ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

FACULTY OF INFORMATICS

DEPARTMENT OF INFORMATION SCIENCE

APPLYING DATA MINING TOOLS AND TECHNIQUES FOR EFFECTIVE CUSTOMER RELATIONSHIP MANAGEMENT OF ETHIOPIA HOTEL

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

BY

HIWOT AMARE GESSESSE

June, 2005

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

APPLYING DATA MINING TOOLS AND TECHNIQUES FOR EFFECTIVE CUSTOMER RELLATIONSHIP MANAGEMENT (CRM) OF ETHIOPIA HOTEL

BY
HIWOT AMARE

Name and Signature of Members of the Examining Board

Prof. B.R. Krishna Rao, Chairman, Examining Board

Dr. B.L. Desai, Advisor

Dr. Kumudha Raimond, Examiner

Chairman, Faculty	Signature	Date
Chairman, Graduate Council	Signature	Date

ACKNOWLEDGMENTS

First and foremost, I should thank God who has given me all the opportunities I need. *"There is nothing possible without you!"*

I would like to thank my advisors Dr. B. L. Desai and Nigusse Tadesse for their constructive comments and support for my work. Especially thanks to Dr. B. L. Desai who has always been there when I needed his help.

Thanks to Dr. Gashaw Kebede who has given me important advices and comment on my work.

I would like to thank the entire stuffs of Ethiopia hotel especially working at the reception department of Ethiopia Hotel, and the general manager.

I wish to thank also my parents Mom, Dad and my sisters and brothers who encouraged me and gave me strength to finish my study.

And my special thanks goes to my Husband, Daniel Elias who is always with me and will ever be.

Thanks all who advised, supported and encouraged and helped me to finish the study.

Table of Content

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

These days there are a lot of hotels built with high standard and better quality of services in Ethiopia. Ethiopia hotel is one of the historical hotels in Ethiopia. To compete with those hotels having higher standards and keep its own customer loyal and profitable, Ethiopia hotel should understand its own customer's data and make use of the information understood. And data mining is a powerful tool for extracting this useful information, particularly, for supporting good CRM by providing important knowledge about the customers. This study aimed at applying data mining technology on Ethiopia hotel's customer data for identifying valuable customer segments and their behavior to support for better CRM (Customer Relationship Management) in the hotel.

In this study, to prepare the data, data preparation tasks including cleaning missing value, smoothing outliers, and transformation and aggregation were made. By using the data mining tool called Knowledge Studio, clustering and classification models were built. The clustering model was used to identify customer segments. From this model five defined and meaningful clusters (customer segments) were identified.

The classification model was built to generate rules used to develop a simple customer classification prototype that can help to classify new customer records to one of customer segments with the description of each customer clusters. The findings of this study would encourage business organizations to work on the application of data mining technology for better customer relationship management, and as a result gain a competitive advantage.

CHAPTER ONE

INTRODUCTION

1.1. Background

The economic figures show that tourism has grown to be an activity of worldwide importance and significance. For countries like Ethiopia, tourism is the largest commodity in international trade (Macintosh, et. al, 1995). Ethiopia, a country endowed with attractive sites that include historical places, national parks and different cultures, is currently working on expanding its tourism industry. In order for this to be realized, the capacity and quality of services of different sectors like the hotel industry, airways and travel agencies have to be significantly expanded with especial emphasis on the hotel industry. Such expansions can lead to healthy competition in the industry and result in improved quality of services and competitive rates.

With the increasing competition, the key to a hotel's survival is its ability to cater for the changing preferences and life styles of ever demanding customers. These preferences may include greater access to amenities, comfortable rooms, fast check-in/check-out and courteous customer services, and reasonable price. Full understanding of such preferences, however, cannot be translated into a competitive advantage unless the industry develops detailed profiles of customers to target valued customers. That is why customer profiling becomes the corner stone of an effective customer relationship management (Min *et al*, 2002).

Hotels, like any business organization, have to take into consideration their customers' behavior so as to be competent in the market. Knowing customers' behavior will help hotels

target their profitable customers and retain loyal customers. This is called customer relationship management (CRM).

CRM is referred to as the business practice that is intended to improve service delivery, build social bonds with customers, and secure customer loyalty by developing a long-term, mutually beneficial relationship with valued customers selected from a pool of customers. It focuses on valued customers who repeatedly purchase a great deal of services and remain committed to their particular hotels (Min, et. al, 2002).

Lejeune (2001) noted that relationship building and customer-oriented management are key factors to which companies' success or failure is closely linked. Customer management requires the collection of a significant amount of information and set up of procedures for interpreting this information.

Business data is increasingly being seen as valuable commodity by itself and not just a by-product of day-to-day transaction processes. An organization's business data represents the current state of an organization's business. When it is combined with historical business data, it can tell an organization where it is and where it is heading. Since business decisions are being made at alarmingly fast rate, managers and executives need information on which to base these decisions (Wobishet, 2002). Therefore, in a highly competitive and information intensive environment of the hotel industry the speed of a decision may be as critical as the decision itself.

The amount of data collected or purchased and warehoused continues to grow at an enormous rate even though stores are already vast. The primary challenge is how to make the database a competitive business advantage by converting enormous stores of seemingly meaningless facts into useful information. That is why a new technology called data mining

has emerged to solve the challenge of converting vast amount of data into an important knowledge (Lejeune, 2001).

Data mining is the search for relationship and global patterns that exists in large databases but are "hidden" among the vast amount of data. Data mining when applied to business data is about understanding customers and prospects, understanding products and markets, and understanding suppliers and partners by applying its tools and techniques on collection of customer data.

The application of data mining techniques to customer databases has significantly facilitated the area of marketing. The uses range from analysis of supermarket point of sale data to predicting who the audience will be for television programs. Other uses include studying the effects of changing the price or distribution channel on sales volume and market share.

1.2. Statement of the problem and justification

Data mining tools and techniques can be applied in many areas including marketing, banking, insurance, and health care and medicine. Particularly it can be applied in CRM for prospecting new customer records, retaining existing customers, increasing customer value, and understanding customer behavior. Studies by different scholars were done on insurance, marketing and medicine. It is also the understanding of the researcher that data mining tools and techniques can be employed for better CRM in the business world.

Ethiopia, being a country of tourism potential can attract a lot of tourists, politicians, and different professionals. This initiates the development of the hotel industry with international standards, which will lead to stiff competition among hotels. To be competent and profitable

in the market, it is important to build a strong CRM, since the customers are the key factors in making profit.

In order to be competitive and profitable in the market, hotels should support CRM by applying data mining tools and techniques so that customers can be segmented based on the profitability, motivation, behavior or others factors as a result each segment can be treated accordingly by applying different marketing strategies. This research particularly aimed to apply clustering and classification of data mining methods on hotel customers' data to attain effective CRM by identifying different customer segments and their respective behavior and how they can be interpreted in terms of marketing concept. This will enable marketing managers to know who are profitable, loyal, not profitable, and not loyal. In general they can have important knowledge about their customers; but the decision of how to retain valuable customers is left to the marketing managers.

This research took Ethiopia hotel as a case to apply the data mining technology. There are lots of raw data accumulated from the beginning of its operation, but no attempt has been made to make use of the data for any marketing research including customer segmentation and hence no clear-cut group/class based marketing orientation. From this customer data, very important and hidden knowledge can be discovered by applying data mining tools and techniques. This will help the hotel to improve its marketing, sales, and customer support operations and gain a competitive advantage through better understanding of its customers. Though there are researches conducted at the department of information science on CRM, all of them are on airlines, banks and insurance. Therefore this research attempted to develop a customer segmentation model by using the data mining techniques, clustering and decision

tree induction learning methods for analysis on Ethiopia hotel customers' data for better CRM.

1.3. Objective of the Study

1.3.1. General objectives

The general objective of this research was to build customer segmentation and classification model that can help for better CRM through implementing different marketing strategies.

1.3.2. Specific objectives

The specific objectives were:

- To review literatures on the application of data mining in CRM, the techniques used, and the concepts of CRM
- To explore the possible significance of CRM in hotel industry
- To identify important parameters for customer segmentation and sort out customer segments on the basis of shared or common attributes.
- To study Ethiopia hotel's customer segments obtained
- To know which domestic and foreign customers are profitable (loyal if any) and how they behave
- To identify which parameter (variable) can most effectively classify new customer records to the identified customers
- To generate rules that will help to take decision on how to assign new customer records to identified customer segments

- To make use of the rules generated for developing a prototype for classifying new customer records to one of customer clusters

1.4. Methodology

In order to build good data mining models for a CRM, there are number of steps that one should follow. Accordingly, in the course of this investigation, the following steps were used which are data mining methodologies supported by Cross –Industry Standard Process for Data Mining (CRISP_DM).

1.4.1. Data collection and data understanding

The hotel has established a computerized system recently. However, due to shortage of skilled manpower who can work with the system, it was not able to proceed with this computerized system. They only worked with the system for two months. From this, two months data were collected and another two months data were also added from manual records.

Two different tables were obtained during data collection; guest registration table and resident summary. The guest registration table is about the demographic data of each customer like name, number of persons, room number, bill number, rooms occupied, nationality, profession, date of arrival, origin and destination of the guest, passport number, duration of stay, and purpose of visit. It contains 4,366 rows (1.2 mega bytes) of data, collected from manual data entry and 5,024 rows (2.24 mega bytes) from the computerized database initially. The resident summary table is about financial data of each customer like daily status of the transfer, presence or check out and payment of charge of each customers containing room number, bill number, room charge, value added tax, service charge, extra

revenue, payment, and total charge. It contains 4,096 rows (972 kilo bytes) of records from manual data entry and 6,614 rows (1.17 mega bytes) of records from computerized database initially.

1.4.2. Data preparation

Microsoft excel was used to organize and preprocess the data because of its ability to make different simple calculations and summarize the data in simple way.

To prepare the raw data collected into a form suitable for data mining analysis, data preparation activities like data selection, data cleaning, and data transformation and aggregation were done. The details of the processes can be seen in later part of the study (section 4.4).

1.4.3. Building and training the models

In order to build the model, data mining techniques such as clustering and decision tree techniques were used. Since clustering is unsupervised learning and helps to identify various customers with similar features, it was selected for building the clustering model. Among the clustering techniques, K-Means was selected; since it is most popular, simple and straightforward clustering technique. In addition, knowledge studio software (Data mining tool developed by Angoss software Corporation) supported the use of K-Means for most situations. But K-Means does not support when there are large amount of missing values.

In another part of the research, decision tree induction technique was applied to generate rules which will help to assign new customers records to the already identified segments. This technique was selected; since it is widely applicable for classification problems, easy to build, very flexible, easily understandable (interpretable), can identify important variables, and can handle both categorical and continuous variables (Gargano and Raggad, 1999).

1.4.4. Evaluating or testing the models

Since clustering is unsupervised learning, there is no need to classify the data into training and testing dataset. Hence, the total dataset was used to build the clustering model. Cluster runs by changing the number of variables, type of variables, and the parameter K (K set 6, 5, or 4) was made to reach to meaningful customer segments. The cluster run, resulting with better patterns discovered was selected. In addition, in the process of selecting good model, the support of domain experts' comments and suggestions who are in the field were vital and hence considered.

For the classification model, the data set containing the cluster index obtained from the clustering model was divided into two groups for evaluation purpose. Seventy percent of the dataset allocated for training and thirty percent of the dataset was used to evaluate the accuracy of the model over subsequent data and to evaluate the impact of pruning. The decision tree generated from knowledge studio was used to identify the most statistically significant attributes. Different experiments were made to determine the contribution of the first splitting attribute for better prediction of the dependent variable, cluster index (cluster segments).

1.5. Scope and Limitation

The output of this research is customer segmentation. And from this, important points like who are profitable domestic customers and international customers, and how they behave were known. It also gave knowledge about who are non profitable domestic and international, and how they behave. Ended by developing simple customer classification prototype. It did not include what should be the marketing strategies that should be

implemented by the hotel. Developing marketing strategies and implementing them is up to marketing department of the hotel.

Important fields like 'income' and 'age', even though they are very important, were not incorporated for the reason that there was no any registered information about these fields. The field 'income' may give information on the relationship between the income and the profitability of customers. If there was an 'age' attribute, which age groups of customers were mostly profitable would be known. In addition, there was one variable incorporated in the data taken for analysis called 'purpose of visit' which explains purpose of visiting the country or the hotel. It has four values, 'tourist', 'business' (either conference, or workshop, or meeting, or investment), 'other' (vacation, visiting relatives or friends), and 'transit' (waiting for another flight). The two values 'business' and 'other' should have been specifically stated to know the behavior of customers in a better way rather than using the general name 'business' and 'other'.

Only four months data were obtained because, converting raw data in to electronic format was time consuming along with the limited budget allocated.

Budget being a restrictive factor, the acquisition of appropriate data mining software from software developing company was not feasible

1.6. Significance of the study

This will help to show how data mining can support CRM by answering some important business questions. Business questions like who are the most loyal customers, which customers are more profitable, and which customers can be best satisfied according to the standard of the hotel can be answered. And different decisions on developing and implementing marketing strategies for retaining profitable customers, attracting new

customer and customers at risk of going to other hotels, and for designing ways to convert non-profitable customers to profitable ones can be taken. This vital information will help the hotel industry to stay competitive and to increase revenue.

It could serve as a basis for further research on customer segmentation and an academic exercise as well helping researcher to acquire knowledge on how to apply data mining tools and techniques in the real world.

CHAPTER TWO

DATA MINING

2.1 Introduction

This chapter reviewed about the general concepts of data mining, and related topics that should be known along with data mining. Specifically, data mining techniques used in this research, the relationship between data mining and CRM, different studies conducted on the application of K-Means and decision tree techniques for better CRM, and finally the application of data mining in hotel industry were discussed.

2.2 Historical Overview of Data mining

In order to know the world and explain natural phenomenon, people have been gathering and analyzing data. Investigating this data has brought different theories, observations, and approaches that could help understand and know the natural world and its laws. Without the aid of machines, people had been analyzing data and looking for patterns (Berry and Linoff, 1997). However, gradually, new technologies have begun to play a vital role to facilitate storage, analysis and processing of data. Specially, the advent of computer technology has revolutionized the way in which data are managed. These new methods of looking into data as well as the keen interest to learn from data have brought disciplines like that of data mining (Thearling, 2000).

Recently data mining has drawn the attention of intellectuals, business article writers and software developers. Although data mining is the evolution of fields with a long history such as statistics, artificial intelligence and machine learning, it evolved to become widely known in the 1990s (ANGOSS, 2003).

Much of the tools and techniques of statistics are adopted in the study of data mining. However, although statistics is very useful technique, it is not capable of addressing all data mining problems (Berry and Linoff, 1997). For instance, some problems may demand learning from experience and statistical methods could not address them. Moreover, statistics usually employs sample data (part of the population data thought to be representative) to build statistical models and this method can miss large body of information about the population (Thearling, 2000).

The second longest family line of data mining is artificial intelligence. This field of study is developed on the basis of heuristics in contrast to statistics. It is an attempt to apply "human-thought-like" approach to statistical problems. The application of artificial intelligence has become pervasive when computers began to provide useful power at affordable prices (Ibid).

The other field of study that contributed a lot to data mining is machine learning, which is more properly described as the hybrid of statistics and artificial intelligence (ANGOSS, 2003). Machine learning attempts to let computer programs learn about the data they study, such that programs make different decisions based on the qualities of the studied data, using statistics for fundamental concepts, and adding more advanced artificial intelligence heuristics and algorithms to achieve its goals (Carbone, 1997). This depicts that the application of machine learning in the study of large volume of data is a radical shift not only from statistics but also from artificial intelligence via merging both fields.

From the foregoing arguments it seems plausible to conclude that data mining, in many ways, is basically the adaptation of machine learning techniques to scientific and business

problems. Data mining is the union of historical and recent developments in statistics, artificial intelligence, and machine learning. The tools and techniques borrowed from these fields of studies used together to extract previously unknown patterns buried in large database. Data mining is becoming popular in science and business areas where there is large amount of data.

2.3 Meaning and Definition of Data Mining

Data mining has become useful over the past decade in business to gain more information, to have a better understanding of running a business, and to find new ways and ideas to extrapolate business to other markets.

According to Han and Kamber (2001) the major reason that data mining has attracted a great deal of attention in the information industry in recent years is the wide availability of data and the immediate need for turning such data into useful information and knowledge. Data mining tools perform data analysis and may uncover important data patterns, contributing greatly to business strategies, knowledge bases, and scientific and medical research.

Piatetsky-Shapiro (2000) defined data mining as:

> *Data mining or knowledge discovery in database (KDD) as it is also known, is nontrivial extraction of implicit, previously unknown, and potentially useful information from data. This encompasses a number of different technical approaches, such as clustering, data summarization, learning classification, finding dependencies networks, analyzing changes, and detecting anomalies.*

Data mining is the process of selection, exploration, and modeling of large quantities of data to discover regularities relations that are at first unknown with the aim of obtaining clear and useful results for the owner of the database (Giudici, 2003).

Generally, data mining technology has become a new paradigm for decision making, with applications ranging from electronic commerce to fraud detection, credit scoring, warranty management, and even auditing data before storing it in a database.

2.4 Data Mining (Knowledge Discovery) Processes

Data mining is more than just applying software, it is a process that involves a series of steps to preprocess the data prior to mining and post processing steps to evaluate and interpret the modeling results (Han and Kamber, 2001; Levin and Zahavi, 1999).

Like any other problem solving methods, it starts with defining the problem. From the definition of the problem what processes are needed to find the solution will be determined (Two crows Corporation, 1999; France, et al, 2002; Jembere, 2003).

Most authors (Han and Kamber, 2001;Zalane, 1999; Berry and Linoff 1997) agree that data cleaning and integration, data selection and transformation, data mining, and finally pattern evaluation and knowledge presentation, are the most common steps employed for the mining processes.

Incomplete, noisy, and inconsistent data are common properties of large real world databases and data warehouses. Incomplete data can occur for a number of reasons. Attribute of interest may not always be available, or other data may not be included simply because it was not considered important at the time of entry, or relevant data may not be recorded due to misunderstanding or because of equipment malfunctioning, or the recording of the history or modification to the data may have been overlooked.

There may also be noisy data due to faulty data collection instrument used, or human or computer errors at the time of data entry, or inconsistencies in naming conventions or data

code used. Therefore, the first step which is data cleaning cleans the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies. After cleaning has taken place, the multiple heterogeneous data sources should be combined and integrated. The issues considered at this point are Schema integration (how can equivalent real-world entities from multiple data sources be matched up?), avoiding redundancies (inconsistencies in naming an attribute or dimension), and avoiding duplication of tuples that should be considered at the time of integrating heterogeneous data sources.

The next important step in data mining process is transformation of the data. Levin and Zahavi (1999) stated "often, the predictive power of data resides in transformation of the data, rather than in the raw data itself." The data should be transformed or consolidated into forms appropriate for mining. There are different issues like smoothing, aggregation, generalization, normalization, and attribute construction, which should be considered in the transformation process (Han and Kamber, 2001).

The next step, which is the data mining step, is the crucial step in which clever techniques are applied to extract patterns which are potentially useful. This step invokes data mining models and tools to interrogate the data and convert it into knowledge for decision-making activities.

The process of building predictive models requires a well-defined training and validation protocol in order to insure the most accurate and robust prediction. A model is built when the cycle of training and testing is completed. Training and testing the data mining model requires the data to be split into at least two groups; one for model training and other for model testing. After the model is generated using the training database, it is used to predict

the test database, and resulting accuracy rate is a good estimate of how the model will perform on future databases that are similar to the training and test database.

Finally the model(s) built should be evaluated as see whether the data mining objectives and business objectives are met in order to make the result useful for decision-making. Levin and Zahavi (1999) argued that evaluation and interpretation of knowledge discovered by the modeling engine is essential for making sure the resulting model is any good, and to convert the model results into useful knowledge for decision making.

Upon the successful completion of the knowledge discovery this is the phase where the discovered knowledge is visually presented to the user. This step uses visualization techniques to help users understand and interpret the data mining results.

But there is no hard and fast rule to follow the above data mining steps one after the other. The details of building and training a model vary from technique to technique and hence there are no blue print procedures that should be followed all the time (Thearling, 2000). During the knowledge discovery process, the iterative process may go from one step to the next and for some reason may comeback to earlier steps. Again after the presentation of the discovered knowledge to the user, in order to enhance the evaluation measures or to further refine the mining output, new data can be selected or further transformed, or new data sources can be integrated. This iterative and dynamic approach can be used in order to get different, more appropriate results.

2.5 Data Mining and Data Warehousing

When beginning work on data mining problems, it is necessary to bring all the data together into a set of instances first. Integrating data from different sources usually presents many challenges. Different departments of an organization may use different styles of record

keeping, different time periods, different degrees of data aggregation, different primary keys, and will have different kinds of error. Thus the data must be assembled, integrated, and cleaned up. The idea of enterprise wide database integration is called Data warehousing (Witten and Frank, 2000).

A data warehouse is a semantically consistent data store that serves as a physical implementation of a decision support data model and stores the information on which an enterprise needs to make strategic decisions by integrating data from multiple heterogeneous data sources. It enables each user not only to share a common, widely distributed, diverse database but also to analytically explore, discover, and better comprehend fundamental trends and relationships using all of the available data quickly and correctly.

Metadata, information concerning data describing the data warehouse, are also an integral part of the system. The warehouse architecture must manage standard information delivery systems and data queries, interfaces with applications developments platforms and executives information systems, and *online analytical processing (OLAP),* in addition to advanced information technology data mining tools (Gargano and Raggad, 1999).

Even though a data warehouse is not the prerequisite for data mining and data analysis, data mining potential can be enhanced if the appropriate data has been collected and stored in a data warehouse. Because during the construction of a data warehouse, many data preprocessing steps has already taken place, it most likely does not need further cleaning in order to be mined. In addition, many of the problems of data consolidation and integration have already been addressed. But one could mine data from one or more operational or transactional databases by simply extracting it into a read-only database. This new database functions as a type of data mart (Two Crows Corporation, 1999)

2.6 Data Mining Techniques

2.6.1 Introduction

In practice, data mining can accomplish about six common tasks namely; classification, estimation, prediction, association, clustering, and description. However, as the investigation at hand is directly related with clustering (segmentation), and is mainly concerned about classification, more emphasis is given to clustering and classification. Because of this, a detailed discussion of this data mining activities is presented in section 2.5 of this chapter.

The main objective of this research is to segment the customers having similar characteristics based on some variables. Because segmentation involves identifying the groups of customers with the similar characteristics, the study focuses on the data mining technique, clustering and a classification learning method called decision tree. Saarenvirta (1998) notes that customer clustering and classification are two of the most important data mining methodologies used in marketing and CRM. Therefore in this study only clustering and decision tree induction (classification) techniques of data mining are discussed.

Clustering is one of the unsupervised learning method that will help to identify segments without prior knowledge about the number of customer segments and to which segment is each customer most likely belongs. Witten and Frank (2000) also supported this when they commented that there is no specified class, clustering is used to group items that seem to fall naturally together. So, by using clustering technique, different group of customer segments will be identified based on the available data. But what is remaining is how to group new customer records after all segments are identified? In order to solve the above problem data mining technique called classification can be used. Classification assumes that there is a set

of objects – characterized by some attributes or features which belong to different classes. The class label is a discrete (symbolic) value and is known for each object. The objective is to build classification models (sometimes called classifiers), which assign the correct class label to previously unseen and unlabeled objects. The mining tool automatically identifies the clusters, by studying the patterns in the training data.

Therefore once the clusters are generated, classification (like decision tree) can be used to generate rules defining how to assign new customer records to the identified segments.

2.6.2. Clustering Techniques

Before going to the process of clustering and its techniques, it is important to have a general understanding of what a cluster is. Clusters represent mixtures of multivariate normal populations. This approach has been adopted as a rationale and basis for the design of clustering algorithms by some researchers. Qin He (1999) describes clusters as continuous regions of a p-dimensional space containing a relatively high density of points, separated from other such regions by regions containing a relatively low density of points. And they should exhibit the properties of external isolation and internal cohesion. External isolation requires that entities in one cluster should be separated with entities in another cluster by fairly empty areas of space. Internal cohesion requires that entities within the same cluster should be similar to each other, at least within the local metric (Milligan, in Qin He, 1999; Han and Kambar, 2001).

Clustering is the processes of grouping a set of physical or abstract objects into classes of similar objects (Han and Kamber, 2001). Clustering method as a multivariate procedure that

starts with a data set containing information about a sample of entities and attempts to reorganize these entities into relatively homogeneous groups.

Clustering unlike classification and prediction is unsupervised learning where the algorithm is provided with just the data points and no labels, and the task is to find a suitable representation of the underlying distribution of the data (Berkhin, 2000).

The ultimate goal of clustering is to find groups that are very different from each other, and whose members are very similar to each other and finally the clusters obtained should be interpreted by a person knowledgeable in the particular business domain so as to make the result obtained practical (Two Crow Corporation, 1999).

Researchers have different views about the classification of clustering methods. But the major clustering techniques can be grouped in to hierarchical and non-hierarchical.

Hierarchical method creates hierarchical decomposition of the given set of data objects forming a dendrogram - a tree that splits the database recursively into smaller subsets. The dendrogram can be viewed in two ways: bottom-up or top-down. The bottom-up approach is also called agglomerative approach starting with each object forming a separate group. It successively merges the object or groups according to some measures like distance between the two centers of two groups and this is done until all of the groups are merged into one (the top most level of hierarchy), or until a termination condition holds. The top-down, also called the divisive approach, starts with all the objects in the same cluster. In each successive iteration, a cluster is split into smaller cluster according to some measures until eventually each object is in one cluster. Earlier hierarchical methods suffer from the use of over-simplified measure to split or merge the clusters (Han, kamber, 2001). They also added that

these methods suffer from the fact that once a step (merge or split) is done, it can never be undone. This is useful in that it leads to smaller computation costs by not worrying about a combinatorial number of different choices. The major problem of such techniques is that they cannot correct erroneous decisions.

In contrast to hierarchical methods, non-hierarchical clustering procedures do not involve the tree-like construction process. Rather, cluster centers are initially selected (formed) and each object is assigned in a cluster based on its proximity to the cluster center. The most well known non-hierarchical clustering algorithm is the K-Means algorithm (Jembere, 2003).

There are many different clustering methods but in the area of data mining two, namely, Kohonen and K-means algorithm are in wide usage. This is because the large class of hierarchical clustering methods require that distances between every pair of data records be stored (there are n*(n-1)/2 such pairs) and updated, which places a substantial demand on memory and resources for the large files common in data mining. Instead clustering is typically performed using the K-means algorithm or an unsupervised neural network method or Kohonen (by SPSS Inc., 2002). But for the case of this research, K-means is reviewed as it is used in the software for building cluster modeling.

2.6.2.1. The K-Means Algorithm

The "K" in K-Means is derived from the fact that for a specific run the analyst chooses (guesses) the number of clusters (K) to be fit. The "means" portion of the name refers to the fact that the mean (or the centroid) of observations in a cluster represents the cluster.

Since the analyst must pick a specific number of clusters to run, typically several attempts are made, each with a different number of clusters, and the results should be evaluated. Since the number of clusters is chosen in advance and is usually small relative to the number of

observations, the K-means method runs quickly. This is because if seven clusters are requested, the program needs only to track the seven clusters. In hierarchical clustering the distance between every pair of observations must be evaluated and intercluster distances recomputed at each cluster step (which is a fairly intensive computational task). For this reason K-means is a popular method for data mining.

A brief description of the K-means method follows. If no starting values for the K cluster means are provided, the data file is searched for K well spaced (using distances based on the set of cluster variables) observations, and these are used as the original cluster centers. The data file is then reread with each observation assigned to the nearest cluster. At completion every record is assigned to some cluster, and the cluster means (centroids) are updated due to the additional observations (optionally the updating can be done as each observation is assigned to a cluster). At least one additional data pass (you can control the number of iterations) is made to check that each observation is still closest to the centroid of its own cluster (recall the cluster centers can move when they are updated, based on addition or deletion of members), and if not, the observation is assigned to the new nearest cluster. Additional data passes are usually made until the clusters become stable (by SPSS Inc., 2002).

As described by Bishop (1995), the process begins by assigning the points at random to K sets and then computing the mean vectors of the points in each set. The algorithm assigns each of the points to the cluster to whose center it is closest in *Euclidean* distance. Next, each point is re-assigned to a new set according to which is the nearest mean vector. The means of the sets are then recomputed. This procedure is repeated until there is no further change in

the grouping of the data points or the objective criterion function E does not change (or minimum), where E is as follows:

$$E = \Sigma_{i=1}^{k} \Sigma_{p \in c_i} [p - \mathbf{m}_i]^2$$

Where E is the sum of square-error for all objects in the database, p is the point in space representing a given objects, and $\mathbf{m}_i$ is the mean of cluster c_i.

In order to form clusters, each record from a database is mapped to a point in 'record space.' The number of dimensions contained in the space corresponds to the number of fields in the records. The value of each field can be geometrically interpreted as a distance from the origin along the corresponding axis of the space. In addition, to ensure the usefulness of this interpretation, the fields must all be converted into numbers and the numbers must be normalized so that a change in one dimension is comparable to a change in another.

2.6.3. Decision Tree

A decision tree is a flow-chart-like structure consisting internal node, branch and leaf node. Each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf node represents classes or class distributions where the top most node in a tree is the root node (Han and Kamber, 2001). Decision tree classify instances by sorting them down the tree from the root to some leaf node, which provide the classification of the instances. An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, then moving down the tree branch corresponding to the value of the attribute. The figure below illustrates a simple decision tree.

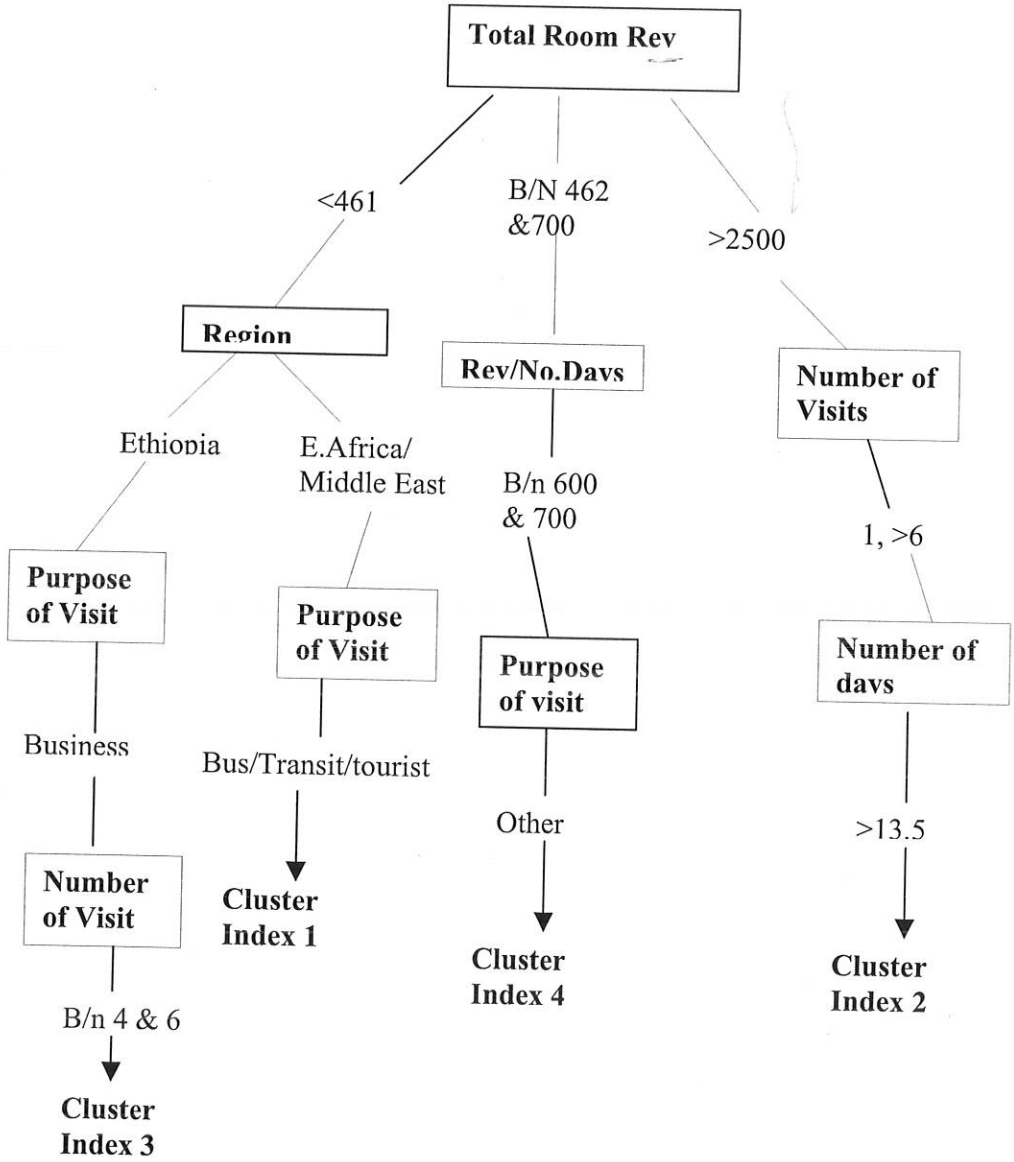


Figure 1:Sample decision tree

Constructing a decision tree is expressed as follows; first an attribute at the most node will be selected and made one branch value for each possible value. Then the dataset become splited into subsets, one for every value of the attribute. This process can be repeated recursively for each branch 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.

But the important thing in constructing decision tree is deciding the best attribute classifier among the attributes in the data set. Even though there are many measures of attribute classifier, the most commonly used one is information gain. It measures how well a given attribute separates the training data according to their classification.

Decision tree learning, according to Michell (1997) is a method for approximating discrete valued target functions, in which a decision tree represents the learned function. But it doesn't mean that continuous valued attributed can't be represented by decision trees. This can be accompanied by dynamically defining new discrete-valued attributed that partition the continuous attribute value into a discrete set of interval.

Most algorithms that have been developed for learning decision trees are variations on a core algorithm that employs a top-down greedy search through the space of possible decision trees. These methods are typically *greedy* in that, while searching through attribute space, they always make what looks to be the best choice at the time. Their strategy is to make a locally optimal choice in the hope that this will lead to a globally optimal solution. Such greedy methods are effective in practice and may come close to estimating an optimal solution (Han and Kamber, 2001; Mitchell, 1997).

2.6.3.1. Decision tree pruning

The decision when to declare a node terminal or to continue splitting is deemed critical for the construction of good decision trees. These pruning methods aim to simplify those decision trees that overfitted the data. A hypothesis overfits the training examples if some other hypothesis that fits the training less well actually performs better over the entire

distribution (Mitchell, 1997). Overfitting is a significant practical difficulty for decision tree learning and many other learning methods.

There are several approaches to avoid overfitting in decision tree learning. These can be grouped into two classes:

- Approaches that stop growing the tree earlier, before it reaches the point where it perfectly classifies the training data,
- Approaches that allow the tree to over-fit the data, and then post prune the tree.

Because the first approach has the difficulty of estimating precisely when to stop growing the tree, the second approach of post pruning overfit trees has been found to be successful. The most common criterion used to determine the correct final tree size is using a separate set of examples, distinct from the training examples, to evaluate the utility of post pruning.

2.7. Data Mining Applications

A wide range of organizations have developed successful applications of data mining because of the substantial contribution it can make.

Many organizations are using data mining to help manage all phases of the customer life cycle, including acquiring new customer, increasing revenue from existing customers, and retaining good customers. By determining characteristics of good customers (profiling), a company can target prospects with similar characteristics. By profiling customers who have bought a particular product it can focus attention on similar customers who have not bought that product (cross-selling). By profiling customers who have left, a company can act to retain customers who are at risk for leaving (reducing churn or attrition), because it is usually

far less expensive to retain a customer than acquire a new one (Two Crows Corporation, 1999).

Data mining can be applied in many companies including Credit Card, insurance, Transportation, medical companies, and airline and hotel industry.

Telecommunications and credit card companies are two of the leaders in applying data mining to detect fraudulent use of their services. Brause et al (1999) applied the data mining techniques, association rule and neural network for credit card fraud detection on one of credit card companies in Germany and obtained higher level of fraud detecting diagnostic rules that can help to develop an online learning diagnostic system.

A diversified transportation company with a vast direct sales force can apply data mining to identify the best prospects for its services. Using data mining to analyze its own customer experience, this company can build a unique segmentation identifying the attributes of high-value prospects.

Data mining can also be used in medical organizations to predict the effectiveness of surgical procedures and medical test or medications (Two Crows Corporation, 1999). Furthermore, hospitals use data mining to analyze consumer behavior and use the information gained to expand market opportunities. For instance, Sinai health system in Chicago used data mining techniques to answer why patients who have received a complete prenatal care didn't deliver at the hospital. Based on the results obtained the health system understood the marketing implications and prepared to improve marketing strategies that can help to attain these patients lost at their delivery time (Rafalski, 2002). The hotel and airline

industry can apply data mining for customer segmentation and customer profiling. So many organizations are applying data mining to solve customer related problems and gain important knowledge from the customers' data.

2.8. Data Mining and CRM

Businesses gather enormous amounts of data in their day-to-day operations. Every interaction with a customer generates data, and the amount of data gathered is rising exponentially. In order to create more successful personalized systems and build more accurate consumer behavior models, firms need to understand their customers better. This includes understanding customers' preferences through collecting more information and, customers' behavior through analyzing their transaction data.

Data mining is used to change this collection of customers' data into useful information and knowledge. As indicated earlier, by using data mining, you can get the customer segmentation model to identify the groups of customers with the similar characteristics which will help you to understand your customers. Eventually it will be used as the foundation of customer analysis and marketing strategy development.

Knowing which customers are potential is a starting point to understand the rapidly changing market environment in the hotel industry. Grouping similar customers together based on their characteristics is essential for the customer understanding and the target marketing for the specific group of customers.

Customer segmentation can be done based on many different criteria. These criteria could be as simple as age, gender, geography or the combination of these variables.

Data mining technology comes in when this criteria becomes more complicated. Decision on which criteria to use is solely depending on what is the objective of your customer segmentation and how to utilize it (Baragoin et.al, 2001).

In data mining, segmentation means clustering (Baragoin, et.al, 2001). The objective of the segmentation produced by data mining is to discover the different groups of customers suggested using the data hold about the customers, rather than by making some judgments about what the most important characteristics are. In other words, it suggests that what the business rules should be, rather than what you think they may be.

When developing customer segmentation using data mining, the most important part is that the result should be meaningful from the business perspective and able to be utilized further in real business environments. One thing to keep in mind is that since the market environment is dynamically changing, the segmentation modeling process should be iterative and the model should evolve as the market changes.

In general, Data mining will, if it is used correctly, allow to identify customer types that had not been recognized before, and will inevitably lead to new ideas about the market segments and the most appropriate way to offer new products and services to the new customer groups that is discovered.

Saarenvirta (1998) noted that businesses could use customers' data to divide customers into segments based on such variables as current customer profitability, a measure of the lifetime value of a customer, and retention probability, which highlight visible marketing opportunities.

Saarenvirta (1998) in his article "mining customer data" had tried to prove the power of data mining techniques for segmenting customers successfully than using statistical methods.

Saarenvirta's research, there is the loyalty group in Canada running an Air Miles Reward program (AMRP) for a coalition of more than 125 Canadian companies in all industry sectors, including finance, credit card, retail, grocery, gas and telecommunications. The coalition partners capture consumer transaction and transmit them to loyalty group, which stores these transactions and uses the data for database marketing initiatives by the coalition partner. But Saarenvirta wanted to prove data mining effectiveness in exploring better customer segments than the loyalty group.

The primary source of the data for Saarenvirta's study consisted of 50,000 customer transactions for a twelve months period directly extracted from the database of loyalty group. He selected Intelligent Miner software as a tool and demographic clustering algorithm. He did cleaning, data transformation, and aggregation for preparing the data for analysis. Different techniques were used for handling missing values when the data is categorical and numeric. For categorical values either filling it with unknown values or discarding the field if large portion of data is missing was done. And either assigning the mean value or assigning zero value were the methods used for handling missing numerical values. There was also ratio variables created for improving the result of data mining analysis by making the data more meaningful. In addition, numeric data were changed to discrete ranges using quintile break points for making the data easy to interpret.

Finally he selected nine variables as an input for the mining tool. Most of the variables can measure the shareholder value of each customer. The result of his analysis indicated nine

different clusters, among them three clusters (cluster 2, 5, 7) contain the best customers generating a high revenue per person than the other clusters, cluster 8 and 4 were found to be the worst clusters, and cluster 1 are customers with new customer records having unknown behavior.

To conclude with, Saarenvirta proved the data mining technique could successfully segment customers than statistical analysis and recommended that different marketing strategies can be developed based on the result obtained from the analysis. As a result, the loyalty group decided to understand several functions of data mining techniques.

Another study by Wiwattanacharoenchai and Srivihok (2003) indicated the importance of the data mining technique, K-Means, to study customer behavior and to segment customers into clusters on electronic banking usage for Thai commercial banks in Thailand.

The target data used or the study was e-banking transactions from web log file for that bank. There were 52,195 transactions collected from 1st to 31st of July 2002. Each transaction gave details of web usage including user accounts of those who accessed the web site, requested web pages and their order, and the period of time pages were viewed.

After all data preparations were completed, those e-banking transactions collected with in the specified period of time were used as an input for segmentation analysis. The technique selected for this study was K-means as the problem was clustering.

They selected five important variables for segmenting customers and the variables were log in time, date, channel [Personal Computer or Automatic Teller Machine (ATM)], language (Thai, English), and type of transactions. Among the five variables two of them

had continuous values. For these continuous values transformation to categorical values was made.

By using the data mining software, Minset and selecting K-Means algorithm to cluster, they found five clusters. The majority of the customers in the five clusters accessed e-banking during the daytime. They suggested that because they can cause a heavy traffic in the daytime, there should be a campaign for using e-banking during evening time or nightime to reduce the peak load. Moreover, they found those personal computers are more popular for e- banking than ATM and few English-speaking customers used e-banking in the specified bank.

Gokturk and Basgoze (2003) used data mining techniques to build customer profiles of one super market. They started the process by initial data collection on both customer factual (demographic) and transactional data. In marketing application based on customer purchasing history, the factual data includes demographic information such as name, gender, birth date address, and salary. Transactional data consist of records of the customer purchase during a specified period like purchase date, product purchased, amount paid, etc.

After the initial data collection, they applied all the necessary steps to prepare the data for analysis. Classification and regression trees were used to discover rules that describe the behavior of individual customers. Finally rules generated after applying different rule validation techniques were used to build the profiles of customers for the super market. They recommended that it is favorable to use customer profiling techniques to make inferences before planning marketing campaigns.

Finally, in the case study at bank by Ahola and Rinta-Runsala (2001), K-Means was also applied to cluster bank customers into clusters with similar paying behavior. Paying behavior refers to the use of cards, cash dispensers, on-line banking terminals, internet banking, direct debiting, bank transfer, withdrawals from branch office, etc. Totally the paying behavior of each customer is described with 11 variables which tell the number of times the customer used each of the paying channels (cards, check, etc) during a period of six months.

Another 11 variables were derived by dividing the number of times the customer used each of the paying channels by the total number of paying channels used. They tried to train the model by using the tool SPSS and selecting k-Means algorithm. Different values of K (K set to 5, 10, and 20) were tried to reach to the appropriate number of clusters. The clustering with K=5, resulted into four homogeneous and one heterogeneous clusters. Cash, web, ATM, card, and miscellaneous paying habits reflected in each clusters respectively. Increasing the number of clusters to 10 resulted in splitting of the miscellaneous clusters into smaller one with more distinct features without touching the remaining four clusters (cash, web, ATM, and cards). With 10 clusters, the average distance of an observation from its cluster center is about the same in all the clusters. But increasing the number of cluster to 20 resulted for each cluster to be divided into two of the clusters obtained with k set to 10. When the numbers of clusters were 20, each clusters contained very few number of customers and there was also unnecessary splitting of clusters. Finally they conclude that the number of clusters should be 10 for indicating that the miscellaneous cluster is not one big cluster of heterogeneous cluster, but in fact a group of smaller but more extreme clusters. They also applied classification and regression trees for classifying customers into each clusters.

Therefore clustering and classification are two of the most important data mining methodologies used in CRM.

2.9. Data Mining and Hotel Industry

The importance of the hotel industry for countries like Ethiopia is very significant. The hotel industry brings visitors and business people to the heart of a community. They can feel a compelling market made by accommodating visitors to area business and institutions. The industry can generate sales for nearby retail and service business, and significant tax revenues while creating many new jobs for local residents.

For a certain industry to be competent in the market, it should be able to understand customers' behavior, distinguish profitable customers from unprofitable customers, and retain loyal and valuable customers so that it can remain competent in the market by developing and implementing appropriate marketing strategies. This needs the integration of CRM and data mining.

As Magninni (2003) noted, in the hotel industry knowing your guests--where they come from, how much they spend, and when and on what they spend can help you formulate marketing strategies and maximize profits. Fueled by the proliferation of centralized reservation and property-management systems, hotel corporations accumulate large amounts of consumer data. This information can be organized and integrated in databases that can then be tapped to guide marketing decisions. However, identifying important variables and relationships located in these consumer-information systems can be a difficult task . Data mining can be instrumental in overcoming such obstacles. Data mining techniques enable the hotel industry to understand changing customer wants and predict future demand trends. By

using data mining, customers can be classified by segment and then reveal clusters within or across segments and spot unexpected changes in purchases with an eye to determining why that occurred- for example, by associating purchase behavior with advertising campaigns or loyalty programs promotions. In addition, data mining techniques can enable to forecast future customer actions based on trends in the data.

Min et al (2002) used data mining technique for discovering the answers for the following questions by the use of the customer data who had stayed at eleven different luxury hotels showing similar characteristics in terms of price, location and service amenities. The questions to be solved were:

- Which customers are likely to return to the same hotel as repeat guest?
- Which customers are at greatest risk of defecting to other competing hotels?
- Which service attributes are more important to which customers?
- How to segment the customer population into profitable or unprofitable customers?
- Which segment of the customers best fits the current service capacity of the hotel?

Some 281 customers attending in one or more of the eleven hotels representing many different nationalities and various demographic sectors were selected. Questionnaires asking for the information about demographic profiles (e.g. age, occupation, nationality), frequency of their hotel visitation, the purpose of their travel, relative importance of service attributes to overall service quality, and the level of customer satisfaction based on service experiences were distributed with the help of surveyors. Finally the response obtained consisted of 85.15% from the target sample size of 281 participants.

After collecting all the necessary data, they prepared the data for analysis and tried to categorize numerical data like age to categorical values during data preparation step.

The purpose of the study was to answer the questions indicated, by generating sets of rules that can easily be understood by the hotel managers. Hence, C5.0 was selected for building decision trees using the data mining software Clementine 6.0.

At last they generated very important "if then" rules that can answer the questions indicated so that hotels can classify the existing customers database into certain types of segmentation and then predict a customer's behavior in selecting a particular hotel.

From this chapter the major concepts of data mining and its application for better CRM were discussed. It can be understood that, data mining is not simply one discipline but a combination of different disciplined, and the process of mining (discovering) knowledge has iterative processes, starting from data collection to knowledge presentation. The importance of building a data warehouse, and the data mining techniques used for this study; clustering and classification particularly, K-Means and decision tree are also discussed in detail. Finally different related literatures on the application of data mining for effective CRM were also reviewed. Generally from this chapter, it can be understood that integrating data mining with CRM is very important in order to make businesses especially the hotel industry competent and profitable. The next chapter discusses what is CRM and related topics.

CHAPTER THREE

CUSTOMER RELATIONSHIP MANAGEMENT

3.1. Introduction

Because the research is about developing customer segmentation and classification model to help in better CRM, it is important to understand what CRM is. In this chapter, an attempt was made to review literatures on the concepts of CRM and key issues in CRM such as components of CRM, customer segmentation, and customer loyalty. In addition, background about the study area (Ethiopia hotel) and what CRM activities the hotel considers were also discussed.

3.2 The meaning of CRM

The term CRM has been applied to almost every element of business that even remotely interacts with a customer. It consists of separate words. Which are "customer", "relationship" and "management"

Customer is the only source of the company's present profit and future growth. However, all customers can't be profitable and those profitable with less resource are always limited. Differentiating which customers are the real customers is the most difficult task in CRM.

Relationship between a company and its customers involves continuous bi-directional communication and interaction. The type of relationship can be short term or long term, continuous or discreet, repeating or one time, and attitudinal or behavioral.

Management covers marketing management, manufacturing management, human resource management, service management, sales management, and research and development management.

From these words together, CRM can simply be understood as the process of managing relationship of customers with the business firm. There are many touch points that customers and the business firm interact. It is concerned with managing direct or indirect contacts of customers with the firm including marketing, manufacturing, customer service, field sales and field services.

Some times people consider CRM as only marketing strategy. But traditional marketing strategies focused on the price, product, promotion and place to increase market share. The main concern of traditional marketing strategies was to increase the volume of transaction between seller and buyer. And hence the good measure of performance was volume of transaction.

But CRM is business strategy that goes beyond increasing transaction volume. Its objective is to increase profitability, revenue and customer satisfaction.

Parvatiyar1 & Sheth (2001) cites that:

> *CRM is a comprehensive strategy and process of acquiring, retaining, and partnering with selective customers to create superior value for the company and the customer through the integration of marketing, sales, customer service, and the supply-chain functions of the organization to achieve greater efficiencies and effectiveness in delivering customer value.*

Moedritscher (2002) also points out that CRM is a customer oriented business philosophy that involves analyzing, planning and controlling customer relationships by means of modern information and communication technology. It is concerned with the creation, development and enhancement of individualization customer relationship with carefully targeted

customers and customer groups. The customer centric view combined with preferential treatment of selected customers should result in maximizing customer relationship in the long run, both for the company and its customers.

The benefits of CRM to customers are, increased convenience and speed of service, and the benefit to organizations is their ability to develop profitable customer-focused strategies. In general, CRM is about integrating marketing strategies, process, technology, and people to improve service delivery, build social bonds with customers and secure customers loyalty by keeping long term, mutually beneficial relationships with valued customers selected from a pool of customers. To attain the objectives, CRM has its basic principles. Basic principles of CRM are: treating customer individually, attaining and acquiring customer loyalty through personal relationship, and selecting good customer instead of non valuable customer. The reason why customers should be treated individually is because customers have different preferences and behavior, and the profitability or value of them also varies. Hence there has to be strategies to keep the right customers who generate the most profits. Through differentiation, a company can allocate its limited resources to obtain better returns. The best customer deserves the most customer care.

All of these measures imply doing a better job acquiring and processing internal data to focus on how the company is performing at the customer level.

3.3. Components of CRM

The components of CRM are:

a). A database of customer activity

b). Analysis of the database

c). Given the analysis, decisions about which customers to target.

d). Tools for targeting the customers

e). How to build relationships with the targeted customers.

f). Metrics for measuring the success of the CRM program

a) Creating a Customer Database

A necessary step to a complete CRM solution is the construction of a customer database or information file. This is the foundation for any CRM activity.

The database should contain information about the following:

- Transactions
- Customer contacts
- Descriptive information
- Response to marketing stimuli

b) Analyzing the Data

The database containing information about daily transactions, customer contacts, descriptive information and response to marketing stimuli should be analysed by using different software for customer segmentation, direct marketing, prediction of risks, cross-selling and different analysis for better CRM

c) Customer Selection

Given the construction and analysis of the customer information contained in the database, the next step is to consider which customers to target with the firm's marketing programs. The results from the analysis could be of various types. If segmentation type analyses were performed on purchasing or related behavior, the customers in the most desired segments (e.g., highest purchasing rates, greatest brand loyalty) would normally be selected first. Other segments could also be chosen depending upon additional factors. For example, if the

customers in the heaviest purchasing segment already purchase at a rate that implies further purchasing is unlikely, a second tier with more potential would also be attractive.

The marketing manager can use a number of criteria such as simply choosing those customers that are profitable (or projected to be). The goal is to use the customer profitability analysis to separate customers that will provide the most long-term profits from those that are currently hurting profits. This allows the manager to "fire" customers that are too costly to serve relative to the revenues being produced.

d) Targeting the Customers

Mass marketing approaches such as television, radio, or print advertising are useful for generating awareness and achieving other communications objectives, but they are poorly-suited for CRM due to their impersonal nature. More conventional approaches for targeting selected customers include a portfolio of direct marketing methods such as telemarketing, direct mail, and, when the nature of the product is suitable, direct sales.

In particular, the new mantra, "1-to-1" marketing, has come to mean using the Internet to facilitate individual relationship building with customers. An extremely popular form of Internet-based direct marketing is the use of personalized e-mails.

e) Building Relationship Programs

While customer contact through direct e-mail offerings is a useful component of CRM, it is more of a technique for implementing CRM than a program itself. Relationships are not built and sustained with direct e-mails themselves but rather through the types of programs that are available for which e-mail may be a delivery mechanism. To keep the selected or targeted customers, different relationship programs like Customer service, Frequency/loyalty programs, Rewards programs should be established.

The overall goal of relationship programs is to deliver a higher level of customer satisfaction than competing firms. There has been a large volume of research in this area. Managers today realize that customers match realizations and expectations of product performance, and that it is critical for them to deliver such performance at higher and higher levels as expectations increase due to competition, marketing communications, and changing customer needs (Winte, 2000). Thus, managers must constantly measure satisfaction levels and develop programs that help to deliver performance beyond targeted customer expectations.

f) Metrics

The increased attention paid to CRM means that the traditional metrics used by managers to measure the success of their products and services in the marketplace have to be updated. Financial and market-based indicators like profitability, market share, and profit margins have been and will continue to be important. However, in a CRM world, increased emphasis is being placed on developing measures that are customer-centric and give the manager a better idea of how CRM policies and programs are working.

Some of these CRM-based measures are the following:

- Customer acquisition costs
- Conversion rates (from lookers to buyers)
- Retention/churn rates
- Loyalty measures.

All of these measures imply doing a better job acquiring and processing internal data to focus on how the company is performing at the customer level.

3.4 Customer Loyalty

With the increased importance being placed on customer satisfaction in today's business climate, many companies are focusing on the customer loyalty as the key to increase market share. Many business managers who are going to implement CRM strategies should be able to understand what customer loyalty means and what should be done to earn the customer loyalty?

Any person in the business world knows that there is a positive relationship between customer loyalty and profitability. The increased profit from loyalty comes from reduced marketing costs, increased sales and reduced operational costs. Loyal customers are less likely to switch because of price and they make more purchases than similar non-loyal customers. This means, companies will be able to segment customers 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.

Then what is customer loyalty and how is it measured? Loyal customers are customers who repurchase from the same service provider whenever possible, and who continues to recommend or maintain a positive attitude towards the service provider (Kandampully & Suhartanto, 2000). Michud (2000) also defines loyalty as when you make personal connection with your customers and let them know that you hear what they are saying and then prove it by being responsive to their needs, you are building loyalty that influences behavior. Loyalty is always going to be on relationships and that is what you want.

There are measurements of customer loyalty, which are behavioral and attitudinal measurements. The behavioral measurement considers, repetitive purchase behavior as an indicator of loyalty. But repeated purchases are not always the result of psychological

commitment towards the brand (Tepeci in Kandampully & Suhartanto, 2000). For example, a traveler may stay at a hotel because it is the most convenient location. When a new hotel opens across the street, they switch because the new hotel offers better services. Hence, repeated purchase does not always mean commitment. The attitudinal measurements use attitudinal data to reflect the emotional and psychological attachment inherent in loyalty. The attitudinal measurements are concerned with sense of loyalty, engagement and allegiance. There are times when a customer holds a favorable attitude towards a hotel, but he or she does not stay at the hotel because of the reason that the hotel was expensive for him or her but still recommends the hotel to others.

Combining the two measurements of customer loyalty is best approach to evaluate the customer who is very likely to remain with the company. Because it combines the two dimensions, measures loyalty by customers' preference, propensity of brand-switching, frequency of purchase, recency of purchase and total amount of purchase (Pritchard and Howard, 1997; Kandampully & Suhartanto, 2000).

There are factors, which lead to customer loyalty. Anderson, et.al (1994) found that customer satisfaction and service quality are prerequisites of customer loyalty. Customer satisfaction is considered to be one of the most important outcomes of all marketing activities in market-oriented firm. The ultimate need for satisfying firm's customer is to expand the business, to gain a higher market share, and to acquire valuable customers, all of which lead to improved profitability.

Gettty and Thompson (1994) based on their findings suggest that customers' intentions to recommend are a function of their perception of both their satisfaction and service quality.

Therefore, it can be concluded that there is a strong relationship between customer satisfaction and customer loyalty.

3.5 Customer segmentation

For most business firms, locating and effectively targeting unique market segments is both a reality and a necessity in today's competitive market place. Creative market segmentation strategies usually afford the business organization a strategic advantage over their competitors and provide marketing efficiencies that greatly improve customer retention and profitability. If a firm can address its markets by way of a creative new vision of how that market is structured and operates, and can uncover the needs and wants of the segments therein, then it has the opportunity to act on that vision to enhance its own profitability, often at the expense of the competition.

Customer segmentation is one of the most important CRM strategies, which is also part of market segmentation. The word segmentation is a way to have more targeted communication with customers. It is the process of putting population into segments according to their affinity or similar characteristics.

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

Customer segmentation describes the division of customers into homogeneous groups, which will respond differently to promotions, communications, advertising and other marketing mix variables. A different marketing mix can target each group, or "segment," because the segments are created to minimize inherent differences between respondents within each segment and maximize differences between each segment. In segmentation, the objective is

to group consumers into relatively homogeneous groups that respond in a similar manner to marketing activities. The chosen segmentation characteristics (wishes, needs, knowledge) of consumers should allow the resulting clusters to be homogeneous within and heterogeneous between the segments.

Variables useful for description of the segments can be used as the basis for segmentation. Kotler (1998) defines the following segmentation variables:

- Geographical
- Demographical
- Psychographical and
- Behavioral

- **Demographic variables** — Age, gender, income, marital status, education, occupation, household size, length of residence, type of residence, etc.
- **Geographic variables** — City, state, zip code, census tract, region, metropolitan or rural location, population density, climate, etc.
- **Psychographics variables** — Attitudes, lifestyle, hobbies, risk aversion, personality traits, leadership traits, magazines read, television programs watched, etc.
- **Behavioral variables** —loyalty, usage level, benefits sought, distribution channels used, reaction to marketing factors, etc.

Kotler (1998) suggests that the approach of segmenting according to behavioral variable is superior in most cases. He also suggests that the main advantage of segmenting according to

behavioral variable is creation of a relatively homogeneity of segments solution according to selected behavioral characteristics.

Segmentation is important of the following reason:

- It is easier to address the needs of smaller groups of customers, particularly if they have many characteristics in common
- Identify under-served or un-served markets. Segmentation can allow also a new company or new product to target less contested buyers and help a mature product seek new buyers.
- More efficient use of marketing resources by focusing on the best segments for your offering— product, price, promotion, and place (distribution). Segmentation can help you avoid sending the wrong message or sending your message to the wrong people.

However there are some difficulties in making good segmentation (Bounsaythip, 2001). These are:

- Relevance and quality of data are essential to develop meaningful segments. If the company has insufficient customer data or too much data, then it can lead to complex and time-consuming analysis. If the data is poorly organized then it is also difficult to extract interesting information. Furthermore, the resulting segmentation can be too complicated for the organization to implement effectively. The use of too many segmentation can also be confusing, resulting in segments which are unfit for management decision making
- Segmentation demands continuous development and updating as new customer data is acquired. In addition, effective segmentation strategies will influence the

behavior of the customers affected by them; thereby necessitating revision and reclassification of customers.

- A segment can become too small or insufficiently distinct to justify treatment as segments.

Therefore, care should be taken in segmentation process to avoid the above difficulties.

3.6. CRM and Ethiopia Hotel

3.6.1 Ethiopia Hotel

Ethiopia hotel is a historical government hotel in Ethiopia. Its construction was initiated by the establishment of the organization of African union in Addis Ababa and was meant to accommodate guests coming to attend meetings and conferences. The hotel was initially constructed to serve as a guesthouse.

There are two branches of Ethiopia Hotel in Addis Ababa namely Harambie Hotel and Sky Restaurant. The hotel has 340 employees out of which 227 are males and the rest 113 are females. The hotel has 147 rooms for renting at an average rate of 200 Birr per room. During the last fiscal year total gross sales of the hotel was 4,113,032 Birr collected from room sales, food and beverage, and from different miscellaneous sales. Ninety percent of this amount was obtained from room renting which is the major source of income for the hotel (2003/2004 budget year annul report by Ethiopia Hotel Enterprise).

With expansion of the hotel industry in the country, the hotel needs to improve the quality and quantity of services so as to remain competent in the business. One of the key factors for the continued survival and proliferation of hotels is their ability to provide services to the changing preferences of customers. These can be achieved through customer segmentation.

Therefore Ethiopia hotel has to start developing of CRM strategies to keep its valuable customers and loyal.

3.6.2. Relationship Programs

One of the main strategies used for effective CRM is adopting different relationship programs. The main objective of these programs is to deliver a higher level of customer satisfaction than competent firms. In Ethiopia hotel customers are usually called guests.

3.6.2.1. Customers Service

Any contact (" touch points") that a customer has with a firm is a customer service and has the potential to gain repeated business and keep CRM. In Ethiopia hotel, when the customer has a problem, the guest can communicate directly or by using the telephone line with front office mangers.

In most of the cases, the customers initially come to the hotel through different travel agents. The marketing department of the hotel has a direct communication with the travel agents. When guests come to the hotel through a travel agent, the agent will be paid a commission which is 15% of the revenue obtained from room sales.

But it doesn't mean that there is no customer who makes direct contact with hotel. Some customers communicate with the hotel at reception directly through telephone to make reservations.

3.6.2.2. Loyalty/ Frequency Programs

These programs provide rewards to customers for repeat purchase (visits). In Ethiopia hotel there is no defined reward program established for motivating customers who repeatedly visit the hotel. What is done is simply, if a guest has come repeatedly to hotel, and if he asks for a bonus to front office manager, they can get the bonus depending on how

much loyal he or she is. But the loyalty of guests is determined only if front office department remembers physically that he or she has come repeatedly. There is no any information system designed to determine the loyalty of the customer.

In this chapter, different concepts of CRM, historical background about Ethiopia hotel, and what CRM activities were considered in the hotel are discussed. From this, it is not possible to conclude that the hotel has established good CRM. The next chapter deals with how the hotel's data are collected, prepared and analyzed for building customer segmentation model. And from the analysis, what kind of customers are there in the hotel, who are profitable and non profitable, and what are the rules used for assigning new customer records to identified customer segments are obtained.

CHAPTER FOUR

EXPERIMENTATION

4.1. Introduction

As the main objective of the research is to apply data mining tools and techniques on Ethiopia hotel's customer data for effective CRM, this chapter is the most important part of the research. In this chapter data collection, preparation and analysis processes were dealt. The process of building clustering and classification models by using the algorithm K-Means and decision tree respectively were also discussed. This research incorporates all the typical stages that characterize data mining process, especially the CRoss-Industry Standard Process for Data Mining (CRISP-DM) process cycle (CRISP-DM, 2000). Therefore, the steps followed were based on CRoss-Industry Standard Process for Data Mining which is mentioned as follows;

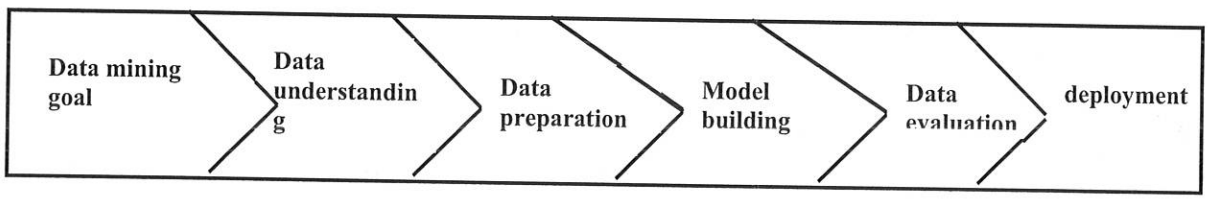


Figure 2: Steps of the crisp-DM processes

4.2. Data Mining Goals

The first data mining goal is to identify the most important variables from the data collected that can be used for clustering model building. As a result, the second data mining goal that is segmenting group of customers with similar characteristics can be achieved by using these variables. This will greatly help for understanding customers. A clear understanding of customers is the basis of taking targeted action to each group of customers.

In order to provide segments that can be explainable to domain experts, emphasis was given to data preparation and an exploratory data analysis. This process allowed the identification of important attributes as input for the model building phase. The last data mining goal is to build a classification model that can generate rules for assigning new customer records to the already identified segments by using a decision tree learning method. This can be achieved by using the cluster index as dependant variables and the rest selected variables as independent variables.

The success criterion for this data mining project is the discovery of customer segments which are meaningful to domain experts and building a classification model for generating rules that can classify new customer records properly to the identified segments.

4.2.1 Data Mining Tool Selection

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

- The data mining tasks that the tool is intended for (clustering and classification)
- The algorithms supported (K-means and decision trees)
- Operating system and hardware requirement: the hardware requirements on which the software runs (Intel 32-bit, 256 RAM) and a MS Windows operating system.
- Data sources: possible formats for the data that is to be analyzed (MS Excel)
- Size: the maximum number of records the software can comfortably handle (up-to 10,000 records)
- Visualization capabilities

Tools that can meet the above criteria were searched and were found to be Clementine, Intelligent Miner data mining tools, and Knowledge studio version 4.1.1. But the cost of

Clementine and Intelligent Miner data mining tools were beyond the budget allocated for this project. So, the researcher selected and used Knowledge studio version 4.1.1, which is freely accessible from the department of information science in Addis Ababa University. Knowledge studio provides a number of ways to visually explore and express patterns in the data. It also helps the whole data mining process- from preparing the data through producing the final graphs and reports.

4.3 Data Understanding

The next step of data mining goal is to understand the collected data. In this study, most of the data are collected from day to day transactions and it is used for administrative purposes. Therefore, in order to fulfill the data requirements for this experiment, the existing data situation should be studied and all the necessary data for the study should be collected carefully

4.3.1. Data collection

In order to achieve the above goals, the data was collected in two steps;

1. Initially only two months data was collected from the database which was maintained through NCR Easy application hotel software in June 2004 which the hotel used only for two months (June and July) due to lack of technical manpower. The data from this software was converted into MS-Excel
2. Another two months data was added by entering manually into MS-Excel (August and September). Totally four months data were collected.

The data collected for this research consists of three sources; which were resident summary, guest entry registration, and derived attributes from the existing data.

1. The guest entry registration form consists of name, number of person, room number, bill number, room occupied, nationality, profession, date of arrival, duration of stay, purpose of visit, where she/he is coming from and going to, passport number, and address.
2. The resident summary consist of daily status of the transfer, presence or check out and payment of charge of customers containing room number, bill number, room charge, value added tax, service charge, extra revenue, payment, and total charge
3. In addition to the above attributes, attributes such as number of days stayed(actual duration of stay), frequency of visit, total room charge and total revenue were derived from the application software.

The derived attributed were obtained in the application software in the following way, the fields obtained on daily residence summary were only available for each day. But the total amount can be obtained by adding the amounts in each row containing the same information about each customer. To know the actual duration of stay, one can count the number of rows with the same information about each customer in resident summary table. If one wants to know the frequency of hotel visit, go through for the different bill number in the resident summary or count the number of rows informing about the same customer in guest entry registration table.

On the other hand, it was only possible to extract residence summary report for each day from the application software. Hence, for the two months data 61 files, one file for each day had been extracted. In order to extract all the important variables from the data, all these files were merged together to form a collection of residence summary report constituting for the

two months period. The guest entry registration table can be totally extracted once to the excel format. So there were two tables obtained from the database; guest entry registration form and residence summary. This process can be shown as the figure below:

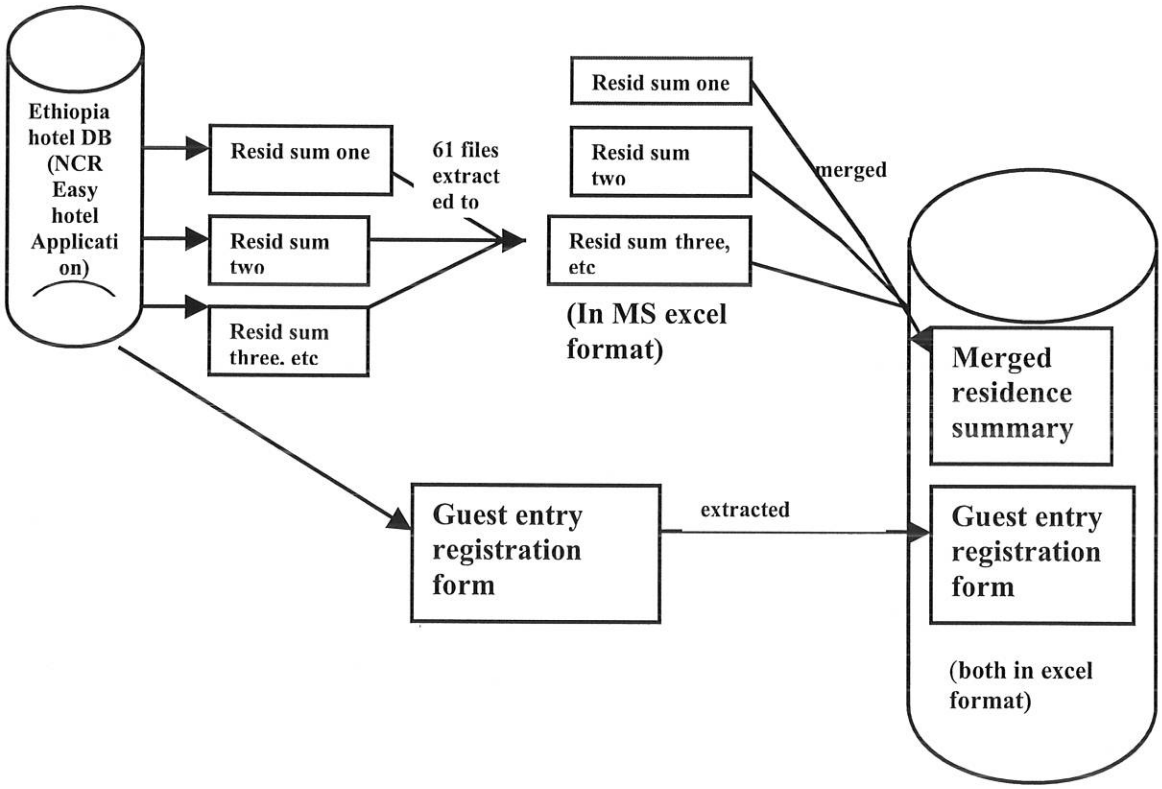


Figure 3:The processes of extracting data from the computerized system part.

Another process taken, was entering data from the manual documents and books to the computer in the excel format. Two tables (worksheets) were created by feeding data from collection of the guest entry registration forms (cards) and daily residence summary book. As a result, guest entry registration database containing demographic data and residence summary containing financial data of customers in electronic format were the final out puts

of this process constituting of two months data (August and September, 2004). The process of data entry is shown in the figure below:

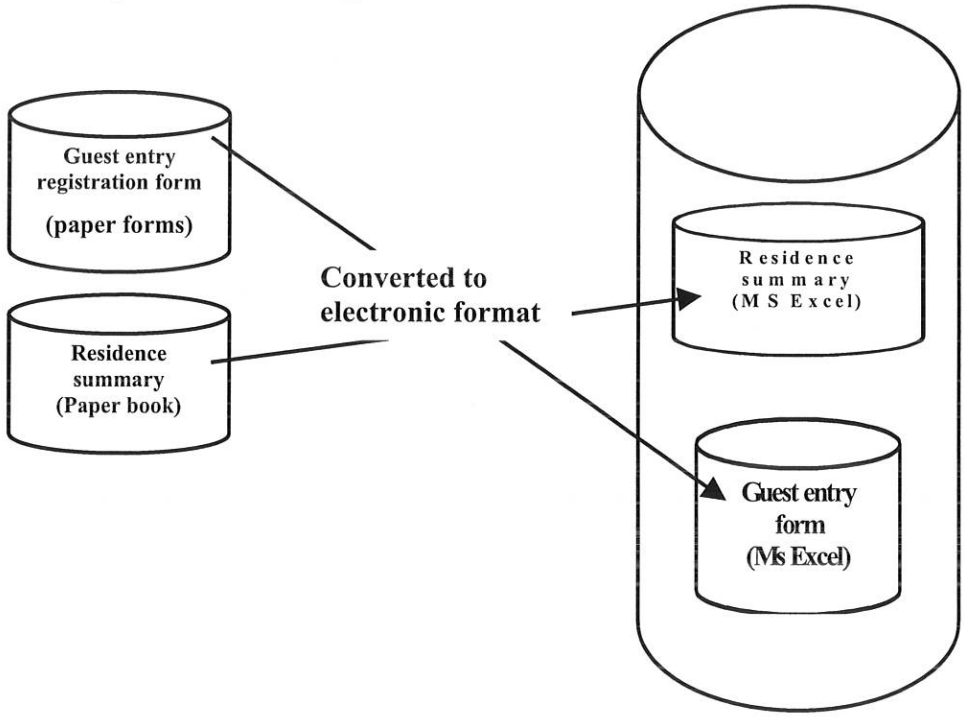


Figure 4:the process of converting manual data to electronic MS Excel format

A total of four months data were collected; two months data from computerized database, and two months data from the manual documents. Here there were two pairs of similar tables obtained from each group. Both groups were merged to form a large collection of data.

To prepare the data for final analysis, the two tables again should be integrated into one table. That means every record should contain both the demographic and financial data in the form of one table. For this purpose the important variables were bill number, date of arrival and name of guest. How these tables were integrated was discussed on data preparation part. But all the variables extracted are not important, the only important fields taken into consideration are bill number, date of arrival, name of guest, nationality, room occupied,

number of persons, duration of stay, purpose of visit, profession, date of arrival, room charge, board charge, extra charge, and number of rows the customer's detail was recorded.

4.3.2. Description of the data

The following tables were obtained from the initial data collected using MS-Excel:

1. Guest registration table: it consisted of the total data of 7113 records of customers with the following attributes stated in the table below.

Attribute/Field name	Data type	Description
Name	Text	The name of the guest
No.Room occp	Number	The number room occupied by each person
No.person	Number	Number of persons number of person staying
Room No.	Text	The room occupied Identification number
BillNo	Text	Bill number
Nationality	Text	Nationality of guests
Profession	Text	Profession of guests
Date arrival	Date	Time of arrival of the guest
Durtn Stay	Number	Number of days to stay
Purpose of visit	Text	The purpose of coming to hotel or country
Origin of the guest	Text	Where is he/she coming from
Destination of the guest	Text	Where is he/she going to

Table 1:Attributes of guest registration worksheet

2. **Resident summary table:** it consisted of the total data of 10,590 records of customers with the following attributes stated in the table below.

Attribute/Field name	Data type	Description
Name	Text	Name of the guest
RoomNo	Text	The room number occupied
Bill No	Text	Bill number
Date arrival	Date	The time guest arrival
Room charge	Number	Revenue from room rent
Extra charge	Number	Additional revenue from other services
Service tax	Number	10% service tax
Sales tax	Number	15% sales tax
Ttl deductn	Number	The sum of sales and service tax
Brought forward	Number	Amount brought from yesterday
Totaltodate	Number	Accumulated total revenue
Payment	Number	Payment of charge

Table 2: attributes of resident summary worksheet

The discrepancy in number of records in guest entry registration table and in resident summary table was because, the software that the hotel had used for two months, was used for transaction processing i.e. every detail of daily transaction were processed as a result the customer detail can appear repeatedly on the same database. Every time the customer visits the hotel, the detail of this customer will appear on the guest entry registration table to number of times he/she made. On resident summary, the detail of this customer will appear number times he/she stayed in the hotel. How this redundancy can be handled is discussed in the data preparation part (4.4.1)

4.3.3. Data Quality Verification

There were some missing values identified in two different ways. Some of the attributes or the entire of the individual detail can be missing. There were a lot of missing values on the attributes "origin of the guest" and "destination of the guest" attributes. To solve this problem, the domain experts in the field informed the researcher that most of the time "origin of the guest" represents "nationality". "Origin of the guest" can be deleted because the information obtained from this field, can be obtained from the nationality field. "Destination of the guest" was also discarded as it contains a lot of missing values.

The researcher found, during the study that there was no consistency in the database. The same customer name found in different places with different spelling Hence, in order to maintain integrity on the database the researcher followed the procedure given below:

1. Sort the database by name
2. If differently spelled but similar names appears in different rows with reference to same bill.No, bill-Date and room number then delete redundant records
3. Else match with profession and nationality, if match then delete redundant records.

In addition, the researcher found that the resident information (100 records) were missing due to un-safe maintaince of records in the hotel. Hence the records with the missing information were deleted.

4.4. Data Preparation

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

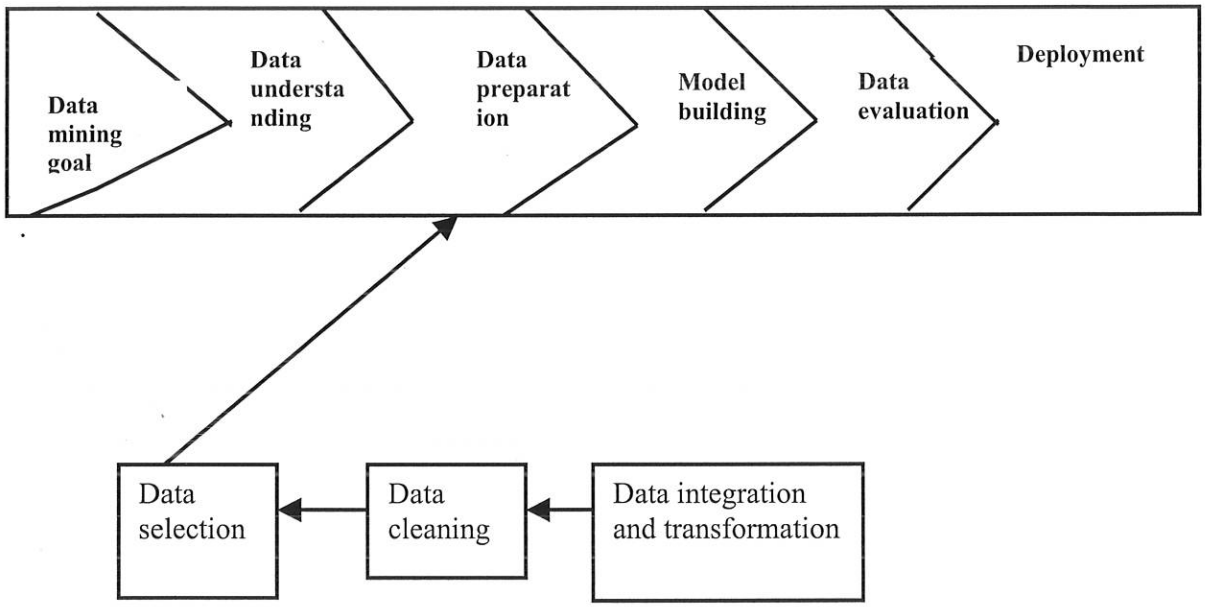


Figure 5: The data preparation phases

4.4.1. Data Selection

Selection decisions were made on the data to be used for analysis. The criteria include relevance to the data mining goals as well as quality constraints. The initial data extracted contains irrelevant attributes like service tax, sales tax, brought forward, and redundant customers' detail information. Service and sales taxes were excluded from the dataset because the two fields depend on the total revenue obtained and they give the same information as the total revenue.

The total amount of revenue generated by each customer was obtained by adding Room charge and extra charge generated during his/her stay at the hotel. Therefore brought forward

field was not important for deriving the total amount of revenue field (attribute) and discarded from the data.

The software that the hotel had used for two months, was used for transaction processing i.e. every detail daily transactions were processed. As a result the customer detail can appear repeatedly on the same database. For example, if customer stays six days visiting the hotel for two times, demographic data (on guest entry registration form) of this customer appears for two times where as financial data (on resident summery) appears for six times in the database. That is because there was discrepancy in number of records in guest entry registration table and in resident summary table. Therefore, care should be taken to eliminate redundant information without losing any information. In order to eliminate redundant information, the researcher adopts the following procedures;

1. On the resident summary table, look for rows informing about the same customer and count the number of rows (this gave duration of stay by each customer). Add total room charge and total extra charge of these rows and put them in a separate column along with one of the rows. And finally delete the rest redundant rows.
2. On the guest entry registration form, count the number of rows of data informing about the same customer. This gives the number of times the customer paid visit (number of visits) and put it together in a separate column with one of the rows by deleting the rest of the rows.

4.4.2. Data cleaning

As Han and Kembar (2001) noted data cleaning helps to clean the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving

inconsistencies. Dirty data (data which are not clean) may cause confusion for the mining procedure, resulting in unreliable output. There are various recommendations as to how to compute missing values of key attributes such as, ignoring the tuples, filling the missing values manually, using global constant, using the attribute mean to fill in the missing values, or using the most probable value to fill in missing values (Han and Kembar, 2001). For this study the researcher, has used two methods; such as using the attribute, mean to fill in the missing values and filling the missing values manually.

There were 300, 300, 207, and 40 missing values on the attributes Profession, Purpose of visit, Room charge, and NoDay respectively. The technique applied for handling the missing values of Profession was, filling the most frequent value of the customers' profession grouped under the same nationality. If all the records' of Profession field's value under the same group of nationality is missing, the technique of filling the most frequent profession value of the total population was filled ("business man"). The purpose of visit attribute has four values; 'transit', 'business', 'tourist' and 'other'. The hotel has fixed room rate when the type of customers are transit, so by looking the room rate, it was possible to identify whether the purpose of visit was transit. So it was found out that 200 records in which purpose of visit value were missing, were known to be 'transit' and the rest 100 records in which purpose of visit value were missing, can be 'tourist', 'business' or 'other'. Among these three values, the most frequent value was 'other' and hence this value was replaced on the place of the rest 100 missing records. On the other hand, for the attribute room charge there were 207 missing records among these 107 records were transit type of customer and hence replaced manually by the fixed price rate set by the hotel. The rest 100 records were replaced by the mean value

of room charge. There were also 40 records in which NoDays has missing values and they were replaced by the values of Drn stay (duration of stay) on guest registration form.

The attributes origin and destination of the guest has a lot of missing values and hence they were not included in the dataset as indicated on the data quality verification part of the study (section, 4.3.3). There were also inconsistencies on the name of customers. The name of the same customer may have different spelling. This was handled by applying the solutions indicated in data quality and verification section of the study.

4.4.3. Data Transformation and Aggregation

Since the data set that Knowledge Studio accepts is a single minable table, the two tables namely, guest entry registration table, and resident summary table should be integrated in a single table. The fact that Knowledge Studio has Open Database Connection (ODBC) facilities enables the researcher to import the data set directly from the data mart, which is constructed using MS excel. In addition, Knowledge Studio has options for choosing which of the attributes to be considered in building a model from the 'single table' dataset.

The guest entry registration form database and the resident summary database were integrated together so as to put the full information of each customer into one table and this was done by, matching the name of guest and the Bill number of both tables. So the final dataset contains 5787 number of records. As indicated on data cleaning phase, some of the records of resident summary matches were not found on guest entry registration form due to unsafe keeping of cards used for recording guest demographic data.

After integrating the two tables (databases) into one, it was found important to generalize the attribute nationality in to higher-level concept "region" referring to the region where

customers are coming from so that it would make enough records to be available in one category of region.

According to Saarenvirta (1998 indicated) creating new attributes from the existing attributes can improve the result of data mining models. So there are derived attributes from the available attributes in the database which are included in the dataset for model building.

The derived attributes were:

1. Total revenue obtained for each person (Rev/No Person) = total revenue/number of person
2. Average revenue obtained for each day stayed (Rev/NO Days) = total revenue/ number of days
3. Average revenue obtained for each visit (Rev/No Visits) = total revenue/ number of visits
4. Average number of days stayed for each visit (No Days/visit) = number of days/ number of visits
5. Average revenue obtained for each room occupied (Rev/ NO Rooms) = total revenue/ number of rooms occupied

To help the algorithm recognize patterns, inputs must be arranged to common ranges usually between 0 and 1 or –1 and 1. Particularly, for building clustering model it is very important to normalize the input data since normalization helps to prevent attributes with initially large ranges from outweighing attributes with initially smaller ranges. Knowledge studio normalizes the input data automatically and for numeric attributes values, scalar method of normalization was selected and for nominal attribute values 1-of-N method was selected from knowledge studio.

Finally, the dataset prepared for analysis contained the attributes shown in the table below:

Attribute/Field name	Data type	Description
Name	Text	Full Name of the guest
Region	Text	The region where they come from
No Person	Number	Number of guests occupied room
Profession	Text	The profession of guests
Room Occupied	Number	Number Name of the guest
Drn of stay	Number	Number of days to stay
Purpose of visit	Text	Purpose of coming to the hotel
Bill No	Text	Bill number
Date arrival	Date	The time of guest arrival
Room Ttl	Number	Total Revenue from room rent
Extra charge Ttl	Number	Total Additional revenue from other services
Ttl Revenue	Number	Total revenue generated
NoDays	Number	Number of days stayed
Ttl Rev/NoDays	Number	Average revenue generated per day
TtlRev/Visit	Number	Average revenue generated per visit
Avg Day/Visit	Number	The average number of days stayed per each visit
Rev/room Occupied	Number	Revenue generated per each room occupied

Table 3: Attributes of the final dataset

4.5 Model Building

4.5.1 Selection of Modeling Technique

The segmentation technique was chosen to identify customer behavior. Segmentation is also known as the clustering technique, which is mainly used for the customer segmentation. Clustering is selected; as it is direct data mining technique (where there is no class to be predicted but rather when the instances are to be divided into natural groups) helps to identify

segments without having prior knowledge about the segments to be discovered. After the segments are identified, new customer records should be classified to the already identified segments by using a widely applicable classification learning method called decision tree. Decision trees have a significant advantage because they can be built manually, easily to explain and operations are completely interactive, and they can benefit from powerful visualization features.

Knowledge studio has two clustering algorithms; K-Means and Expectation Maximization algorithms. K-Means was selected, as it is the most popular simple and straightforward clustering technique. In addition, knowledge studio software (data mining tool developed by Angoss software Corporation) supported the use of K-means for most situations except when there are large amount of missing values.

For decision tree learning the software supports two algorithms KnowledgeSeeker and HeatSeeker. Knowledge Seeker was selected, as it is a powerful, flexible algorithm that is especially good for exploration purposes and manual tree building. It can also handle a large amount of variables with either a continuous or discrete dependent variable. But HeatSeeker is good for automatically generating a tree and performs better with fewer variables.

Attributes with initially large ranges outweigh attributes with initially smaller ranges when the input data are not normalized in the case of K-means. For this problem the Knowledge Studio can automatically normalizes data for both numeric and non-numeric attributes. Because outliers may affect the performance of k-Means algorithm, care should be taken to remove outliers in the dataset.

For decision trees even though it can handle both continuous values and discrete values, the accuracy will be improved if all the variables are discrete. Based on this logic, the

continuous values are converted in to discrete values after building the clustering model for classification model building.

4.5.2 Test Design

Building clustering model does not need testing dataset because clustering is unsupervised learning where the algorithm is provided with the data points without labels, the task is to find a suitable representation of the underlying distribution of the data. But on every result of cluster run, the domain experts' opinions along with the understandability of each segment are considered.

To find representative rules from the decision tree, the total dataset was divided into two. The first 70% of the dataset was allocated for training, the rest 30% of the dataset was used for evaluating the impact of the pruning and the accuracy of the model over subsequent data.

4.5.3 Clustering Modeling

After the preparation of the dataset, the next step was to build the clustering model using the selected tool. The basic parameters available in Knowledge Studio for K-means clustering include:

- Number of clusters: the number of clusters (k in K-means) that need to be created. This value has to be manually input into the system.
- Number of iterations: This parameter indicates the maximum number of times the algorithm will read the data.
- The variables to be used for building the clusters

K is a user-defined number. Initially, different values of K, which ranged between 6 and 4, were randomly used. Saarenvirta (1998) advises, the number of clusters chosen should be

driven by how many clusters the business can manage. So the researcher consulted domain experts to determine appropriate number of clusters and as to how they are interpreted.

The following three approaches were employed to understand the clusters using Knowledge Studio:

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

To experiment how the algorithm assigned the distribution of variables from cluster to cluster, the data overview report and the dataset chart from Knowledge Studio was the basis for determining the threshold values of the attributes for the analysis of the results. The report provides a summary of the minimum, maximum, mean and standard deviation values for the different data sets, and the dataset chart shows the distribution of the data.

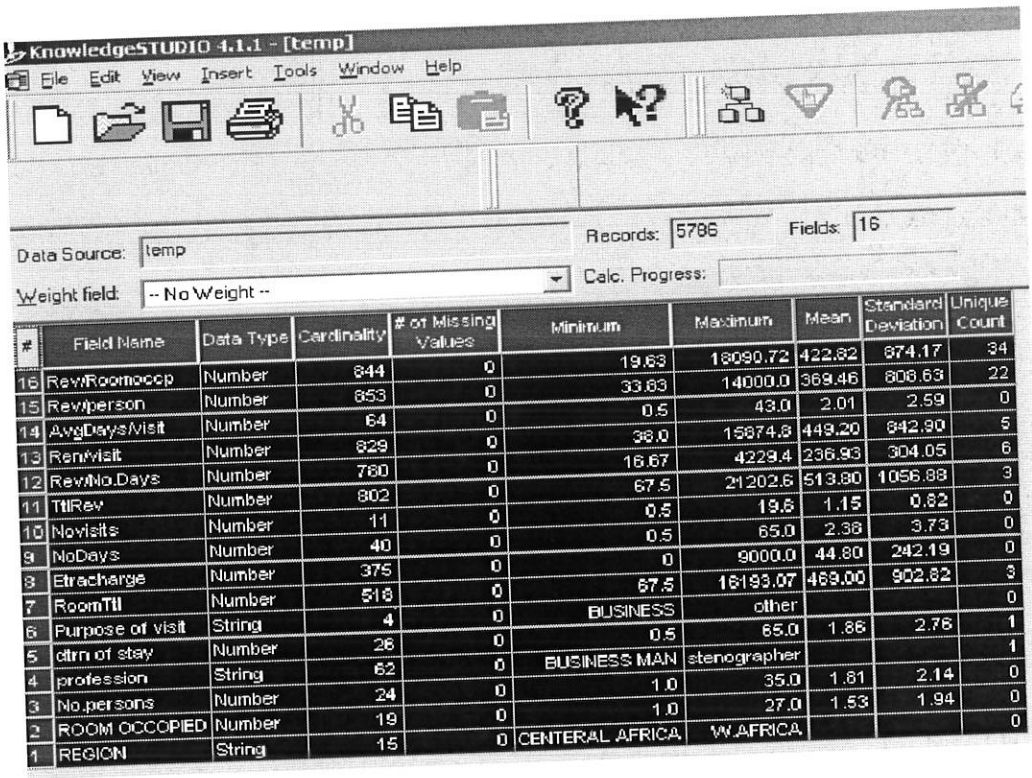
The following experiments with K set to 6, 5, and 4 were conducted to determine the number of clusters and to see their patterns discovered:

Experiment 1

All the variables were the input of the first cluster run except BillNo, Name, RoomNO since they don't provide useful information but may reduce the accuracy of the algorithm.

Custer Run	Variables	N0 of records	No of clusters	No of iteration
1st	NoPerson, Nationality, NoPerson, Profession, purpose of visit, RoomTtl, extracharge, NoDays, Novisits, TtlRevenue, Rev/Visit Rev/NoDays, AvgDay/visit, Rev/NoPerson	5786	6	10,000

Table 4: Summary of input parameters for the first cluster run



#	Field Name	Data Type	Cardinality	# of Missing Values	Minimum	Maximum	Mean	Standard Deviation	Unique Count
16	RevRoomoccp	Number	844	0	19.63	18090.72	422.82	874.17	34
15	Revperson	Number	653	0	33.83	14000.0	369.46	806.63	22
14	AvgDaysVisit	Number	64	0	0.5	43.0	2.01	2.59	0
13	RenVisit	Number	829	0	38.0	15874.8	449.20	842.90	5
12	RevNo.Days	Number	780	0	16.67	4229.4	236.93	304.05	6
11	TtlRev	Number	802	0	67.5	21202.6	513.80	1056.88	3
10	Novisits	Number	11	0	0.5	19.5	1.15	0.82	0
9	NoDays	Number	40	0	0.5	65.0	2.38	3.73	0
8	Etracharge	Number	375	0	0	9000.0	44.80	242.19	0
7	RoomTtl	Number	518	0	67.5	16193.07	469.00	902.82	3
6	Purpose of visit	String	4	0	BUSINESS	other			0
5	dtrn of stay	Number	26	0	0.5	65.0	1.86	2.76	1
4	profession	String	62	0	BUSINESS MAN	stenographer			1
3	No.persons	Number	24	0	1.0	35.0	1.81	2.14	0
2	ROOM OCCOPIED	Number	19	0	1.0	27.0	1.53	1.94	0
1	REGION	String	15	0	CENTERAL AFRICA	W.AFRICA			0

Figure 6: Overview report of the dataset

From figure 6, it is possible to know each variables mean, maximum, minimum, data type, cardinality, and missing values if any. Some of these values (mean, maximum, and minimum) help the researcher to set the threshold values of each variable for the analysis of the results.

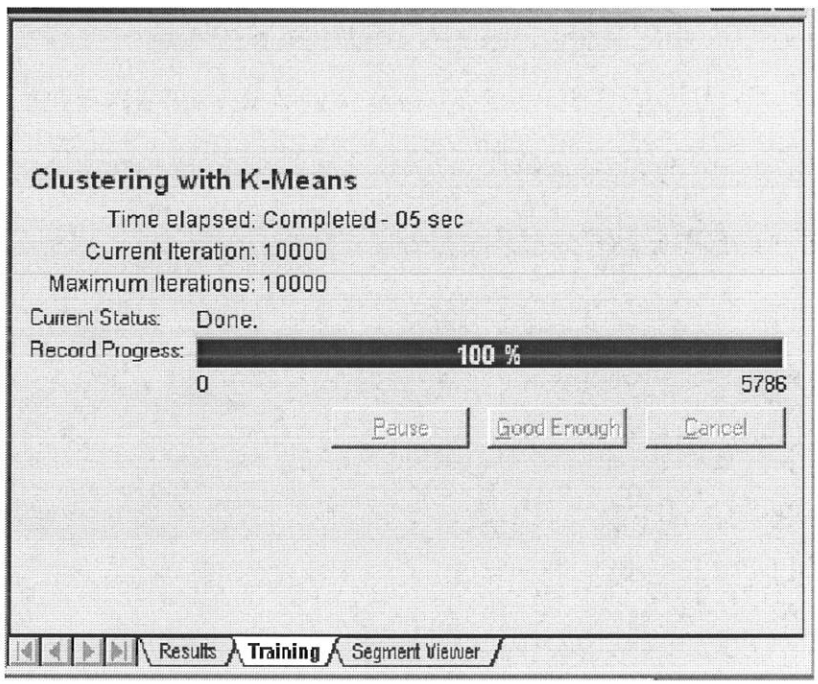


Figure 7: Training the first cluster run

Figure 7 indicates the process of training a clustering model including the time taken to train and the number of iteration used.

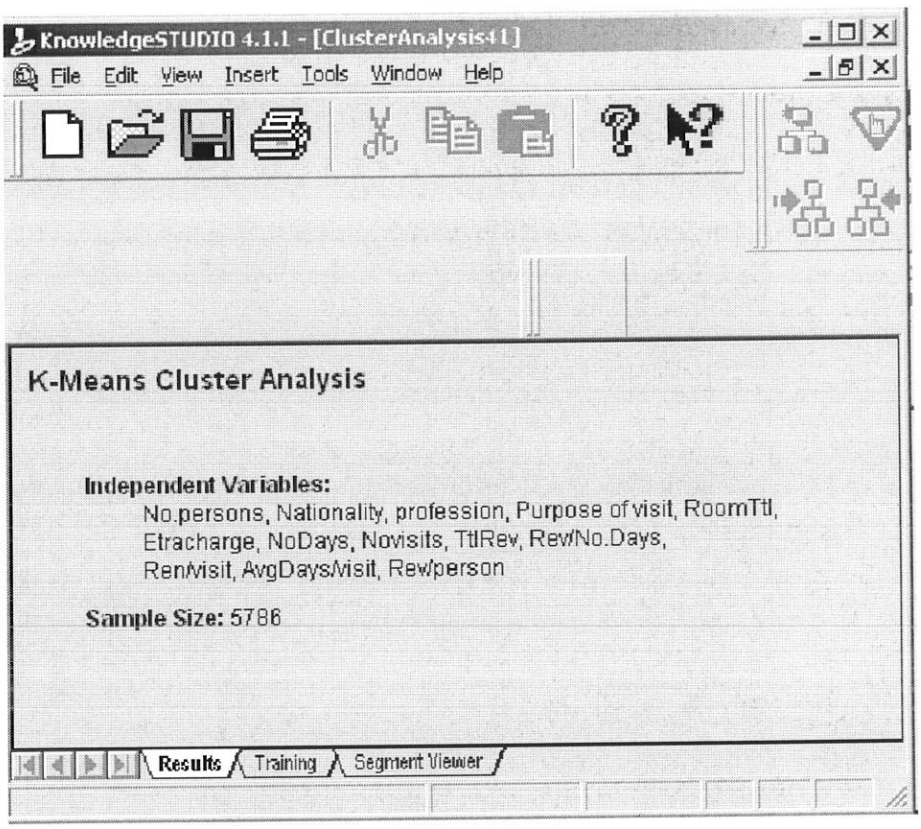


Figure 8: Result of the first cluster run training

Figure 8 indicates the result of training the model, with the variables and the number of records used for model building.

To label all the records in the dataset with the appropriate cluster apply **score** from the **tools menu.** Then a decision tree with cluster index as dependent variable is the output of scoring. The decision tree provides a descriptive classification model of the cluster, which can be used to explore the clusters and detect the characteristics of each cluster.

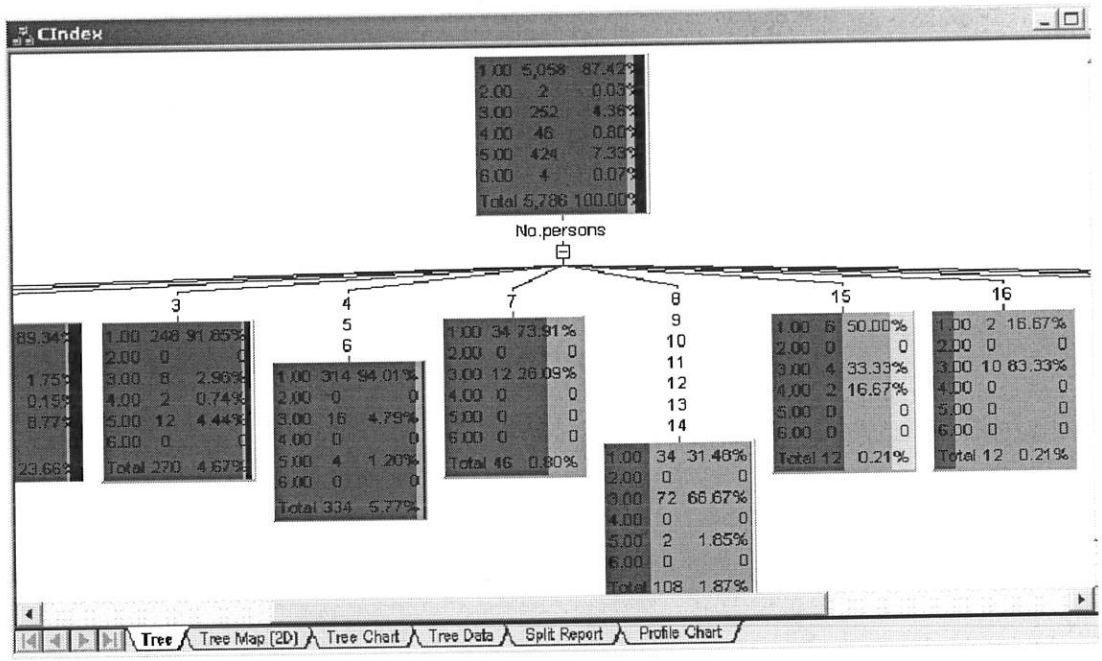


Figure 9: output of scoring the first cluster run

From figure 9 above, it is found that cluster two and six have few records that the researcher suspects the presence of outliers in the dataset, which are far from the normal distribution. Algorithm may treat these outliers as customers behaving differently from other clusters and put them in separate clusters

Then the scored dataset was assessed visually for the presence of outliers. And it was found that there were errors on records having values which were very far from normal distribution of the dataset in cluster two on revenue related attributes and in cluster 6 on the NoPerson attribute. These outliers were removed manually from the dataset.

It is also important to identify the most statistically significant attributes from the dataset and to use these attributes only in model building process to increase the performance of the algorithm. These attributes can be obtained by inserting decision tree from the software and selecting cluster index as dependent variable and the rest as independent variable. Finally the

variables found at the top most of the decision tree were considered for the second experiment

Second Experiment

The variables found to be statistically significant were:

- Region,
- Purpose of Visit,
- RoomTtl (toal revnue generated from room renting)
- NoDays,
- NoVisit,
- TtlRev(total revenue generated)
- Rev/NoDays

Having the above attributes along with the dataset with removed outliers and with same number of K (6), the second experiment was conducted

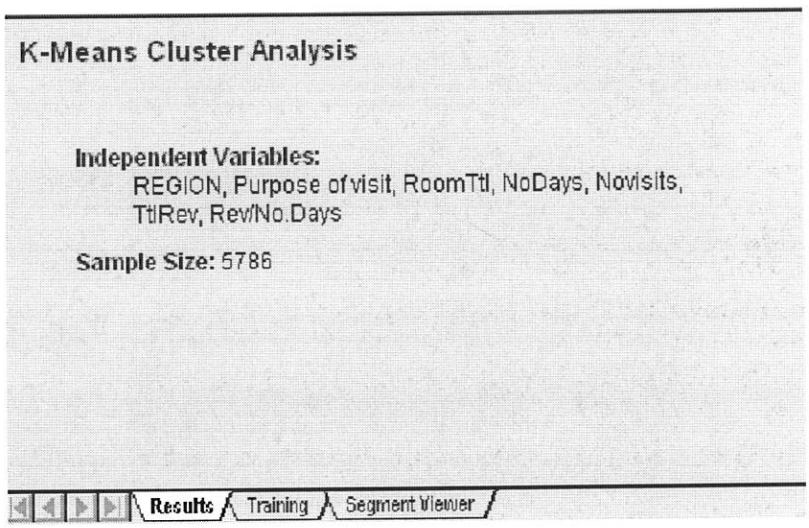


Figure 10: Result of second experiment with the variables used

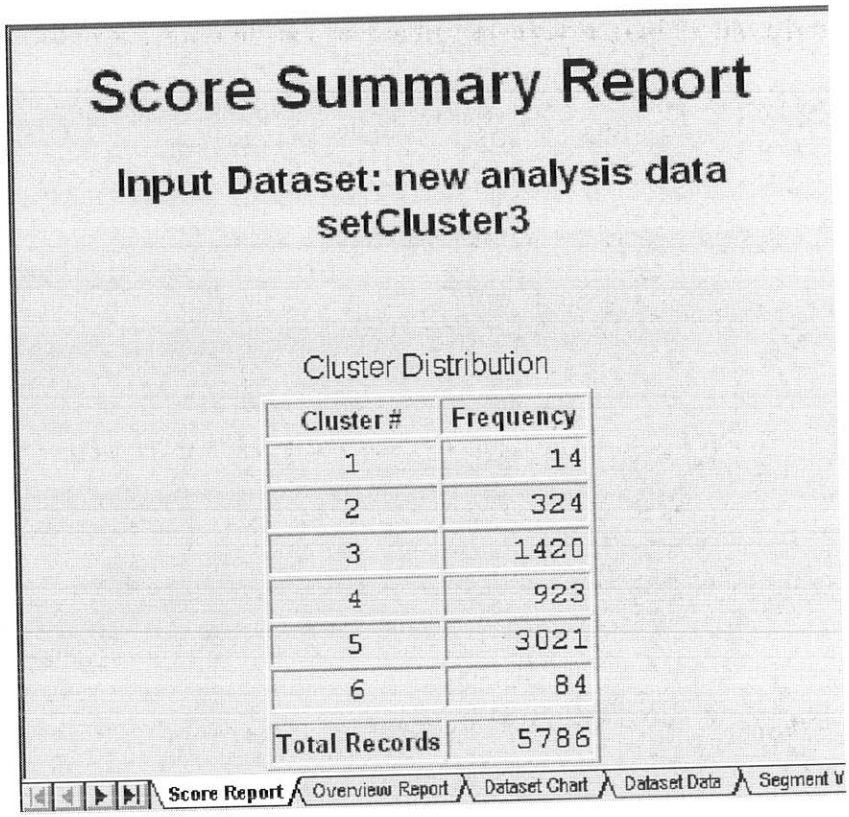


Figure 11:Result of the second experiment scoring

In order to see what patterns are discovered, the researcher used the dataset chart and the overview report from the tool along with the suggestion of domain experts' experience at the hotel for determining the threshold values as follows:

1). **RoomTtl (RT)** – the revenue obtained from room renting

If **RT** is less than or equal to 461 then **RT** is categorized as "low"

If **RT** is greater 461 and less than or equal to 700 then **RT** is categorized as "medium"

If **RT** is greater than 700 and less than or equal to 2500 then **RT** is categorized as "high"

If **RT** is greater than 2500 then **RT** is categorized as "very high"

2). **TtlRevenue (TR)** – the total revenue obtained from each customers (room charge and extra charge)

If **TR** is less than or equal to 511 then **TR** is categorized as "low"

If **TR** is greater 511 and less than or equal to 750 then **TR** is categorized as "medium"

If **TR** is greater than 750 and less than or equal to 3300 then **TR** is categorized as "high"

If **TR** is greater than 3300 then **TR** is categorized as "very high"

3). **NoDays (ND)** – the total number of days stayed by each customers

If **ND** is less than or equal to 4 then **ND** is categorized as "short duration"

If **ND** is greater than 4 and less than or equal to 13 then **ND** is categorized as "intermediate duration"

If **ND** is greater than 13 and less than or equal to 18 then **ND** is categorized as "long duration"

If **ND** is greater than 18 then **ND** is categorized as "very long duration"

4). **NoVisits (NV)** – the total number of visits made by each customers

If **NV** is less than or equal to 1 then **NV** is categorized as "infrequent"

If **NV** is greater than 1 and less than or equal to 3 then **NV** is categorized as "moderately frequent"

If **NV** is greater than 3 and less than or equal to 6 then **NV** is categorized as "frequent"

If **NV** is greater than 6 then **NV** is categorized as "very frequent"

The researcher used "L" for "low", "M" for "medium", "H" for "high", "SD" for "short duration", "ID" for "intermediate duration", "LD" for "long duration", "VLD" for "very long duration", "IF for "infrequent", "MF" for "moderately frequent", "F" for "frequent" and "VF" for Very frequent".

The result of the second experiment is summarized in the table below:

Cluster Index	Freq. records	RoomTtl	Ttlrevenue	NoDays	NoVisit	Purp. Visit	RoomOcp	Remark
1	14	50% H 50%VH	50% H 50% VH	80% ID	VF	Bus & Other	One	95% Domestic
2	324	56% H 24% L 20% VH	53%H 26%L 21%VH	60%ID 22%SD 18%VLD	18% F 82% IF	30% bus 12% tran 51% 0ther	One	76% Domestic
3	1420	85% L	78%L	91%SD	95%IF	Other	One	87% Domestic
4	923	83%L	84% L	85%SD	95%IF	Bus, Other, transit	One	Internation al
5	3021	81% L	79%L	SD	IF	95% tran	one	90% internatio nal
6	84	67% VH 33% H	66% VH 34% H	42% VLD 58% SD	IF	Tran &Bus	80%(8-27)	

Table 5: Summarized result of the second experiment

From the result of the second experiment indicated in table 5, there are very high profitable customers in both cluster 1 and cluster 2, and non profitable customers in cluster 4 having similar patterns with cluster 5. If this model were a good clustering model, these customers showing similar pattern would be together. In addition the number of customers in cluster 1 were also very few (14). Bounsaythip (2001) noted that a cluster should contain

enough customers to develop separate marketing strategy. Domain experts supported the idea of Bounsaythip. Because of the points mentioned, another experiment was conducted by reducing K to 5.

However, the result of the second experiment could still show that domestic customers with duration of stay of more than 4 days are profitable (cluster 2)

Third Experiment

On this experiment, k was set to five and the variables and the training dataset were the same as experiment 2 and number of iteration was also 10,000.

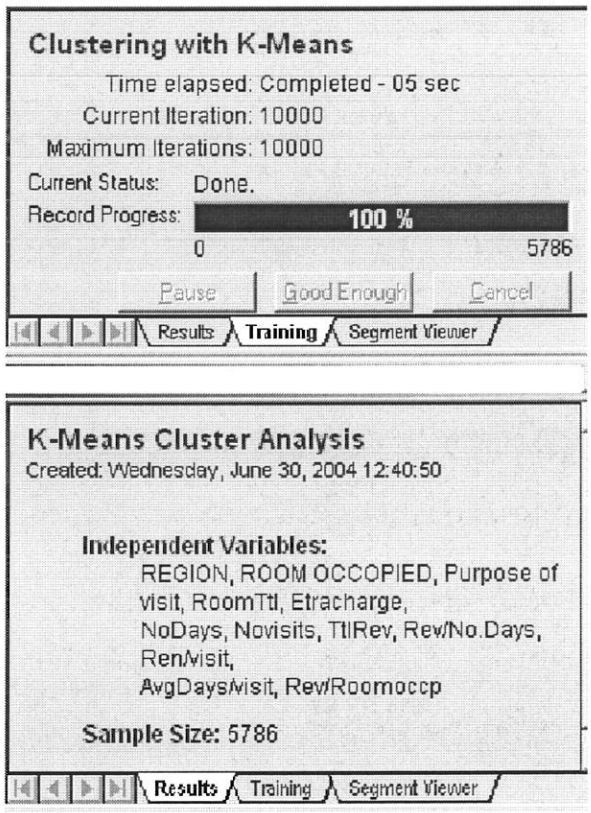


Figure 12:Training and result of training the third experiment

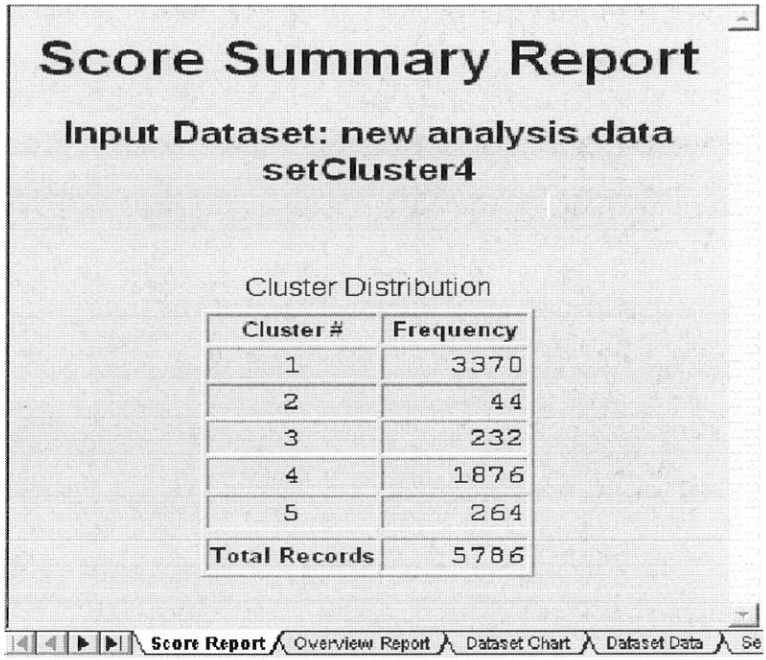


Figure 13: Result of the third experiment scoring

The result of this experiment is summarized in the table below:

Cluster Index	Freq. records	RoomTtl	Ttlrevenue	NoDays	NoVisit	Purp. Visit	RoomOcp	Remark
1	3370	96% L	95%L	SD	IF	92% tran	90% 0ne	International customers
2	44	VH	VH	60%VLD 35%SD	85% IF	50% bus 20% other 27% tran	26% (8-27)	
3	232	80% H	79%H	64%ID 36% VLD	68%IF 32% F	43% Other 67% bus	One	81% Domestics
4	1876	78%L	80% L	85%SD	IF	80 other	One	89% Domestics
5	264	81% H	83%H	86SD	IF	70% tran	88% (2-20)	Internatio nal custmers

Table 6: Summarized result of the third experiment

As indicated, most of the customers were domestic, and international whose purpose of visit was transit. From the table above, it is possible to identify which of domestic and transit international customers were profitable and which were not. Clusters 1 contained customers who were not profitable and international customers whose purpose of visit was transit, cluster 3 contained customers who were profitable and domestic customers, cluster 4

contained not profitable domestic customers, cluster 2 contained different kinds of customers who were highly profitable, and cluster 5 contained profitable international customers.

The result of this experiment looks satisfactory; as it showed different groups of customer segments, and the drawbacks indicated in experiment two were solved (those customers having similar pattern but the algorithm puts in different cluster in the second experiment, were put together in this experiment). .

But to check if reducing the number of clusters may further produce better result, the next experiment with k set to 4 was conducted.

Experiment Four

The dataset, number of iteration, and number of variables (attributes) were the same but k was set to four.

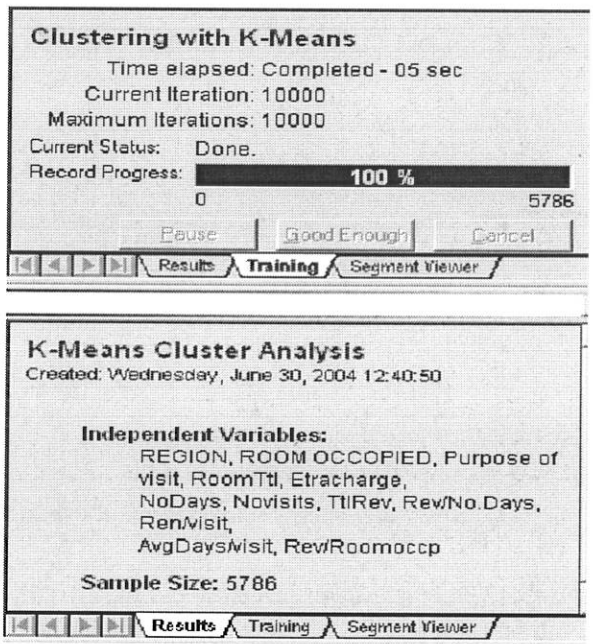


Figure 14: Training result & result of training the fourth experiment

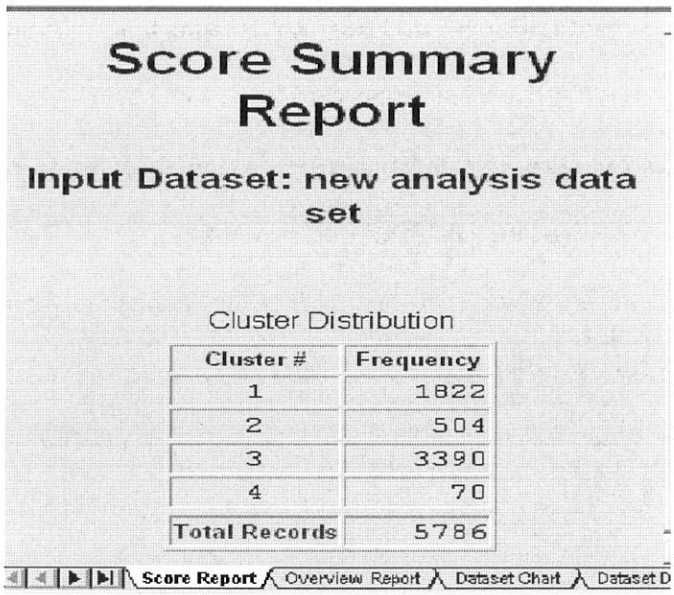


Figure 15: Result of the fourth experiment scoring

The result of the fourth experiment is summarized in the table below:

Cluster Index	Freq. records	RoomTtl	Ttlrevenue	NoDays	NoVisit	Purp. Visit	RoomOcp	Remark
1	1822	76% L	77%L	82%SD	IF	66% other 27% bus	One	90% Domestic
2	504	40%L 42% H	41%L 48% H	74%SD	83% IF	25% bus 34% other 35% tran	40% (8-20)	
3	3390	79%L	77%L	95%SD	IF	83% tran	90% One	81% internation al
4	70	80%VH	76%VH	54%VLD 26%ID 24%SD	20%VF 46%IF	39%bus 22% transit	23%(8-27) 77% one	

Table 7: Summarized result of the fourth experiment

From the above table it can be understood that cluster 2 did not indicate fully defined relationships on the revenue generated. That means 40% of them were grouped under "low" and 42% under "High" and the rest "medium". Further more, even though it is possible to identify non-profitable domestic customers (cluster 1) and international customers whose purpose of visit is transit (cluster 3). However, it is not possible to know profitable domestic and transit international customers.

Choosing the best clustering model

Four experiments were conducted to reach to the appropriate segmentation model. The first experiment indicated that very few numbers of records were assigned in cluster 2 and 6.

Because of these the researcher suspected outliers and outliers were found. In addition the patterns discovered were not meaningful and satisfactory.

The second experiment with the same number of K (6) but with removed outliers was conducted, and the result was indicated in table 5. From the summarized result in the table, there were customers showing similar patterns but categorized or grouped in different clusters. For example, there were profitable customers (who were intermediate duration, occupying one room, and domestic) belonging together but were grouped some of them in cluster 1 and the rest in cluster 2. In addition the first cluster contained very few customers. For this cluster it is not feasible to develop a separate marketing strategy.

The third experiment by reducing K to five was conducted and again the result was summarized in table 6. As it could be seen from table 6, there were five clusters behaving differently. There were defined separations or differentiations among the five identified segments and homogeneity or similarities with in the clusters. All the segments were meaningful to understand customers and give information for developing target market strategies.

Finally, the last experiment with K set to four was conducted. And the result was not satisfactory as of the third experiment. Cluster 2 as indicated in table 7, did not indicate fully defined relationships on the revenue generated. And it was not possible to identify profitable domestic and transit international customers.

So the researcher, along with the comments and suggestion of domain experts decided that appropriate numbers of clusters were five and hence the third experiment was the selected experiment showing good segmentation of Ethiopia Hotel customers. This clustering model

was used further for decision tree building input dataset. The interpretation of the result of the third experiment (when K=5) is summarized in the next section.

Interpretation of the Clusters Obtained from the Third Experiment

- **Cluster 1:** most of the customers were under this cluster, 95% of the customers generated low revenue, stay short duration, 92% of the customers' purpose of visit were 'transit' and Room occupied by these customers was one. In general these customers can be categorized as international transit non-profitable customers.
- **Cluster 2:** contained very small number of customers (44) but they are very profitable type of customers staying very long duration and 80% of these customers occupied 8-27 rooms (coming in group). In this cluster their purpose of visit can be either "transit" or "business" or "other". The meaning of this cluster as the domain experts supported is when the purpose of visit is "business" or "other", they stay very long duration and when their purpose of visit is "transit" the number of room occupied by each customer was usually between 8 and 27(customers coming in groups) and they are usually international customers. These were very profitable type of both domestic and international customers.
- **Cluster 3:** most of the customers in this cluster (80%) generated high revenue and 64% of them stayed, 5 up to 13 number of days (intermediate duration), some of them (32%) visited the hotel repeatedly (very frequent), their purpose of visit were 'other' and 'business', and 81% of them were domestic and occupy only one room (coming in single). In summary they were profitable domestic customers.
- **Cluster 4**: customers in these group constituted most of the population, generated low revenue, stay short duration, their purpose of visit was either 'other' or' 'business'

and 81% of them were domestics customers. In general they were non-profitable domestic customers.

- **Cluster 5:** 83% of them generated high revenue, stay short duration, 75% of them were transit and 88% of them occupied more than one rooms (coming in group). They can be generalized as profitable international transit type of customers.

In summery almost all customers don't show the characteristic of loyalty except in cluster 3 where some of them visit the hotel repeatedly (32% of the cluster). The hotel management should examine why customers don't visit the hotel regularly.

4.5.4 Classification Modeling

Building clustering model can do the task of segmenting the potential customers successfully. But the remaining is how to classify new customer records to different identified groups. This specifies the problem of classification based on a defined class. The decision tree can help us to solve this classification problem.

During this phase the dataset with cluster index generated from the clustering model where the records are segmented into meaningful groups, was used to build a decision tree. Decision tree was used to derive rules explaining the assignment of new records to the correct cluster using the tool. The cluster index was used as the dependent variable and the rest variables were used as independent variables.

The total database was used to construct the decision tree. The decision tree has been implemented in Knowledge Studio by partitioning the database into training, and testing datasets. Therefore, 70% of dataset was used for training and 30% of the dataset was used for testing.

Total dataset	Training set (70%)	Testing or Evaluation set (30%)
5786 records	4050 records	1736 records

Table 8: Partitions of datasets used

There were to two issues considered while building decision tree classification model. These were:

- Handling continuous values
- Determining how to construct a decision tree

4.5.4.1 Handling continuous values

As it is shown in section 4.3.2, in table 1 of data description part, there were some attributes with continuous values. Even though decision tree can handle both continuous and discrete values it will be effective and the rules generated from the decision tree can be more meaningful and understandable provided if these continuous values were adjusted into discrete values.

For building the classification model, the variables used for building the clustering model and additional variables (region, profession, and purpose of visit) from the guest registration form were included. Including as many variables as possible will help make the rules more understandable and improve the accuracy of classifying the new records to the identified cluster indexes. The objective of the classification model was to generate rules that explain the assignment of new records to the already identified customers. The attributes used for building classification model were:

Field Name	Data Type
Region	Text
Profession	Text
Purpose of visit	Text
RoomTtl	Number
Rev/NoDays	Number
NoDays	Number
Novisit	Number
Room Occupied	Number

Table 9: attributes used for classification model building

In addition to the observed relationship of the variable with variables with the segment types obtained, the researcher has discussed the issue with the experts in the field and adjusted the attribute the attribute values into discrete values. As a result the attribute RoomOccp has three values; '1', 'between 2 & 4', 'greater than 4'. The attribute Number of Stay (duration of stay) has four values; 'less than or equal to 4', 'between 5 & 13', 'between 13.5 & 18', 'greater than 18'. Rev/NODays has four values 'greater than 600', 'between 600 & 201', between 200 & 102, 'less than 101. Finally, the attributes RoomTtl and NoVisit are adjusted according to the ranges used for building clustering model.

4.5.4.2 Determination of decision tree construction

Knowledge studio supports three ways of constructing decision tree:

- By automatically growing a tree – it helps to automatically grow the complete classification tree while freeing you to do other things while the tree is growing. However, we must be careful while using this feature. KnowledgeSTUDIO only gives and then operates on the strongest statistical

split that it finds on any given node; this may or may not be the information that is required.

- Step by step procedure of constructing a tree through find split or force a split to guide constructing a tree to the information searched for
- Combining both - The automatic feature can be invoked anytime and it will carry on the tree development from where we left off. Any previous Find Split or Force Split results above the current node will remain in the tree and influence the tree's growth.

The researcher first built models by growing a full decision tree by forcing the first splitting variable into one of the different variables used for building the classification model using the 70% of the training data set. Second, the same fully grown decision tree as a result of automatic grow was then manually improved either by pruning some of the irrelevant nodes or adding nodes by forcing split to guide the growth of the tree to the information required.

Then validate it with the evaluation set which was the 30% of the dataset allocated. Finally by comparing the accuracy of the validation set which is automatically grown tree and a tree with some of nodes pruned and some nodes added, the better one was selected for generating classification rules. Therefore, eight models by altering the splitting variables of the dataset used and another eight respective models by improving the nodes either by pruning some of the irrelevant nodes or adding important nodes, a total of sixteen models were constructed.

Nodes were pruned here;

- When it is required to guide the construction of the decision tree to the pattern or relationship required,
- When the lower splitting node generate the same pattern with the node above,
- When a certain node is extremely splitted without increasing any new pattern, and
- When the same splitting variable is repeatedly used in the decision tree.

And nodes are added

- When it is important to guide the construction of decision tree to the pattern or relationships required.

Trained decision tree with splitting variable 'RoomTtl' and its validation with the evaluation set are shown in the figure below for illustrative purpose.

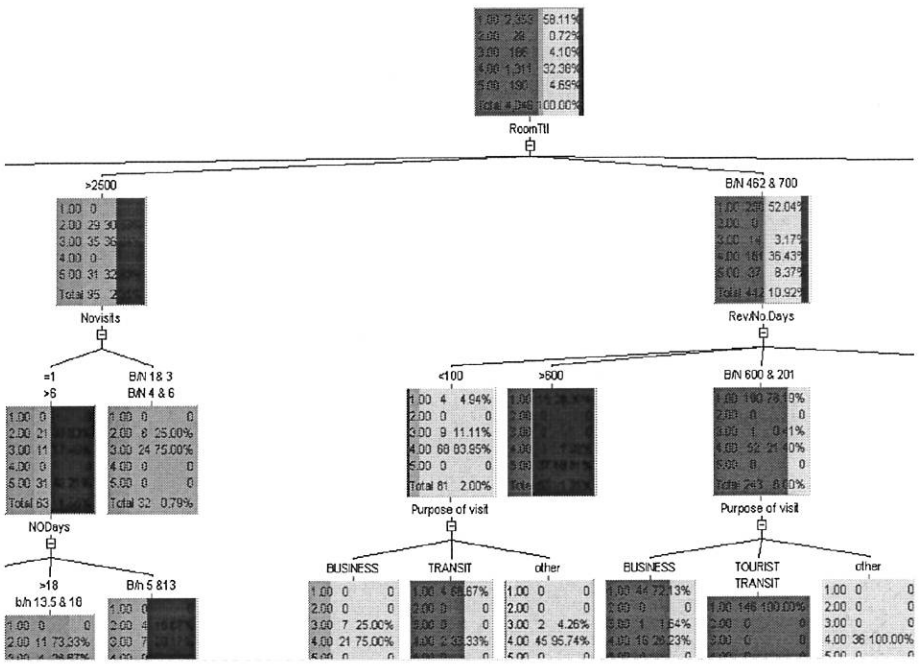
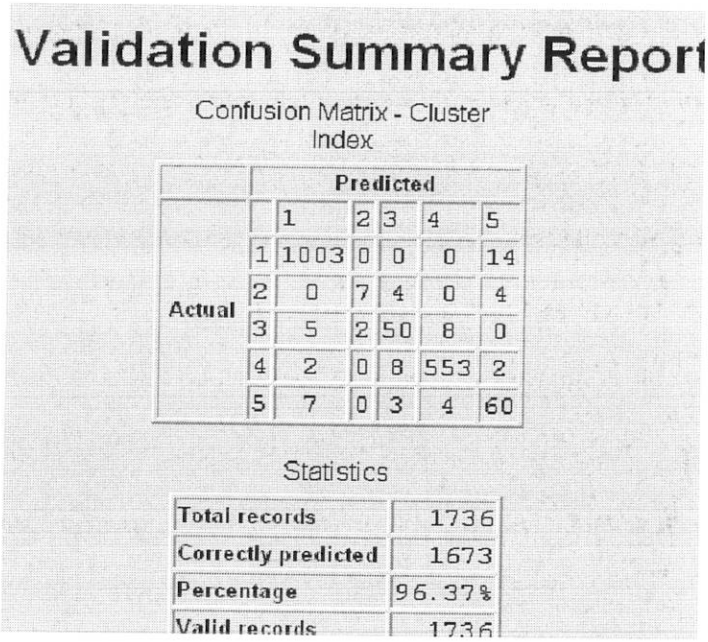


Figure 16: Partial View of trained decision tree with the splitting variable 'RoomTtl'



Validation Summary Report

Confusion Matrix - Cluster Index

		Predicted				
		1	2	3	4	5
Actual	1	1003	0	0	0	14
	2	0	7	4	0	4
	3	5	2	50	8	0
	4	2	0	8	553	2
	5	7	0	3	4	60

Statistics

Total records	1736
Correctly predicted	1673
Percentage	96.37%
Valid records	1736

Figure 17: Confusion matrix obtained as a result of validation of the decision tree model with splitting variable 'RoomTtl'

The above figure is presented to indicate the accuracy of one of the decision tree model with splitting variable 'RoomTtl' among the sixteen models.

The result of the first eight models with different splitting variables obtained by automatically growing the tree is summarized in the table below:

Model	Splitting variable	Pruned nodes	Added nodes	Testing accuracy
1	Region	-	-	95.16%
2	Purpose of visit	-	-	95.62%
3	Number of days	-	-	96.25%
4	Room occupied	-	-	95.39%
5	Number of visit			95.51%
6	Room total			96.37%
7	Renenue/No.Days			95.54%
8	Profession	-	-	89.86%

Table 10 Summary of results of training and testing of decision tree modeling

The above fully-grown decision tree as results of automatic grow was then manually improved either by pruning some of the irrelevant nodes or adding nodes by forcing split. The result of validating these models' result is indicated in the table below:

Model	Splitting variable	Pruned nodes	Added nodes	Testing accuracy
1	Region	22	25	93.32%
2	Purpose of visit	13	20	95.77%
3	Number of days	15	26	96.41%
4	Room occupied	8	20	96.08%
5	Room Total	17	14	96.72%
6	Renenue/No.Days	15	12	95.62%
7	Number of visits	26	29	95.62%
8	Profession	9	20	89.90%

Table 11:Summary of results of training and validation of pruned decision tree modeling with evaluation set

As it can be seen form the above table the same model grown automatically when some irrelevant nodes are pruned and important nodes are added improved the accuracy except the model with splitting variable 'Region'. That means decision tree grown manually can help to classify customers to identified customer segments better than the automatically grown decision tree except the model with the splitting variable 'Region'.

As it indicated on table 11, Model 5 with splitting variable 'RoomTotal' shows good accuracy (96.72%). As Mitchell(1997) noted evaluating a learned hypothesis is very important to consider the accuracy with which it will classify future instances. So the researcher selected the best model that shows greater testing or evaluation accuracy, which is the model with the first splitting variable "Room Total" which is about the total revenue obtained from room rent by each customer. In addition domain experts also agreed that the

variable 'room total' is a more determining variable for measuring customer profitability. Hence, this model is used to generate rules for assigning new customer records to the identified segments. Some of the rules generated are annexed for reference.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1. Conclusion

In this research the different concepts of CRM were reviewed, and one can understand that CRM is the best marketing strategy for acquiring, retaining and partnering with selected customers to create super value for the hotel. Data mining techniques also plays significant role for effective CRM especially in extracting important knowledge about customers.

Data mining, processes and principles were employed on Ethiopia Hotel customers' data to identify the different customer groups. In this research data understanding, data preprocessing, and modeling were undertaken. The software Knowledge studio was used for this purpose. Both clustering and classification algorithms were available in the software specified. The K-Means algorithms and decision tree algorithm were used for this study. Since there was no defined number of k for best clusters, different attempts of K were made i.e. K=6, K=5, K=4. Best result was found when K was 5. The five identified clusters were meaningful and different as indicated on the interpretation of clusters part of this chapter. In summary:

- The first cluster was characterized by international, transit, non-profitable customers
- The second cluster was found to be very profitable type of both domestic and transit customers
- The third cluster was characterized by domestic, profitable customers
- Th fourth cluster was found to be domestic non-profitable customers

- The fifth cluster is found to be international, transit, profitable customers

Based on the result, different marketing strategies like customer retention and promotion strategies can be developed. Important variables for identifying valuable customers were also known but the best variable obtained was "RoomRevenue" (revenue obtained from room renting).

After identifying the group of customers by assigning different cluster index (1, 2, 3, 4, 5), decision tree was trained so that rules used for assigning new customer records to the already identified segments were generated and some sample rules are indicated in the appendix. The rules generated are then used for a simple classification prototype development. The first important splitting variable was found to be "RoomRevenue" which ensures that the most important variable for customer clustering was the total revenue obtained from room renting.

To conclude, data mining techniques especially clustering and classification techniques were employed to create customer segments and to generate classifying rules for new customer records to these segments. The result gives vital information about the customers' behavior. The behavior of the first cluster can be explained as customers who stayed in the hotel starting from half day to four days only, coming to the hotel individually (number of rooms occupied was one), and the purpose of visiting the country (hotel) was "Transit"(waiting for another flight). They were generally international, non-profitable customers. The second cluster was a group of different kinds (either domestic or international) of customers who were highly profitable. In this cluster there are customers, staying very long period of time with high average revenue for each day, when the purpose of visiting the country (hotel) is either "other" or "business", and staying few number of days

(up to 4), and generating very high average revenue for each day coming in group (occupying many rooms (above 8)), when the purpose of visiting the country (the hotel) is "transit". The third cluster's behavior was explained as domestic customers staying many days in the hotel, coming individually (occupying one rooms), showing some degree of loyalty (32% of them visited the hotel repeatedly), purpose of visiting the country or the hotel was either "business" or "transit", and generating high total revenue. The fourth cluster was a group of domestic customers staying very few days with no history of repeated visits to the hotel, purpose of visiting the country (hotel) is "other", coming individually (occupying one rooms), and generating low revenue. And finally the behavior of the fifth cluster is explained as international customers who stayed in the hotel starting from half day to four days only, purpose of visiting the country (hotel) is "transit", coming in group (occupying many rooms (above 8)), and generating high revenue with high average revenue for each day.

If the demographic variables such as income and age would have been incorporated, segments could give better information about the value of customers. In conclusion, data mining can, if it is used correctly, indicates different customer segments, more importantly customer types that had not been recognized before. As a result, important strategies can be developed to meet the needs of customers to increase their satisfaction and to make them loyal and profitable.

5.2. Recommendation

As it is explained in the objective part of the thesis, the research is conducted for academic purpose, but at the same time it has significance to the business organization especially to hotels. It can be a way or hint as to how data mining can be used to support CRM.

Hotels like any business organization have to take in to consideration their customers' behavior so as to be competent in the market. Knowing customers' behavior will help hotels differentiate profitable from unprofitable customers, target their profitable customers, and retain loyal customers.

The following recommendation made by the researcher in relation with (CRM):

- Ethiopia Hotel being a large hotel has a lot of customers and hence should start to implement CRM strategies. There is no behavioral segmentation even though attempt is made to cluster customers based on only demographic data.
- Relationship programs like long duration or reward programs should be supported by integrated customer data warehouse and touch points (touch points refers to the many ways in which customers and the hotel interact) should be integrated with recent technology like Internet.
- Separate department should be assigned to deal with different customer complaints.

Any business organizations have a lot of business transactions undertaken on a daily basis. The data available is an important resource of the business organization. It contains and reflects activities and facts about the organization. By applying data mining very important knowledge that can contribute a lot in decision making for business organization can be extracted from data. The researcher suggests the following relating with data mining:

- The hotel should build an integrated, time-variant and subject oriented data warehouse. It should include guest history, guest reservation, sales information database, visit history, and guest complaint data.
- The hotel should invite researchers who can integrate data mining with CRM so that effective marketing strategies can be developed.
- The hotel should consider what important CRM strategies can be developed based on the result obtained in this research.

Further research should be conducted by including additional factors that the researcher could not include due to some limitations. In connection with this, the researcher gives the following suggestions:

- The datasets used for model building should be large enough
- The result obtained by using K-Means algorithms should be validated by other clustering algorithms.
- Variables like age and income of each customer should be incorporated and the purpose of visit of each customer should give specific information about their real purpose of visit.
- Different data mining tools which are mostly used for customer segmentation analysis like Clementine and Intelligent Miner should be reviewed and used for analysis.

6.3 The Customer Classification System: A Prototype

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

In this study an attempt was made to develop a simple application prototype named Customer Classification System that uses the classification rules generated from the decision tree learner in the classification sub-phase of this chapter. The prototype is used to classify a customer into one of the customer clusters, search for a customer and find the cluster where the customer belongs, and also provides with the description of each customer clusters. The Customer Classification System contains MS access database, the MS visual basic program hosting the classification rule. The prototype interface and the rules used for building it are annexed in the appendices.

BIBLIOGRAPHY

Ahola, J and Rinta-Runsala, E.(2001) Data Mining Case Studies in Customer Profiling. *Version 1. VTT Information Technology.*

Anderson, E.W., Fornell, C. & Lehmann, D.R. (1994)Customer satisfaction, market share, and profitability: *Journal of Marketing, 58: 53-66.*

Baraggoin,C., Andersn,C., Bayerl,S., Bent,G., Lee,J &Schommer,C.(2001) Mining Your Own Business in Telecoms: Using DB2 Intelligent Mine for Data. IBM Corp. http://www.ibm.com/

Basgoze, A and Gokturk, M. (2003) Building Customer Profiling Using data mining techniques. *International Turkish symposium on Artificial Intelligence and neuralNetwork-Taining*

Berkhin,P. (2000) Survey of Clustering Data mining Techniques. *Accrue Software, Inc.*

Berry, M.J and Linoff,G(1997) *Data Mining Techniques For Marketing Sales and Support.* 2nd edn, New York: John Wiley & Sons

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

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

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

Bowen, J. T. and Shang-Lih, C. (2001). The relationship between customer loyalty and customer satisfaction. *International Journal of Contemporary Hospitality Mangement.13 (59): pp1-7*

Brause, R., Langsderf, T and Heep, M. (1999). Neural Data Mining for Credit Card Fraud Detection. *J.w. Goethe-university, Frankfurt a.m.*

Carbone, P. (1997). Data Mining or "Knowledge Discovery In Databases": An Overview. http://www.mitre.org/pubs/data_mgt/Papers/DMHdbk.pdf

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

Ethiopia Hotel enterprise. *2001/2-budget year annual report*

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

France, T., Yen, D., Wang, J-Cand Chang, C-M (2002). *Information managing and computer security.10 (5): pp242-254*

Garano, L.M& Raggad, G.b(1999). *OCLC systems and service* .15 (2): PP 81-90 http://www.twocrows.com.

Getty, J.M. and Thompson, K.N. (1994). The relationship between quality, satisfaction, and recommending behavior in lodging decision. *Journal of Hospitality and Leisure Marketing.2 (8): pp 2-22*

Giudici, P. (2003). Applied Data Mining. New York: John Wiley & Sons, Inc.

Goebel, M. and Gruenwald.L (1999). A Survey of Data Mining and Knowledge Discovery Software Tools. SIGKDD Explorations. June, 1999. *Online. Internet http://www.kdnuggets.com*

Han, J. and Kamber, M. (2001) *Data Mining: Concepts and Techniques.* San Francisco: Morgan Kaufmann Publishers.

Holshenra & Siebes (1994). *Introduction to data mining*

Jembere, D (2003). The application of Data Mining to support CRM at Domestic Airlines. Addis Ababa University. *Unpublished Master's Thesis*

Kandampully Jay and Suhartanto Dwl (2000). Customer loyalty in the hotel industry: the role of customer satisfaction and image. *Journal of Cotemporary Hospitality management.12(6). Pp: 346- 351*

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

Lejeune, A.P.M. measuring the impact of data mining on churn management. *http://www.mcbup.com/research-register*

Levin, N. and Zahavi.J (1999) Data Mining.Online. Internet. http://www.kdnuggets.com.

Magnini,V. Data mining for hotel firm: Use and Limitation. Focus on research. *Cornel Hotel and Restaurant Administration quarterly, 2003.*

McDonald, M. *(2001)* Advantage Marks 20th birthday. *Travel Weekly. Online. Internet. httlp://www.findarticles.com*

Mcintosh, R., Goeldner, R.c. & Ritchie, B.J. (1995) *Tourism: principles, practice philosophies.*

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

Min,H.K., Min, H.S & Emam, A. (2002). *International Journal of cotemporary Hospitality management. 14(6):pp: 274*

Mitchell, T. (1997). *Machine learning.* Singapore, McGraw-Hill International Edition.

Parvatiyar, A and Sheth, J (2001). Customer relationship management. Emerging practice, Process, and Discipline. *Journal of Economic and social Research .3(2): pp 1-34.*

Piatetsky-Shapiro, G. (2000). *Knowledge Discovery in Databases*. Online. Internet. http://www.kdnuggets.com/gpspubs/sigkdd-explorations-kdd-10-years.html

Pritchard, M.P, and Howard, D.R (1997). The loyal traveler: examining a typology of service patronage. *Journal of Traveler Research.35 (4): pp 2-11.*

Qin He (1999). A review of clustering algorithm as applied in IR. University of Illinon at Urbana.Champaign

Rafalski, E. (2002). Using data mining/data repository methods to identify marketing opportunities in health care. *Journal of consumer marketing.19 (7): pp 607-613.*

Reichheld, F and Sasser, W.E (1990). Zero defections: quality comes to service. *Harvard Business Review. 11;pp105-111*

Ryals, L. (2003). Journal of Targeting, measuring and analysis for marketing.11 (4); pp 343-349.

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

SPSS Inc.(2002). Data mining modeling. *Online*
Internet. http://www.spss.com

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

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

Witten, I.H. and Frank, E. (2000) *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* San Francisco: Morgan Kaufmann Publisher

Wiwattanacharoenchai, S and Srivihok, A. (2003). Data Mining of Electronic Banking in Thailand: Usage Behavior Analysis by Using K-Means Algorithm. *Part IV Internet and Information Technology for Greater Mekong sub region (GMS) business*

Wobishet, H (2002). The application of Data Mining to support customer relationship management at Domestic Airlines. *Addis Ababa University. Unpublished Master's Thesis*

Rule #6:

If Revenue from room charge is equal to B/N 462 & 700 and Average revenue for each day is B/N 600 & 201 and Purpose of visit is equal to BUSINESS or other and REGION is equal to ETHIOPIAN then Cluster Index will be 4.

Rule # 7

If Revenue from room charge is B/N 462 & 700 and Average revenue for each day is B/n 200 & 120 and REGION is equal to CENTERAL AFRICA, E.AFRICA, S.AMERICA, SOUTHERN AFRICA or W.AFRICA then Cluster Index will be 1.

Rule # 8:

If Revenue from room charge is B/N 462 & 700 and Average revenue for each day is B/N 600 & 201 and Purpose of visit is equal to TRANSIT and REGION is equal to CENTERAL AFRICA, CHINA, E.AFRICA, EUROPE, INDIAN, MIDDLE EAST, N.AMERICA, S.AMERICA, SOUTHERN AFRICA or W.AFRICA then Cluster Index will be 1

Rule # 9:

If Revenue from room charge is B/N 701 & 2500 and REGION is equal to ETHIOPIAN Average revenue for each day is B/n 200 & 120 and Purpose of visit is equal to other and NODays is >18 or b/n 13.5 & 18 then Cluster Index will be 3,

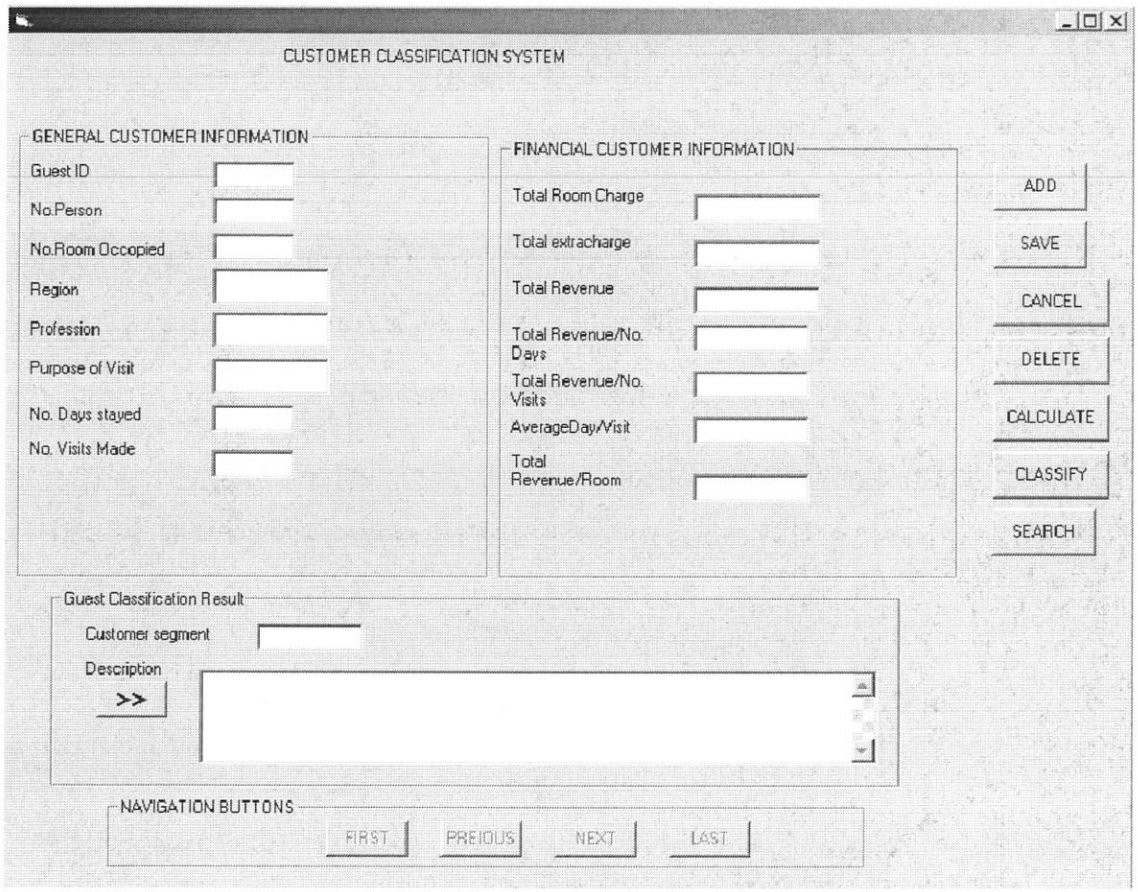
Rule # 10:

If Revenue from room charge is >2500 and REGION is ETHIOPIA and Novisits is >1 and NODays is equal to >13.5 then Cluster Index will be 2.

Rule # 11:

If Revenue from room charge is <=461 and REGION is equal to ETHIOPIAN and Purpose of visit is equal to TOURIST or other or business and NODays is equal to <=4 then Cluster Index will be 4.

An Interface of the Classification Prototype



A procedure used to calculate total revenue, average revenue per day, average revenue per visit, average revenue for each room, and average duration of stay per each visit (AVNODAYS)

```
Private Sub CMDCALACULATE_Click()
    Dim EXTRACHARGE As Currency
    Dim ROOMCHARGE As Currency
    Dim TOTALREVENUE As Currency
    Dim AVERAGEREVENUE, AVNODAYS, REVPERROOM, RENPERVISIT As
    Currency
If Val(Txtnnodays.Text) = 0 And Val(Txtnovisits.Text) = 0 And Val(Txtroomoccp.Text) = 0
     And Val(Txtroomcharge.Text) = 0 Then
    MsgBox " make sure that these values(No.Days Stayed, No.Visits Made, Total Room
    Charge)are not zero", vbInformation, " Input Information"
Else
    EXTRACHARGE = Val(txttextracharge.Text)
    ROOMCHARGE = Val(Txtroomcharge.Text)
    TOTALREVENUE = EXTRACHARGE + ROOMCHARGE
    AVERAGEREVENUE = TOTALREVENUE / Val(Txtnnodays.Text)
     RENPERVISIT = TOTALREVENUE / Val(Txtnovisits.Text)
    AVNODAYS = Val(Txtnnodays.Text) / Val(Txtnovisits.Text)
    REVPERROOM = TOTALREVENUE / Val(Txtroomoccp.Text)
    txttotalrevenue.Text = TOTALREVENUE
    txtrevenuepervisit.Text = RENPERVISIT
```

```
    Txtaverageday.Text = AVNODAYS
    Txtavgrevenue.Text = AVERAGEREVENUE
    Txtrevenueperroom.Text = REVPERROOM
End If
End Sub
```