 percent chance that Cluster Index will be 1, a 78.125 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 0 percent chance that Cluster Index will be 4.

English Language Rule # 40:
If NTENURE is equal to >20 and REV_TEN_RANGE is equal to 893-1785.99 and REVENUE_ RANGE is equal to 23047-46093.99 then there is a 0 percent chance that Cluster Index will be 1, a 100 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 0 percent chance that Cluster Index will be 4.

English Language Rule # 41:
If NTENURE is equal to >20 and REV_TEN_RANGE is equal to 893-1785.99 and REVENUE_ RANGE is equal to >46094 then there is a 0 percent chance that Cluster Index will be 1, a 100 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 0 percent chance that Cluster Index will be 4.

English Language Rule # 42:
If NTENURE is equal to >20 and REV_TEN_RANGE is equal to 893-1785.99 and REVENUE_ RANGE is equal to 23047-46093.99 and SHIP_TEN_RANGE is equal to 0.048-0.15 then there is a 0 percent chance that Cluster Index will be 1, a 100 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 0 percent chance that Cluster Index will be 4.

English Language Rule # 43:
If NTENURE is equal to >20 and REV_TEN_RANGE is equal to 893-1785.99 and REVENUE_ RANGE is equal to 23047-46093.99 and SHIP_TEN_RANGE is equal to <0.048 then there is a 0 percent chance that Cluster Index will be 1, a 100 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 0 percent chance that Cluster Index will be 4.

English Language Rule # 44:
If NTENURE is equal to >20 and REV_TEN_RANGE is equal to 893-1785.99 and REVENUE_ RANGE is equal to 23047-46093.99 and SHIP_TEN_RANGE is equal to >0.32 then there is a 0 percent chance that Cluster Index will be 1, a 100 percent chance that Cluster Index

will be 2, a 0 percent chance that Cluster Index will be 3 and a 0 percent chance that Cluster Index will be 4.

English Language Rule # 45:
If NTENURE is equal to >20 and REV_TEN_RANGE is equal to 893-1785.99 and REVENUE_ RANGE is equal to 23047-46093.99 and SHIP_TEN_RANGE is equal to O.16-0.32 then there is a 0 percent chance that Cluster Index will be 1, a 100 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 0 percent chance that Cluster Index will be 4.

English Language Rule # 46:
If NTENURE is equal to >20 and REV_TEN_RANGE is equal to <=267 and REVENUE_ RANGE is equal to <6913.99 then there is a 26.3158 percent chance that Cluster Index will be 1, a 0 percent chance that Cluster Index will be 2, a 46.6165 percent chance that Cluster Index will be 3 and a 27.0677 percent chance that Cluster Index will be 4.

English Language Rule # 47:
If NTENURE is equal to >20 and REV_TEN_RANGE is equal to >1786 and REVENUE_ RANGE is equal to >46094 then there is a 0 percent chance that Cluster Index will be 1, a 100 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 0 percent chance that Cluster Index will be 4.

English Language Rule # 48:
If NTENURE is equal to >20 and REV_TEN_RANGE is equal to >1786 and REVENUE_ RANGE is equal to >46094 and SHIP_TEN_RANGE is equal to 0.048-0.15 then there is a 0 percent chance that Cluster Index will be 1, a 100 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 0 percent chance that Cluster Index will be 4.

English Language Rule # 49:
If NTENURE is equal to >20 and REV_TEN_RANGE is equal to >1786 and REVENUE_ RANGE is equal to >46094 and SHIP_TEN_RANGE is equal to <0.048 then there is a 0 percent chance that Cluster Index will be 1, a 100 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 0 percent chance that Cluster Index will be 4.

English Language Rule # 50:
If NTENURE is equal to >20 and REV_TEN_RANGE is equal to >1786 and REVENUE_ RANGE is equal to >46094 and SHIP_TEN_RANGE is equal to >0.32 then there is a 0 percent chance that Cluster Index will be 1, a 100 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 0 percent chance that Cluster Index will be 4.

English Language Rule # 51:
If NTENURE is equal to >20 and REV_TEN_RANGE is equal to >1786 and REVENUE_ RANGE is equal to >46094 and SHIP_TEN_RANGE is equal to O.16-0.32 then there is a 0 percent chance that Cluster Index will be 1, a 100 percent chance that Cluster Index will be 2, a 0 percent chance that Cluster Index will be 3 and a 0 percent chance that Cluster Index will be 4.

DECLARATION

This thesis is my original work, hasn't been presented for a degree in any other university and that all sources of materials used for the thesis have been duly acknowledged.

Kumneger Fikre Wolde

July 2006

The thesis has been submitted for examination with my approval as university advisor.

Prof. Bandaru Rama Krishna Rao

July 2006