FLIGHT REVENUE INFORMATION SUPPORT SYSTEM FOR ETHIOPIAN AIRLINES

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 Gobena Mikael

May, 2000

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION STUDIES FOR AFRICA

FLIGHT REVENUE INFORMATION SUPPORT SYSTEM FOR ETHIOPIAN AIRLINES

BY

GOBENA MIKAEL

Name and Signature of Members of the Examining Board

Ato Getachew Birru, Chairman, Examining Board Getachew Birru

Ato Tesfaye Biru, Advisor

Dr. John Cowell, External Examiner John R Cowell

DEDICATION

I would like to dedicate this paper to:

My father and mother, Mikael Imru and Almaz T/Hawariat, who have taught me the value of education throughout my life;

my wife, Lili, who has continuously supported me in pursuing it; and

my son, Rasselas, for whom I hope to pass on the culture.

ACKNOWLEDGEMENTS

I would like to thank Ato Tesfaye Biru for his encouragement and constructive comments and who has been a very stimulating and challenging supervisor; Ato Yared Tilahun and Wrt. Netsanet Admassu for their invaluable assistance in data preparation and testing; Wro. Senait Tilahun for her unfailing care, support and patience throughout my study; Ato Samson Yemaneab and Ato Girma Desalegn for their assistance in searching and providing the necessary software used in this study; and Wrt. Beza Asfaw and Wrt. Wubalem Tefera for whom I am indebted to in so many ways.

TABLE OF CONTENTS

List of Figures

List of Appendices

ABSTRACT

Ethiopian Airlines is a profit-oriented business organization whose objective is to provide the maximum value to its customers, consistent with the need to make some return on each transaction. One of the major primary activity at the airline is Sales. In addition, because commercial organizations only survive by identifying and satisfying the market, Marketing Services is also regarded as a major primary activity.

This study focuses on the revenue process within the Sales and Marketing operation. In particular, it aims at understanding the critical business functions and processes involved in the flight revenue process, to identify and assess the availability of revenue data elements and develop a model for Ethiopian Airlines that will support information on revenue realized by flight and forecast revenue by flight; accurately and timely.

Ethiopian Airlines has numerous state-of-the-art application systems and as a result retains a vast amount of data in its different databases. However, it has failed to make good use of this data and has not been able to use it to create competitive advantage. As a result, the revenue information model has been developed using data mining techniques. In particular, the neural network model was used to train, test, validate and develop the prototype model. The ultimate objective being to find out the suitability of data mining applications to the Ethiopian Airlines problem.

Since the scope of the study is limited to a single organization, the major method that has been used to assess revenue information needs of users is case study; implemented through interviews (planned discussion), questionnaires, observation and document analysis.

After reviewing the various areas that are affected by the Sales and Marketing operation; Sales, Scheduling, Pricing, Revenue Management and Airport Operations have been identified as the critical functions in the revenue process. As a result the focus of the study has been on these functions.

Survey results reveal that of the 5 most important information required by the concerned airline managers, revenue related information ranks on top with 31% of respondents ranking it first. In addition, 84% of respondents rate flight revenue information as either one of the most or the most critical information, 88% as either very or extremely strategic, and 94% as one that would provide opportunity to gain competitive advantage.

The major revenue related data elements identified during the study are advanced booking data, post departure data, schedule data, and revenue data. These revenue related data elements are available within the existing system, but are scattered in the various application systems. Over one year's historic advance booking data is available, over two years' post-departure data is available, and historical flight revenue data since April, 1997 is available.

After selecting a suitable software to build a revenue information model, the revenue related data elements identified were collected for 8 flights and a comprehensive testing was conducted. The test included 6 different experiments using the back propagation network and radial basis function neural network models, 3 different sets of independent variables and a multitude of training parameters. The experiments produced 327 different models which were compared and evaluated and finally one was selected to represent the revenue information model. The developed model, with an average of 33-37% error rate, is only a preliminary or initial step towards, hopefully, more detailed work in this area.

I am confident that through a selection of more fields and with more historical data, the error may be able to be reduced to users' requirement of 5-10%. It is, therefore, my belief that this research has some contribution to further research in this area. It has been able to successfully demonstrate that data mining applications can be an alternative approach to build information systems; especially for complex problems having vast amount of data and high interaction among non linear variables. Others can pursue similar research using different types of data mining applications, including other neural network models. I hope that some of the problems I encountered and the methodologies I used will help to shed light and guide others undertaking similar studies.

Chapter 1

INTRODUCTION

1.1 Background

1.1.1 Ethiopian Airlines

Ethiopian Airlines (ETHIOPIAN) is an international commercial air transport company involved in the carriage of goods and passengers. It was conceived through a commercial agreement signed between the Ethiopian Government and Transcontinental and Western Airlines, today renamed and known as Trans World Airlines (TWA), on the 8th of September, 1945. It was formally founded with an inaugural flight from Addis Ababa to Cairo on the 8th of April, 1946. The following year Aden and Bombay were added to Cairo as International destinations (Ethiopian Airlines: Bringing Africa Together, 1988).

Although ETHIOPIAN started with a meager authorized capital of USD 2.5 million (ETB 5 million) and 1 Douglas airplane, it was a significant leap in integrating the nation's political and economic goals. Today its capital is ETB 688,605,000 and its annual revenue is nearly ETB 2 billion with a strong working force of 3600 (Ethiopian Airlines: Annual Report, 1998).

ETHIOPIAN's primary objective has always been to emerge as Africa's leading carrier and foster speedy and effective communication between its peoples for the ultimate development of the continent. In line with this objective, route expansion and fleet modernization have been a continuous strategy for the airline.

Consistent with its strategy, its fleet size has steadily grown to include, today, 21 passenger and 3 cargo aircraft.

By 1955, two Douglas DC6s and three 36-seater convair 240s had been added. December, 1962 marked a historical landmark for ETHIOPIAN as it entered into the jet age through the purchase of two Boeing 720s providing the possibility for long range operations. Boeing 727s, 767s and 757s were added to the fleet in the 1970s, 1980s, and 1990s respectively. Its latest addition is a 767-300 aircraft acquired on a lease basis in 1999.

Likewise, its route expansion has grown from three international destinations in 1947 to include today 50 international and 26 domestic destinations. Its latest destinations, incorporated in 1999, include Copenhagen, Zanzibar and Maputo.

Region	Destinations	Weekly Frequencies
Africa (30)	Abidjan, Accra, Addis Ababa, Bamako, Brazzaville, Bujumbura, Cairo, Dakar, Dar-es-salaam, Dire Dawa, Djibouti, Entebe, Harare, Johannesburg, Kampala, Khartoum, Kigali, Kilimanjaro, Kinshasa, Lagos, Lilongwe, Lome, Louanda, Lusaka, Maputo, Nairobi, Ndjamena, Niamey, Zanzibar	66
America (2)	New York, Washington	4
Europe (5)	Athens, Copenhagen, Frankfurt, London, Rome	14
Gulf, M.E & Asia (13)	Bahrain, Bangkok, Beijing, Beirut, Delkhi, Dubai, Jeddah, Kharachi, Mumbai (Bombay), Muscat, Riyadh, Sanaa, Tel Aviv	34

Table 1.1: Current International destinations served by ETHIOPIAN and weekly frequencies (Source: ETHIOPIAN world wide timetable)

Throughout the 1980s and during most of the 1990s, ETHIOPIAN has remained true to its objective and its motto 'Bringing Africa Together' to emerge as the leading and most respected airline in Africa. It currently operates the largest network in Africa with 30 destinations, flies to more destinations in Africa than any other airline in the world, is the only airline with a daily flight across Africa from East to West and operates to all corners of Africa - East, West, South, North and Central.

Since recent times, however, the airline has also been focusing on strengthening services outside of Africa and accordingly has modified its motto to "Linking Africa to the Rest of the World." This is characterized by its hub and spoke operation with Addis Ababa as the hub and the various African, Asian, European, and American destinations as the spokes. Today, ETHIOPIAN operates 118 international flights every week and carries about 600,000 international passengers annually, of which over 60% transit through Addis to other destinations. Its nature of operation puts ETHIOPIAN in direct competition with all sorts of airlines - small, large, powerful, advanced.

ETHIOPIAN's mission statement emphasizes on safety, profitability, route expansion, service, customer relationship, technology and development. As a result, the general objectives of the airline focus on:

- Enhancing customer service
- Expanding its network
- Raising and maintaining employee morale
- Upgrading corporate image and technology
- Improving profitability

ETHIOPIAN, therefore, can be characterized as a profit-oriented business organization whose objective is to provide the maximum value to its customers, consistent with the need to make some return on each transaction.

Figure 1.1 depicts the organizational chart of the senior management of the airline. Divisions shown on the chart are further divided each into several departments; which in turn are divided into sections.

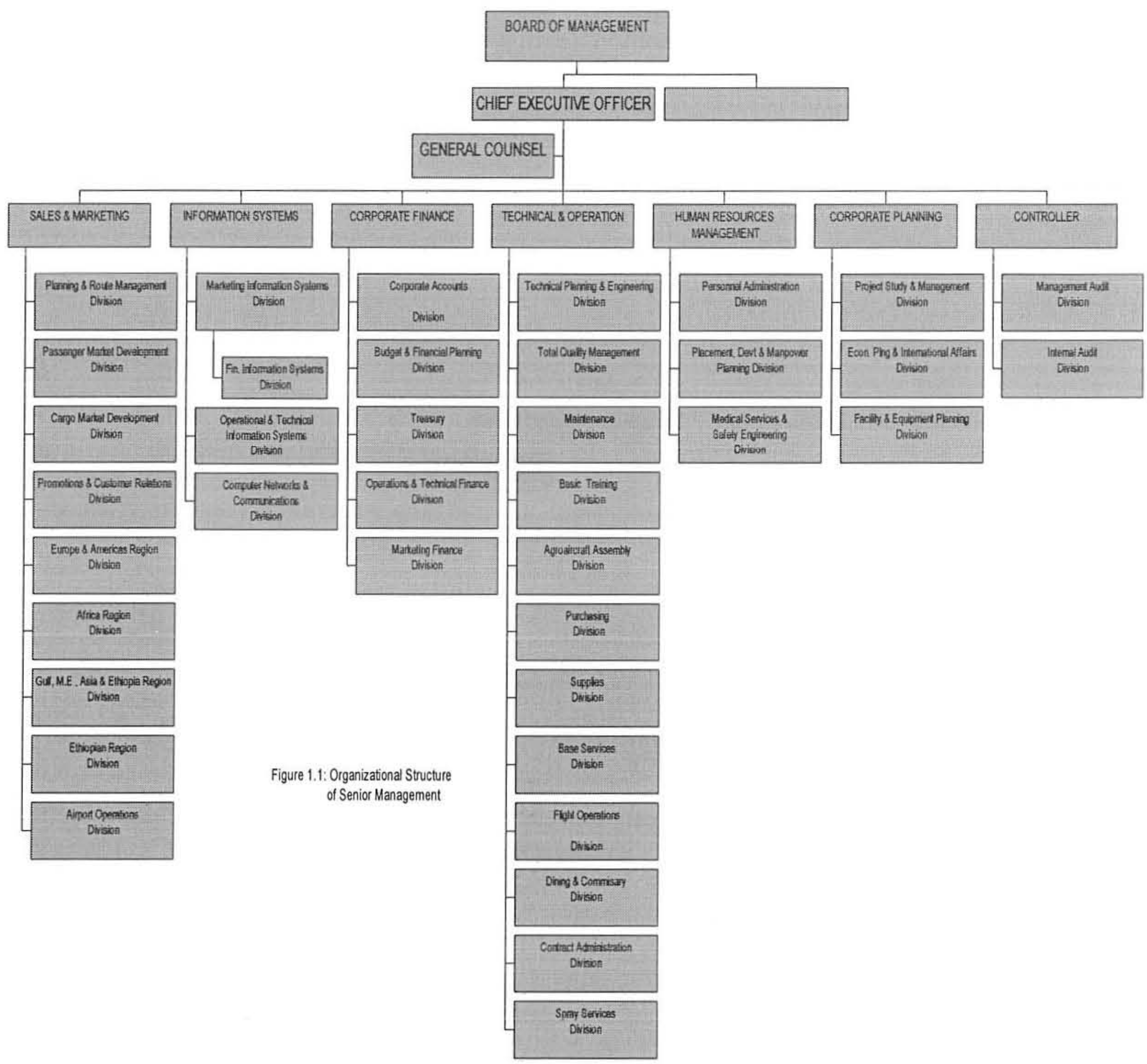


Figure 1.1: Organizational Structure of Senior Management

1.1.2 Information Systems in Organizations

Information systems are only one of a myriad variety of systems which may operate within an organization. An important prerequisite to understanding the specific role and operation of information systems in an organization is an appreciation of the significance of the term system and of the major components of any system; irrespective of whether the system relates to information or some other resource in the organization.

Bhattacharya (1994) defines the term system as, "a complex entity made up of functionally and structurally related components (parts) of various orders, functioning in a recognizable environment; in order to achieve a definite objective; being equipped with a feedback mechanism meant to regulate its productivity." Rowley (1990:1) adds that a system, "is concerned with taking inputs or resources, executing some form of regulated change and achieving results or outputs."

Broadly, therefore, a system is an integrated organization of resources with an input, a process, and an output, functioning within a certain boundary, to achieve the same objective.

The four main components of a system are, therefore:

Inputs: Resource material which might take the form of people, finance, energy, or information (data) that the system receives.

Process: Activity that takes place as a response to the inputs received.

Outputs: The modified resources, typically an enhanced value of the resource inputs.

Boundary: Context or Framework within which a system operates.

An Information System is a particular type of system whose input consists of capturing and gathering raw data; process consists of manipulating, storing and transforming the raw data; output consists of producing and disseminating useful information; and feedback consists of mechanism that modifies the input or processing activities to achieve the desired goal. The goal is to provide useful, accurate, timely, reliable and affordable information that will help people in problem-solving, decision-making and coping with life in general.

Information systems transform data into useful information. In the context of an organization, Whitten et al. (1998:38) state that, "an information system is an arrangement of people, data, processes, interfaces, and geography that are integrated for the purposes of supporting and improving the day to day operations in a business, as well as fulfilling the problem-solving and decision making information needs of business managers." An information system is typically concerned with providing data on a particular situation or problem of concern to the user or organization.

Organizations essentially comprise of functions, activities, processes and structures to carry out their objectives. To assist in performing these activities and to manage the various functions, processes and resources involved, organizations have developed a series of procedures, tools and techniques. In addition, organizations also develop a variety of structures, by dividing the organization's functions, to improve the effectiveness and efficiency of resource utilization and management of the processes.

The activities of the organization can be broadly divided into primary activities and support activities. Primary activities are those that are directly concerned with the production and delivery of the product or service to the customer or client, and support activities are those that are not directly involved in the production process but play an essential role in facilitating and supporting the primary activities.

One of the major primary activity at ETHIOPIAN is Sales. In addition to Sales, service delivery activities such as Airport Services, In-flight Services and Customer Relations are also regarded as primary activities since they directly deal with the customer. Because commercial organizations only survive by identifying and satisfying the market, Marketing Services are also regarded as a primary activity.

Information has always been a source of competitive advantage in the business context. The real change that enhances the potential value of information in an organization is the ability of organizations to exploit this source of advantage through the use of new technologies. An effective information system in an organization is one which minimizes the amount of raw data that is processed without becoming information; that is, one which reduces or eliminates the contents of the data box in figure 1.2 and increases the contents of the information box.

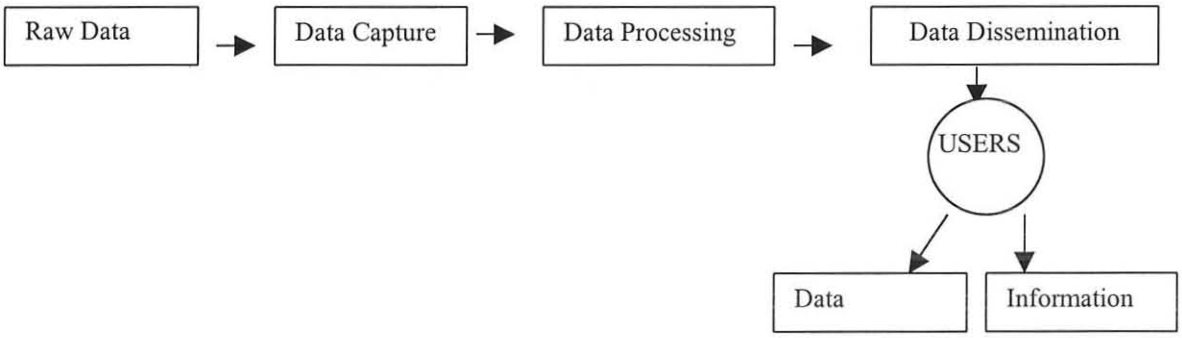


Figure 1.2 Simplified model of the data processing process (Eardley, et al. 1995)

The conversion of data to information will occur only when the data received by the user enhances his or her knowledge or understanding of a situation. Users would need to recognize that the data received is relevant to their particular situation and they must have sufficient prior knowledge and experience to be able to use the data to enhance their knowledge and understanding. Hence, value of information in an organization is a function of user understanding and relevance of data to the user. One can categorize the value of information as shown in Table 1.2

Level of user understanding	Relevance/importance to user's situation	Information value
None	None	None-remains as data
None	High	None-remains as data
None	Critical	None-remains as data
Some	None	Low value
Some	High	Moderate value
Some	Critical	High value
Good	Some	Moderate value
Good	High	High value
Good	Critical	Maximum value

Table 1.2: Value of information

In the Sales and Marketing area, ETHIOPIAN has acquired Passenger reservations, cargo reservations, ticketing, fare quotation, travel information, hotel booking, car rentals, departure control, baggage tracing, yield management and frequent flyer program systems. In the Operations and Technical area, the major systems include maintenance, engineering, inventory management, crew planning, flight planning, crew tracking, crew planning and weather reception systems. For finance and administration it has acquired passenger revenue accounting, general accounting, human resources management, cargo revenue accounting and payroll systems. The headquarters is connected with 11.5 kms of 10 base T cables and a Fiber optic backbone. This local area network (LAN) has a collision detection mechanism and serves 650 clients and 10 servers. In addition, these systems can be accessed by ETHIOPIAN offices outside the headquarters. Addis city offices, selected domestic stations, and all international city and airport offices are connected through a Wide Area Network (WAN) with modern modems and routers, fiber optics, and satellite links. Over 400 workstations and numerous servers are connected on the WAN.

1.2 Statement of the Problem

The airline industry today can aptly be characterized as highly volatile and competitive, with erratic demand, variable pricing, downward pressure on yield, and demanding customers. In the earlier days, this industry was very stable with high regulation in prices and standards. However, the deregulation of the US market in the late 1970s seems to have changed this forever. New airlines mushroomed to challenge the established ones; price wars and cut-throat competition ensued. The effect of the deregulation quickly affected the European carriers as they slowly started deregulating their markets too.

Eventually, the whole world was affected and ETHIOPIAN was not immune to this roller coaster ride.

As the competition intensified, complex fare structures and complex schedules developed. In addition, players in this industry found themselves trying to balance fixed capacity spread across a network with variable price and variable demand and they had not only shorter time frame to make these decisions but also incurred greater risk from their decisions. Gone were the days with fixed prices, point-to-point schedules and flexible decisions.

With increasing complexity, typically, the airlines which survived were those that were able to innovate and quickly adapt to the changes. In the 1970s computerized reservations systems were developed which greatly assisted airlines to effectively manage their inventory. For survival in the 1980s new ideas were in demand. As a result, concepts such as yield management and frequent flyer programs emerged to manage the complexity. Such new concepts, however, required immense amount of data that were to be processed with computerized information systems.

These and many other systems became critical in improving the bottom line of a company, and in fact its survival. Consequently, ETHIOPIAN has been at the forefront to introduce new information technology and, as a result, today has numerous state-of-the-art information systems and communications networks to run its various operations effectively. Most of the information systems acquired by ETHIOPIAN, however, are applications systems typically designed to serve specific purposes, such as reserving passenger seats, rewarding customer sales, quoting itinerary fares, issuing automated tickets, optimizing fare-class bookings, checking-in passengers at the airport, or to make sales accounting.

During the preliminary survey, users have strongly expressed that the information inputs from such application systems are very limited and inadequate to make some critical decisions such as optimal resource allocation, establishment of competitive prices, rationalization of schedules and restructuring marketing mix. A severe handicap is the unavailability of timely information on the performance of a flight in terms of passenger revenue realized and contribution to the bottom line.

Currently, information on revenue is provided in two ways. One is through a daily flight performance report that shows passenger load carried by each flight departed on the previous day. However, it is not possible to estimate, within a reasonable margin of error, the total revenue generated by a flight using this data. Users' requirements regarding acceptable error margins for flight revenue information as expressed during the survey was a maximum error of 10% for the forecast and 5% for the realized revenue information.

The other revenue information source is the monthly financial report which provides aggregate revenue information of all flights. This report, according to users, is not to the detail required, i.e. by flight, and is always 3-5 months late, after all the necessary processes and accounting transactions have been completed. During the initial survey, conducted in October, 1999, the latest report available on hand with the airline was that of April, 1999 activity. All of the decision makers interviewed and questioned during the preliminary survey agreed that the acceptable delay to make strategic use of this information is a maximum of one week after flight departure and that it should be at flight level.

Another problem that I came across while investigating the current system for which the airline currently had no information on is forecasting passenger flight revenue. If management was provided the forecast of revenue to be realized by flight one month or more ahead of departure it could proactively act to modify some of its marketing strategies.

In view of the foregoing, the study aims at tackling the following three major problems:

1. Untimely or delayed provision of information on the passenger revenue generated by each flight.
2. Inadequacy of the detail of information provided on passenger revenue.
3. Unavailability of forecast information on passenger flight revenue that would enable management to assess performance ahead of time.

In the process of analyzing the current system several problems were identified. The pricing decision process is especially an interesting one and I have, therefore, included a discussion surrounding this issue in appendix 1.

1.3 Justification

We have seen that the airline industry is a highly competitive industry. In such a business environment airline managers must be equipped with the right tools and skills to gather information to help them develop, price, promote, distribute and sell their products in line with customers' requirements and satisfaction and to increase market share and revenues.

Passenger flight revenue information provides a means to quickly understand the effect of certain problems in order to timely establish the relevant causal factors and take the necessary corrective measures. However, in order to determine that data has been successfully converted into information, i.e. information has high value, the flight revenue information must be relevant to the particular situation of the users and they must be able to use the information to enhance their decision.

Out of the 5 most important information required, revenue related information ranks on top with 31% of respondents ranking it first. In addition, 84% of respondents rate flight revenue information as either one of the most or the most critical information, 88% as either very or extremely strategic, and 94% as one that would provide opportunity to gain competitive advantage. Rating the usefulness, 79% said that, if provided the day after departure, revenue realized by flight information is either high or very high; 92% gave the same weight to the revenue forecast information. In addition, 63% rated the current revenue information provided three months late as either useless or of very low use.

Referring to Table 1.2, the value of the flight revenue information to the identified user group can be categorized as having high to maximum value. The response to the survey shows that the level of user understanding to enhance their knowledge through this information is good and relevance to their situation is high to critical. Users have noted many problems that have been encountered due to the absence of such information. These include, misconception on revenue generated by using old yield information, inability to react to low performance in time to avoid loss and take corrective actions, inability to properly evaluate performance, inability to organize the necessary support base and make accurate operational decisions.

Flight revenue information is an invaluable resource for supporting the short and medium-term tactical decision involved in the ongoing operations of the airline. It also provides a potential source of added value to its clients and customers. In addition, this information also contributes significantly to the airline's strategic or long-term development. Obviously, if management can obtain information on potential developments and trends and future forecasts, it will go a long way in assisting to reduce risk elements which long term planning entails. The majority of the respondents believe that flight revenue information can create opportunities for optimal capacity and resource allocation, better planning, proactive actions, redefining sales and marketing strategies, and maximizing revenue.

The development of a flight revenue forecasting system is a potentially important strategic planning tool that would enable ETHIOPIAN to adapt its sales and marketing strategies to the more competitive global environment within which it now operates. ETHIOPIAN no longer operates in a stable environment with simple fare structures and schedules. Revenue information provides an opportunity to gain a competitive advantage because strategies are not easily observed and copied by competitors. A key aspect of the model we are attempting to build, therefore, is that it has the potential to reshape strategies in pricing, scheduling, distribution, revenue management, and sales.

With increasingly complex competitive schedules from the development of "hub and spoke" systems and the use of average cost pricing with multi-level pricing structures, ETHIOPIAN management needs to rely on decision support tools, such as a flight revenue information system, to manage information and provide timely and accurate analysis. Forecast flight revenue information and immediate information on revenue realized by flight have not only been identified by users as crucial for strategic planning, but also as a problematic area that needs an urgent solution. According to the Chief Information Officer, the managing board of Ethiopian Airlines and the CEO have repeatedly raised their concern on the delay and inadequacy of information provided on passenger revenue. Hence, high management support is expected for such a research.

Eardly, et al. (1995) note that the overall use of information within any organization is for recording transactions, decision making, planning, performance measurement, control and communication. The importance of revenue information in terms of these criteria can also be observed.

- The entire decision-making process is based on capturing and processing information. Hence, the importance of flight revenue information to the decision-making process is evident. Revenue information, in general, supports the decision-making process.

- The measurement of actual performance against planned or budgeted performance is a fairly common technique which is used in most of the standard planning approaches. Realized flight revenue information helps fulfill this role.

- In the reactive approach to control, the organization's management responds to existing deviations from the planned activities or performance levels. Managers will act to correct the deviations, or to prevent those deviations from damaging the future interests of the organizations. Realized flight revenue information helps fulfill this role.

- In the proactive approach, management consciously explores the potential sources of future deviations with a view to instituting the necessary avoidance or damage limitation measures in advance. Forecast flight revenue information helps fulfill this role.

- Flight revenue information will provide the ability to control a situation. Since symptoms of the deviation from planned or anticipated performance targets will be clearly signaled in time, the necessary trigger to initiate the decision-making activities within the control process will be effected.

In a nutshell, therefore, the role of revenue information is to provide input to:

- Improve efficiency in resource usage
- Reduce cost
- Develop added value services
- Gain competitive advantage
- Improve effectiveness
- Enhance profitability

Finally, I believe this study will help to open the way for more work and opportunity in the provision of all types of information to marketing management and provide a model for similar carriers. Furthermore, the prototype model to be developed on the basis of the findings of this study can be further developed into a full fledged operational system that can greatly assist in the efforts of the airline in competing more effectively with more resourceful and wealthier carriers of the developed countries. In his response to my questionnaire, the Sales Manager for Lufthansa at Addis Ababa noted that their airline receives passenger flight revenue realized information within one month of flight departure and forecast on flight revenue three months ahead of departure. However, their system is purely for in-house purpose and is not commercially available.

1.4 Objective

Marketing research, whose key objective is to provide reliable information for effective marketing decisions, is a vital function to the survival of such an organization as ETHIOPIAN. "Marketing research", Chisnell (1991:6) asserts, "is concerned with the systematic and objective collection, analysis, and evaluation of information about specific aspects of marketing problems to help management make effective decisions." To this end, the overall goal of this research is to develop a flight revenue model that will assist the marketing research process at Ethiopian Airlines.

The flight revenue model that is to be developed through this research will try to minimize the amount of data passing through the hands of the user without becoming information as shown in figure 1.2. The goal of the model is to minimize redundant data and increase the value of the information by using data produced by the various existing application systems sitting in their databases and turning them into information – valuable flight revenue realized and forecast information.

Instead of setting out to develop a system through the conventional manner following the systematic procedures prescribed by modern system analysis and design methodologies, I decided to use data mining techniques and see how I can use available data lying around in the company to reach the same objective and maybe discover new relationships or offer enhanced functionalities.

This research has two parts – a business and a technical part – and hence two general objectives.

i. To understand the critical business functions and processes involved in the flight revenue process, survey the availability of data, and assess any problems and opportunities that will affect the development of a flight revenue information model.

ii. To develop an information system or a model for ETHIOPIAN that will support information on revenue realized by flight and forecast revenue by flight accurately and timely.

Specific objectives include the following:

- To investigate the revenue information needs of users.
- To survey the business processes involved in generating flight revenue information.
- To assess the available data at ETHIOPIAN that will support flight revenue.
- To review data mining technologies and tools (software) with a view to select an appropriate algorithm and architecture for the research problem.
- To train, test and evaluate the selected data mining algorithm using different training parameters.
- To develop a model (a prototype system) that would help to forecast flight revenue information, and provide timely information on realized revenue by flight.

1.5 Scope and Limitations

Because ETHIOPIAN is a large organization, the scope of this study has been limited to the airline's marketing function. This function has been selected because of its importance to the overall objective of the airline. "Marketing," Lavin (1992:398) notes, "requires an enormous amount of up-to-date information for planning and decision-making purposes...Marketing is all-encompassing, covering every function connected with bringing a product or service to the customer, from setting prices to determining distribution channels. The need for current accurate information appears at every stage of the marketing process."

Due to the vastness of marketing research and the information required, the study is limited to passenger flight revenue information requirement, henceforth referred to as simply revenue information. Within marketing the business research has focused primarily on those critical functions which contribute directly to the revenue of flights and the generation of flight revenue information. These functions are Scheduling, Pricing, Distribution, Sales, Revenue Management, and Airport Services. Furthermore, due to time limitations, data on only eight flights has been collected and comprehensive testing was conducted on only one representative flight.

The development of an information support system has been limited to the development of a prototype model using data mining tools. The research has been limited by access to data mining tools. The search for a suitable software to build the desired model was time consuming. A relatively satisfactory tool was finally obtained with the assistance of colleagues living abroad.

Another limitation was access to certain databases. Collection of some pertinent data was restricted and required special approvals. Furthermore, historical data was limited to one year reducing the total amount of records available with which to train the model. Lastly, when administering questionnaires to other airlines operating to Addis, the response was very unsatisfactory. Most did not want to share their experience in this area as they thought it to be strategic. As a result, comparative study could not be made.

1.6 Methodology

In-line with the objective of the study, the procedure that has been used in conducting the research is descriptive and applied approach. The following data sources and methods have been employed.

a) Data Source and Subjects

Large amount of the data has depended on primary sources. Concerned executive managers, marketing senior, operational and middle managers of Ethiopian Airlines and staff involved in the revenue process have been included. In addition, relevant books, manuals, reports, the Internet and other documents have been consulted.

b) Data Gathering Tools and Instruments

Since the scope of the study is limited to a single organization, the major method that has been used to assess information needs of users is *case study*. The case study was implemented through interviews (planned discussion), questionnaires and document analysis. The questionnaire was administered in different forms; for the different managerial levels. For a more in-depth analysis, non-structured interview and close observation of selected users and operational areas have been conducted to gather additional data.

Details of the questionnaire and interview guides, sample responses, and the response rate are found under appendix 2. Although all the questions administered are noted under appendix 2a, no single group was asked to respond to all the questions. Different subsets of these questions were administered to the four different groups.

c) Sampling and Analysis Procedures

There are nine executive managers (including the CEO), nine senior marketing divisional managers, about sixty marketing area sales managers and about thirty middle marketing departmental managers in the airline.

Questionnaires and Interviews have been distributed to 30% of the candidates in the population using purposeful, quota and cluster/area sampling methods. This consists of two executive managers, nine senior divisional managers, and twenty three area sales and middle managers. One representative flight number has been selected for training and building the models and seven more for scoring and validating the models using cluster and simple random sampling methods.

Data obtained from the respondents has been analyzed using basic statistical techniques.

d) Systems Development Methodology

Data mining techniques and in particular neural networks using multi layer perceptron back propagation network and radial basis function architectures are used to train, test, validate and develop the prototype model.

1.7 Organization of the Thesis

The thesis is organized into five chapters. The following chapter discusses the business survey conducted. The major component of flight revenue information, fare classes (which are the primary source data or the fields used to build the revenue information model) are discussed. This part also covers the results of the questionnaires and interviews, type and availability of data necessary for the research, and overview of the critical functions involved in the revenue process.

In chapter three a discussion on the technology used to develop the prototype system, namely data mining, is presented. An overview of data mining, the concept, the process and the various models are discussed. Neural networks, in particular, are presented in more detail.

The technical part of the research is covered in the fourth chapter. The collection, cleansing and preprocessing of the required data is explained. Using different architectures of a neural network and a multitude of training parameters, the training and learning process used for experimenting is described. The chapter concludes with an analysis of the data and discussion of results. Finally, a concluding chapter is presented with a concluding remark and recommendations.

Chapter 2

FLIGHT REVENUE: BUSINESS SURVEY

2.1 General

This business survey is intended to provide an overview of the prevailing revenue environment at ETHIOPIAN. The purpose is to describe the concept of flight revenue, conduct an analysis of the current flight revenue process, identify the critical functions and activities involved, and identify and assess the availability of data that can support the revenue information model I am attempting to build. The focus of the business survey shall, as a result, concentrate only on those sales and marketing activities that have significant relevance to the revenue process.

The major element to successfully conduct the survey is to obtain information by collecting facts about user requirements and the environment; mostly through interviews. Bingham and Davies (1984), however, caution that "although interviewing is the most important fact–finding technique, it is not the only one, and consideration must be given to the other sources of information available to a systems analyst."

Whitten et al. (1998) define fact-finding as, "the formal process of using research, interview, questionnaire, sampling, and other techniques to collect information about systems requirements, and preferences." Fact–finding techniques include sampling of existing documentation, forms, and databases; Research and site visits; Observation of the work environment; Questionnaires; Interviews; Rapid Application Development (RAD); Joint Application Development (JAD) and Prototyping (Whitten et al. 1998).

The appropriateness of a particular fact-finding technique highly depends on the type of project or research. As a result, all except the last three techniques listed above have been applied. In addition, detailed discussions were conducted with all levels of staff, as well as representatives of all departments that interact with or support the revenue generation process, either directly or indirectly.

The facts and opinions that needed to be collected and from whom were first determined. Based on the above, free format, multiple choice, rating and ranking questions were used in developing the questionnaire format which was administered to middle and lower levels of management. The Interview was constructed using both open-ended and closed-ended questions and was administered at a higher level to include eleven senior managers. The total response rate was around 88%.

A pilot questionnaire was first developed and distributed to a large sample of the population. For candidates located at distant locations I used the e-mail through the ETHIOPIAN WAN. Two problems were identified; the first was that the response rate was very low and the second was that many of those who answered misunderstood quite a lot of the questions. I attribute the problems to following four factors.

1. My follow-up was very weak: I made the mistake of assuming the candidates would answer without urging on my part.

2. The questions were not properly designed: I had too many open-ended questions. I found out that candidates prefer to answer rating, ranking and multiple choice questions.
3. There were too many questions. In addition, for some candidates the majority of the questions were irrelevant.
4. I never prepared the candidates ahead of time of my intention.

This preliminary exercise, however, gave me an invaluable experience when designing the main questionnaire. As a result, for my main questionnaire and interview, I made sure to hold appointments with candidates ahead of time and formally discussed the subject matter with them before administering the interview or mailing the questionnaire. I also grouped the candidates into four groups with similar characteristics. I designed my questions to be precise and minimized open-ended questions. Thirty eight questions were finally developed; which were grouped relevant to each group and individual. As a result, no one candidate had over 15 questions to answer. I shortened the interview time of those higher up in the hierarchy to ensure their co-operation. Under appendix 2, I have attached a sample of the instruments used to conduct the survey.

In this chapter the results of the case study is presented. Since the findings are a combination of discussions, interviews, observations and document analysis, the results are reported in a form of presentation and discussion throughout this chapter.

Following is a discussion of the important results of my interviews, questionnaires and observations.

2.2 Fare Classes

Flight revenue is calculated as that portion of the fare that is shared between the different flights involved in the itinerary of the passenger. The sharing is done by a combination of mileages and other special rules.

Calculating flight revenue is not simply a matter of multiplying the fare from point A to point B by the number of seats occupied. If that were the case both the revenue realized and forecast would have been easily calculated manually. However, the situation is somewhat more complex.

To begin with, there are a multitude types of fares offered for the same seat from point A to point B. This price discrimination is to satisfy the different market behavior and requirements of customers. In addition, passengers occupying seats between points A and B could originate from a multitude of points other than point A and could be continuing their journey to numerous points beyond point B.

To make matters even more complicated many fares are coded on the ticket and the actual fare cannot be readily determined. In effect, therefore, there could be hundreds, and even thousands, of different fares that a particular seat can be sold for travel between points A and B.

As an example let us consider the following flight: Flight ET730, a 200 seater Boeing 767 flight operating from ADDIS to ROME to LONDON. (ET730 ADD-ROM-LON 200 seater).

Each of the 200 seats may be sold as a combination of a sector and any of the fare types as marked by the 'Xs' in Table 2.1

	SECTOR		FARE TYPES							
Origin	ADD-ROM	ADD-LON	Full Fare	APEX	CHD	STD	EXCUR	SP1	SP2	SP3
ADD	X		X							
ADD	X			X						
ADD	X				X					
ADD	X					X				
ADD	X						X			
ADD	X							X		
ADD	X								X	
ADD	X									X
ADD		X	X							
ADD		X		X						
ADD		X			X					
ADD		X				X				
ADD		X					X			
ADD		X						X		
ADD		X							X	
ADD		X								X

Table 2.1: A scenario of various combinations of fares for ET730 ADD-ROM-LON

As can be observed from Table 2.1 there are 16 combinations of fares for passengers travelling from ADD-ROM and ADD-LON on flight ET730 originating from ADD alone. There exist another hundreds of possibilities for passengers travelling beyond ROM and LON (eg. Stockholm, Turin, Paris, Geneva, etc..). However, taking only the most common beyond destinations about 10 can be identified. We saw that for two destinations, ROM and LON, there are 16 different fare combinations, or 8 combinations per destination. Another 10 destinations would mean another 80 combinations. In addition, this is for one way or prorated one-way fares only. There are an equal number of combinations for the round trip fares. Hence, one origin, ADD, can have as many as (16+80)x2 or 192 possible fares on flight ET730.

This flight, however, is not only sold by ADD. Most common other origins include Nairobi, Kilimanjaro, Dar-es-Selam, Bujumbura, Kigali, Luanda, Johannesburg, Lilongwe, Maputo, and Entebbe. Each of these can have the same number of fare combinations as ADD, i.e. 192. This brings the total possible fares that can be sold on flight ET730 to 192+(192x10) or 2112.

But this is not all. Passengers on this flight could have originally commenced their trip from ROM, LON or the 10 beyond points identified and could be returning on this flight from any of the 11 originating points including ADD. These form another 2112 possible fare combinations bringing the total to 4224. There are also other rare but realistic itineraries that have not been considered but which can theoretically bring the possible fare combinations into the hundreds of thousands. More realistically, however, let us consider only the calculated 4224 combinations and assume only a maximum of 5% will occur on a specific flight. This amounts to 211 different fares; which translates into the possibility of selling each seat of an aircraft of a flight at a different price or fare. Some of these fares are straightforward, such as one-way ADD-ROM normal full fare. But some are very complex such as a special fare NBO-ADD-LON-STO in which the ADD-LON portion is calculated through mileages and special airline terms called provisos.

The above discussion has been presented to stimulate the imagination as to how complex the concept of flight revenue is and how difficult it is to timely calculate or forecast the revenue realized on a flight.

The major components of flight revenue are point of sale, fare, fare basis or fare type, prorated amount, and class of service as can be seen from the above example. Calculating the fare from these components is a difficult task. However, Reservation Booking Designators (RBDs) or fare-classes, are believed to be able to accurately represent flight revenue and could serve as a very good representation of all the components mentioned above. As a result, I have found it appropriate to dedicate a section to present the survey made in regard to the concept of fare-classes.

RBDs or fare classes are classes of service created for booking purposes. They are not physical classes on the aircraft but rather ranges of fares in which different types of fares are booked in. Since fare classes represent fare ranges their values are the primary contributors to the overall revenue of the airline. Of course, due to the limited number of fare classes that a reservation system can handle, the ranges will not be a perfect representation of all the possibilities.

Fare stratification is the process of creating ranges of fare classes that could capture or represent the value of different fares. The representations of the stratified fares represent the fare-class fields for the database that will be used to build our revenue information model. The above implication to the revenue process and flight revenue information is tremendous. Unless fare classes properly represent the fare ranges it is doubtful that a reliable model to forecast and report flight revenue information can be developed.

The stratified fares represent the major fields for our databases; they are our primary data sources and it is important, therefore, to understand their concept. To provide a better understanding of the pricing issues involved in stratification of fare classes, a detailed discussion and comparison of how the airline industry has traditionally and more recently viewed the structuring of fares follows.

- Traditional Approach:

  - Airlines have traditionally established a commonality of fare types in buckets (classes) based on the percentage off the full fare (S) class in each market. Fares are stratified based on a specific market yield and not the entire network.

  - The key determinant for placing fares in the hierarchy is to maximize market revenue.

- Filing fares by product type and or by percentage from full fare results in proper stratification at the market level only, and not at system (company network-wide) level. Optimal results will occur for single leg flights only if all seats are occupied by local passengers. Flight legs with beyond or connecting passengers will result in sub-optimal network revenue optimization.

- Filing fares by product type and or by percentage from full fare results in multiple revenue values across a network filed in the same fare class. This makes system revenue management difficult because:

    - It is difficult to differentiate revenue values and product demand.
    - It lowers average fare values and can constrain the flow of high revenue passengers.
    - It provides no control of route network passenger flows.

Figure 2.1 below shows the relationship of yields, fare rules and fare classes in the traditional approach.

Fare-Class

S Fewer Restrictions (fare rules), Higher Yields

B

H

K

L

M

Q

T More Restrictions (fare rules), Lower Yields

Figure 2.1 Traditional Fare Stratification, By Flight and Leg

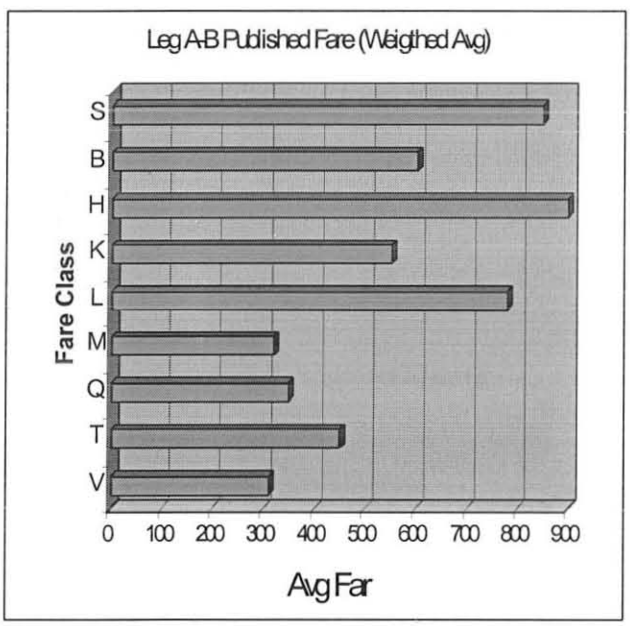
The implication of this figure is that inventory is being managed with all similar fare basis codes, which is governed by rules and <u>not</u> fare amount, in one fare class. Table 2.2 shows an example of this implication.

Class	Market	Route	Fare Basis	Fare Value
S	AAABBB	A-B	Y3	800
S	AAACCC	A-B-C	Y3	1200
S	AAADDD	A-B-D	Y3	1200
B	AAABBB	A-B	BEE1M	600
B	AAACCC	A-B-C	BHAP1M	1000
B	AAADDD	A-B-D	BHAP1M	1000
H	AAACCC	A-B-C	QHAP3M	900
H	AAADDD	A-B-D	QHAP3M	900
Q	AAABBB	A-B	MEE2M	400
Q	AAACCC	A-B-C	MHAP3M	700
Q	AAADDD	A-B-D	MHAP3M	700

Table 2.2 Example of what a traditional fare structure looks like.

- In the above example, the closing of Q fare-class (fare value 400) on flight leg A-B will restrict the availability to the higher revenue through-traffic AAACCC (700) and AAADDD (700) in Q fare-class. Since all connecting passengers must transit leg A-B, by closing Q fare-class to restrict the lower revenue local passengers over leg A-B there is the potential to turn away higher revenue connecting passengers in Q fare-classes (fare value 700)for lower revenue local passengers in B fare-class (Fare value 600).

When fare structures are built using market fare structures instead of system fare structure it leads to a wide range of fare values within a single fare-class. The fare values for all fare-classes overlap, making it difficult to analyze the revenue impacts of the different fare classes or collecting reliable data on fare values by fare-class. The result is graphically represented in figure 2.2.



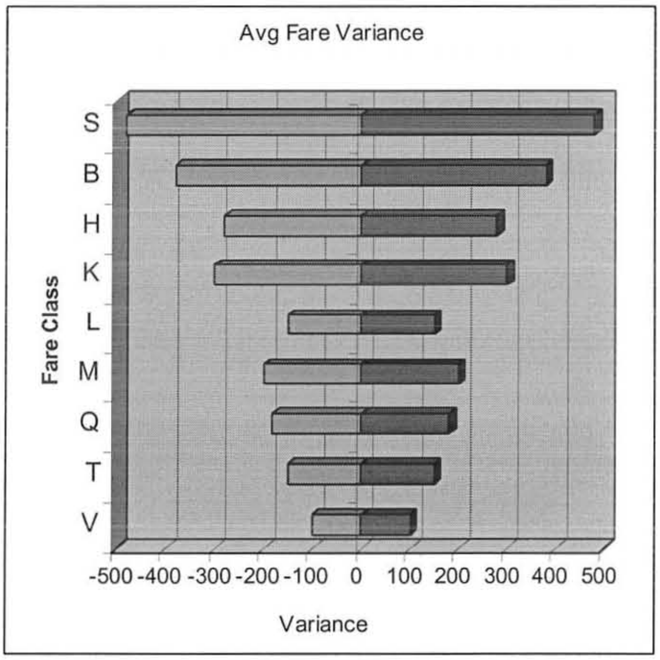


Figure 2.2 Example of a fare structure that is <u>not</u> stratified properly.

To effectively manage revenue in a network (hub and spoke) environment, an airline needs to know the system revenue generated by each fare-class regardless of the market fare structure. Therefore, airlines had to re-think the stratification of fares to encompass a system viewpoint and not a market viewpoint (Connolly 1996). In general, the variation within each fare class can be improved in order to support revenue optimization.

In order to solve the problem of maximizing revenue over a network, the structuring of all market fare structures based on network revenue values need to be accomplished. Here below, is a discussion of how the structure of fares developed based on system, that is network, revenue values rather than market revenue values.

- Modern Approach:
  - System revenue could be increased by regrouping fares so that the lower revenue fares become unavailable before the higher revenue fares.
  - The key determinant to properly stratified fares is higher system network revenue. In the traditional approach, the fare stratification focus was at the market level.
  - The resulting fare structure has multiple product types in each fare class (or multiple fare classes for each product type). The variance of fares in each fare class across the fare structure will be reduced enabling better control of higher revenue passenger traffic and better source data for the revenue information model we aim to build.
  - The proper approach to fare stratification will result in combining similar fare values into the same fare class (regardless of the restrictions or yields).

Figure 2.3 shows the new relationship between revenue or fare values, fare rules and fare classes. This is a critical factor for our research as without such stratification developing a reliable model for our prototype system would not be possible.

S Higher System Revenue Value (irrespective of fare rules)

B

H

K

L

M

Q

T Lower System Revenue value (irrespective of fare rules)

Figure 2.3 Fare-class Proper Fare Stratification, By Route Network

The implication here is that each fare-class will properly and accurately represent the fare value (and not the fare basis or rule as in the previous approach). Table 2.3 shows how the new fare structure compares with the old.

Traditional (old) Fare-Class	Current (New) Fare-Class	Market	Route	Fare Basis	Fare Value
S	S	AAACCC	A-B-C	Y3	1200
S	S	AAADDD	A-B-D	Y3	1200
B	B	AAACCC	A-B-C	BHAP1M	1000
B	B	AAADDD	A-B-D	BHAP1M	1000
H	H	AAACCC	A-B-C	QHAP3M	900
H	H	AAADDD	A-B-D	QHAP3M	900
S	H	AAABBB	A-B	Q3	800
Q	Q	AAACCC	A-B-C	MHAP3M	700
Q	Q	AAADDD	A-B-D	MHAP3M	700
B	Q	AAABBB	A-B	MEE1M	600
Q	T	AAABBB	A-B	HEE2M	400

Table 2.3 Comparison of New and Traditional Fare Structure

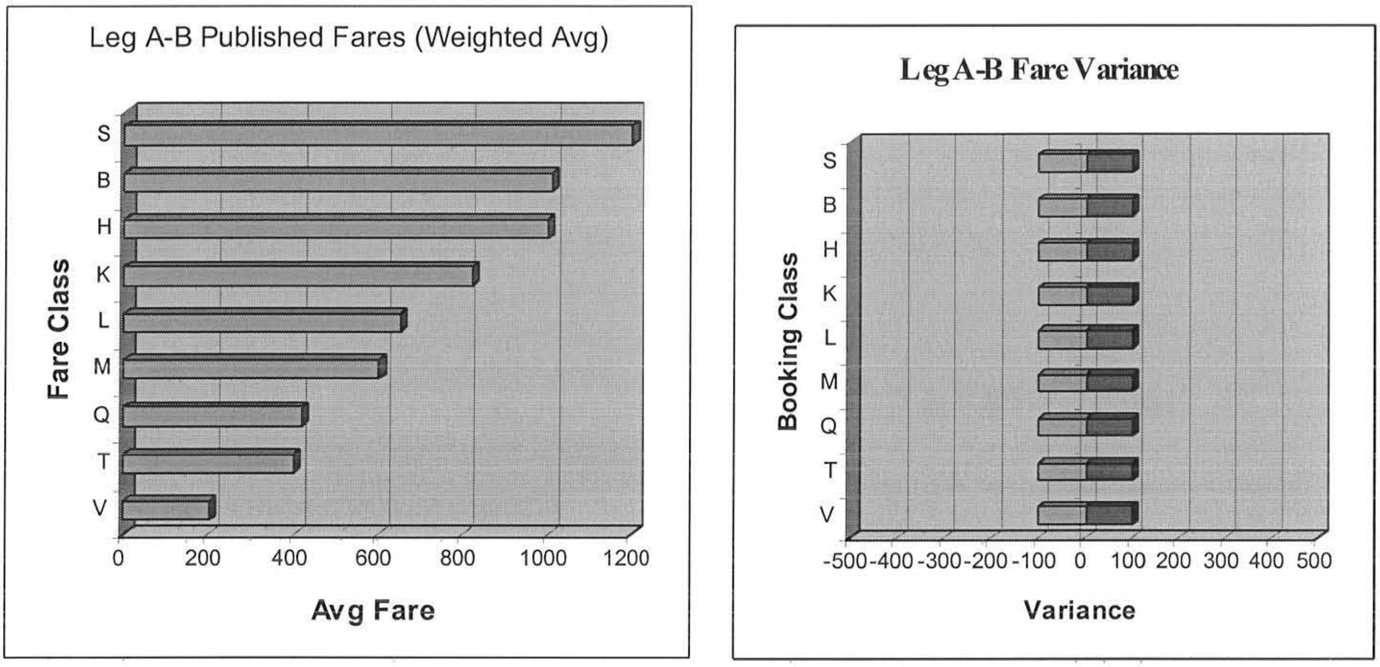


Figure 2.4 Example of a fare structure that is properly stratified.

When the fare structure is accomplished by fare revenue value, system fare structures exhibit less variance of fare values within a fare class. There are fewer overlaps of fare values between fare classes, and fare inversions do not occur. This will result in very reliable fare classes and a very reliable source data for our revenue information model.

Ethiopian Airlines' objective in its fare stratification strategy is to support the optimization of both flight and network revenue through establishing a genuine revenue hierarchy in the fare structure. ETHIOPIAN uses fare stratification at a system level and as a result collects very reliable data represented in the fare classes to estimate revenue. The resulting network fare structure of ETHIOPIAN, adjusted by a certain factor to protect confidentiality, is shown under Table 2.4.

Stratified Bucket	Revenue Range (X currency)
Cloud nine class cabin	
C	>= 39,997
D	<39,997
Economy class cabin	
S	> 19740
B	15557-19740
H	13912-15556
K	12502-13911
L	11327-12501
M	9917-11326
Q	8507-9916
T	<8507

Table:2.4 ETHIOPIAN network fare structure (fare class based on one-way flight prorate value)

The relationship between fare classes is also important to note. One fare class is related to another fare class through a concept known as nesting. For illustrative purpose, the following is a simple example of a linear nesting concept which has been implemented at ETHIOPIAN. Linear nesting consists of parent-child fare class relationship. An example of a linear inventory nesting structure, assuming only four fare classes, i.e. S,B,H,K, from higher to lower fare value, respectively, is given in Figure 2.5

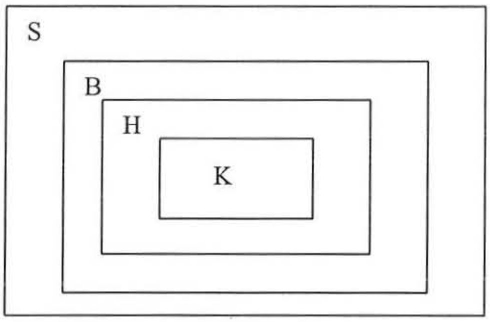


Figure 2.5: Linear inventory nesting structure using four fare classes

In the example shown in figure 2.5 :

- **S** fare-class is the superior of **B** fare-class;

- **B** fare-class is the superior of **H** fare-class AND is the subordinate of **S** fare-class;

- **H** fare-class is the superior of **K** class AND is the subordinate of **B** fare-class;

- **K** fare-class is the subordinate of **H** fare-class.

In all cases of linear nesting, the seat availability count in the subordinate fare-class can never be greater than the seat availability count in its parent class. The objective of linear nesting is to conservatively account for the availability of discounted fare-classes based on a superior-subordinate relationship among all the fare-classes. A superior fare-class always has total access to its own inventory and the inventory of its subordinate fare-class(es). However, seats are never taken away from the subordinate fare-class, unless one of two things happens:

1. Seats are sold from the subordinate fare-class; or
2. Seats are sold from the superior fare-class that subsequently result in a reduced amount of total inventory that can be accessed by either a superior or subordinate fare-class.

An illustration of how this nesting structure will operate in the reservations system under various sequences of booking requests is provided in Appendix 3. This concept of linear nesting and the mechanism of the reservations system in supporting the structure is important to the research since the nesting relationship governs the possible values of each fare class which in turn make up the data set for the revenue information model.

2.3 Availability of Data

One of the purpose of the survey is to identify and assess the data required to develop the revenue information model. After reviewing the key areas that are affected by the revenue process, I believe that the revenue related data elements are available within the existing system, but are scattered in the various application systems. The main category of raw data identified for use in the modeling of the flight revenue information system are the following.

- Advanced booking data: This is data on the forecast of passenger demand on future flights.
- Post departure data: This is the count of the actual passengers who have boarded a flight (that has departed).

- Revenue data: This is revenue of a flight that has departed, which is currently provided to the airlines 3 - 4 months late.
- Schedule data: This data pertains to all details of a flight schedule such as routing, type of aircraft, available seats, aircraft configuration, mileage, meal service, traffic right sectors, etc.

The data are available at a flight, day of week, segment, fare-class level. While revenue data is the dependent variable, all the other are independent. Within the existing system, the data capture of all the raw data described above is done automatically through various application systems and stored in the yield management, departure control, revenue accounting and reservations systems databases, respectively.

Building an appropriate revenue information model requires more than just analysis of current data. The availability of historical data is indispensable to train the model. The following is a summary of the historical data elements currently available at ETHIOPIAN.

Advance booking data represents forecast of passenger demand and prediction of what the fare class booking combination of a given flight will be at the time of flight departure. Over one year's historic data is available in the airline's yield management system database. Post departure data is captured by the airport staff immediately after flight departure in the Departure Control System. Over two years' post-departure data is available at the fare-class level in spreadsheet reports. Historical flight revenue data since April, 1997 is available in the revenue accounting system. Historical schedule data is not of concern as our interest with schedule data is either current or future, not historical; nevertheless, historical schedule data is available on hard copy.

2.4 Critical Functions and Processes

The purpose of this part of the survey is to review the organizational, technical, and business process under which the functions that lead to flight revenue are practiced and to make recommendations that would support timely revenue information.

Figure 2.6 shows a data flow diagram representing the revenue process and the data elements required for the Revenue Information Model.

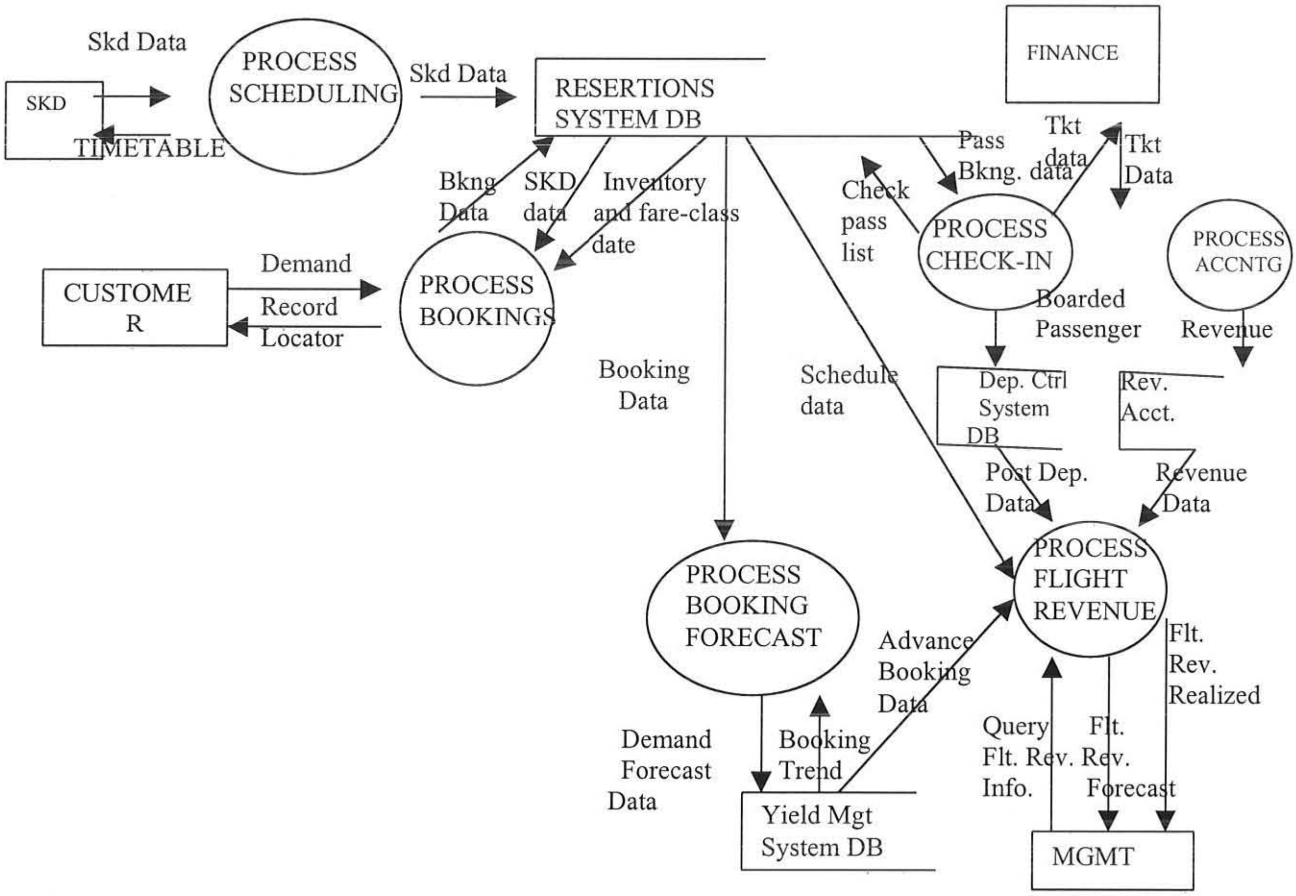


Figure 2.6 Data Flow Diagram of the revenue process.

The functional organization for the revenue process, at the highest level, reports to the Executive Officer Sales and Marketing, whose primary goal is to maximize profitability and maintain customer satisfaction. As a result, one of its objectives is to improve the revenue performance of its flights and routes, where there are many different fare types. The Sales and Marketing organization is divided into nine main divisions; six for Marketing and four for Sales.

Marketing	Sales
1. Planning and Route Management	1. Europe and America Region
2. Market Development	2. Ethiopia Region
3. Promotions and Customer Svcs.	3. Africa Region
4. Airport Operations	4. Gulf, Middle East and Asia Region
5. Cargo Marketing	
6. Marketing Information Systems	

Figure 2.7 below shows the organizational structure of Sales and Marketing.

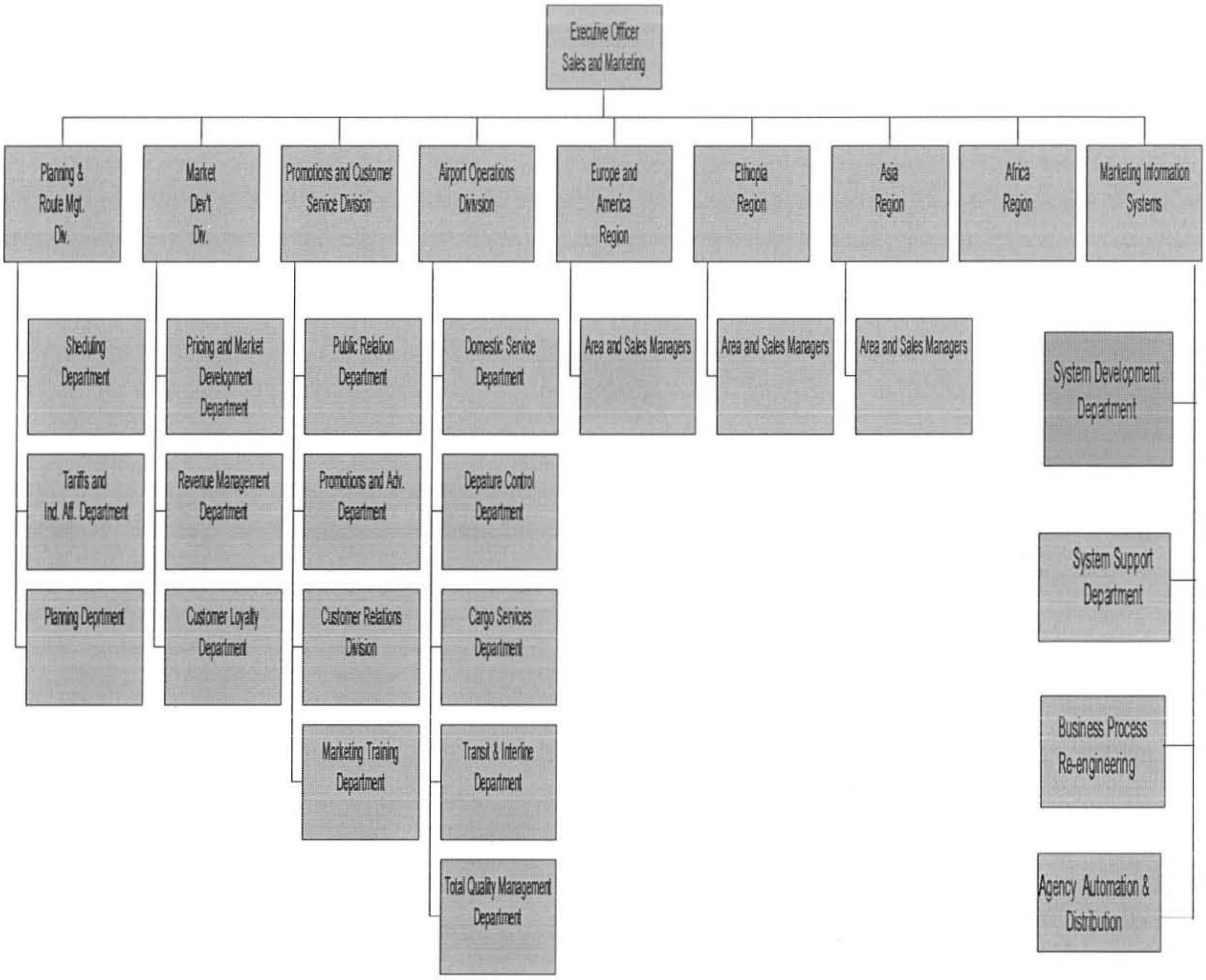


Figure 2.7: Organizational Structure of Sales and Marketing Management.

(Source: ETHIOPIAN Organization Manual)

- As can be seen from Figure 2.7, the Sales function is divided into four Regional divisions responsible for sales and revenue generation. Each of these regions have a number of Area Sales managers reporting to the Regional Managers.

- The Tariffs, Planning and Scheduling departments report to the division manager of Planning and Route Management.
- The Pricing, Revenue Management and Customer Loyalty departments report to the division manager of Market Development.
- System Development, Electronic Distribution and Agency Automation, System Support and Business Process Reengineering report to the division manager of Marketing Information Systems.
- Public Relations, Promotions and Advertisement, Customer Relations and Training report to the division manager of Promotions and Customer Services.
- The Departure Control, Cargo Services, Transit and Interline, and Domestic Services report to the division manager Airport Services.

The survey identified Sales, Scheduling, Pricing, Revenue Management and Airport Operations as the critical functions in the revenue process. As a result, the survey has focused on these functions.

2.4.1 Sales

The Area Managers are responsible for sales and revenue in their district of jurisdiction and are based on site at the various on-line destinations where ETHIOPIAN operates. The importance of these offices in the revenue process is critical. They are the primary contact with the customers and the source of the airline's revenue. What makes them particularly important to our research is that they are the source for one of our primary data we are going to use in building our revenue information model.

When a customer approaches one of the point of sales of ETHIOPIAN, the sales agent makes the sale and reserves the passenger in the reservations system in the appropriate fare-class. Every night the booking data of all passengers on all flights are downloaded to the yield management system. Based on these bookings, over time, the yield management system builds a historical trend and forecasts the demand or future bookings by fare-class. However, it does not forecast revenue. Based on these future expected demand, each fare-classes of a flight is allocated seats for the sales offices to sell.

2.4.2 Scheduling

Scheduling is equivalent to the product manufacturing function of a manufacturing company. The schedule, which includes routings, timings, aircraft type, meal services, class, configuration, etc., is essentially the primary product of the airline. Respondents claim that the ETHIOPIAN route and schedule structure is complex. In general the international route system is scheduled such that most flights are two and three-leg flights with an interconnecting network through Addis Ababa.

Major schedule changes are completed twice a year, October and March, to adjust the schedule for peak and off-peak seasons. Major schedule changes are decided upon one to two months prior to implementation and loaded into the reservation system one month prior to implementation. Major schedule changes are those that review the entire network, make changes that affect most flights and are published in a timetable. Minor schedule changes are the ad hoc changes made on specific flights. Minor schedule changes are decided upon one week prior to implementation or less, and can be loaded into the reservation system with as little as one days notice. Minor schedule changes occur as required.

Once a major schedule change has been effected, the schedule's performance is analyzed on a day-to-day basis. Recommendations to combine flights or overfly stops are regularly made depending upon demand. These schedule adjustments occur late in the booking cycle and can occur often, thus creating problems with any forecasting or predicting model. There is no consistency of schedules. Users in the Scheduling department believe that flight revenue forecast information can eliminate the need to make these daily equipment flow adjustments.

2.4.3 Pricing

All published tariffs are filed with IATA (International Air Transport Association) which distributes the information to the various global distribution systems. Pricing levels are based on analysis of operating costs as well as negotiations. Fares are directional, as ETHIOPIAN takes into consideration differences in currencies and income levels from country to country. Respondents claim that the percentage of passengers that pay the published tariffs is so low that they are mainly used to determine rules by which other reduced fares or special fares are to be applied.

Once the regional sales staff ask for special pricing, the pricing expert in the pricing department analyzes the competitive situation, yields, and costs. The special pricing usually takes the form of a discount off the IATA tariff or a net fare. A tremendous amount of pressure is exerted by the regional sales offices on the pricing expert to approve the fare levels. This is due to the fact that flight performance is measured only after departure with the major emphasis being on load factors. According to 86% of the respondents using load factors for performance measurement is not effective in markets with multiple products using variable or average cost pricing. This is the case in most international markets that Ethiopian Airlines serves.

Respondents believe that flight revenue information is critical to develop competitive prices that will maximize revenue and would extremely help in evaluating pricing decisions. In addition over 50% of respondents claim that such information is the best indicator of flight performance and can greatly assist to decentralize the approval process to achieve quicker reaction to market situations.

2.4.4 Revenue Management

Revenue Management is responsible for overbooking of flights and their fare mix allocation, and reports to the Market Development Division Manager. Overbooking is accomplished by setting pre-defined profiles that allow a varying level of seat availability (authorization level) above the physical compartment capacity during the booking cycle. The fare stratification approach discussed in section 2.2 is very important to optimize fare mix allocation.

ETHIOPIAN currently operates in a multi-cabin, multi-class environment for its international service. Two cabins are designated as C (Cloud nine) and S (Economy). The C cabin contains one primary class (C class) and one subset class (D). The S cabin contains one primary class (S class) and seven subset classes (B, H, K, L, M, Q, and T).

When there are multiple fare products being used, it is very important that multiple fare classes are used for most bookings over each flight leg. As a result, ETHIOPIAN is able to restrict low fare bookings in favor of higher fare bookings, thus ensuring substantial incremental revenue and to collect reliable fare class data that properly represents fare values.

Once a month, the Revenue Management department prepares monthly updates on route/region flight performance and group/tour performance. These reports examine trends and competitive environments. About 84% of the respondents feel there are problems with these reports. In order of priority, the problems expressed include insufficiency of revenue information, untimely provision, difficulty in usage and information provided is not to the required level of detail.

2.4.5 Departure Control

Departure control is that function of airport operations which checks in and boards passengers. It is a critical function in the revenue process because it supplies one of the data types, post departure data, that is used to develop the revenue information model.

The boarding information is provided electronically through the departure control system. These airline standard format close-out messages allow for the reporting of boarding information by fare class. The reporting mechanisms at Addis Ababa, the major hub, appear to be sound. The check-in to close-out process is very well controlled and organized. Numerous checks are required to ensure that all passengers boarded are accounted for. The checks and balances include ticket count, immigration count and boarding pass count. It appears that much effort is exerted to reconcile these counts to produce accurate boarding statistics.

Stations report information at the fare class level. However, the data can be misleading because post departure passenger counts are prorated across the fare classes based on the day of departure bookings. The fare class data is the "best guess" of the station personnel and not an accurate count of actual fare class boardings.

boarding pass count. It appears that much effort is exerted to reconcile these counts to produce accurate boarding statistics.

Stations report information at the fare class level. However, the data can be misleading because post departure passenger counts are prorated across the fare classes based on the day of departure bookings. The fare class data is the "best guess" of the station personnel and not an accurate count of actual fare class boardings.

Chapter 3

DATA MINING AND NEURAL NETWORKS

3.1 General

The primary objective of an information system, we have seen, is to transform data into useful information. . The major activities involved in the process are illustrated in figure 3.1 below; as related to the research problem under investigation. In this process, users are a critical element in the overall information system. They are actually the primary focus of the data processing activities, as they alone can determine the usefulness of the information, validate the conversion of data to information, and actually convert information to action. Eardley, et al. 1995 note that it is at the level of the user that the information system actually provides benefit or value to the organization; prior to this the data processing has only been incurring costs.

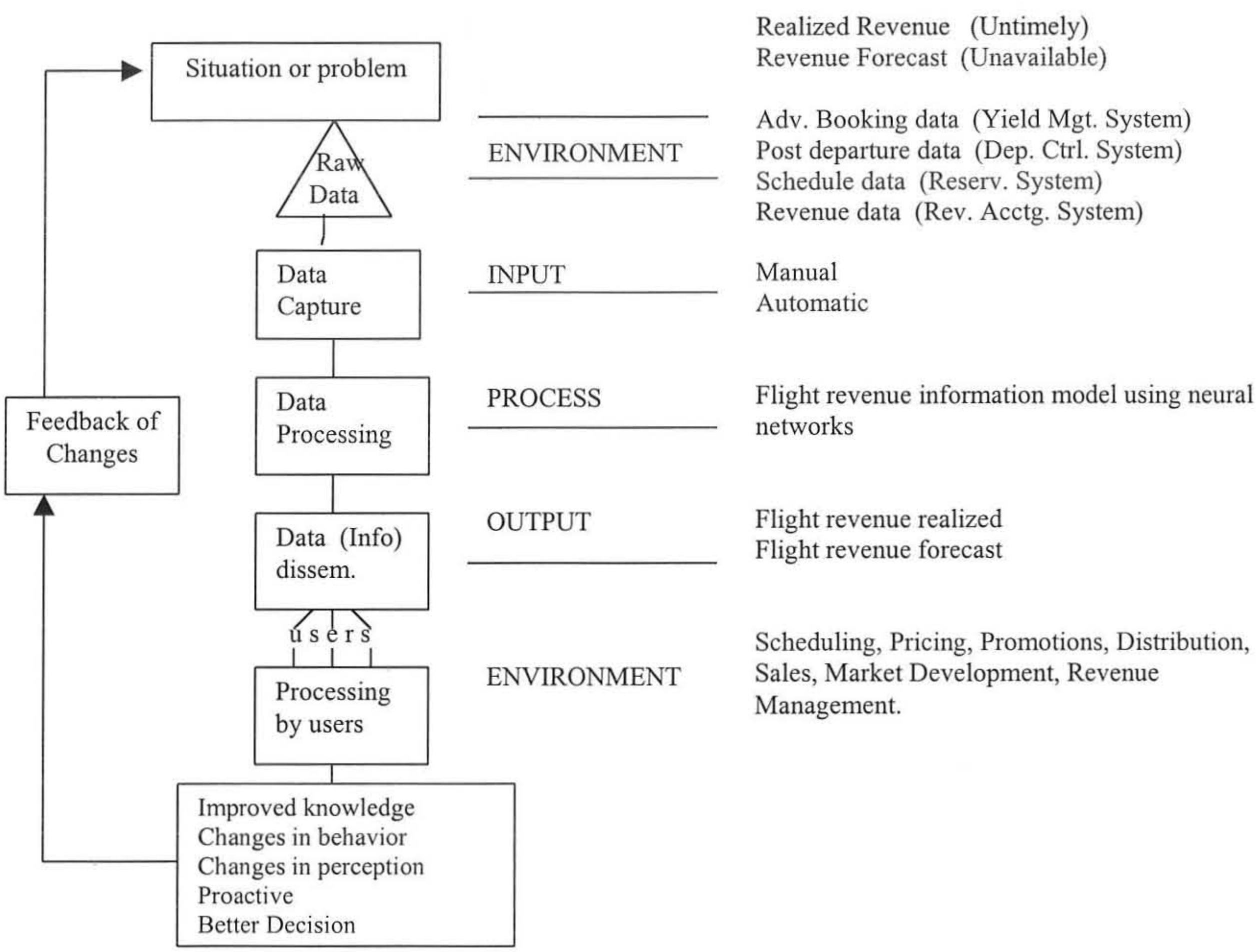


Figure 3.1: The information system: an outline model

Computer-based information systems means that the major part of the system uses computers to capture and store data, produce and disseminate reports and handle enquiries. Over the years, in response to increasing user requirements, the use of computer-based information systems in organizations has evolved from the automation of certain business operations into the integrated end-to-end automation of all major business functions and processes we know today.

To support development of these information systems, methodologies, likewise have evolved. A methodology is a group of techniques, tools, and procedures based on a philosophical view, that guide the systems developer to systematically develop and implement information systems.

There are a number of methodologies developed for use in systems development. However, they are all based on the concepts of abstraction (modeling) and decomposition. Those that decompose on the basis of process or data alone come under the structured approach (nowadays considered traditional); examples include SSADM, Jackson, and Yourdon. Those that decompose on the basis of objects, where data is combined with the operations of the object, come under object-oriented; examples include OMT, OOSE and UML. An object, as defined by Singer (1996:48) is, " a concept abstraction, or thing with crisp boundaries and meaning for the problem at hand."

Some systems are permanent systems, continuously and regularly collecting data on a particular situation: The yield management, reservations, departure control and revenue accounting systems are such examples . These systems have accumulated a vast amount of data through the years as a by-product of transactions. However, ETHIOPIAN is not making effective use of this data. Increasingly, business data is being seen as a valuable commodity in its own right, not just as a by-product of processing the day's transactions. It is advisable, therefore, that ETHIOPIAN management leverage its investments in business data, to use it as an aid in decision making, and to turn it into operational applications.

However the methodologies described earlier have not been able to address these requirements. The structured approach works well for a wide variety of problems and has become the standard technique used by programmers worldwide but it is not suited for such problems. In object-oriented, although the focus is on objects and their behavior rather than on problem decomposition, it is still just another approach to writing algorithms for digital computers (Bigus 1996). In addition, there are many cases where traditional programmed applications cannot be developed, since no one in the business understands how the data relates well enough to design or write an algorithm to capture those relationships. In such cases new methods are required.

In this connection, for the purpose under investigation, that is building a revenue information model, I have considered an approach based on data mining. Upon exploring the possibilities of using data mining tools for this problem, I found it a worthwhile undertaking; primarily because it represented a new concept to experiment for both the airline and myself, and secondarily because it would effectively address the human element; a serious problem mentioned in appendix 1. This approach would eliminate the need to develop new procedures or hire additional staff; there would be no need to change mentalities, or hire extra staff; extra costs or extra work for the airline would be eliminated since the data mining tools would use data already existing in the airline's many databases. As indicated in the preceding chapter, these data are created as a result of other transactions which had to be performed in any case.

Another exciting prospect was that I would not actually need to write algorithms and codes to develop the system and could concentrate on the business research, data preparation, training and testing the data mining tools and develop a model for my objective.

ETHIOPIAN retains so much data in its different data bases but has not been able to use it to create competitive advantage. This, together with the points noted above, is the reason why data mining was found to be compelling. To this end, in this chapter, I intend to briefly introduce data mining technology, and as a further background, to the techniques employed in my experiment.

3.2 Data Mining Overview

Data mining is a process of discovering meaningful new correlations and patterns by going through large amounts of data. Data mining techniques, such as neural networks, have the advantage of building applications without understanding how the data relate. Information systems, in this approach, are built by using the models build during the iscovery process. For example, starting with a selection of prepared booking and revenue data, a data mining algorithm can be used to discover relationships in the data and build a revenue forecast model. Once this model is created, it can serve as an information system to predict revenue. Figure 3.2 shows an outline of how such a system is developed.

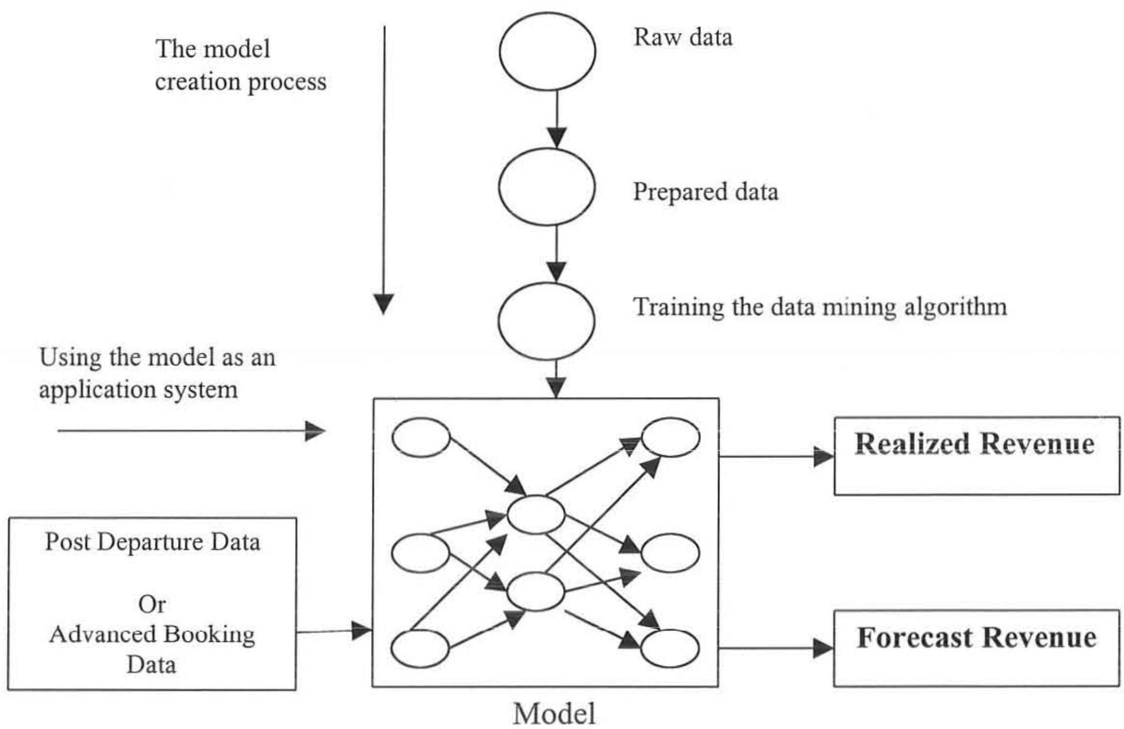


Figure 3.2: Data Mining for application development

Applications built using a data mining algorithm have the benefit in that the application is generated automatically as a by-product of the data mining process. For example, once the model to forecast revenue is built, applying it to discover the realized revenue is a by-product of this process. By providing both descriptive (what has happened) and predictive (what is likely to happen) insight into data structures, this technology helps individuals make better decisions, and helps organizations react quickly to changes in their operating environments.

Many definitions of data mining are found in published literature. Bigus (1996:9) defines it "as the efficient discovery of valuable, nonobvious information from a large collection of data." Another definition is "the process of discovering meaningful correlations, patterns and trends by sifting through large amounts of data stored in repositories, and using pattern recognition technologies as well as statistical and mathematical techniques (http://KDNuggets.com)." The core concept behind all of them seems to be the same. Data mining centers on the automated discovery of new facts and relationships in data.

Data mining is also known as knowledge discovery in databases (KDD). The term data mining is usually applied to the operational procedure for obtaining new insights, whereas KDD is used by the research community developing new techniques (Chung and Paul 1999).

Data mining is not the same as Decision Support Systems (DSS). "Decision Support Systems (DSS) use deductive reasoning. Data mining techniques use inductive reasoning. Deductive reasoning form a conclusion based on a set of general rules, inductive reasoning is when a conclusion is drawn for observations" (http://KDNuggets.com).

Nor is data mining an on-line analytical processing (OLAP) tool. The key difference is that OLAP is user-driven; the analyst generates an hypothesis and uses the OLAP tool to verify the hypothesis. In contrast, in data mining the tool is used on the data to generate a hypothesis. Edelstien (1997) confirms that " when users employ OLAP and other query tools to explore data, they guide the exploration. However, when users employ data mining tools to explore data, the tools perform the exploration."

In the evolution from business data to business information, each new step has built upon the previous one. Data mining is now ready for application in the business community because it is supported by three technologies that are now sufficiently mature: massive data collection, powerful multiprocessor computers, and data mining algorithms.

Data Mining is a continuous iterative process. It involves data preparation, data processing and data analysis. It requires the use of algorithms, software and sound methodology. Once the problem and the required data is defined the data mining process follows the following iterative steps.

1. Data preparation: selection, collection, cleansing (handling missing and noisy data), and preprocessing (scaling and representing).
2. Data processing : choosing the model (regression, classification, etc.), choosing the algorithm (neural network, decision trees, etc.) and building the model (actual data mining).
3. Data analysis : interpretation of what the model learned, evaluation of results, and monitoring the model.

Small (1997) claims that much of the difficulty in applying data mining comes from the same data-organization issues that arise when using any modeling technique. These include, data preparation tasks such as deciding which variables to include and how to encode them and deciding how to interpret and take advantage of the results.

Data mining is generally used to build six types of models aimed at solving business problems: classification (making distinctions between items), clustering (dividing similar things into groups), regression (learning relationships between variables), association analysis (associating two or more things), modeling (learning to predict outcomes based on examples), time-series forecasting (predictors into the future) and sequence discovery.

Classification, regression, and modeling are primarily applied to prediction, while association and sequence discovery are primarily used to describe behavior that is captured in the database. Clustering may be used for either forecasting or description. Many techniques have been used to construct these common data mining models. The techniques, also called algorithms, range from statistics to neural networks. Although to assess the various techniques is beyond the scope of the research , identifying some major ones helps put neural network in perspective. A white paper produced by the data intelligence group (DIG) in 1995 describes the most commonly used techniques in data mining as follows.

- Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.

- Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset.
- Genetic algorithms: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.

- Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the K records(s) most similar to it in a historical dataset. It is sometimes also called the k-nearest neighbor technique.

- Rule induction: The extraction of useful if-then rules from data based on statistical significance.

According to Bigus (1996), the type of data mining model we are trying to construct, the application we are trying to perform and the quality and quantity of data available combine to specify which data mining algorithm should be used. Table 3.1 adapted from Bigus (1996:12) shows a list of the most common data mining models, the corresponding data mining algorithms, and their typical applications.

Data Mining Models	Algorithms	Application Examples
Association	Statistics, set theory	Market basket analysis
Classification	Decision trees, neural networks	Target marketing, quality control, risk assessment
Clustering	Neural networks, statistics	Market segmentation, design reuse
Predictive Modeling or regression	Linear and nonlinear regression, curve fitting, neural networks	Ranking/scoring customers, pricing models, process control
Time-Series forecasting	Statistics ARMA models, Box Jenkins, neural networks	Sales forecasting, interest rate prediction, inventory control
Sequential patterns	Statistics, set theory	Market basket analysis over time

Table 3.1: Data Mining Functions

3.3 Neural Networks

3.3.1 Historical Development

Neural networks fall in the division of artificial intelligence which claim that our symbol processing forebrain processes information that has already been processed at a sub-symbolic level by the body senses – a process of pattern recognition at a subconscious level. It claims massive parallelism is a fundamental aspect of intelligence.

A paper written by Intelligent Technologies (Intelligent Technologies 1999) states that, "neural networks are one of a group of intelligence technologies for data analysis that differ from other classical analysis techniques by learning about your chosen subject from the data you provide them, rather than being programmed by the user in a traditional sense."

Although work on neural networks commenced in the 1950s, its development progressed at a very slow pace due to the early successes of the other group; which claims that 'physical symbol' can represent intelligent action (Expert Systems).

By the mid-1980s, however, the situation began to be reversed when progress of rule-based expert systems and other symbol–based artificial intelligence was not to the expected or desired level in solving some of the fundamental problems in developing intelligent software systems. Assisted by the emergence of massive storage capabilities and high processing powers, work on neural networks re-emerged. Table 3.2 traces the major historical developments of neural networks (Bigus 1996).

1943	McCullough and Pitts Binary Neuron
1950s	Two major schools of thought on Intelligent computers developed
1960	Widrow's Adaline model of neural networks was developed
1962	Rosenblatt's Perceptron model of neural networks was developed
1969	Minsky and Pappert wrote the book perceptrons which criticized neural networks and served to kill neural network
1970s	Small group of researchers carry work on neural network
1980s	Research on neural network reemerges
1986	PDP (Parralel Distributed Processing) research group publish a two-volume manifest on neural networks
1987	First International Conference on neural networks held
1990s	Emergence of commercial neural network applications
1990s/ 00s	Research to create neural network machines with humanlike intelligence or behavior continues

Table 3.2: Historical development of Neural Networks

3.3.2 Biological Inspiration

Solving problems with neural networks is quite similar to the way people naturally solve problems. A neural network learns to solve problems by being given data, examples of the problem, and its solution. Neural networks are data models which simulate the structure of the human brain. Like the brain, neural networks learn from a set of inputs and adjust the parameters of the model according to this new knowledge to find patterns in data (http:www//spss.com).

Clark (1997) describes neural networks as an information processing paradigm that was inspired by the way biological nervous systems, such as the brain, process information. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons. A brief description follows.

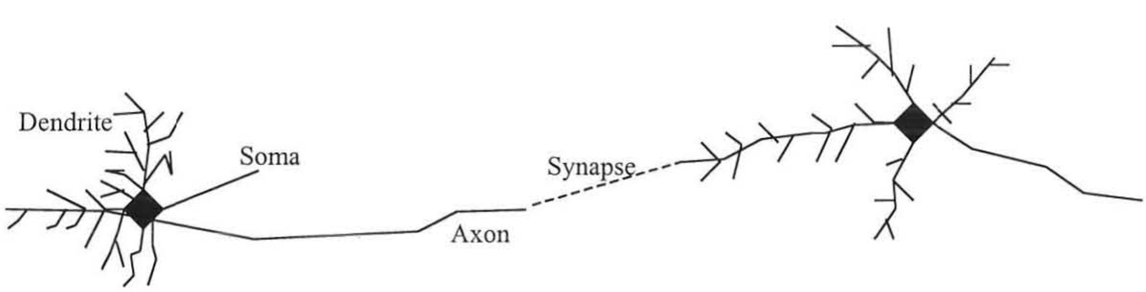


Figure 3.3: Biological Neural Network

The elementary building block of biological neural systems is the neuron. The single cell neuron consists of the cell body, or soma, the dendrites, and the axon (Figure 3.3). The function of the neuron is to integrate the input it receives through its synapses on its dendrites and either generate an action or not. When the action reaches a synapse at the end of the axon, it undergoes several processes and an electrical signal is conveyed along the dendrite to the soma of the post-synaptic neuron.

The dendrites receive signals from the axons of other neurons and conduct impulses towards the soma. The axon conducts impulses away from the soma. The small space between the axon of one neuron and the dendrite of another is the synapse. The function of the synapse is to covert electrical signal received from the axon of a neuron into a chemical signal; and reconvert the chemical signal into an electrical signal at the other end of the synapse to be conveyed to the soma of the other neuron through its dendrites. A synapse can either be excitory or inhibitory. Input from an exitory synapse increases the internal activation level of the neuron, while input from an inhibitory synapse reduces it. Biological neural systems are highly distributed and highly interconnected. A single neuron may receive input from as many as 8×10^4 neighboring neurons (Clark 1997).

Unlike the digital computer, neural networks and neural computers are based on a model of the brain. A neural processing element receives inputs from other connected processing elements. These input signals or values pass through weighted connections, which either amplify or diminish the signals. Inside the neural processing element, all of these input signals are summed together to give the total input to the unit. This total input value is then passed through a mathematical function to produce an output or decision value ranging from 0 to 1. Varying degrees of similarity are represented by the intermediate values. In a very real sense, neural networks are analog computers, which deal with the data represented in a continuous form.

3.3.3 Overview

In my review of the major concepts of neural networks below, I have deliberately excluded detailed mathematical treatment of how neural networks operate as it is beyond the scope of this study. The discussion rather focuses on how neural networks are used to build a model.

It has been established that neurons work in a parallel, multiply-connected manner and that singly, they perform a simple function; when connected, they are massively intelligent (Hormby 1992). A visual representation reflects the idea that each input is connected to the various outputs, or possibly to intermediate processing nodes, as shown in figure 3.4 below (Pham and Liu 1995). Each of these connections is weighted, reflecting the fact that some inputs will have more predictive value than others. Layers of these interconnected processing nodes interact in a multidimensional, parallel structure described as a network. Each component of the network performs a simple, nearly mechanical function. It is the linkage of multiple layers of multiple processing nodes that allows for the complexity and accuracy of this approach. The weights linking the layer of neurons are the "synapses" of an artificial neural network . The output values can serve any number of functions: a transformed pattern, a complex equation, a set of estimates, or a predicted value. It is this latter function that is of interest to us in regression or forecasting.

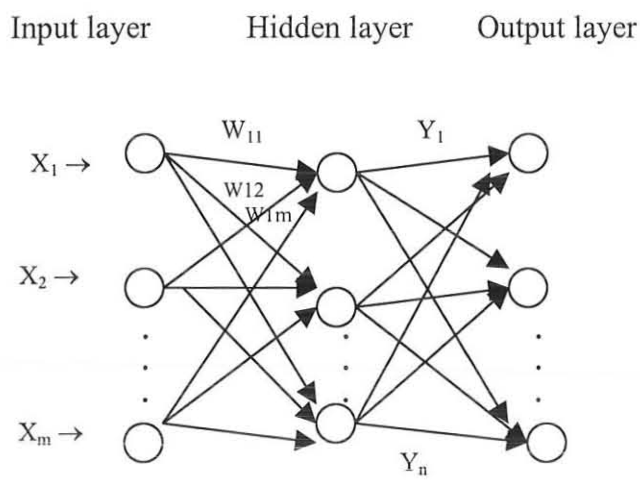


Figure 3.4 A multi-layer perceptron with three layers

The way the weights are derived is that the system is able to incorporate feedback (in the form of accuracy of the forecast), and to use this feedback to adjust the connection weights in an adaptive fashion. Each connection between an input and a processing node, or a processing node and an output node has a connection weight. These weights can be modified based on feedback processed through a learning algorithm which fine-tunes the network. The algorithm uses the accuracy information to make minor adjustments in the connection weights. During training, these minor adjustments move the state of the network toward an optimal solution.

Building neural network applications is similar to training a knowledge worker. Clear-cut extreme cases are given as examples to the neural network and sufficient data is provided so that the neural network can learn to accurately make decisions. The neural network will learn from experience.

In a nutsshell, examples are presented to a neural network, the network makes a prediction, and the connection weights are adjusted so that the output corresponds more closely to the desired output. This adjustment process is done automatically by the learning algorithm being used to train the network. The process of learning in neural networks is to use feedback to adjust internal connections. By making connections stronger or weaker, reinforcing or inhibiting, the artificial neural network is mimicking the behavior of the synapses of the brain, which undergo physical changes in response to input patterns and feedback.

When neural networks are used as the data mining algorithm, the output of the process is a trained model. This model can be used to process transactions. Whatever the function, the trained neural network is the application module. When input data is presented to it and the results from the output units are retrieved, the module becomes a transaction processing system.

The research problem revolves around building a model that predicts revenue. Modeling involves making a static one-time prediction based on current information. Hence, when presented with actual booking data of a departed flight, my assumption is that the model will predict the realized revenue and when presented with a forecasted booking pattern of a future flight, it will predict the forecast revenue value of that data set. The forecast booking pattern is currently being performed by the yield management system.

According to Bigus (1996),much of the application development process in neural networks follows the same phases as other programming projects. However, since we are using a commercial neural network development tool to model the problem of our research, there is no coding phase. This is replaced with the training and testing phase.

Angoss Corporation (http://www.Angoss.com), a data mining software developer, believes that neural network technology delivers two key business benefits. For ETHIOPIAN, the benefits translate as follows.

i. An enabler in the context of defined business objectives, to automatically explore, visualize and understand data, and to identify patterns, relationships and dependencies that impact on business outcomes (such as revenue growth and profit improvement) – a descriptive function.

ii. An enabler to express relationships, uncovered and identified through the data mining process, as business rules, or predictive models. These outputs can be communicated in traditional reporting formats (presentations, reports, electronic information sharing) to guide business planning and strategy. Also these outputs, expressed as programming code, can be reused in business operating systems to generate predictions of future outcomes, based on newly generated data, with higher accuracy and certainty – a predictive function.

These are the two functions of interest to our research. In particular, realized revenue serving as a descriptive function and forecast revenue as a predictive function.

Edelstein (1997) warns us, however, that neural networks have two problems. He states that, "first, one of the most common arguments against neural nets is their 'opaque' quality; the factors leading to a prediction are not obvious. The second problem is that neural nets are prone to overfitting; they become very good at predicting the test data at the expense of accuracy on new data." In order to tackle the first problem one needs to exhaustively test hundreds of models with different training parameters to be able to identify some of the factors leading to a prediction. To tackle the second problem the training data set should be properly managed with different proportions of learning and testing sets.

3.3.4 Neural Networks versus Traditional Forecasting Methods

With enough experience in a variety of marketing situations, and with sufficient time to react to new situations, we know it is possible for a human being to make forecasts that are better than random. However, human expertise is limited to a small number of market situations, and people are not such good optimizers.

Hormby (1992), explains that database systems enable the use of summaries and exception reports to focus the manager's attention on those specific problems with the most revenue at risk. In addition, databases provide the historical resources to allow the manager to research a particular trend or past activity. Databases help to organize information, but still rely on human judgment for inference and forecasting.

Statistics and operations research have developed a whole range of methods to spot trends and regularities in a body of data. Some of these techniques have been in use for many years for forecasting and projecting future activity on the basis of historical trends. Varieties of regression techniques and even simple graphs can reveal hidden patterns (Small 2 000).

Companies have been using related quantitative techniques in many parts of their businesses for a long time. Neural networks are just one more advance in a research process that has been ongoing for many decades. Small (2000:4) states that "Neural nets are a special case of projection pursuit regression which were developed in the 1940s. CART (Classification and Regression Trees) methods were used by social scientists in the 1960s. K-nearest neighbor, a form of density estimation, has been used for a half-century. All these methods-just like regression techniques – model relationships between a set of profile variables and an outcome."

Systems in use today, however, have limited flexibility and require programmer intervention when the marketing environment changes. Hormby (1992) gives an example… "an exponential smoothing model, which is essentially a moving average with some trend and seasonal effects added in, can take several weeks to recognize a large change in market behavior." Additionally, these traditional models perform well only when the input-output relationships are fairly clear, and when the combinations of predictive factors can be simply represented.

What is new about neural networks is that they can be applied to more general business problems, thanks to the increased availability of data and high processing power. As a result neural networks can better deal with massive data in which relationships are unclear and quickly react to changes in market behavior. In addition, recent interest in data mining has made neural networks more available to business experts through user-friendly software.

Traditional statistical models which rely purely on trend analysis to predict future outcomes will not respond to changes in the marketing environment until those changes have a chance to affect the trend. For problems where there is a long delay between changes in predictive inputs and final outcomes, such as a competitor airline changing its pricing strategy, I believe neural network models will be able to react much more quickly than other methods.

Additionally, a neural network model provides both a forecasting tool and a simulation environment in the same model. By manipulating the predictive inputs, alternative outcomes can be easily generated. "What if" analyses are possible by manipulating demand, no-shows, capacity, fare classes, schedule changes and price levels. Thus, a neural network model is able to respond quickly to changes in the environment, and it allows a user to evaluate the impact specific changes would have on the final outcome.

According to Edelstein (1997), the complexity of a problem will determine how difficult it is to extract meaningful relationships from the data. Problems increase in complexity as the amount of data increases. Other contributors to problem complexity are the level of interaction among variables being examined and nonlinearity in the variables and parameters. Furthermore, as the patterns become more subtle and the need for accuracy rises, finding the right patterns becomes more difficult. **Our revenue problem deals with vast amount of data and high interaction among nonlinear variables which classifies it as a very complex problem.**

For example, the number of days prior to departure is a factor that a model must use in its predictions for our problem. But because of the bi-modal booking pattern exhibited by business and leisure market segments, as shown in figure 3.5, this variable is non-linear. According to Hormby (1992), neural networks can casily incorporate these non-linearities, while also using data that are discrete or categorical.

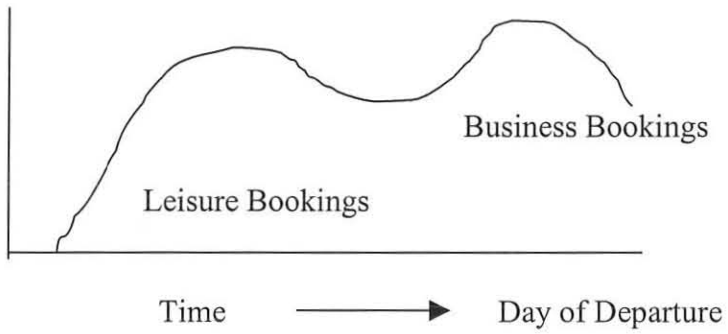


Figure: 3.5: Non-Linear Bi Modal Booking patterns

Neural networks are powerful solutions to such problems because you do not program a neural network – it is trained just like a human being. While other statistical models, such as regression, may be able to incorporate interactions between many factors if identified to them, neural networks automatically define these interactions. Although it is theoretically possible for statistical methods to model all the possible interactions among input factors, this approach is impractical for problems of any complexity.

Hormby (1992) claims that , "two of the many beneficial features of neural networks are that these models can accommodate non-linear data and higher-order interactions among the various input factors. Neural networks share the objective of artificial intelligence – to model human decision making - but it does so in a way that is less likely to involve erroneous assumptions and to depend on the availability of an expert in the field."

A key benefit of neural networks is that you can use them to build a model of the system or subject you are interested in, from just the data you provide them. One may know the inputs and outputs that are important but may not know what happens internally - the neural network will model this system simply by identifying the relationship of the data.

Hormby (1992), summarizes the advantages of using neural networks to current methods of analysis as follows. "They can successfully:

- Deal with non-linearities.
- Be developed from data without an initial system model.
- Handle noisy or irregular data.

- Quickly provide answers to complex issues.
- Be easily and quickly updated.
- Interpret information from tens or even hundreds of variables or parameters.
- Readily provide generalised solutions."

Choosing Data mining, and in particular neural networks, as the forecasting technology for my research was mainly due to the reasons discussed above. However, the following other factors have also contributed.

- Thesis on using neural networks for pattern recognition are quite common, but I was not able to find any applications similar to flight revenue developed using these techniques. A recent survey by Tow Crows Corporation, however, shows that the top three end users of data mining are in the marketing area (Edelstein 1997).
- Interview with the Chief Information Officer revealed that data mining is new to ETHIOPIAN and no application has been built using this technique. In addition, the airline would be very enthusiastic to test such a development and would support the endeavor.

3.3.5 Choosing the Neural Network Model

As we have already seen, neural networks generally consist of a number of interconnected processing elements or neurons. How the inter-neuron connections are arranged and the nature of the connections determine the structure of the network. How the strengths of the connections are adjusted or trained, to achieve a desired overall behavior of the network, is governed by its learning algorithm. Neural networks can be classified according to their structures and learning algorithms.

In terms of structure, neural networks can be divided into three types: Feed forward, limited recurrent, and recurrent networks. The structure of the neural network defines how data flows between the input, hidden and output processing units. In feed-forward networks the neurons are generally grouped into layers. Signals flow from the input layer through to the output layer via unidirectional connections, the neurons being connected from one layer to the next, but not within the same layer. Pham and Liu (1995) state that the feedforward network, "performs static mappings between an input space and an output space: the output at a given instant is a function only of the input at that instant." This means that data flows through the network in one direction, and the answer is based solely on the current set of inputs as shown in figure 3.6

Figure 3.6 Unidirectional Connection of Feedforward Networks

In limited recurrent networks, information about past inputs is fed back into and mixed with the inputs through recurrent or feedback connections for hidden or output units. Fully recurrent networks provide two-way connections between all processors in the neural network.

In terms of learning algorithms, neural networks can also be divided into three types: supervised, unsupervised and reinforcement learning. In supervised learning, the strengths or weights of the inter-neuron connections are adjusted according to the difference between the desired and actual network outputs corresponding to a given input. It requires a 'supervisor' to provide the desired output signals. In unsupervised learning there is no 'desired output'. Unsupervised learning is described by Intelligent Technologies (1999:3) as, "a process in which the network is able to discover statistical regularities in its input space and automatically develops different modes of behavior to represent different classes of inputs (in practical applications, some 'labelling' is required after training, since it is not known at the outset which mode of behavior will be associated with a given input class)". In reinforcement learning, examples of the problem or case is known, but we do not have the exact answer, or at least not immediately.

Bigus (1996) summarizes the situations under which these structures and learning paradigms are used and the circumstances by which to choose the neural network topology, as follows.

Feedforward : Used in situations when we can bring all of the information to bear on a problem at once, and we can present it to the neural network.

Limited recurrent: Used in situation when we have current information to give the network, but the sequence of inputs is important, and we need the neural network to somehow store a record of the prior inputs and factor them in with the current data to produce an answer.

Fully recurrent: Used primarily for optimization problems and as associative memories.

Supervised: Used when you have a database of examples that contain both problem statements and the answer.

Unsupervised: Used in cases where we have lots of data but we do not know the answer.

Reinforcement: Used when problem involves some time sequential process or when the exact feedback is not available and only secondary signals are visible.

Chapter 4

EXPERIMENT

4.1 General

This part of the paper is the core of the entire research. Basically, it seeks to develop a neural network model that will help in predicting flight revenue by testing different training parameters, variable values and architectures.

The goal of this experiment is to compare the results of the various training parameter combinations and find out which, if any, of the neural network structure will best represent our model. It also wants to test whether fare classes are sufficient representatives of revenue and discover which other critical fields are required by our model. Since there are many approaches to develop a system, this experiment aims also at testing data mining applications and in particular the suitability of neural networks as a feasible, cost effective tool for building application systems.

Building models is not a simple sequence of steps, because the best choices are not always obvious. Consequently, data collection, data transformation, model building, and interpretation are iterative tasks that require vigorous tests.

Booking, schedule and revenue data of eight flights, namely, ET650, 661, 730, 731, 921, 920, 553, 562, representing Asia, Africa, Europe and America were collected from the various databases and cleansed. This process took eight man–weeks. The flights were selected based on good booking data, that had values in all fare classes and more or less complete data on round trip flights.

Using the data of flight ET650 random preliminary testing using different fields, training parameters and variables were conducted using two neural network development software, namely Knowledge Studio and Polyanalyst. Preprocessing of the data was performed by the software.

Using data of flight ET661 random secondary testing was done using the chosen software, Knowledge Studio, to select the best fields to build the model. The preliminary and secondary testing took two man-weeks. Finally, using the data of flight ET730 comprehensive and systematic training, testing and validating with a variety of combinations, was conducted using two types of algorithms and two different databases. This was the core of the experiment and took around six man-weeks. The training parameters of a single model (out of 327 test cases) was selected to represent the flight revenue information model. Finally, the selected model was scored or tested in a live environment using input data of two separate days (not used in the training) of flight ET730.

4.2 Software Selection

My purpose here was to acquire a suitable software with which to conduct my experiments and build a flight revenue model. Edelstein (1997) tells us that, "you can not categorize data mining tools into simple labels such as high-end or low-end because the products are too rich in functionality to be divided along just one dimension." As a result, my selection criteria included the following factors.

- Availability of Software: Willingness of vendor to provide software for evaluation or possibility of download.
- Data access: Easy access to external data sources.

- Capability of handling large files.
- Models: Capability of handling multiple neural network models and algorithms.
- Provision of graphs to display data.
- Ease of handling data preparation (cleaning, preprocessing).
- Availability of integrated graphics feature to present findings.

Finding the appropriate software for the purpose of this experiment was not an easy task. It took two man-weeks of research and correspondence to finally acquire and select two software which more or less fulfilled the above criteria. These software are Knowledge Studio version 2.0 and Polyanalyst version 4.0. Knowledge studio is a product produced by a Canadian company called Angoss Corporation (http://www.Angoss.com) and Polyanalyst is produced by a company named Megaputer Intelligence Inc. After going through their site and not being able to establish contact, I used the connection of a colleague in the USA to call the company and obtain a trial version of the software. Although it is not the goal of this research to evaluate software, a preliminary evaluation of the software was necessary to continue work with only the best one. The preliminary testing mentioned earlier was conducted to choose the software with which to build our model. The selection criteria included:

- Speed: Both how fast a model is built and how fast a deployed model can evaluate new data.
- Accuracy: Measured in the error rate of the model.
- Interpretation: How easily results are interpreted.
- Presentation: Alternative representation of results.

Data from ET650 (upto 74 records) was used for this evaluation. Over 90 tests were conducted on Knowledge Studio using different combinations of the values of the training parameters. These included:

- Partition sample size for learning and testing: 60%, 70%, 80%.
- Predictive model type: MLN, PPN, RBF.
- Training records: 55, 20, 65, 74.
- Number of neurons in hidden layer: 10, 20, 100,500.
- Root mean square error: 0.05.
- Learning type: Logistic, hyperbolic, linear.

80 tests were conducted on Polyanalyst using different parameters. These included:

- Multi Layer Network (MLN).
- Polynominal order.
- Maximum missing revenue value.
- Input variables.

Sample of the testing environment and results of knowledge STUDIO and Polyanalyst are shown in Table 4.2 and Tables 4.3a and 4.3b, respectively. Knowledge Studio was finally chosen as the software to be used for the experiment as it performed better in the above tests. Table 4.1shows the final results from a scale of 1 to 5.

S/W	Access	Handling large files	Algorithms	Graphs	Data handling	Interpret	Speed	Accuracy	Average
KS	4	3	4	4	3	4	4	4	3.75
Poly	4	3	3	3	3	1	1	2	2.75

Table 4.1 Comparison of Knowledge STUDIO and Polyanalyst

4.3 Model Selection

As earlier discussed, the network structure or topology combined with the learning paradigm and learning algorithm defines the neural network model. A number of models exist, including back propagation network, Kohenen feature maps, recurrent back propagation, probabilistic, radial basis function, adaptive resonance theory and others from which to choose our models. The best model for a given application or data mining function depends on the data and the function required.

Test No.	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
	TRAINING																RESULT			
4	L=% V=Rec	L=Random V=Random	L=60% V=38	None	MLN	Rev	d	20	Random	30	1	25	10	2000	0.05	Logistic	47.62%	04min	38	42.10%
5	L=% V=Rec	L=Random V=Random	L=60% V=38	None	MLN	Rev	e	20	Random	30	1	25	10	2000	0.05	Logistic	82,61%	04min	36	47.36%
7	L=% V=Rec	L=Random V=Random	L=60% V=38	None	MLN	Rev	g	20	Random	30	1	25	10	2000	0.05	Logistic	91.30%	05min	53	50.00%
8	L=% V=Rec	L=Random V=Random	L=60% V=38	None	MLN	Rev	h	20	Random	30	1	25	10	2000	0.05	Logistic	84.78%	09min	318	44.73%
11	L=% V=Rec	L=Random V=Random	L=70% V=28	None	MLN	Rev	b	65	Random	30	1	25	10	2000	0.05	Logistic	68.00%	06min	55	46.42%
13	L=% V=Rec	L=Random V=Random	L=70% V=28	None	MLN	Rev	d	65	Random	30	1	25	10	2000	0.05	Logistic	37.04%	03min	29	42.85%
14	L=% V=Rec	L=Random V=Random	L=70% V=28	None	MLN	Rev	e	65	Random	30	1	25	10	2000	0.05	Logistic	40.00%	03min	31	42.85%
16	L=% V=Rec	L=Random V=Random	L=70% V=28	None	MLN	Rev	g	65	Random	30	1	25	10	2000	0.05	Logistic	66.00%	05min	54	64.28%
20	L=% V=Rec	L=Random V=Random	L=80% V=19	None	MLN	Rev	b	74	Random	30	1	25	10	2000	0.05	Logistic	74.32%	04min	40	73.68%
23	L=% V=Rec	L=Random V=Random	L=80% V=19	None	MLN	Rev	e	74	Random	30	1	25	10	2000	0.05	Logistic	13.00%	03min	33	42.10%
27	L=% V=Rec	L=Random V=Random	L=80% V=19	None	MLN	Rev	i	74	Random	30	1	25	10	2000	0.05	Logistic	37.50%	03min	35	47.36%
29	L=% V=Rec	L=Random V=Random	L=60% V=38	TTL Bkg	MLN	Rev	b	55	Random	30	1	25	20	2000	0.05	Logistic	63.16%	02min	24	42.10%
34	L=% V=Rec	L=Random V=Random	L=60% V=38	TTL Bkg	MLN	Rev	g	55	Random	30	1	25	20	2000	0.05	Logistic	86.84%	07min	69	55.26%
35	L=% V=Rec	L=Random V=Random	L=60% V=38	TTL Bkg	MLN	Rev	h	55	Random	30	1	25	20	2000	0.05	Logistic	76.32%	09min	110	55.26%
36	L=% V=Rec	L=Random V=Random	L=60% V=38	TTL Bkg	MLN	Rev	i	55	Random	30	1	25	20	2000	0.05	Logistic	66.67%	04min	43	47.36%
38	L=% V=Rec	L=Random V=Random	L=70% V=28	TTL Bkg	MLN	Rev	b	65	Random	30	1	25	20	2000	0.05	Logistic	67.69%	04min	42	64.28%
41	L=% V=Rec	L=Random V=Random	L=70% V=28	TTL Bkg	MLN	Rev	e	65	Random	30	1	25	20	2000	0.05	Logistic	67.44%	04min	44	60.71%
43	L=% V=Rec	L=Random V=Random	L=70% V=28	TTL Bkg	MLN	Rev	g	65	Random	30	1	25	20	2000	0.05	Logistic	86.05%	08min	91	50.00%
47	L=% V=Rec	L=Random V=Random	L=80% V=19	TTL Act Bkg	MLN	Rev	b	74	Random	30	1	25	20	2000	0.05	Logistic	86.49%	06min	65	47.36%
50	L=% V=Rec	L=Random V=Random	L=80% V=19	TTL Act Bkg	MLN	Rev	e	74	Random	30	1	25	20	2000	0.05	Logistic	36.84%	03min	31	47.36%
52	L=% V=Rec	L=Random V=Random	L=80% V=19	TTL Act Bkg	MLN	Rev	g	74	Random	30	1	25	20	2000	0.05	Logistic	63.16%	04min	42	47.36%

Table 4.2 Sample (30 results out of 92 cases) of preliminary Testing Environment of Knowledge STUDIO using ET 650

Where : -

a = A/C, Date, Segment,TTL Bkg,NS,NR,TTL Act Bkg,Bkg by Class,Mileage.
b = A/C,Date,NR,NS,Segment,TTL Bkg
c = A/C,Date,TTL Act Bkg,Bkg by Class,Segment
d = Date,TTL Act Bkg,Bkg by Class,Segment
e= Date,TTL Bkg,NS,NR,Segment
f = Segment,Bkg by Class
g = Date,Segment,TTL Bkg,NS,NR,TTL Act Bkg
h = TTL Bkg,NS,NR
i = TTL Act Bkg,Bkg by Class

ET 650 Polyanalyst Training Description						ET Polyanalyst 650 Training result							
Test No.	Exploration Engine.	Target Learning Field	Exploration Engine's order type.	Exploration engine's minimal part of predicted values(%)	Dependent fields during learning	Training time elapsed	Significance index.	Standard error	R-squared	Standard deviation	No of network layers	No of network nodes
1	PolyNet predictor	Revenue	3rd order	60	Except Z & V.	14 sec.	8.136	0.52714	0.7222	3.29E+08	1	3
2	PolyNet predictor	Revenue	3rd order	50	Except Z & V, Flt No & A/C	01 sec.	7.423	0.5271	0.7222	3.29E+08	1	3
3	PolyNet predictor	Revenue	3rd order	60	Except Flt No, A/C, DCP22, NS & NR	03 sec.	7.063	0.5271	0.7222	3.29E+08	1	3
4	PolyNet predictor	Revenue	3rd order	40	Except Flt No, A/C, NS, NR, Z & V	02 sec.	9.077	0.5271	0.7222	3.29E+08	1	3
5	PolyNet predictor	Revenue	3rd order	70	Except A/C	02 sec.	7.423	0.5271	0.7222	3.29E+08	1	3

Table 4.3a Sample preliminary test environment of Polyanalyst using ET650

Flt. No.	Date	A/C	Segment	TTL Bkg	NS	NR	TTL Ack Bkg	C	D	Z	S	B	H	K	L	M	Q	T	V	Miles	Revenue	PN Revenue	Error %
ET650	09/17/99	"B767"	"ADD-BKK"	178	13	0	127	4	0	0	1	1	5	15	3	21	77	0	0	4186	278083	118440	57.0
ET650	10/15/99	"B763"	"ADD-BOM"	113	38	0	59	4	0	0	2	0	9	5	8	12	17	2	0	2381	110558	126933	15.0
ET650	10/15/99	"B763"	"ADD-BKK"	116	16	0	86	4	0	0	0	0	1	12	11	28	30	0	0	4186	178063	146855	17.5
ET650	08/27/99	"B767"	"ADD-BKK"	111	4	0	108	4	0	0	0	3	0	6	12	15	68	0	0	4186	221656	237176	7.0
ET650	05/14/99	"B767"	"ADD-BOM"	97	15	0	80	4	1	0	1	0	0	11	8	36	12	7	0	2381	150246	191375	27.4
ET650	08/20/99	"B767"	"ADD-BOM"	61	17	0	48	4	0	0	4	0	5	6	3	20	6	0	0	2381	84566	142479	68.5
ET650	02/18/00	""	"ADD-BKK"	48			40	4	0	0	0	0	9	17	0	10	0	0	0	4186		-64020.9	
ET650	09/03/99	"B757"	"ADD-BOM"	42	3	0	29	4	0	0	7	0	0	0	8	3	7	0	0	2381	56920	145688	155.0
ET650	01/28/00	"B767"	"ADD-BOM"	98	10	0	85	4	0	0	5	0	10	11	23	17	15	0	0	2381		-20644.5	
ET650	01/28/00	"B767"	"ADD-BKK"	42	3	0	43	4	0	0	4	0	0	12	13	10	0	0	0	4186		30841.9	
ET650	07/02/99	"B767"	"ADD-BKK"	50	1	0	44	5	1	0	3	0	2	11	2	20	0	0	0	4186	121676	116791	4.0
ET650	01/14/00	"B767"	"ADD-BOM"	109	9	0		5	1	0	0	0	4	3	20	22	20	2	0	2381		25294.4	
ET650	09/17/99	"B767"	"ADD-BOM"	69	7	0	50	5	0	0	1	0	8	1	12	14	8	1	0	2381	122011	146508	20.1
ET650	12/24/99	"B767"	"ADD-BKK"	57	14	0	42	5	0	0	1	0	3	6	4	1	22	0	0	4186	141563	68675.3	51.5
ET650	07/23/99	"B767"	"ADD-BOM"	107	7	0	98	5	0	0	3	0	3	11	23	31	19	3	0	2381	172127	164185	4.6
ET650	03/03/00	""	"ADD-BKK"	70			61	5	0	0	0	0	0	16	22	18	0	0	0	4186		-141738	
ET650	07/16/99	"B763"	"ADD-BOM"	149	7	0	123	5	0	0	1	0	0	12	11	52	31	11	0	2381	198327	250022	26.1
ET650	10/08/99	"B763"	"ADD-BKK"	132	11	0	108	5	0	0	1	1	5	4	6	52	34	0	0	4186	223170	163654	26.7
ET650	02/11/00	"B763"	"ADD-BOM"	96	8	0	66	5	3	0	2	0	12	7	11	13	13	0	0	2381		-70562.2	
ET650	09/24/99	"B767"	"ADD-BOM"	76	44	0	41	5	0	0	1	0	1	2	3	10	18	1	0	2381	71641	133629	86.5
ET650	09/10/99	"B767"	"ADD-BKK"	152	5	0	139	5	0	0	1	0	6	16	3	15	93	0	0	4186	268149	-184595	168.8
ET650	08/13/99	"B767"	"ADD-BOM"	66	1	0	54	6	5	0	4	0	3	3	7	7	17	1	1	2381	86444	144514	67.2
ET650	10/01/99	"B763"	"ADD-BKK"	129	10	0	109	6	0	0	2	0	3	15	6	43	34	0	0	4186	287373	168300	41.4
ET650	10/22/99	"B763"	"ADD-BOM"	111	4	0	92	6	4	0	2	4	7	8	14	21	24	2	0	2381	180005	128838	28.4
ET650	03/10/00	""	"ADD-BOM"	81			66	6	0	0	1	0	1	8	19	12	19	0	0	2381		-165108	
ET650	01/07/00	"B757"	"ADD-BOM"	86	12	0	70	7	0	0	1	0	2	4	22	11	23	0	0	2381		48145	
ET650	07/02/99	"B767"	"ADD-BOM"	139	28	0	119	7	1	0	7	0	5	2	36	47	2	12	0	2381	199247	116118	41.7
ET650	10/22/99	"B763"	"ADD-BKK"	78	14	0	56	7	0	0	1	0	3	17	9	4	15	0	0	4186	153589	125895	18.0
ET650	07/16/99	"B763"	"ADD-BKK"	97	4	0	91	7	0	0	3	3	1	22	16	7	32	0	0	4186	187322	257283	37.3
ET650	07/30/99	"B763"	"ADD-BKK"	135	4	0	126	7	0	0	0	0	0	15	13	16	75	0	0	4186	277476	265542	4.3
ET650	02/25/00	""	"ADD-BOM"	93			65	8	0	0	0	0	6	4	22	10	15	0	0	2381		-122888	
ET650	12/03/99	"B767"	"ADD-BOM"	157	6	0	103	8	1	0	0	0	1	2	22	28	41	0	0	2381	194530	161383	17.0
ET650	03/03/00	""	"ADD-BOM"	93			79	8	2	0	0	0	2	9	24	14	16	4	0	2381		-149959	
ET650	02/11/00	"B763"	"ADD-BKK"	48	3	0	46	9	0	0	0	0	7	8	7	15	0	0	0	4186		-29603.8	
ET650	12/10/99	"B757"	"ADD-BOM"	102	20	0	79	9	0	0	6	1	0	5	13	12	32	1	0	2381	124301	117980	5.1
ET650	02/18/00	""	"ADD-BOM"	80			64	9	4	0	1	0	4	14	11	10	11	0	0	2381		-100348	
ET650	04/30/99	"B757"	"ADD-BOM"	108	16	0	81	10	0	0	0	0	9	4	20	22	14	2	0	2381	141215	245566	73.9
ET650	05/07/99	"B763"	"ADD-BOM"	89	7	0	68	10	0	0	2	1	7	1	8	19	13	7	0	2381	141326	216358	53.1
ET650	11/19/99	"B763"	"ADD-BOM"	175	11	0	107	10	3	0	0	0	0	4	11	36	42	1	0	2381	159007	170446	7.2
ET650	08/06/99	"B767"	"ADD-BKK"	138	0	1	134	11	0	0	1	1	2	16	11	42	50	0	0	4186	294654	307622	4.4
ET650	10/01/99	"B763"	"ADD-BOM"	109	12	0	92	11	0	0	2	0	10	6	11	19	33	0	0	2381	161649	164883	2.0
ET650	07/30/99	"B763"	"ADD-BOM"	83	5	0	78	11	0	0	3	0	2	3	14	19	17	9	0	2381	134777	150807	11.9
ET650	09/10/99	"B767"	"ADD-BOM"	56	1	0	54	12	1	0	0	0	1	5	5	8	20	2	0	2381	95044	140572	47.9
ET650	05/21/99	"B767"	"ADD-BOM"	122	12	0	89	13	1	0	4	0	4	4	16	24	19	4	0	2381	182653	260977	42.9
ET650	12/17/99	"B757"	"ADD-BOM"	135	16	0	106	14	0	0	5	0	2	9	15	35	26	0	0	2381	184037	88882	51.7
ET650	07/09/99	"B767"	"ADD-BOM"	125	4	0	112	14	1	0	2	0	2	4	12	34	37	6	0	2381	211637	311819	47.3

Table 4.3b Result of Polyananlyst

In this connection, the situation of the research problem can be summarized as follows.

- All the information bearing on the problem can be brought at once and sequence may not be important.
- Our database contains both fields for the problem and the output, revenue.
- The application or function we are trying to model is predictive in nature.

On the basis of the foregoing and the discussion presented in section 3.3.3 of this report, it was felt that the research problem be modeled using a supervised training paradigm, a feedforward topology and a predictive modeling function. Using Table 4.4, provided by Bigus (1996), the models that seem to best represent our problem are the Back Propagation Network (BPN) and the Radial Basis Function (RBF).

Model	Training Paradigm	Topology	Function/Application
Adaptive Resonance Theory	Unsupervised	Recurrent	Clustering
ARTMAP	Supervised	Recurrent	Classification
Back Propagation	Supervised	Feed forward	Classification, predictive modeling, time-series
Radial Basis Function networks	Supervised	Feed forward	Classification, predictive modeling, time-series
Probabilistic neural networks	Supervised	Feed forward	Classification
Kohenen feature maps	Unsupervised	Feed forward	Clustering
Recurrent back propagation	Supervised	Limited recurrent	Predictive modeling, time-series
Temporal difference learning	Reinforcement	Feed forward	Time-series

Table 4.4: Neural Network Models and Their Functions

Back propagation network (BPN) uses a feed forward topology, supervised, and the back propagation learning algorithm. According to Bigus (1996), it is the most commonly used model but expensive in terms of computational requirements for training.

The basic back propagation algorithm consists of the following three steps.

- Input pattern is presented to the input layer of the network.
- Inputs are propagated through the network until they reach the output units producing the actual or predicted output pattern.
- Actual outputs are subtracted from the desired outputs and an error signal is produced. (The error signal is the basis for the back propagation step).

The errors are passed back through the neural network by computing the contribution of each hidden processing unit and deriving the corresponding adjustment needed to produce the correct output. The connection weights are then adjusted and the neural network is said to have just 'learned' from an experience.

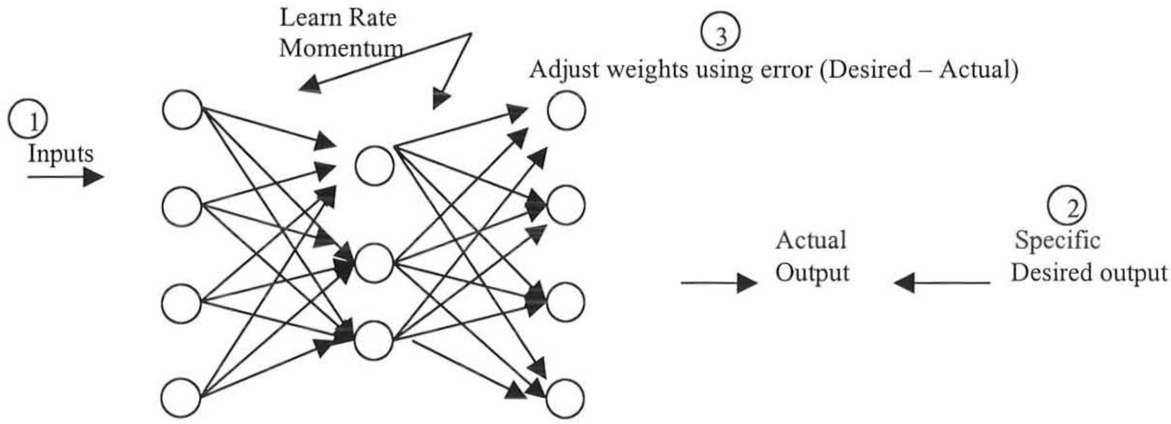


Figure 4.1: Back Propagation Network

The radial basis function (RBF) uses a feed forward topology, supervised, and a radial basis function learning algorithm. It is typically configured with a single hidden layer of units whose activation function is selected from a class of functions called basis functions.

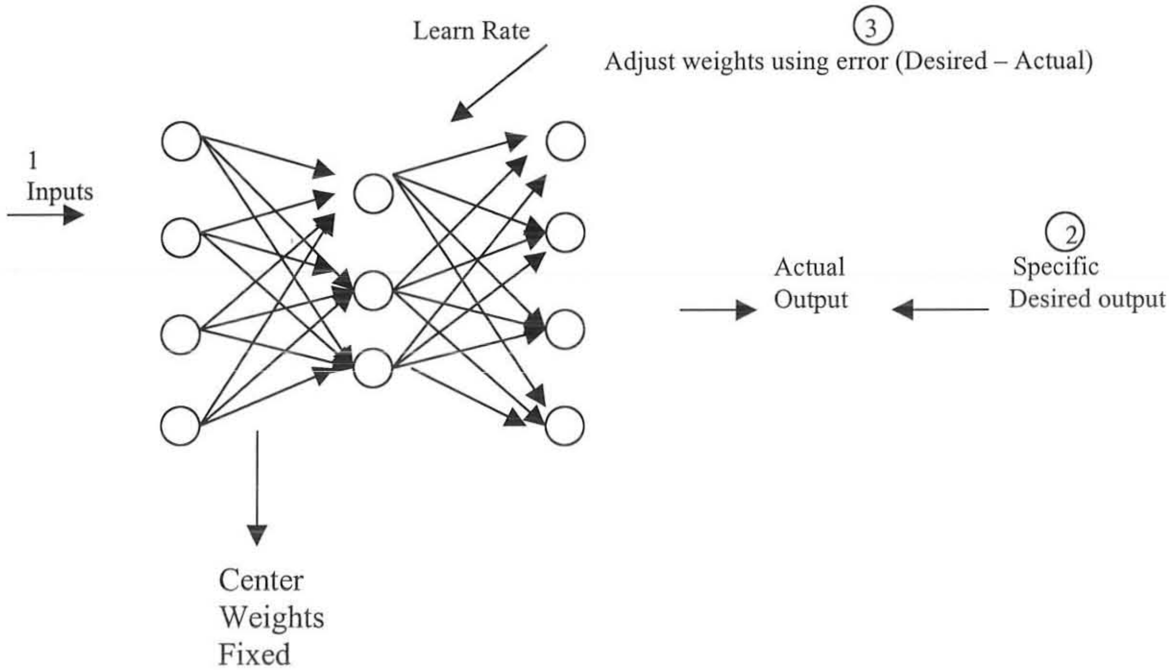


Figure 4.2: Radial Basis Function Network

Bigus (1996) notes that while BPN are similar in many respects to RDF, the latter has the advantage in that it usually trains much faster than BPN. The major difference is the behavior of the single hidden layer. Rather than using the sigmoidal or S-shaped activation function as in BPN, the hidden units in RBF networks use a Gaussian or some other basis function.

4.4 Building the Model

4.4.1 General

The process of building a model involves finding the connection weights that produce the most accurate results by "training" the neural network with data. After enough passes through the training data have been made, the neural network typically becomes a good predictor. Our goal is to minimize the prediction error.

Training a neural network and building a good model is the hardest part of using neural networks for data mining. It is the equivalent step of writing the algorithm of a program, then coding and testing it using a conventional programming language.

Following is an outline of the major steps in building the model using the selected software, knowledge STUDIO, together with the screen used in each step. All screens are taken from the system provider's guide and do not represent my data. They have been included only to better present the graphical user interface and to help visualize some of the menu.

1. Importing and transforming the collected and stored data from the source database stored in MS Excel 97.
   - Figure 4.3 shows the screen after the data has been imported. To transform the data we choose a field and specify the required properties.

2. Cleansing the data.
    - Table 4.6 shows the screen of the data set that has been transformed. Based on the values shown cleansing is performed. The question marks show missing values.
3. Preprocessing.
    - Table 4.5 shows the fields used to represent the data. Based on the target type, length, etc. specified by us, the preprocessing activities are internally and automatically performed by the software.
4. Partitioning the data set between learning and validating data sets.
    - Figure 4.4 shows the screen used for this activity.
5. Selecting the type of model (BPN, RBF, etc.) to be used for training
    - Figure 4.5 is the screen where the type of neural network model that is to be used is specified.
6. Selecting the dependent and independent variables (Figure 4.6).
7. Partitioning the learning data set into training and testing data sets to minimize overfitting (Figure 4.7).
8. Specifying training parameters for the selected model (Figure 4.8).
9. Presenting the training pattern to the neural network.
    - Figure 4.8 shows the screen in which different training parameters are selected. After the parameters are selected and the 'next' button is pressed the software provides a summary of all the values input upto this step. If we 'OK' the screen the training process will commence. Figure 4.9 shows the status of the model during the training process.

10. Validating the training by using the validating set produced in step 4.

    - Figure 4.10 shows the validation screen. If we are not satisfied with the result we retrain the model; if we are satisfied, we save the model and the weights are locked.

11. To use the model developed in a live environment, we recall the saved model and present the data using the score facility in Figure 4.10

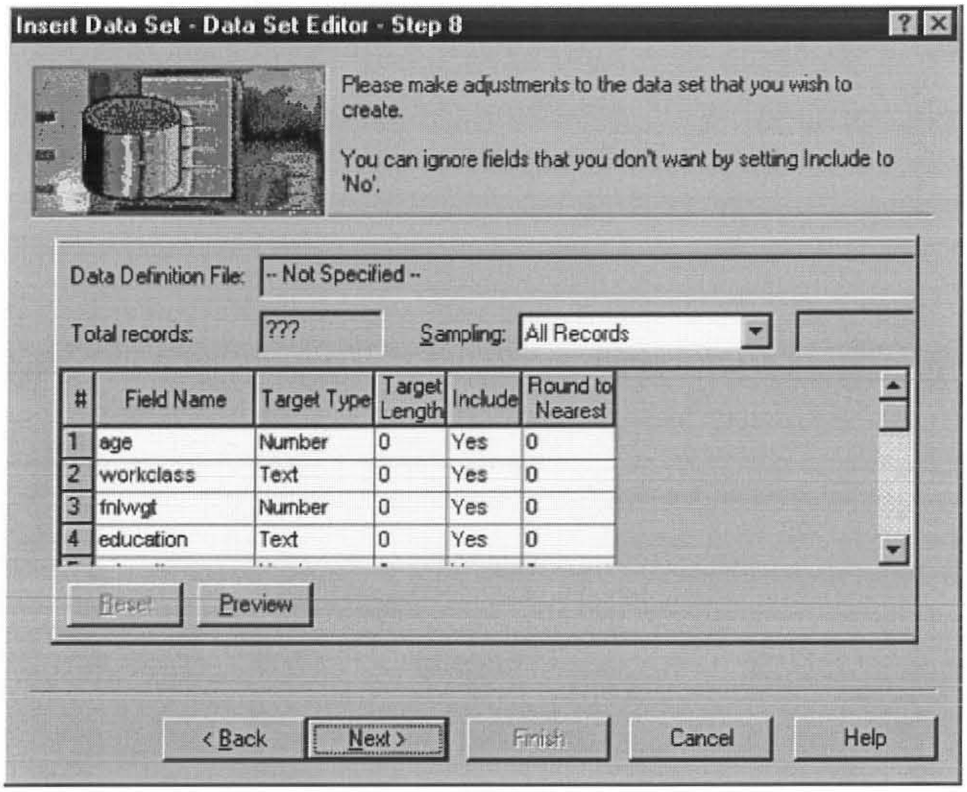


Fig 4.3 Data Transformation Screen

#	Field Name	Target Type	Target Length	Include	Round to Nearest
1	Flight No	Text	0	Yes	0
2	Date	Date	0	Yes	0
3	Segment	Text	0	Yes	0
4	C	Number	0	Yes	0
5	D	Number	0	Yes	0
6	S	Number	0	Yes	0
7	B	Number	0	Yes	0
8	H	Number	0	Yes	0
9	K	Number	0	Yes	0
10	L	Number	0	Yes	0
11	M	Number	0	Yes	0
12	Q	Number	0	Yes	0
13	T	Number	0	Yes	0
14	Revenue	Number	0	Yes	0

Table 4.5 Data Representation

	Date	A/C	Segment	TTL Bkg	NS	NR	TTL Act Bkg	C	D	S	B	H	K	L	M	Q	T	ASK	RSK	LF by Flt	Yield	Mileage	Revenue
1	Date	A/C	Segment	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	???	B757	LHR-FCO	8	2	0	8	1	0	0	0	0	0	1	1	0	5	???	???	???	???	897	7615
3	???	B757	LHR-ADD	29	2	0	29	0	0	0	0	0	0	1	7	21	0	???	???	???	???	3676	70603
4	???	B757	FCO-AD	32	4	0	35	1	0	0	0	1	1	4	19	9	0	???	???	???	???	2782	81338
5	???	???	???	???	???	???	???	???	???	???	???	???	???	???	???	???	???	856515	325887	0.38	0.49	???	159556
6	???	B757	LHR-FCO	2	0	2	6	0	0	3	0	0	0	0	0	0	3	???	???	???	???	897	12514
7	???	B757	LHR-ADD	44	8	0	32	8	0	0	0	0	0	2	8	14	0	???	???	???	???	3676	196232
8	???	B757	FCO-ADD	47	2	0	46	3	2	2	3	3	0	2	23	6	0	???	???	???	???	2782	183675
9	???	???	???	???	???	???	???	???	???	???	???	???	???	???	???	???	???	856515	412123	0.48	0.95	???	392421
10	???	B757	LHR-FCO	9	3	0	0	0	0	0	0	0	0	0	0	0	0	???	???	???	???	897	2913
11	???	B757	LHR-ADD	26	4	0	24	2	0	0	0	3	0	7	5	7	0	???	???	???	???	3676	83911
12	???	B757	FCO-ADD	28	10	0	26	2	0	0	0	1	1	3	9	10	0	???	???	???	???	2782	49763
13	???	???	???	???	???	???	???	???	???	???	???	???	???	???	???	???	???	921492	230874	0.25	0.59	???	136587
14	???	B763	LHR-FCO	0	0	3	3	3	0	0	0	0	0	0	0	0	0	???	???	???	???	897	14195
15	???	B763	LHR-ADD	45	24	0	25	0	0	0	1	0	0	6	11	7	0	???	???	???	???	3676	78412
16	???	B763	FCO-ADD	13	1	0	14	2	0	0	0	0	0	2	8	2	0	???	???	???	???	2782	31544
17	???	???	???	???	???	???	???	???	???	???	???	???	???	???	???	???	???	1435401	215856	0.15	0.58	???	124151
18	???	B757	LHR-FCO	4	0	4	4	1	0	0	0	0	0	0	0	0	3	???	???	???	???	897	1992
19	???	B757	LHR-ADD	50	4	0	50	0	2	1	0	2	1	11	9	21	3	???	???	???	???	3676	129660
20	???	B757	FCO-ADD	33	0	0	28	0	2	0	0	1	1	7	8	9	0	???	???	???	???	2782	131652
21	???	???	???	???	???	???	???	???	???	???	???	???	???	???	???	???	???	856515	575583	0.67	0.46	???	263304
22	???	B757	LHR-FCO	13	3	0	12	0	0	0	0	0	0	2	1	0	9	???	???	???	???	897	14237
23	???	B757	LHR-ADD	36	4	0	24	0	0	0	0	2	0	5	7	10	0	???	???	???	???	3676	36927
24	???	B757	FCO-ADD	36	1	0	29	0	0	0	0	0	0	1	19	9	0	???	???	???	???	2782	59617
25	???	B757	???	???	???	???	???	???	???	???	???	???	???	???	???	???	???	921492	492384	0.53	0.22	???	110781
26	???	B757	LHR-FCO	11	2	0	10	0	0	0	2	0	0	2	0	0	6	???	???	???	???	897	10944
27	???	B757	LHR-ADD	34	7	0	28	0	0	0	1	1	1	4	12	9	0	???	???	???	???	3676	81410
28	???	B757	FCO-ADD	29	0	1	30	0	0	0	0	1	0	6	19	4	0	???	???	???	???	2782	76243
29	???	???	???	???	???	???	???	???	???	???	???	???	???	???	???	???	???	1080981	349515	0.32	0.48	???	168597
30	???	B757	LHR-FCO	3	0	0	3	0	0	0	1	0	0	0	0	0	2	???	???	???	???	897	5116

Table 4.6 Visualization of Data Set

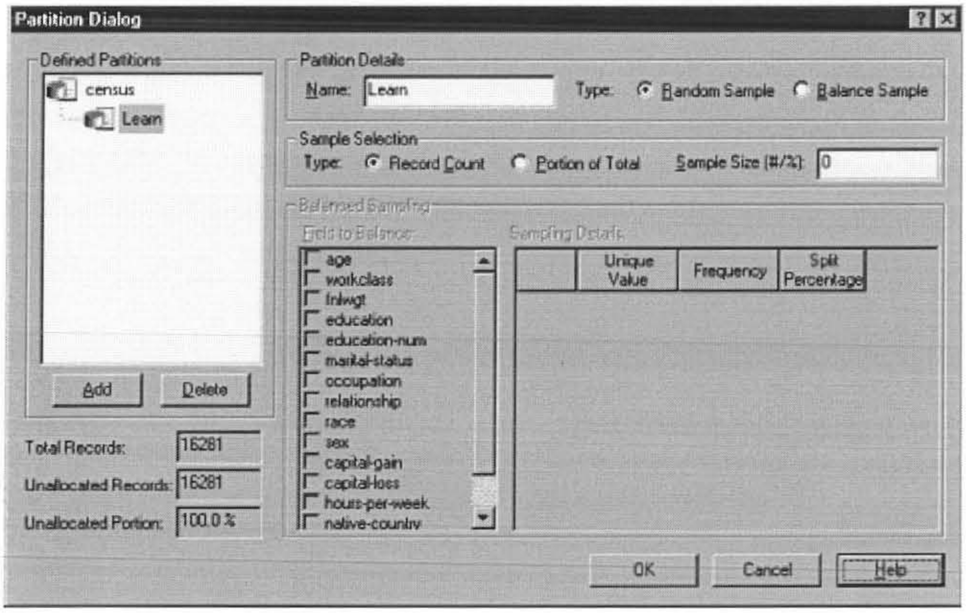


Fig 4.4 Data Set Partitioning Screen

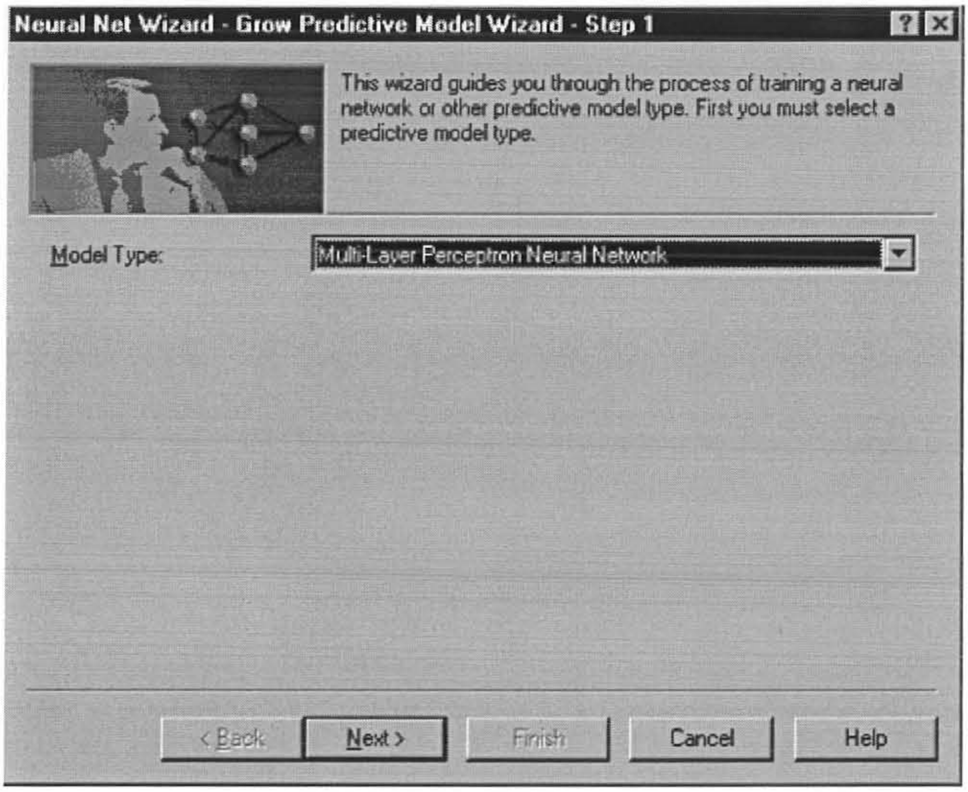


Fig 4.5 Model Selection Screen

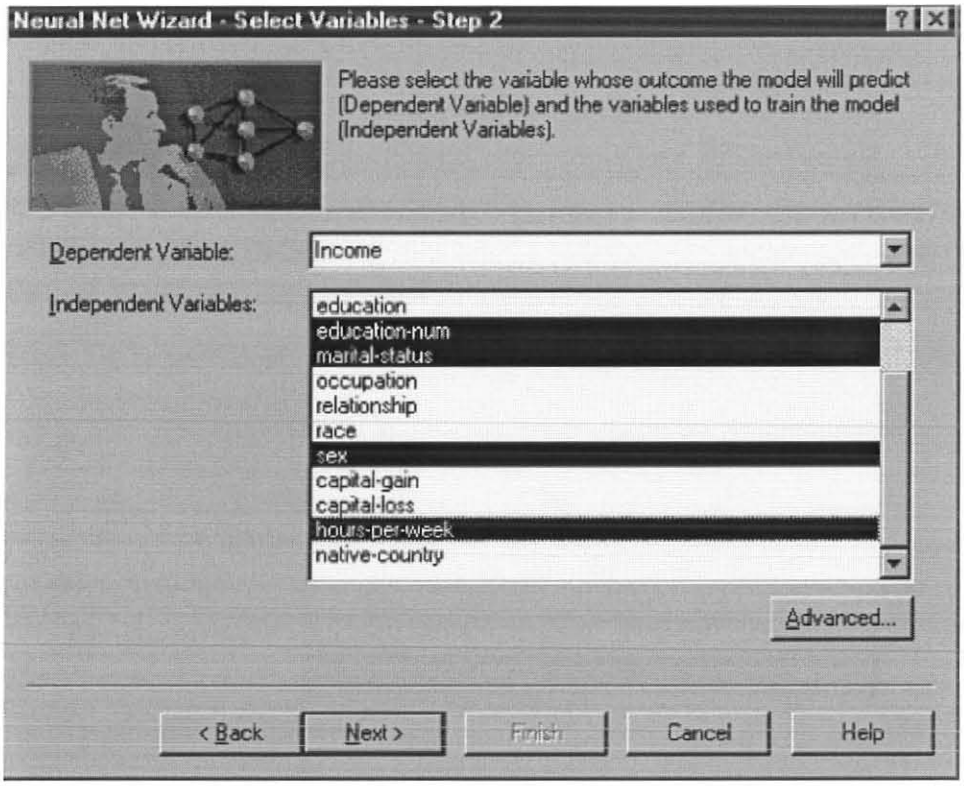


Fig 4.6 Variable type Selection Screen

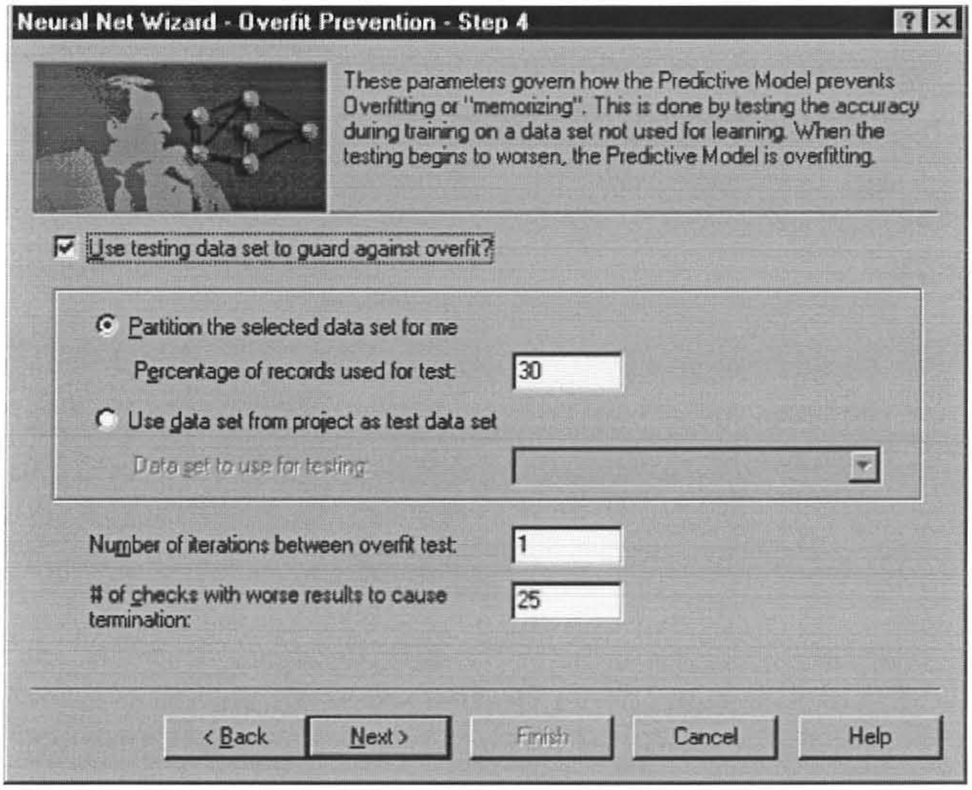


Fig 4.7 Learning Data Set Partitioning Screen

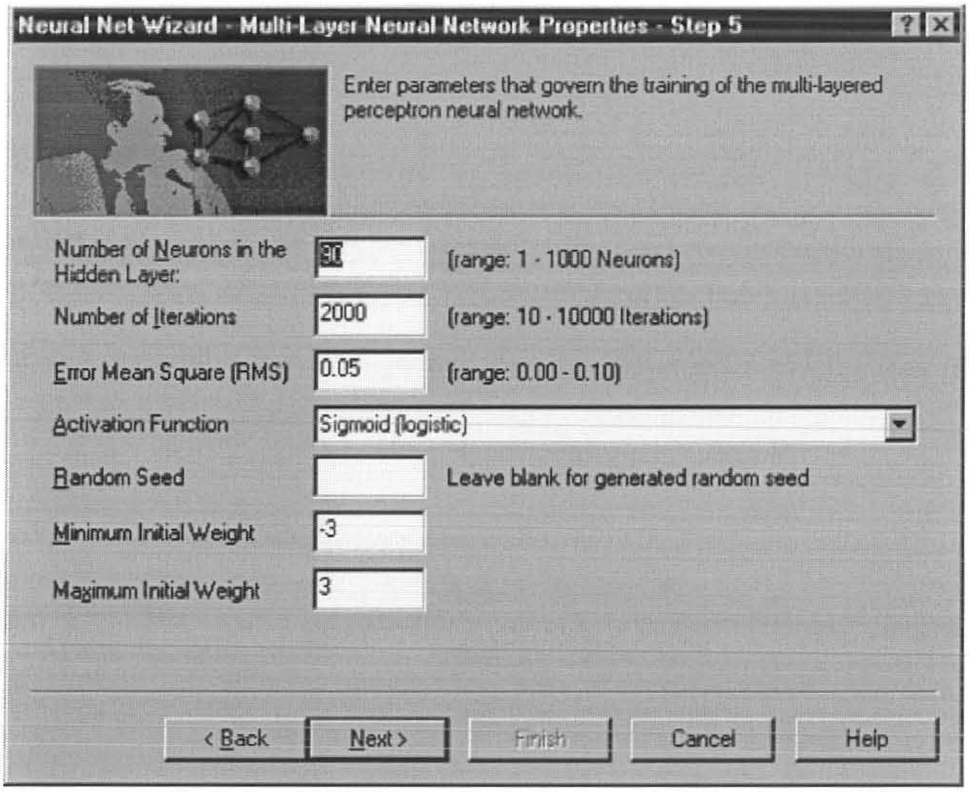


Fig 4.8 Training Parameter Selection Screen

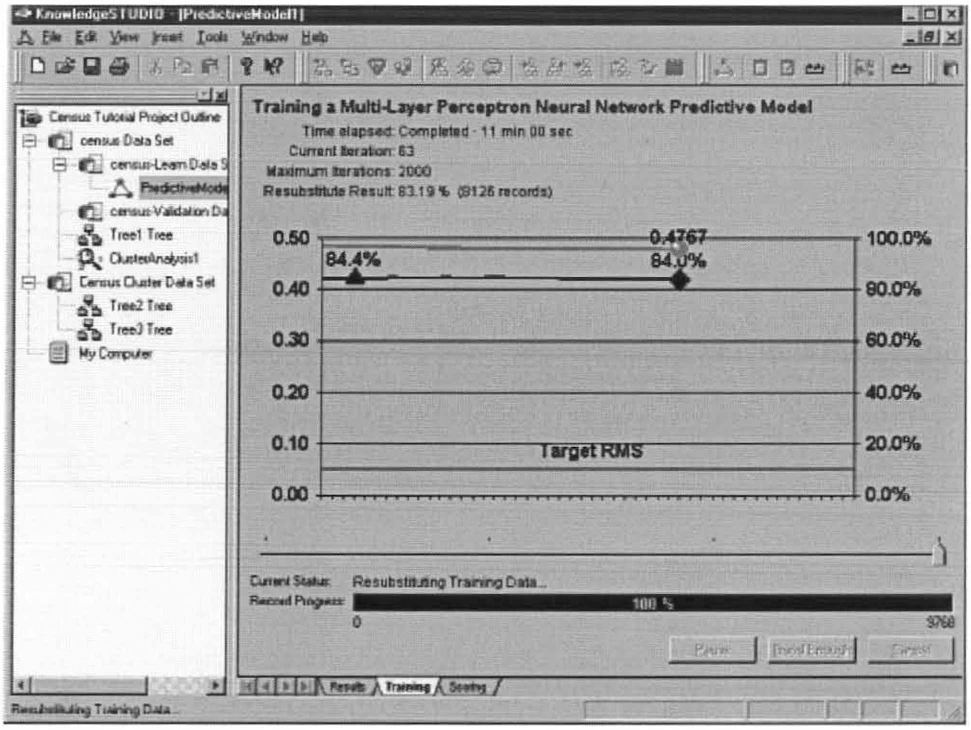


Fig 4.9 Training Status Screen

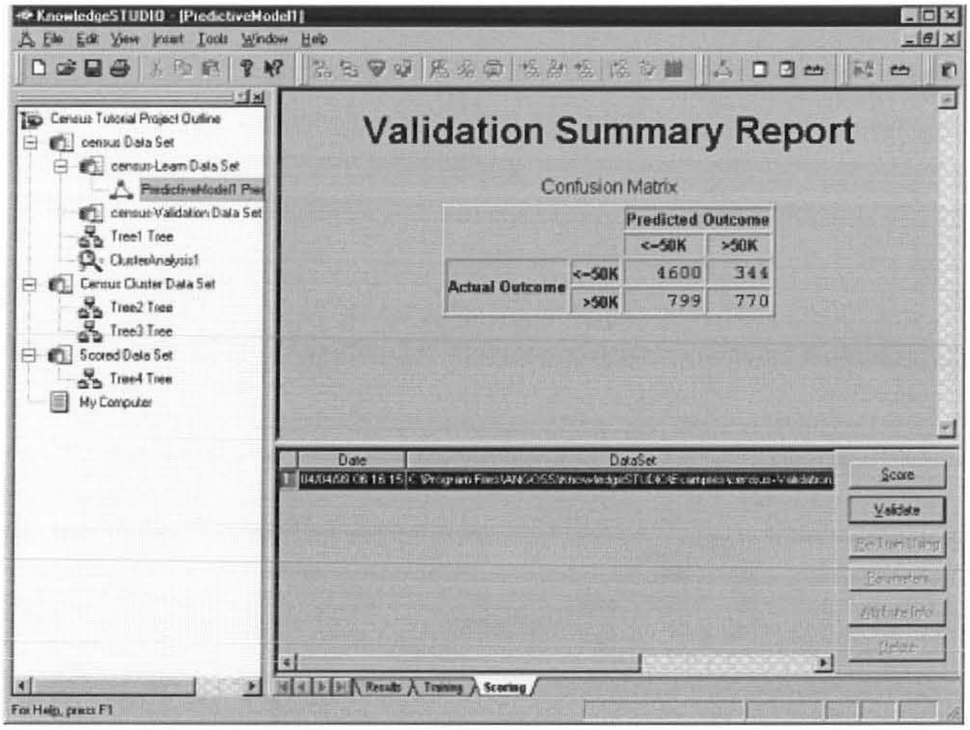


Fig 4.10 Validation and Scoring Screen

4.4.2 Data Preparation

Data preparation is one of the major phases in data mining. It consists of collecting, cleansing, and preprocessing data.

a) Collection

Data capture involves the processes of finding the relevant data, recording it and converting it into a suitable form for further processing. The quality of the recording and the conversion processes used are crucial both to the quality of the data provided and to the consequent diagnoses and remedial actions.

The initial fields identified pertinent to flight revenue were 19 (including the output field; revenue). These fields include date, aircraft, segment, total advance booking, no show, no record, total actual booking, 10 fare classes, mileage and revenue. For flight ET730 alone, a year's data (20 January, 1999 to 12 January, 2000) was collected. This consisted of 132 records and amounted to about 65KB of data. This data was collected for all the 8 flights selected over a period of about one year. Table 4.7a shows a sample of this database (DB1) for flight ET730.

Another database (DB2) was derived from the values of DB1 and from new fields that were obtained from spreadsheet reports on hard copy. The segment and mileage fields were omitted from DB1 and 4 new fields; availed seat kilometer, revenue seat kilometer, load factor and yield were added to DB2. Appendix 4.7b shows the entire records and fields of DB2.

Flight No	Date	A/C	Segment	TTL Bkg	NS	NR	TTL Act Bkg	C	D	S	B	H	K	L	M	Q	T	Mileage	Revenue
ET730	20-Jan-99	B757	ADDFCO	68	0	0	38	3	0	1	1	2	2	13	21	9	2	2782	113,943
ET730	20-Jan-99	B757	ADDLHR	62	3	0	31	4	0	0	0	2	0	4	18	32	1	3676	111,000
ET730	20-Jan-99	B757	FCOLHR	21	0	0	5	0	0	0	0	0	0	0	0	0	21	897	4,532
ET730	3-Feb-99	B763	ADDFCO	75	2	0	31	3	0	1	1	2	2	13	21	9	2	2782	74,359
ET730	3-Feb-99	B763	ADDLHR	62	0	1	51	4	0	0	0	1	0	3	14	26	1	3676	105,915
ET730	3-Feb-99	B763	FCOLHR	21	0	1	5	0	0	0	0	0	0	0	0	0	20	897	4,152
ET730	3-Mar-99	B757	ADDFCO	68	5	0	41	0	0	1	1	2	2	13	21	9	2	2782	114,890
ET730	3-Mar-99	B757	ADDLHR	57	3	0	44	0	0	0	0	1	0	3	14	26	1	3676	90,583
ET730	3-Mar-99	B757	FCOLHR	21	0	1	4	0	0	0	0	0	0	0	0	0	20	897	2,068
ET730	10-Mar-99	B757	ADDFCO	68	11	0	48	0	0	1	1	2	2	13	21	9	2	2782	77,076
ET730	10-Mar-99	B757	ADDLHR	57	12	0	54	0	0	0	0	1	0	3	14	26	1	3676	129,408
ET730	10-Mar-99	B757	FCOLHR	21	0	0	17	0	0	0	0	0	0	0	0	0	20	897	1,848
ET730	17-Mar-99	B757	ADDFCO	68	7	0	56	0	0	1	1	2	2	13	21	9	2	2782	142,864
ET730	17-Mar-99	B757	ADDLHR	57	10	0	42	0	0	0	0	1	0	3	14	26	1	3676	121943
ET730	17-Mar-99	B757	FCOLHR	21	2	0	5	0	0	0	0	0	0	0	0	0	20	897	3,009
ET730	24-Mar-99	B767	ADDFCO	68	16	0	48	0	0	1	1	2	2	13	21	9	2	2782	202,843
ET730	24-Mar-99	B767	ADDLHR	57	3	0	104	0	0	0	0	1	0	3	14	26	1	3676	145,259
ET730	24-Mar-99	B767	FCOLHR	21	1	0	1	0	0	0	0	0	0	0	0	0	20	897	658
ET730	28-Apr-99	B757	ADDFCO	33	4	0	29	0	0	2	0	2	1	5	9	8	0	2782	78,409
ET730	28-Apr-99	B757	ADDLHR	70	15	0	55	0	0	0	0	2	2	4	23	14	9	3676	102,638
ET730	28-Apr-99	B757	FCOLHR	14	6	0	8	0	0	0	0	0	0	0	1	0	7	897	7,318
ET730	5-May-99	B757	ADDFCO	50	4	0	36	1	0	0	0	1	2	5	8	21	0	2782	86,608
ET730	5-May-99	B757	ADDLHR	62	6	0	51	2	1	1	0	4	0	3	18	17	1	3676	142,367
ET730	5-May-99	B757	FCOLHR	51	3	0	50	0	0	0	0	0	0	2	0	0	48	897	31,362
ET730	12-May-99	B757	ADDFCO	56	5	0	40	1	1	1	1	5	2	10	10	12	0	2782	112,721
ET730	12-May-99	B757	ADDLHR	77	9	0	61	2	1	0	1	1	0	12	26	17	4	3676	154,316
ET730	12-May-99	B757	FCOLHR	7	0	0	8	0	0	0	0	0	0	0	0	0	8	897	4,352
ET730	19-May-99	B757	ADDFCO	36	1	0	32	4	0	0	0	0	0	4	14	7	4	2782	76,549
ET730	19-May-99	B757	ADDLHR	44	6	0	38	1	1	2	0	0	1	6	14	15	2	3676	89,712
ET730	19-May-99	B757	FCOLHR	7	0	0	7	0	0	0	0	0	0	0	0	0	7	897	4,632
ET730	26-May-99	B763	ADDFCO	36	6	0	29	2	4	1	0	0	0	3	13	8	2	2782	67,583
ET730	26-May-99	B763	ADDLHR	71	10	0	58	0	2	1	0	2	0	10	27	21	2	3676	131,527
ET730	26-May-99	B763	FCOLHR	8	0	2	9	0	0	0	0	0	0	0	0	0	9	897	5,472
ET730	2-Jun-99	B757	ADDFCO	65	0	4	41	0	0	2	2	1	1	15	20	10	0	2782	91,654
ET730	2-Jun-99	B757	ADDLHR	70	4	0	63	6	2	6	0	0	1	6	20	11	0	3676	152,524
ET730	2-Jun-99	B757	FCOLHR	17	0	0	14	0	0	0	0	0	0	0	0	1	13	897	9,468
ET730	9-Jun-99	B757	ADDFCO	26	3	0	23	0	0	0	2	0	1	0	11	4	1	2782	69,326
ET730	9-Jun-99	B757	ADDLHR	73	4	0	64	0	0	2	9	0	0	12	25	14	0	3676	125,919
ET730	9-Jun-99	B757	FCOLHR	12	1	0	8	0	0	0	0	0	0	0	0	0	8	897	4,247
ET730	16-Jun-99	B767	ADDFCO	52	3	0	71	0	0	3	4	2	2	8	23	7	19	2782	165,059
ET730	16-Jun-99	B767	ADDLHR	58	5	0	59	0	0	0	0	0	1	8	33	16	0	3676	120,053
ET730	16-Jun-99	B767	FCOLHR	16	0	1	19	0	0	0	0	0	0	1	0	0	18	897	10,500
ET730	23-Jun-99	B757	ADDFCO	50	5	0	46	0	0	1	0	0	1	4	15	21	0	2782	120,291
ET731	23-Jun-99	B757	ADDLHR	96	10	0	65	0	0	0	1	1	2	7	24	28	0	3676	135,518
ET730	23-Jun-99	B757	FCOLHR	28	1	0	27	0	0	0	0	0	0	1	0	0	26	897	14,309
ET730	30-Jun-99	B757	ADDFCO	75	7	0	63	6	0	1	0	4	2	8	28	21	0	2782	158,531
ET730	30-Jun-99	B757	ADDLHR	113	3	0	92	2	1	0	0	2	2	10	49	26	0	3676	175,121
ET730	30-Jun-99	B757	FCOLHR	32	2	0	30	0	0	0	0	0	0	1	1	0	28	897	17,150
ET730	7-Jul-99	B757	ADDFCO	58	4	0	48	2	0	0	0	0	3	7	24	14	0	2782	134,214
ET730	7-Jul-99	B757	ADDLHR	99	9	0	87	5	1	3	0	0	6	7	39	27	4	3676	194,753
ET730	7-Jul-99	B757	FCOLHR	36	0	1	37	0	0	0	0	0	0	1	0	0	36	897	54,189
ET730	14-Jul-99	B757	ADDFCO	63	10	0	50	3	0	0	1	1	1	10	20	23	1	2782	124,952
ET730	14-Jul-99	B757	ADDLHR	107	8	0	87	6	0	3	0	0	1	12	34	30	1	3676	217,513
ET730	14-Jul-99	B757	FCOLHR	43	0	4	44	0	0	0	0	0	0	5	0	0	39	897	171,233
ET730	21-Jul-99	B767	ADDFCO	62	4	0	33	3	0	0	1	1	1	10	20	13	1	2782	100,115

Table 4.7a Sample Database (DB1)

Flight No	Date	A/C	TTL Bkg	NS	NR	TTL Act Bkg	C	D	S	B	H	K	L	M	Q	T	ASK	RSK	LF	Yield	Revenue
ET 730	20-Jan-99	B757	151	3	0	74	7	0	1	1	4	2	17	39	41	24	856,515	360,228	0.42	0.64	229,475
ET 730	3-Feb-99	B763	158	2	2	87	7	0	1	1	3	2	16	35	35	23	1,376,331	447,064	0.32	0.41	184,426
ET 730	3-Mar-99	B757	146	8	1	89	0	0	1	1	3	2	16	35	35	23	791,538	449,000	0.57	0.46	207,541
ET 730	10-Mar-99	B757	146	23	0	119	0	0	1	1	3	2	16	35	35	23	874,236	558,029	0.64	0.37	208,332
ET 730	17-Mar-99	B757	146	19	0	103	0	0	1	1	3	2	16	35	35	23	791,538	505,701	0.64	0.53	267,816
ET 730	24-Mar-99	B767	146	20	0	153	0	0	1	1	3	2	16	35	35	23	1,016,004	830,419	0.82	0.42	348,760
ET 730	28-Apr-99	B757	117	25	0	92	0	0	2	0	4	3	9	33	22	16	856,515	466,053	0.54	0.40	188,365
ET 730	5-May-99	B757	163	13	0	137	3	1	1	0	5	2	10	26	38	49	856,515	533,999	0.62	0.49	260,337
ET 730	12-May-99	B757	140	14	0	109	3	2	1	2	6	2	22	36	29	12	856,515	550,687	0.64	0.49	271,389
ET 730	19-May-99	B757	87	7	0	77	5	1	2	0	0	1	10	28	22	13	921,492	377,615	0.41	0.45	170,893
ET 730	26-May-99	B763	115	16	2	96	2	6	2	0	2	0	13	40	29	13	1,435,401	485,209	0.34	0.42	204,582
ET 730	2-Jun-99	B757	152	4	4	118	6	2	8	2	1	2	21	40	22	13	856,515	575,583	0.67	0.44	253,646
ET 730	9-Jun-99	B757	111	8	0	95	0	0	2	11	0	1	12	36	18	9	921,492	492,384	0.53	0.41	199,492
ET 730	16-Jun-99	B767	126	8	1	149	0	0	3	4	2	3	17	56	23	37	1,075,074	693,290	0.64	0.43	295,612
ET 730	23-Jun-99	B757	174	16	0	138	0	0	1	1	1	3	12	39	49	26	850,608	628,412	0.74	0.43	270,118
ET 730	30-Jun-99	B757	220	12	0	185	8	1	1	0	6	4	19	78	47	28	1,140,051	868,230	0.76	0.40	350,802
ET 730	7-Jul-99	B757	193	13	1	172	7	1	3	0	0	9	15	63	41	40	921,492	781,660	0.85	0.49	383,156
ET 730	14-Jul-99	B757	213	18	4	181	9	0	3	1	1	2	27	54	53	41	921,492	800,649	0.87	0.64	513,698
ET 730	21-Jul-99	B767	216	11	7	181	9	0	3	1	1	2	27	54	43	41	921,492	793,740	0.7	0.47	374,364
ET 730	28-Jul-99	B767	247	8	2	197	14	0	0	13	4	13	27	52	31	43	1,140,051	896,763	0.79	0.44	395,390
ET 730	4-Aug-99	B767	247	13	0	191	5	0	0	0	3	4	15	69	43	52	1,140,051	835,757	0.73	0.42	354,149
ET 730	11-Aug-99	B767	272	32	3	232	6	3	0	0	1	2	27	88	26	75	1,140,051	918,121	0.81	0.43	391,137
ET 730	18-Aug-99	B767	266	9	0	243	14	0	2	2	0	1	30	84	60	54	1,140,051	1,064,763	0.93	0.25	266,054
ET 730	25-Aug-99	B767	230	12	0	205	12	0	1	9	6	6	26	89	42	14	1,140,051	980,933	0.86	0.60	590,263
ET 730	1-Sep-99	B763	249	21	0	207	19	4	1	0	0	3	49	94	25	12	1,329,075	1,002,192	0.75	0.52	519,748
ET 730	8-Sep-99	B767	164	13	0	165	2	0	1	13	2	3	29	76	30	9	1,145,958	8,472,026	0.74	0.04	349,560
ET 730	15-Sep-99	B767	213	36	0	140	2	5	0	0	1	1	27	50	33	21	1,145,958	700,296	0.61	0.49	342,859
ET 730	22-Sep-99	B767	207	12	0	149	10	4	0	2	5	2	21	49	33	23	1,145,958	732,369	0.64	0.48	352,319
ET 730	29-Sep-99	B767	229	12	1	197	6	1	0	14	31	1	13	50	43	38	1,145,958	965,847	0.84	0.58	561,645
ET 730	6-Oct-99	B757	190	10	12	154	6	8	0	0	2	3	20	37	43	41	927,399	696,092	0.75	0.53	370,954
ET 730	13-Oct-99	B757	200	14	0	140	3	3	1	6	2	1	9	35	37	43	921,492	646,300	0.7	0.53	344,832
ET 730	20-Oct-99	B757	159	11	0	143	5	3	3	0	4	0	18	27	44	39	921,492	684,179	0.74	0.49	334,567
ET 730	27-Oct-99	B757	186	14	1	134	8	0	0	2	1	3	10	31	39	40	844701	564202	0.67	0.48	272,275
ET 730	3-Nov-99	B757	17	11	1	112	7	6	1	0	0	0	8	34	41	15	927,399	542,411	0.58	0.56	306,446
ET 730	10-Nov-99	B757	16	11	0	125	7	3	1	1	2	0	15	30	36	17	921,492	596,842	0.65	0.51	306,927
ET 730	17-Nov-99	B757	201	11	0	134	5	4	0	0	9	2	7	62	53	12	921,492	665,023	0.72	0.49	325,773
ET 730	24-Nov-99	B757	157	20	2	98	1	2	2	5	2	5	7	21	38	15	921,492	474,997	0.52	0.48	226,742
ET 730	1-Dec-99	B757	196	19	0	127	5	0	0	0	0	0	0	0	0	23	921,492	591,368	0.64	0.49	289,747
ET 730	8-Dec-99	B757	194	11	0	151	5	3	0	0	2	5	28	49	47	12	921,492	763,673	0.83	0.50	378,633
ET 730	15-Dec-99	B767	211	12	1	185	2	2	0	3	3	1	20	94	34	13	1,140,051	855,816	0.75	0.50	424,584
ET 730	22-Dec-99	B763	271	10	0	214	5	11	0	1	4	5	14	80	51	41	1,429,494	967,548	0.68	0.48	466,625
ET 730	29-Dec-99	B767	258	17	0	190	3	20	0	3	4	2	6	76	34	42	1,169,586	806,822	0.69	0.45	363,484
ET 730	5-Jan-00	B763	163	15	11	241	1	9	0	0	2	3	29	131	33	33	1,329,075	1,123,901	0.85	0.21	241,226
ET 730	12-Jan-00	B763	250	29	9	228	10	0	1	7	7	9	70	81	40	3	1,329,075	1,191,564	0.9	0.54	638,700

Table 4.7b Entire Database (DB2)

The data were collected from four system databases; reservations, yield management, revenue accounting, and departure control systems; and from load performance spreadsheet hard copy reports. The data were prepared in MS Excel format which constituted the source database for Knowledge Studio.

b) Cleansing

Small and Edelstein (1997) warn that in all data mining applications, extra variables can introduce noise and reduce the accuracy of the model. At the same time, important variables that can increase accuracy or simplify the application of the model to new data should not be left out.

When operational data from transactions is loaded into the database, it often contains missing or inaccurate data. In data mining applications, to overcome this problem, the operational data must go through a 'cleansing' process, which takes care of missing or out–of–range values. In addition, the data must usually be refined and processed before it undergoes the data mining process. This process involves selecting specific records of data, selecting appropriate fields of data, filling missing data, etc. Finally, the data would need to be represented in specific format, depending on the data mining algorithm involved.

In order to minimize inaccurate values, the data items were evaluated against the expected ranges of data in the respective field. Data recorded was checked and rechecked and domain experts were consulted to check and edit the values.

To check for missing values and inconsistencies, the software has a visualization technique. Table 4.6 shows a sample of this view for flight ET731. All missing values of flight ET730 (about 20%) were properly filled either through statistical means, omitting, or filling in the correct value.

c) Selection

In order to select the appropriate fields only those elements that have relative contribution to revenue were first selected. In order to select the major contributors, a secondary test, mentioned earlier, was conducted. Data of flights ET661 and ET731 were used for this purpose. Tables 4.8a and 4.8b show samples of the testing environment used to select fields. All the rows of the data set were selected for training because of the time embedded element of the flights. The key to the training parameters and the result field codes (A,B,C, … T) is found in appendix 4.

	TRAINING																Result			
Test No	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	L=% V=Rec	L=Random V=Random	L=60% V=38	None	MLN	Rev	a	55	Random	30	1	25	20	2000	0.05	Logistic	80.00%	02min	19	20.63%
2	L=% V=Rec	L=Random V=Random	L=60% V=38	None	MLN	Rev	b	55	Random	30	1	25	20	2000	0.05	Logistic	65.45%	02min	15	21.05%
3	L=% V=Rec	L=Random V=Random	L=60% V=38	None	MLN	Rev	c	55	Random	30	1	25	20	2000	0.05	Logistic	80.00%	12min	307	52.00%
4	L=% V=Rec	L=Random V=Random	L=60% V=38	None	MLN	Rev	d	20	Random	30	1	25	20	2000	0.05	Logistic	75.00%	11min	204	54.00%
5	L=% V=Rec	L=Random V=Random	L=60% V=38	None	MLN	Rev	e	20	Random	30	1	25	20	2000	0.05	Logistic	50.00%	01min	2	26.31%
6	L=% V=Rec	L=Random V=Random	L=60% V=38	None	MLN	Rev	f	20	Random	30	1	25	20	2000	0.05	Logistic	88.33%	01min	1	82.50%
7	L=% V=Rec	L=Random V=Random	L=60% V=38	None	MLN	Rev	g	20	Random	30	1	25	20	2000	0.05	Logistic	46.67%	01min	4	28.94%
8	L=% V=Rec	L=Random V=Random	L=60% V=38	None	MLN	Rev	h	20	Random	30	1	25	20	2000	0.05	Logistic	79.49%	05min	43	65.00%
9	L=% V=Rec	L=Random V=Random	L=60% V=38	None	MLN	Rev	i	20	Random	30	1	25	20	2000	0.05	Logistic	75.00%	01min	3	73.00%
10	L=% V=Rec	L=Random V=Random	L=70% V=28	None	MLN	Rev	a	65	Random	30	1	25	20	2000	0.05	Logistic	32.31%	02min	12	25.80%
11	L=% V=Rec	L=Random V=Random	L=70% V=28	None	MLN	Rev	b	65	Random	30	1	25	20	2000	0.05	Logistic	66.15%	02min	17	48.00%
12	L=% V=Rec	L=Random V=Random	L=70% V=28	None	MLN	Rev	c	65	Random	30	1	25	20	2000	0.05	Logistic	60.00%	02min	21	50.53%
13	L=% V=Rec	L=Random V=Random	L=70% V=28	None	MLN	Rev	d	65	Random	30	1	25	20	2000	0.05	Logistic	80.00%	02min	15	63.00%
14	L=% V=Rec	L=Random V=Random	L=70% V=28	None	MLN	Rev	e	65	Random	30	1	25	20	2000	0.05	Logistic	66.15%	05min	45	49.00%
15	L=% V=Rec	L=Random V=Random	L=70% V=28	None	MLN	Rev	f	65	Random	30	1	25	20	2000	0.05	Logistic	85.00%	04min	41	82.30%
16	L=% V=Rec	L=Random V=Random	L=70% V=28	None	MLN	Rev	g	65	Random	30	1	25	20	2000	0.05	Logistic	75.00%	01min	11	43.00%
17	L=% V=Rec	L=Random V=Random	L=70% V=28	None	MLN	Rev	h	65	Random	30	1	25	20	2000	0.05	Logistic	61.54%	03min	31	55.91%
18	L=% V=Rec	L=Random V=Random	L=70% V=28	None	MLN	Rev	i	65	Random	30	1	25	20	2000	0.05	Logistic	83.08%	06min	59	75.26%
19	L=% V=Rec	L=Random V=Random	L=80% V=19	None	MLN	Rev	a	74	Random	30	1	25	20	2000	0.05	Logistic	63.51%	02min	15	15.78%
20	L=% V=Rec	L=Random V=Random	L=80% V=19	None	MLN	Rev	b	74	Random	30	1	25	20	2000	0.05	Logistic	78.38%	02min	21	21.05%
21	L=% V=Rec	L=Random V=Random	L=80% V=19	None	MLN	Rev	c	74	Random	30	1	25	20	2000	0.05	Logistic	63.63%	09min	109	60.00%
22	L=% V=Rec	L=Random V=Random	L=80% V=19	None	MLN	Rev	d	74	Random	30	1	25	20	2000	0.05	Logistic	55.65%	12min	402	50.00%

Table 4.8a **Sample Secondary Testing (to select fields)**
Environment of Knowledge Studio Using ET 661

Where : - a = A/C, Date, Segment,TTL Bkg,NS,NR,TTL Act Bkg,Bkg by Class,Mileage.
b = A/C,Date,NR,NS,Segment,TTL Bkg
c = A/C,Date,TTL Act Bkg,Bkg by Class,Segment
d = Date,TTL Act Bkg,Bkg by Class,Segment
e= Date,TTL Bkg,NS,NR,Segment
f = Segment,Bkg by Class
g = Date,Segment,TTL Bkg,NS,NR,TTL Act Bkg
h = TTL Bkg,NS,NR
i = TTL Act Bkg,Bkg by Class

	TRAINING																RESULT			
Test No.	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	L=I% V=Rec	L=Random V=Random	L=70% V=45	No Weighting	MLN	Revenue	1	103	Random	30	1	25	100	2000	0.05	Logistic	60.71%	03min	33	33.33%
2	L=I% V=Rec	L=Random V=Random	L=70% V=45	No Weighting	MLN	Revenue	2	103	Random	30	1	25	100	2000	0.05	Logistic	56.31%	02min	28	24..44%
3	L=I% V=Rec	L=Random V=Random	L=70% V=45	No Weighting	MLN	Revenue	3	103	Random	30	1	25	100	2000	0.05	Logistic	75.32%	03min	31	66.66%
4	L=I% V=Rec	L=Random V=Random	L=70% V=45	No Weighting	MLN	Revenue	4	103	Random	30	1	25	100	2000	0.05	Logistic	65.00%	01min	11	62.36%
5	L=I% V=Rec	L=Random V=Random	L=70% V=45	No Weighting	MLN	Revenue	5	103	Random	30	1	25	100	2000	0.05	Logistic	7917%	08min	149	52.03%
6	L=I% V=Rec	L=Random V=Random	L=70% V=45	No Weighting	MLN	Revenue	6	103	Random	30	1	25	100	2000	0.05	Logistic	86.67%	01min	4	62.01%
7	L=I% V=Rec	L=Random V=Random	L=70% V=45	No Weighting	MLN	Revenue	7	103	Random	30	1	25	100	2000	0.05	Logistic	57.65%	03min	32	44.44%
8	L=I% V=Rec	L=Random V=Random	L=70% V=45	No Weighting	MLN	Revenue	8	103	Random	30	1	25	100	2000	0.05	Logistic	60.00%	02min	28	53.07%
9	L=I% V=Rec	L=Random V=Random	L=70% V=45	No Weighting	MLN	Revenue	9	103	Random	30	1	25	100	2000	0.05	Logistic	66.07%	01min	7	51.01%
10	L=I% V=Rec	L=Random V=Random	L=70% V=45	No Weighting	MLN	Revenue	10	103	Random	30	1	25	100	2000	0.05	Logistic	82.35%	02min	25	55.55%
11	L=I% V=Rec	L=Random V=Random	L=70% V=45	No Weighting	MLN	Revenue	11	103	Random	30	1	25	100	2000	0.05	Logistic	74.11%	01min	6	65.23%
12	L=I% V=Rec	L=Random V=Random	L=70% V=45	No Weighting	MLN	Revenue	12	103	Random	30	1	25	100	2000	0.05	Logistic	90.24%	04min	44	73.05%
13	L=I% V=Rec	L=Random V=Random	L=80% V=30	No Weighting	MLN	Revenue	1	118	Random	30	1	25	100	2000	0.05	Logistic	65.25%	04min	44	31.11%
14	L=I% V=Rec	L=Random V=Random	L=80% V=30	No Weighting	MLN	Revenue	2	118	Random	30	1	25	100	2000	0.05	Logistic	68.64%	03min	39	42.22%
15	L=I% V=Rec	L=Random V=Random	L=80% V=30	No Weighting	MLN	Revenue	3	118	Random	30	1	25	100	2000	0.05	Logistic	84.62%	01min	5	68.09%
16	L=I% V=Rec	L=Random V=Random	L=80% V=30	No Weighting	MLN	Revenue	4	118	Random	30	1	25	100	2000	0.05	Logistic	84.62%	02min	17	58.09%
17	L=I% V=Rec	L=Random V=Random	L=80% V=30	No Weighting	MLN	Revenue	5	118	Random	30	1	25	100	2000	0.05	Logistic	55.08%	02min	23	50.00%
18	L=I% V=Rec	L=Random V=Random	L=80% V=30	No Weighting	MLN	Revenue	6	118	Random	30	1	25	100	2000	0.05	Logistic	84.62%	01min	4	52.09%
19	L=I% V=Rec	L=Random V=Random	L=80% V=30	No Weighting	MLN	Revenue	7	118	Random	30	1	25	100	2000	0.05	Logistic	61.76%	02min	29	20.00%

Table 4.8b **Sample Secondary Testing (to select fields) Environment of Knowledge STUDIO Using ET 731**

Where :-

1=Date,A/C,Segment,TTL Bkg,NS,NR,TTL Act Bkg,Fare Classes,ASK,RSK,LF BY FLT,Yield,Mileage
2=Date,Segment,TTL Act Bkg,Fare Classes,ASK,RSK,LF BY FLT
3=Date,Segment,TTL Act Bkg
4=Date,Segment,Fare Classes
5=Date,Segment
6=Date,A/C,Segment,TTL Act Bkg
7=Date,A/C,Segment,TTL Bkg,NS,NR,ASK,RSK,Lf by flt,Mileage
8=Date,Segment,TTL Act Bkg,ASK,RSK
9=Segment,TTL Bkg,NS,NR,ASK,RSK
10=Date,TTL Act Bkg,ASK,RSK
11=Segment,Fare Classes,ASK,RSK
12=A/C,TTL Bkg,Fare Classes,ASK,RSK

d) Preprocessing

In this final phase of data preparation, the data needs to be transformed into a form that is acceptable as input to the neural network. This includes such processes as turning a data into a day of the week or day of the year, scaling of data, normalizing, representing, etc. This preprocessing phase was automatically done by the software.

4.4.3 Training and Testing

In addition to the data preparation, the other important part in building a neural network model is the training and testing phase. The data set used to build the model was that of ET730 from both DB1 (Table 4.7a) and DB2 (Table 4.7b).

The data sets of DB1 and DB2 were split into three. One subset was used for learning, another to test the accuracy of the neural network and the third to validate the results. The network used the first two subsets to train, by switching the learning and testing data. The third subset was never shown to the network during training.

Three different combinations were used to manage the data set as shown in figure 4.11.

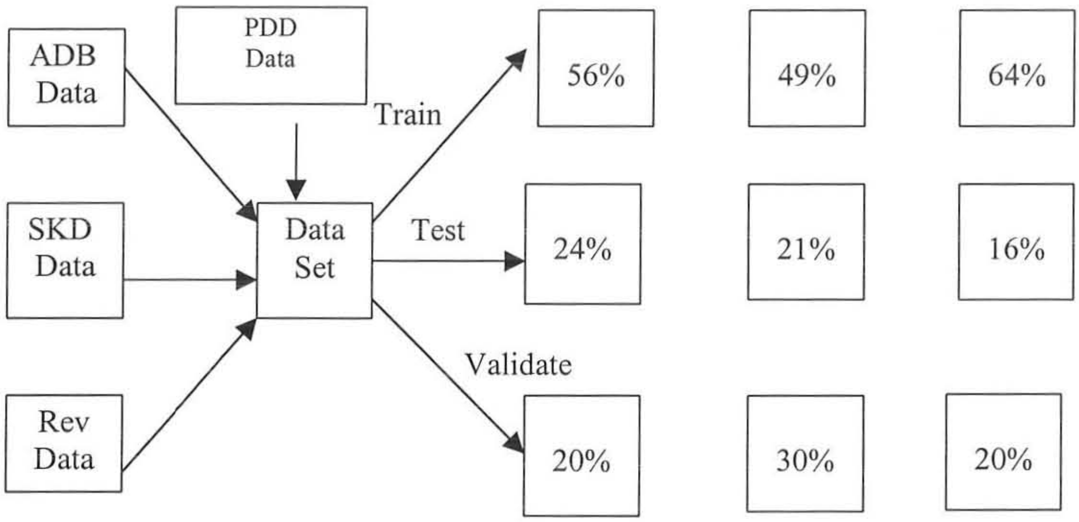


Figure 4.11: Data Set Management

A comprehensive training parameter combination was developed which included different combinations of the following training parameters:

- Learning Sample Size: The portion of the data set that will be used for training and testing (the other portion is the validation sample size).
- Predictive model type: type of neural network algorithm used.
- Percentage (%) of records for testing out of the learning set.
- Number of processing units: number of neurons in the hidden layer.
- Root Mean Square (RMS) error: Acceptable weighted average error of the predicted output.
- Activation function type: Function by which the weights are adjusted.
- Radial basis unit: neurons in the hidden layer of an RBF algorithm.
- Number of iterations: number of times the network calculates the error, adjusts the weights and feeds the error back through the network.

- Number of overlap: Scale factor (multiplier) applied to the variance of the basis function.

a) The parameters and the values used for BPN include the following.

Learning Sample Size: 80%, 70%

Predictive model type: Multi Layer Network - BPN

Percentage % of records for testing out of the learning set: 30%, 20%

Number of processing units: 20, 50, 100

Root Mean Square (RMS) error: 0.1, 0.05

Activation function type: Logistic/Sigmoid, Linear

Three different experiments were conducted using the BPN model.

Experiment 1 included BPN with DB1 (independent variables are the fare classes and segment). In this experiment, 46 different combinations or test cases were produced. Table 4.9 shows a sample of this training environment. The left hand side shows the training parameters and the right hand side the results.

Experiment 2 included BPN with DB2 (independent variables are fare classes). In this experiment 46 test cases were also produced. Table 4.10 shows a sample of this training environment.

Experiment 3 included BPN with DB2 (independent variable is total actual boarded by flight). In this experiment also 46 test cases were produced. Table 4.11 shows a sample of this training environment.

	TRAINING																RESULT			
Test No.	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
27	L=% V=Rec	L=Random V=Random	L=70% V=51	NONE	MLN	Revenue	fare classes & segment	117	Random	30	1	25	50	2000	0.1	Logistic	87.18%	1min	5	66.66%
Retrain																	94.02%	2min	10	
Retrain																	89.74%	2min	10	70.58%
Retrain																	94.05%	1min	4	76.47%
Retrain																	91.67%	1 min	5	68.62%
Retrain																	90.48%	1min	4	72.54%
Retrain																	96.43%	1min	5	78.43%
28	L=% V=Rec	L=Random V=Random	L=70% V=51	NONE	MLN	Revenue	fare classes & segment	117	Random	30	1	25	50	2000	0.1	Linear	82.91%	3min	31	52.94%
Retrain																	83.76%	4min	38	56.86%
Retrain																	87.72%			
29	L=% V=Rec	L=Random V=Random	L=70% V=51	NONE	MLN	Revenue	fare classes & segment	117	Random	30	1	25	50	2000	0.05	Logistic	82.05%	3min	29	64.70%
Retrain																	78.63%	4min	42	58.82%
Retrain																	79.10%	4min	38	60.78%
Retrain																	82.09%	4min	45	60.78%
Retrain																	77.61%	4min	41	58.82%
30	L=% V=Rec	L=Random V=Random	L=70% V=51	NONE	MLN	Revenue	fare classes & segment	117	Random	30	1	25	50	2000	0.05	Linear	70.09%	4min	39	39.21%
Retrain																	61.54%	3min	34	43.13%
Retrain																	50.00%	3min	33	39.21%
Retrain																	58.57%	4min	38	45.09%
31	L=% V=Rec	L=Random V=Random	L=70% V=51	NONE	MLN	Revenue	fare classes & segment	117	Random	30	1	25	20	2000	0.1	Logistic	89.74%	1min	5	84.31%
Retrain																	95.12%	2min	11	86.27%
Retrain																	91.46%	1min	7	78.43%
Retrain																	91.46%	1min	9	84.31%
Retrain																	89.02%	1min	9	86.27%
32	L=% V=Rec	L=Random V=Random	L=70% V=51	NONE	MLN	Revenue	fare classes & segment	117	Random	30	1	25	20	2000	0.1	Linear	83.76%	5min	48	68.62%
Retrain																	77.78%	6min	57	64.70%
Retrain																	62.75%	3min	28	56.86%
33	L=% V=Rec	L=Random V=Random	L=70% V=51	NONE	MLN	Revenue	fare classes & segment	117	Random	30	1	25	20	2000	0.05	Logistic	83.76%	4min	47	58.82%
Retrain																	76.92%	3min	33	54.90%
Retrain																	93.48%	4min	48	66.66%
Retrain																	75.00%	2min	17	58.82%
34	L=% V=Rec	L=Random V=Random	L=70% V=51	NONE	MLN	Revenue	fare classes & segment	117	Random	30	1	25	20	2000	0.05	Linear	60.68%	4min	37	62.74%
Retrain																	64.10%	4min	42	60.78%
Retrain																	68.29%	5min	51	60.78%
Retrain																	56.10%	3min	31	49.01%
35	L=% V=Rec	L=Random V=Random	L=70% V=51	NONE	MLN	Revenue	fare classes & segment	117	Random	20	1	25	100	2000	0.05	Logistic	81.20%	3min	32	50.98%
Retrain																	91.45%	4min	42	49.01%
Retrain																	86.21%	6min	66	58.82%
Retrain																	65.52%	3min	27	41.17%
36	L=% V=Rec	L=Random V=Random	L=70% V=51	NONE	MLN	Revenue	fare classes & segment	117	Random	20	1	25	100	2000	0.05	Linear	58.12%	2min	23	43.13%
Retrain																	61.54%	2min	23	49.01%
Retrain																	61.64%	3min	34	37.25%
37	L=% V=Rec	L=Random V=Random	L=70% V=51	NONE	MLN	Revenue	fare classes & segment	117	Random	20	1	25	100	2000	0.1	Logistic	92.31%	1min	11	60.78%
Retrain																	82.05%	1min	12	52.94%
Retrain																	95.59%	1min	9	82.35%
Retrain																	88.24%	1min	5	80.39%
38	L=% V=Rec	L=Random V=Random	L=70% V=51	NONE	MLN	Revenue	fare classes & segment	117	Random	20	1	25	100	2000	0.1	Linear	82.91%	2min	20	47.05%
Retrain																	83.76%	2min	19	49.01%
Retrain																	85.45%	1min	8	64.70%

Table 4.9 Sample BPN with DB1 (Fare Classe/Segment) ; 46 Combinations (Experiment 1)

TRAINING																	RESULT			
Test No.	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	L=% V=Rec	L=Random V=Random	L=80% V=12	None	MLN	Revenue	fare classes	44	Random	30	1	25	100	2,000	0.1	Logistic	86.36	02min	18	25.00
Retrain																	75.00	01min	3	25.00
Retrain																	75.00	01min	0	50.00
Retrain																	75.00	01min	1	41.66
2	L=% V=Rec	L=Random V=Random	L=80% V=12	None	MLN	Revenue	fare classes	44	Random	30	1	25	100	2,000	0.1	Linear	45.45	02min	12	0.00
Retrain																	52.27	02min	12	0.00
Retrain																	25.00	01min	3.00	0.00
3	L=% V=Rec	L=Random V=Random	L=80% V=12	None	MLN	Revenue	fare classes	44	Random	30	1	25	100	2,000	0.05	Logistic	59.09	04min	41	0.00
Retrain																	59.09	04min	42	0.00
Retrain																	100.00	12min	2000	0.00
Retrain																	100.00	12min	2000	0.00
4	L=% V=Rec	L=Random V=Random	L=80% V=12	None	MLN	Revenue	fare classes	44	Random	30	1	25	100	2,000	0.05	Linear	20.45	02min	10	0.00
Retrain																	25.00	02min	12	8.33
Retrain																	0.00	12min	2000	0.00
5	L=% V=Rec	L=Random V=Random	L=80% V=12	None	MLN	Revenue	fare classes	44	Random	30	1	25	50	2,000	0.1	Logistic	93.18	02min	19	8.33
Retrain																	75.00	02min	19	8.33
Retrain																	100.00	01min	01	25.00
Retrain																	100.00	01min	01	33.33
Retrain																	100.00	01min	01	25.00
6	L=% V=Rec	L=Random V=Random	L=80% V=12	None	MLN	Revenue	fare classes	44	Random	30	1	25	50	2,000	0.1	Linear	61.36	03min	26	0.00
Retrain																	43.18			
7	L=% V=Rec	L=Random V=Random	L=80% V=12	None	MLN	Revenue	fare classes	44	Random	30	1	25	50	2,000	0.05	Logistic	70.48	04min	40	8.33
Retrain																	81.82	02min	21	8.33
Retrain																	100.00	12min	2000	100.00
Retrain																	100.00	12min	2000	100.00
8	L=% V=Rec	L=Random V=Random	L=80% V=12	None	MLN	Revenue	fare classes	44	Random	30	1	25	50	2,000	0.05	Linear	40.91	02min	16	0.00
Retrain																	25.00	01min	04	0.00
Retrain																	25.00	01min	03	0.00
Retrain																	50.00	01min	05	0.00
9	L=% V=Rec	L=Random V=Random	L=80% V=12	None	MLN	Revenue	fare classes	44	Random	30	1	25	20	2,000	0.1	Logistic	79.55	01min	07	25.00
Retrain																	88.64	02min	21	33.33
Retrain																	75.00	01min	0	25.00
Retrain																	50.00	01min	5	0.00
10	L=% V=Rec	L=Random V=Random	L=80% V=12	None	MLN	Revenue	fare classes	44	Random	30	1	25	20	2,000	0.1	Linear	52.27	03min	27	16.66
Retrain																	45.45	03min	29	8.33
Retrain																	100.00	01min	05	8.33
Retrain																	100.00	01min	04	16.66
11	L=% V=Rec	L=Random V=Random	L=80% V=12	None	MLN	Revenue	fare classes	44	Random	30	1	25	20	2,000	0.05	Logistic	61.36	04min	44	16.66
Retrain																	47.73	03min	30	25.00
Retrain																	100.00	12min	2000	100.00
Retrain																	100.00	12min	2000	100.00
12	L=% V=Rec	L=Random V=Random	L=80% V=12	None	MLN	Revenue	fare classes	44	Random	30	1	25	20	2,000	0.05	Linear	34.09	03min	32	8.33
Retrain																	18.18	03min	26	-
Retrain																	29.55	04min	35	0.00
Retrain																	29.55	04min	37.00	8.33

Table 4.10 Sample BPN with DB2 (Fare Classes) ; 46 Combinations (Experiment 2)

TRAINING																	RESULT			
Test No.	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	L=% V=Rec	L=Random	L=80%	None	MLN	Revenue	TTL Boarded Passengers	44	Random	30	1	25	100	2000	0.1	Logistic	93.18	07min	92	66.66
Retrain																	88.64	4min	36	83.33
2	L=% V=Rec	L=Random V=Random	L=80% V=12	None	MLN	Revenue	TTL Boarded Passengers	44	Random	30	1	25	100	2000	0.1	Logistic	63.64	2min	5	58.33
Retrain																	70.45	3min	7	75
3	L=% V=Rec	L=Random V=Random	L=80% V=12	None	MLN	Revenue	TTL Boarded Passengers	44	Random	30	1	25	100	2000	0.1	Logistic	56.82	03min	32	58.33
																	56.82	3min	32	58.33
Retrain																	61.36	4min	30	83.33
4	L=% V=Rec	L=Random V=Random	L=80% V=12	None	MLN	Revenue	TTL Boarded Passengers	44	Random	30	1	25	100	2000	0.05	Linear	31.82	3min	6	33.33
5	L=% V=Rec	L=Random V=Random	L=80% V=12	None	MLN	Revenue	TTL Boarded Passengers	44	Random	30	1	25	50	2000	0.1	Logistic	88.64	3min	37	58.33
Retrain																	88.64	4min	100	66.66
6	L=% V=Rec	L=Random V=Random	L=80% V=12	None	MLN	Revenue	TTL Boarded Passengers	44	Random	30	1	25	50	2000	0.1	Linear	59.09	1min	7	75
Retrain																	61.36	2min	10	66.66
7	L=% V=Rec	L=Random V=Random	L=80% V=12	None	MLN	Revenue	TTL Boarded Passengers	44	Random	30	1	25	50	2000	0.05	Logistic	61.36	3min	32	83.33
Retrain																	70.45	10min	223	66.66
Retrain																	63.64	4min	110	75
8	L=% V=Rec	L=Random V=Random	L=80% V=12	None	MLN	Revenue	TTL Boarded Passengers	44	Random	30	1	25	50	2000	0.05	Linear	36.36	3min	3	50
9	L=% V=Rec	L=Random V=Random	L=80% V=12	None	MLN	Revenue	TTL Boarded Passengers	44	Random	30	1	25	20	2000	0.1	Logistic	93.18	2min	69	83.33
10	L=% V=Rec	L=Random V=Random	L=80% V=12	None	MLN	Revenue	TTL Boarded Passengers	44	Random	30	1	25	20	2000	0.1	Linear	65.91	3min	8	50
11	L=% V=Rec	L=Random V=Random	L=80% V=12	None	MLN	Revenue	TTL Boarded Passengers	44	Random	30	1	25	20	2000	0.05	Logistic	63.64	2min	27	58.33
12	L=% V=Rec	L=Random V=Random	L=80% V=12	None	MLN	Revenue	TTL Boarded Passengers	44	Random	30	1	25	20	2000	0.05	Linear	43.18	4min	5	16.66
13	L=% V=Rec	L=Random V=Random	L=80% V=12	None	MLN	Revenue	TTL Boarded Passengers	44	Random	20	1	25	100	2000	0.1	Logistic	90.91	5min	67	91.66
14	L=% V=Rec	L=Random V=Random	L=80% V=12	None	MLN	Revenue	TTL Boarded Passengers	44	Random	20	1	25	100	2000	0.1	Linear	65.91	7min	8	58.33
Retrain																	68.18	6min	4	41.66
15	L=% V=Rec	L=Random V=Random	L=80% V=12	None	MLN	Revenue	TTL Boarded Passengers	44	Random	20	1	25	100	2000	0.05	Logistic	70.45	4min	151	58.33
16	L=% V=Rec	L=Random V=Random	L=80% V=12	None	MLN	Revenue	TTL Boarded Passengers	44	Random	20	1	25	100	2000	0.05	Linear	31.82	3min	8	33.33
17	L=% V=Rec	L=Random V=Random	L=80% V=12	None	MLN	Revenue	TTL Boarded Passengers	44	Random	20	1	25	50	2000	0.1	Logistic	95.45	10min	275	83.33
18	L=% V=Rec	L=Random V=Random	L=80% V=12	None	MLN	Revenue	TTL Boarded Passengers	44	Random	20	1	25	50	2000	0.1	Linear	65.91	3min	6	50
Retrain																	77.27	3min	4	58.33
19	L=% V=Rec	L=Random V=Random	L=80% V=12	None	MLN	Revenue	TTL Boarded Passengers	44	Random	20	1	25	20	2000	0.1	Logistic	90.91	15min	254	75
Retrain																	84.09	5min	46	91.66
20	L=% V=Rec	L=Random V=Random	L=80% V=12	None	MLN	Revenue	TTL Boarded Passengers	44	Random	20	1	25	20	2000	0.1	Linear	34.09	3min	6	25
Retrain																	61.36	4min	4	41.66
Retrain																	56.82	3min	4	58.33
21	L=% V=Rec	L=Random V=Random	L=80% V=12	None	MLN	Revenue	TTL Boarded Passengers	44	Random	20	1	25	20	2000	0.05	Logistic	61.36	6min	29	41.66
Retrain																	59.09	13min	26	58.33
22	L=% V=Rec	L=Random V=Random	L=80% V=12	None	MLN	Revenue	TTL Boarded Passengers	44	Random	20	1	25	20	2000	0.05	Linear	36.36	5min	7	33.33
23	L=% V=Rec	L=Random V=Random	L=70% V=17 Rec	None	MLN	Revenue	TTL Boarded Passengers	39	Random	30	1	25	100	2000	0.05	Logistic	61.54	4min	29	58.82
Retrain																	66.67	6min	36	58.82
24	L=% V=Rec	L=Random V=Random	L=70% V=17 Rec	None	MLN	Revenue	TTL Boarded Passengers	39	Random	30	1	25	100	2000	0.05	Linear	28.21	3min	4	23.52
Retrain																	23.08	3min	3	35.29
25	L=% V=Rec	L=Random V=Random	L=70% V=17 Rec	None	MLN	Revenue	TTL Boarded Passengers	39	Random	30	1	25	100	2000	0.1	Logistic	87.18	5min	28	82.35

Table 4.11 **Sample BPN with DB2 (Total Actual Boarded) ; 46 Combinations (Experiment 3)**

b) The parameters and their values used for RBF include:

Learning sample size = 80%, 70%

Predictive model type = RBF

Radial basis unit = 10, 35,60

Number of iterations = 100, 500, 1000, 2000, 5000, 10000

Number of overlap = 40, 50, 60

Three different experiments were conducted using the RBF model.

Experiment 4 included RBF with DB1 (independent variables are the fare classes and segments). In this experiment 108 different test cases were produced. Table 4.12 shows a sample of this training environment.

Experiment 5 included RBF with DB2 (independent variables are the fare classes). In this experiment 27 cases were produced. Table 4.13 shows this training environment.

Experiment 6 included RBF with DB2 (independent variable is total actual passengers boarded). In this experiment 54 test cases were produced. Table 4.14 shows a sample of this training environment.

Table 4.15 gives a summary of the experiment types conducted.

Test No.	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
	TRAINING												**RESULT**			
1	L=% of Total V=Rec.Count	L=Random V=Random	L=70% V=51	None	RBF	Revenue	fare classes & Segment	117	Random	10	2000	40	76.07%	05min	2000	45.09%
Retrain													78.63%	05min	2000	39.21%
Retrain													89.23%	04min	2000	70.58%
Retrain													92.31%	05min	2000	68.62%
Retrain													92.31%	03min	2000	72.54%
Retrain													92.31%	04min	2000	70.58%
2	L=% of Total V=Rec.Count	L=Random V=Random	L=70% V=51	None	RBF	Revenue	Fare classes & Segment	117	Random	10	2000	50	83.76%	04min	2000	62.74%
Retrain													81.20%	04min	2000	64.70%
Retrain													76.00%	05min	2000	60.78%
Retrain													66.67%	03min	2000	54.90%
Retrain													70.67%	04min	2000	50.98%
Retrain													72.00%	03min	2000	58.82%
Retrain													72.00%	04min	2000	49.01%
3	L=% of Total V=Rec.Count	L=Random V=Random	L=70% V=51	None	RBF	Revenue	Fare classes & Segment	117	Random	10	2000	60	80.34%	05min	2000	60.78%
Retrain													82.05%	05min	2000	66.66%
Retrain													93.33%	04min	2000	84.31%
Retrain													90.00%	03min	2000	84.31%
4	L=% of Total V=Rec.Count	L=Random V=Random	L=70% V=51	None	RBF	Revenue	Fare classes & Segment	117	Random	10	5000	40	79.49%	08min	5000	54.90%
Retrain													71.79%	08min	5000	54.90%
Retrain													83.10%	08min	5000	80.39%
Retrain													84.51%	06min	5000	80.39%
Retrain													87.32%	07min	5000	78.43%
Retrain													83.10%	06min	5000	82.35%
Retrain													87.32%	07min	5000	76.47%
Retrain													84.51%	08min	5000	80.39%
5	L=% of Total V=Rec.Count	L=Random V=Random	L=70% V=51	None	RBF	Revenue	Fare classes & Segment	117	Random	10	5000	50	88.03%	07min	5000	70.58%
Retrain													87.18%	06min	5000	68.62%
Retrain													78.49%	05min	5000	74.50%
Retrain													76.34%	05min	5000	78.43%
Retrain													70.97%	07min	5000	68.62%
Retrain													74.19%	08min	5000	74.50%
Retrain													73.12%	07min	5000	72.54%
Retrain													72.02%	06min	5000	76.47%
6	L=% of Total V=Rec.Count	L=Random V=Random	L=70% V=51	None	RBF	Revenue	Fare classes & Segment	117	Random	10	5000	60	81.20%	09min	5000	62.74%
Retrain																
Retrain													82.91%	06min	5000	62.74%
Retrain													58.93%	07min	5000	68.62%
Retrain													67.86%	05min	5000	62.74%
Retrain													57.14%	06min	5000	68.62%
Retrain													58.93%	04min	5000	68.62%
7	L=% of Total V=Rec.Count	L=Random V=Random	L=70% V=51	None	RBF	Revenue	Fare classes & Segment	117	Random	10	10000	40	82.91%	10min	10000	45.09%
Retrain													82.05%	10min	10000	33.33%
Retrain													90.00%	08min	10000	68.62%

Table 4.12 Sample of Test Cases RBF with DB1 (Fare Classes/Segment) ; 108 Combinations (Experiment 4)

Training													Result			
Test No.	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	L=% V=Rec	L=Random V=Random	L=80% V=12	None	RBF	Revenue	fare classes	44	Random	10	100	40	45.45	03 min	100	0
Retrain													Error	Error	Error	Error
2	L=% V=Rec	L=Random V=Random	L=80% V=12	None	RBF	Revenue	fare classes	44	Random	10	100	50	50	03min	100	0
3	L=% V=Rec	L=Random V=Random	L=80% V=12	None	RBF	Revenue	fare classes	44	Random	10	100	60	54.55	03min	100	8.33
Retrain													Error	Error	Error	Error
4	L=% V=Rec	L=Random V=Random	L=80% V=12	None	RBF	Revenue	fare classes	44	Random	35	100	40	56.82	04min	100	8.33
Retrain																
Retrain													52.27	04min	100	16.66
Retrain													Error	Error	Error	Error
5	L=% V=Rec	L=Random V=Random	L=80% V=12	None	RBF	Revenue	fare classes	44	Random	35	100	50	54.55	04min	100	8.33
Retrain													59.09	05min	100	8.33
Retrain													Error	Error	Error	Error
6	L=% V=Rec	L=Random V=Random	L=80% V=12	None	RBF	Revenue	fare classes	44	Random	35	100	60	61.36	05min	100	8.33
Retrain													63.64	05min	100	0
Retrain													Error	Error	Error	Error
7	L=% V=Rec	L=Random V=Random	L=80% V=12	None	RBF	Revenue	fare classes	44	Random	60	100	40	Error	Error	Error	Error
8	L=% V=Rec	L=Random V=Random	L=80% V=12	None	RBF	Revenue	fare classes	44	Random	60	100	50	Error	Error	Error	Error
9	L=% V=Rec	L=Random V=Random	L=80% V=12	None	RBF	Revenue	fare classes	44	Random	60	100	60	Error	Error	Error	Error
10	L=% V=Rec	L=Random V=Random	L=80% V=12	None	RBF	Revenue	fare classes	44	Random	10	500	40	65.91	06min	500	8.33
Retrain													47.73	04min	500	8.33
Retrain													Error	Error	Error	Error
11	L=% V=Rec	L=Random V=Random	L=80% V=12	None	RBF	Revenue	fare classes	44	Random	10	500	50	52.27	05min	500	8.33
Retrain													52.27	05min	500	8.33
Retrain													Error	Error	Error	Error
12	L=% V=Rec	L=Random V=Random	L=80% V=12	None	RBF	Revenue	fare classes	44	Random	10	500	60	59.09	06min	500	0
Retrain													56.82	06min	500	0
Retrain													Error	Error	Error	Error
13	L=% V=Rec	L=Random V=Random	L=80% V=12	None	RBF	Revenue	fare classes	44	Random	35	500	40	75	07min	500	0
Retrain													63.64	07min	500	0
Retrain													Error	Error	Error	Error
14	L=% V=Rec	L=Random V=Random	L=80% V=12	None	RBF	Revenue	fare classes	44	Random	35	500	50	70.45	06min	500	0
Retrain													54.55	05min	500	0
Retrain													Error	Error	Error	Error
15	L=% V=Rec	L=Random V=Random	L=80% V=12	None	RBF	Revenue	fare classes	44	Random	35	500	60	43.18	03min	500	0
16	L=% V=Rec	L=Random V=Random	L=80% V=12	None	RBF	Revenue	fare classes	44	Random	60	500	40	Error	Error	Error	Error
17	L=% V=Rec	L=Random V=Random	L=80% V=12	None	RBF	Revenue	fare classes	44	Random	60	500	50	Error	Error	Error	Error
18	L=% V=Rec	L=Random V=Random	L=80% V=12	None	RBF	Revenue	fare classes	44	Random	60	500	60	Error	Error	Error	Error

Table 4.13 Sample Test Cases RBF with DB2 (Fare Classes) ; 27 Combinations (Experiment 5)

	TRAINING												RESULT			
Test No.	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	L=%	L=Random	L=80%	None	RBF	Revenue	TTL Paxs	44	Random	10	2000	40	68.18	7min	2000	50.00
	V=Rec	V=Random	V=12				boarded									
Retrain													72.73	6min	2000	50.00
Retrain													75.00	6min	2000	66.66
2	L=%	L=Random	L=80%	None	RBF	Revenue	TTL Paxs	44	Random	10	2000	50	77.27	8min	2000	50.00
	V=Rec	V=Random	V=12				boarded									
Retrain													63.64	8min	2000	50.00
Retrain													70.45	7min	2000	50.00
3	L=%	L=Random	L=80%	None	RBF	Revenue	TTL Paxs	44	Random	10	2000	60	75.00	8min	2000	41.66
	V=Rec	V=Random	V=12				boarded									
Retrain													75.00	5min	2000	41.66
4	L=%	L=Random	L=80%	None	RBF	Revenue	TTL Paxs	44	Random	10	5000	40	65.91	5min	5000	75.00
	V=Rec	V=Random	V=12				boarded									
Retrain													65.91	5min	5000	75.00
Retrain													65.91	3min	5000	75.00
5	L=%	L=Random	L=80%	None	RBF	Revenue	TTL Paxs	44	Random	10	5000	50	72.73	6min	5000	41.66
	V=Rec	V=Random	V=12				boarded									
Retrain													70.45	4min	5000	41.66
6	L=%	L=Random	L=80%	None	RBF	Revenue	TTL Paxs	44	Random	10	5000	60	70.45	9min	5000	58.33
	V=Rec	V=Random	V=12				boarded									
Retrain													70.45	8min	5000	50.00
Retrain													68.18	8min	5000	58.33
Retrain													70.45	4min	5000	66.66
7	L=%	L=Random	L=80%	None	RBF	Revenue	TTL Paxs	44	Random	10	10000	40	65.91	8min	10000	58.33
	V=Rec	V=Random	V=12				boarded									58.33
Retrain													63.64	7min	10000	
Retrain													68.18	6min	10000	
8	L=%	L=Random	L=80%	None	RBF	Revenue	TTL Paxs	44	Random	10	10000	50	61.36	8min	10000	50.00
	V=Rec	V=Random	V=12				boarded									
Retrain													65.91	6min	10000	58.33
9	L=%	L=Random	L=80%	None	RBF	Revenue	TTL Paxs	44	Random	10	10000	60	68.18	9min	10000	41.66
	V=Rec	V=Random	V=12				boarded									
10	L=%	L=Random	L=80%	None	RBF	Revenue	TTL Paxs	44	Random	35	2000	40	65.91	2min	2000	83.33
	V=Rec	V=Random	V=12				boarded									
11	L=%	L=Random	L=80%	None	RBF	Revenue	TTL Paxs	44	Random	35	2000	50	63.64	4min	2000	66.66
	V=Rec	V=Random	V=12				boarded									
Retrain													70.45	5min	2000	50.00
Retrain													70.45	3min	2000	66.66
12	L=%	L=Random	L=80%	None	RBF	Revenue	TTL Paxs	44	Random	35	2000	60	65.91	5min	2000	66.66
	V=Rec	V=Random	V=12				boarded									
Retrain													70.45	9min	2000	66.66
13	L=%	L=Random	L=80%	None	RBF	Revenue	TTL Paxs	44	Random	35	5000	40	68.18	9min	5000	58.33
	V=Rec	V=Random	V=12				boarded									
Retrain													72.73	8min	5000	66.66
14	L=%	L=Random	L=80%	None	RBF	Revenue	TTL Paxs	44	Random	35	5000	50	70.45	5min	5000	41.66
	V=Rec	V=Random	V=12				boarded									
15	L=%	L=Random	L=80%	None	RBF	Revenue	TTL Paxs	44	Random	35	5000	60	79.55	6min	5000	66.66
	V=Rec	V=Random	V=12				boarded									
Retrain													70.45	7min	5000	83.33
Retrain													77.27	3min	5000	75.00
16	L=%	L=Random	L=80%	None	RBF	Revenue	TTL Paxs	44	Random	35	10000	40	65.91	6min	10000	58.33
	V=Rec	V=Random	V=12				boarded									
Retrain													65.91	5min	10000	75.00
Retrain													70.45	4min	10000	41.66
Retrain													72.73	3min	10000	66.66

Table 4.14 **Sample Cases RBF with DB2 (Total Boarded) ; 54 Combinations (Experiment 6)**

Experiment	BPN	RBF	Test Cases	Fare class & segment	Fare Class	Total boarded	Table
1	X		46	X			4.9
2	X		46		X		4.10
3	X		46			X	4.11
4		X	108	X			4.12
5		X	27		X		4.13
6		X	54			X	4.14

Table 4.15: Summary of the Experiment types

4.5 Results

4.5.1 Discussion of Results

When data mining is used for decision support applications, creating the neural network model is not the last part of the process. When using neural networks as models for information support systems, the most important issue is whether the weights in the neural network accurately capture the model needed for the application.

Even if we do not have a mathematical formula for the function, we can still learn a great deal about what the models learned by varying the training parameters and seeing what the effect is on the output. To determine which tests have accurately captured the essence of the problem, the results have been recorded and evaluated.

The crucial thing to learn is which parameters are most important in delivering the best model. The evaluation criteria to assess these parameters, in order of priority, are as follows:

- RMS error or prediction accuracy.
- % of records correctly predicted.
- No of tests conducted.
- No of records used for training.
- Time elapsed to complete the training.
- Retraining and iterations.

Following is a summary of the findings for each experiment.

i) Experiment 1

	0-20%	21-30%	31-40%	41-50%	51-60%	61-70%	71-80%	81-90%	91-100%
Test case % of records correctly predicted during training			4,24		22,30, 34, 42	8,12, 16,32, 35,36	15,18, 23,33, 38,45, 46	3,6,11, 20,21, 26,28, 31,37, 44	1,2,5,7, 9,10, 13,14, 17,19, 25,27, 39,40, 41, 43
Test case % of records correctly predicted during validation		8, 42	24, 36	4,12, 16,18, 22,30, 34,35, 38,46	7,20, 23,28, 29,32, 33	2,3,6, 11,15, 19,21, 40,41, 44,45	1,5,9, 10,14, 17,27, 37,43	25, 26, 31	13, 39

Table 4.16 Correctly predicted record percentage by test case (experiment 1)

The average % of records correctly predicted, training time, iterations and retrainings is 62%, 15 minutes, 140 times and 4.6 times respectively. The values in Table 4.16 show a normal curve with mean of records correctly predicted within the specified RMS at 60-70%.

Although we notice that many test cases yield very good results during testing (56% are in the over 80% range) they fail during validation (only 11% are over the 80% range). This tells us that there is a lot of overfitting and over training. In most test cases the network has memorized the training data set and fails to recognize a new pattern.

The best models developed during this experiment are test cases 13 and 39. In both these cases over 90% of the records correctly validated to within a 10% error. (RMS was specified as 0.1). Two other very good results were test cases 21 and 41. Although they had less records correctly predicted (70%), their RMS was 0.05.

In contrast the worst models are test cases 8 and 42. It is not surprising that in both cases the RMS was 0.05. However, one test case, 18, had a poor result even with an RMS of 0.01.

The most evident and influential factor that contributes to the success is the activation function. In all the best cases the function is a sigmoid function and in all the worst cases the function is a linear function. Table 4.17 shows the summary of the results of experiment 1.

	Best cases				Worst cases		
Test case	39	13	21	41	8	42	18
Learning Sample	70%	80%	80%	70%	80%		80%
Number of processing units	50	20	20	50	50	50	50
Number of records	117	134	134	117	134	117	134
Number of records correctly predicted	0.91	0.94	0.71	0.71	0.26	0.29	0.41
Activation function	Sig	Sig	Sig	Sig	Linear	Linear	Linear
RMS	0.1	0.1	0.05	0.05	0.05	0.05	0.1
Training time	5m	13m	21m	41m	40m	10m	19m
Retraining	X 3	X 3	X 3	X 3	X 8	X 3	X 4
Iterations	31	41	179	146	222	102	117

Table 4.17: Summary of Results of Experiment 1

ii) Experiment 2

	0-10%	11-20%	21-30%	31-50%	51-60%	61-70%	71-80%	81-90%	91-100%
Test case % of records correctly predicted during training,	4,14, 20		2,12, 18,24 30,34, 36,42, 46	40,45, 8,9, 21,38, 44	26,28, 41	6,19, 23,35	1,29, 32,33	25,27, 31	3,5,7, 10,11, 13,15, 16,17, 22,37, 39,43
Test case % of records correctly predicted during validation	2,3,4, 6,8,9, 12, 13, 14,15, 18,19, 20,22, 24,26, 29,30, 32,33, 35,35, 36,38, 40,42, 46	17, 21,23, 27,28, 31,37, 39,41, 43,45	5,10, 11,1625, 25	1,7					

Table 4.18 Correctly predicted record percentage by test case (experiment 2)

The average percentage of records correctly predicted, training time, iterations and retraining is under 10%, 17 minutes, 825 times, and 3 times respectively. The values in table 4.18 show a decreasing curve with most values concentrated at the worst results.

In general, the outputs of this experiment are very unreliable and show very bad results. Sample of the results can be seen from the sample in Table 4.10.

iii) Experiment 3

	0-20%	21-30%	31-40%	41-50%	51-60%	61-70%	71-80%	81-90%	91-100%
Test case % of records correctly predicted during training		24,34 46	4, 8, 16 22, 30	12, 36	20,21,26 32, 40	3, 6, 7, 10 11, 14, 15 23, 28, 29 35, 38, 45	2, 18, 33 42	1, 5, 19, 25, 37 41	9, 13, 17 27, 31, 39 43
Test case % of records correctly predicted during validation		30, 36	4, 16, 22 24, 33, 34	8, 10, 14 40, 46	11, 15, 18 20, 21, 23 28, 35 45	5, 6, 26 32, 44	2, 7, 29 38, 42 43	1, 3, 9, 17,25, 37	13, 19, 27 31, 39, 41

Table 4.19 Correctly predicted record percentage by test case (experiment 3)

The average percentage of records correctly predicted, training time, iterations and retrainings is 59%, 10 minutes, 191 times, and 1.8 times respectively. The values in table 4.19 resemble a normal curve with the mean records correctly predicted at 50-60%.

The best models developed during this experiment are test cases 13, 19, 27, 31, 39, 41 and the worst ones are 12 and 44. Typically the best results iterated between 200 and 300 times and the worst less than 10. The underlying difference again, however, is the activation function with all the best cases using a sigmoid function and the all the worst case a linear one.

Table 4.20 shows the summary of the results of experiment 3.

	Best cases						Worst cases	
Test case	13	19	27	31	39	41	12	44
Learning Sample	80%	80%	70%	70%	70%	70%	70%	70
Number of processing units	100	20	50	20	50	50	20	20
Number of records	44	44	39	39	39	39	44	39
Number of records correctly predicted	92%	92%	94%	94%	94%	94%	17%	12%
Activation function	Sig	Sig	Sig	Sig	Sig	Sig	Linear	Linear
RMS	0.1	0.1	0.1	0.1	0.1	0.05	0.05	0.1
Training time	5m	20m	10m	11m	20m	8m	4m	10m
Retrainin g	X 1	X 2	X 2	X 2	X 2	X 1	X 1	X 2
Iterations	67	300	272	205	258	27	5	10

Table 4.20: Summary of Results of Experiment 3

iv) Experiment 4

	0-20%	21-30%	31-40%	41-50%	51-60%	61-70%	71-80%	81-90%	91-100%
Test case % of records correctly predicted during training					6	95	2, 5, 12 16, 21, 27,28, 33, 38 41, 44, 45, 46, 56, 59 61, 65, 66,76 78, 86, 88, 93, 98, 99	3, 4, 8, 10, 13, 17, 19, 22, 23, 26, 29, 30, 31, 32,34, 35, 36, 37, 42 43, 47 54, 55 57, 58, 60,62, 69,71, 72,74, 77,82, 83, 84 85, 87, 89 92, 96	1, 7, 11 14, 15 18, 20, 24 25, 39, 40 48, 49 50, 51, 52 53, 63, 64 67, 68, 70 73, 75, 79 80, 81, 90 91, 94, 97
	0-10%	21-30%	31-40%	41-50%	51-60%	61-70%	71-80%	81-90%	91-100%
Test case % of records correctly predicted during validation				2, 23, 27 44, 98	10, 22, 25 30, 32, 37 42, 45, 47 51, 61, 71 72, 76, 80 81, 83, 85 86, 95	6, 9, 12, 13, 16, 18,19, 24,28, 33, 35 36, 38, 39,41, 43, 46 48, 50, 53,54, 57, 59, 62,63, 67, 68, 75,77, 78, 79 87, 89, 90,92, 93, 94 96, 97, 99	1, 4, 5 7, 8, 14 15, 17, 20 21, 31 34, 55, 56 58, 60, 64 65, 66, 70 73, 74, 84 88, 91	3, 11 26, 29 52, 69 82	40, 49

Table 4.21 Correctly predicted record percentage by test case (experiment 4)

The percentage of records correctly predicted, training time, iterations, and retrainings is 68%, 29 minutes, 2000 times and 5 times respectively. The values in table 4.21 show a normal curve with mean records correctly predicted between 60-70%.

The best cases are test cases 40 and 49 and the worst test cases 2, 23, 27, 44 and 98. Table 4.22 shows the summary of the results of experiment 4. Although the test cases have yielded the best results yet (67% predict over 80% accurate), during validation only 8% meet this accuracy. The results do not give a good indication as to the parameter that influences the best models.

	Best cases		**Worst cases**				
Test case	40	49	2	23	27	44	98
Learning Sample	80%	80%	70%	70%	70%	80%	80%
Number of radial basis	35	60	10	60	60	35	60
Number of records	134	134	117	117	117	134	134
Number of records correctly predicted	92%	92%	49%	47%	51%	47%	50%
Number of overlap	40	40	40	50	60	50	50
RMS	0.1	0.1	0.05	0.05	0.05	0.05	0.05
Training time	21m	15m	27m	11m	15m	25m	29m
Retraining	4	3	7	2	2	4	5
Iterations	5000	5000	2000	5000	10000	10000	500

Table 4.22: Summary of Results of Experiment 4

v) Experiment 5

	0-10%	11-40%	41-50%	51-60%	61-70%	71-80%	81-90%	91-100%
Test case % of records correctly predicted during training			1, 2, 10 15	3, 4, 5, 11, 12, 14 20, 21, 22, 4	6, 13, 19,23			
Test case % of records correctly predicted during validation	1, 2, 3, 5 6, 10, 11 12, 14, 15 19, 20, 21 22, 23, 24							

Table 4.23 Correctly predicted record percentage by test case (experiment 5)

The % of records correctly predicted, training time, iterations and retrainings is 5%, 9 min, 500 times and 1.5 times. This is the worst experiment so far with the best cases having as high as 90% of the records predicting wrongly. No good model are found from this experiment as can be seen from the sample results shown in Table 4.13.

vi) Experiment 6

	0-10%	11-20%	21-40%	41-50%	51-60%	61-70%	71-80%	81-90%	91-100%
Test case % of records correctly predicted during training					29, 32 43	2, 4, 5 6, 7, 8 9, 10, 12 14, 17 18, 28, 30 31, 33 35, 42, 45	1, 3 11, 13, 15,16, 36, 37 38, 39, 40 41, 44	34, 43	
Test case % of records correctly predicted during validation	35, 36, 42	45		2, 3, 5 9, 14 18, 30, 40 41	3, 34, 39 44	1, 6, 7 11, 12, 13 16, 28, 29 32, 33, 37 38	4, 15	10, 31	17

Table 4.24: Correctly predicted record percentage by test case (experiment 6)

The average percentage of records correctly predicted, training time, iterations, and retrainings is 54%, 8 min., 5000 times and 2.5 times respectively.

Only 9% of the test cases predicted within 80% accuracy and with only one case above 90%. The best case is an RBF using 80% learning samples with 35 units, 10,000 iterations and 50 overlap. The worst cases typically use 70% learning samples and only 10 units.

4.5.2 Summary of Results

In order to build a model that could well represent a suitable and accurate system to forecast revenue and predict the realized revenue of a flight we have conducted 327 test cases using BPN and RBF neural network algorithms and a multitude of training parameters. The results of these experiments have yielded several interesting findings.

It has been observed consistently that DB2 produces very poor results. This could be in part due to the smaller amount of records it has, the less information on a specific flight it contains and the inappropriateness of the fields selected. On the average DB1 (experiments 1 and 4) used 135 records whereas DB2 (experiments 2,3,5,6) used 45 records only.

Experiments 1 and 4 have a commonality in that both their independent variables are fare class by segment. This tells us something about the importance of these two variables. We have discovered that fare classes alone cannot represent all the components of flight revenue and give accurate estimates of the revenue. Experiments 2 and 5 use fare class only as their independent variables but consistently showed the worst results. Nevertheless, when fare class is coupled with segment, as we have seen in experiments 1 and 4, the result is a robust model.

Although experiment 4 yielded the best average result, it only gave two best models out of a test case of 108. In addition, it had difficulty with training many cases (40%). This leads us to prefer the consistent results of experiment 1.

In addition, the following critical observations have been made:

- For the amount of record presented to them, the test cases had difficulty building a good model when the RMS was 0.05. They gave much better models when the RMS was increased to 0.1.
- In all cases the best cases were when the activation function was a sigmoid function.
- Comparing the results, it seems that time, retraining and iteration are a function of the RMS value; as the RMS gets larger the others decrease. Hence, these are not good criteria for evaluation.
- Too many processing units seemed to confuse the network so keeping it at a lower level is preferred.

In summary, the model, selected from the 327 test cases, that is best suited to our problem has the following features:

- Algorithm: MLN
- RMS: 0.1
- Learning sample: 75%
- Testing sample: 25% of the learning sample
- Activation function: sigmoid
- Independent variable : fare classes and segment
- Processing units: 35
- Number of records: >135

The model was scored, that is, tested in a live environment, using fresh data of flights ET730/19JAN00 and ET 730/26JAN00.

The result is shown under Table 4.25. The total estimate for ET730/19JAN00 is ETB 303,632 and for ET730/26JAN00 is ETB 148,532. The actual revenue realized for these flights is ETB 453,711 and ETB 236,021, respectively. Hence, the prediction or estimate errors were 33% and 37%, respectively

Flight No	Date	Segment	C	D	S	B	H	K	L	M	Q	T	Revenue Prediction ETB)
ET730	19-Jan-00	ADDFCO	7	0	6	2	1	2	10	34	6	5	
ET730	19-Jan-00	ADDLHR	6	0	0	5	2	0	3	47	21	0	287,375
ET730	19-Jan-00	FCOLHR	0	0	0	0	0	0	0	0	0	11	16,257
ET730	26-Jan-00	ADDFCO	1	0	0	2	0	3	13	16	8	4	
ET730	26-Jan-00	ADDLHR	3	0	1	1	0	0	5	19	22	1	132,590
ET730	26-Jan-00	FCOLHR	0	0	0	0	0	0	0	0	0	12	15,942

Table 4.25: Scoring Results of ET730/19/26JAN00

Chapter 5
CONCLUSION AND RECOMMENDATIONS

5.1 Summary and Conclusion

During the preliminary survey of this research we had identified three problems. We had seen that ETHIOPIAN was at a considerable disadvantage because users were not being provided appropriate flight revenue information. The revenue report was untimely; the report was not to the flight level; and there was no information on forecast of flight revenue.

To this end, one of our objective was to conduct a business survey to understand the user requirements, critical business functions and process involved in flight revenue, and availability of data. The business survey conducted confirmed these preliminary findings. As a result it is believed that it is important for the airline to develop a full fledged information system that will completely address these problems.

The survey also indicated that a multi-class inventory nesting structure supported by a corresponding hierarchical fare structure are fundamental requirements to build an accurate model for flight revenue. Furthermore, the necessary data required to build the model are in place although sufficient historical data is not available and the quality of the data is questionable.

The business survey findings on the flight revenue process can be summarized by the functional diagram below (figure 5.1). The central role of revenue management in promoting revenue enhancement makes it an especially important function in the revenue process, as can be seen in figure 5.1. The benefit of revenue information can only be realized when there exists active communication and feedback between revenue management and the key areas that provide valuable input to the revenue process. The survey shows that this process is currently fragmented with little interaction between departments.

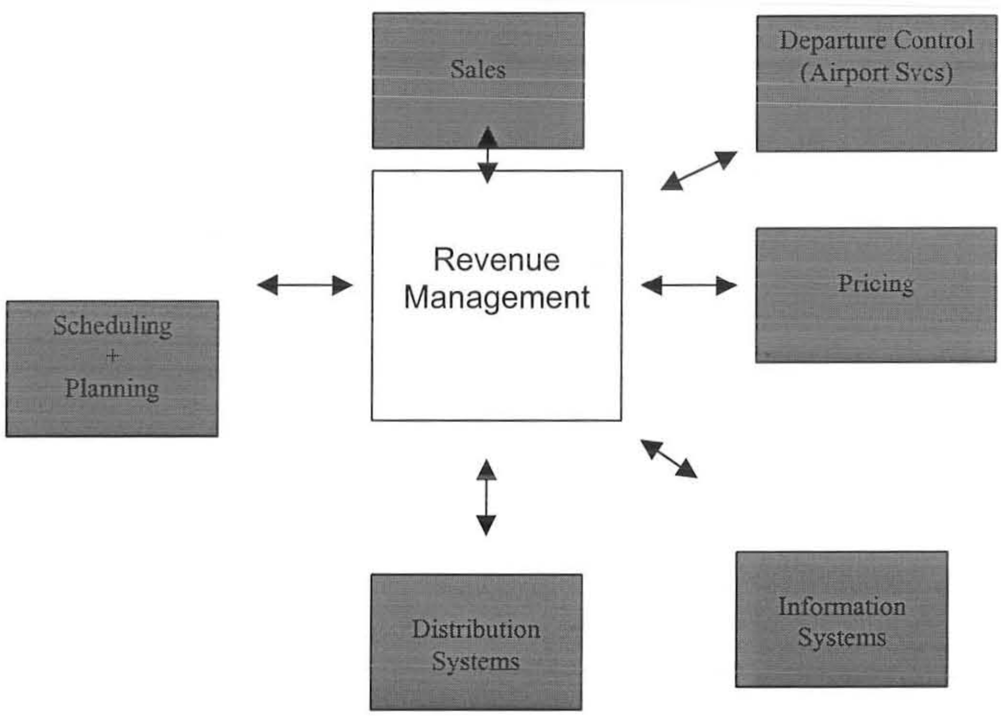


Figure 5.1: Central Role of Revenue Management

The area sales departments are responsible to generate revenue through the sale of inventory. Revenue management determines the number of seats that will be available for sale in each fare class.

The Scheduling department determines the schedules necessary to achieve the objectives of Sales and Marketing and determines the product available in the marketplace and Revenue Management, on the other hand, determines the viability of the product.

Pricing is responsible for establishing fares for scheduled flights based on market economics and competitive factors. Revenue Management is responsible for determining the optimal mix of seats at those fares in such a manner as to maximize revenues on each flight departure.

The critical factor is that the key departments discussed above work closely together and in concert with each other. Interaction between these departments is essential in order to maximize the use of the realized and forecast information on flight revenue. In addition, the Information Systems department has to be more user-oriented and focus its efforts towards solving actual users' needs rather than refining applications.

The other objective of this research was to experiment and find out the suitability of data mining applications to solve our problem. The results of the experiments conducted give sufficient reason to continue work in this area. Several very good models have been developed that attempt to address the research problems. We have tested whether fare classes are sufficient representatives of revenue and conclude that unless they are coupled with segment they are not.

Although the estimate error, which ranges between 30-37% is not to the expectations of the users (which tolerate only upto 10% error), the result is encouraging.

Although the results are encouraging this model, however, is a preliminary or initial step to, hopefully, more detailed work in this area. The preliminary work accomplished through this research is only an eye opener to those interested in data mining applications - there is still much to be done. For example, cargo revenue, mail, exess baggage, in-flight sales, etc. must be integrated in the model; data of all the airline's flights must be incorporated and all sorts of other data must be searched and exhaustively tested to find out their significance. One year's historical data does not produce sufficient records to substantially reduce error. If possible over three years' data, if available, should be collected. In addition, due to time limitation I could only use limited amount of values for the training parameters. This is reflected in the error of 30-37% encountered in the live test.

The weakness of neural networks is that it is very difficult to understand what it learned, how it learned it and the how the weights are set. This is important because I have found it very difficult to analyze results which have similar parameter settings but were giving very different results. In addition, the neural network requires immense amount of data to predict accurately. I believe that many models could not predict accurately because of this. Due to time constraint performance over time could not be measured.

In addition, my own limited experience in this area has probably contributed to some poor results. In collecting data and selecting fields, and even in testing and training their have necessarily been some subjective choices to be made that may have influenced some of the results.

5.2 Recommendations

Several recommendations can be made as a result of the findings of this research.

In the business area, effective implementation of a revenue system requires more than just an automated system. It requires an environment capable of supporting the entire process. Appropriate and cohesive policies that promote effective inter-departmental communication and integrity of data should be developed, implemented, and enforced. For example debit memos should be issued to sales agents who do not book passengers in the correct fare class.

The pricing process must especially be streamlined. With the availability of better revenue information, it is possible to confidently loosen the current tight control on the pricing approval process.

The important observation made in regard to the business process is that the critical functions in the revenue process are disparate and disjoint. I would recommend that pricing and revenue management be combined into one department and report to the same head as that of scheduling. In addition, the communication with other Sales and Marketing departments should be improved through better procedures and through more efficient use of available information technology.

It is critical that the basic data elements required for input into the revenue model (advanced booking, post departure and schedule) be available in a timely and accurate manner. A policy which states that revenues, rather than yields or load factors, constitute the yardstick by which performance is measured should be established in order to make effective use of such a system.

It should also be noted that this model, even after refinement, needs to be monitored on a continuous basis and retrained. Market behaviors change over time; external variables completely change the data elements that were used as inputs to build the model and render the information output useless or misleading.

A number of system interfaces will need to be developed to support the integration of the flight revenue model into the Ethiopian Airlines revenue information system. These interfaces are directly related to the need to be able to transfer information to and from the revenue information model as follows:

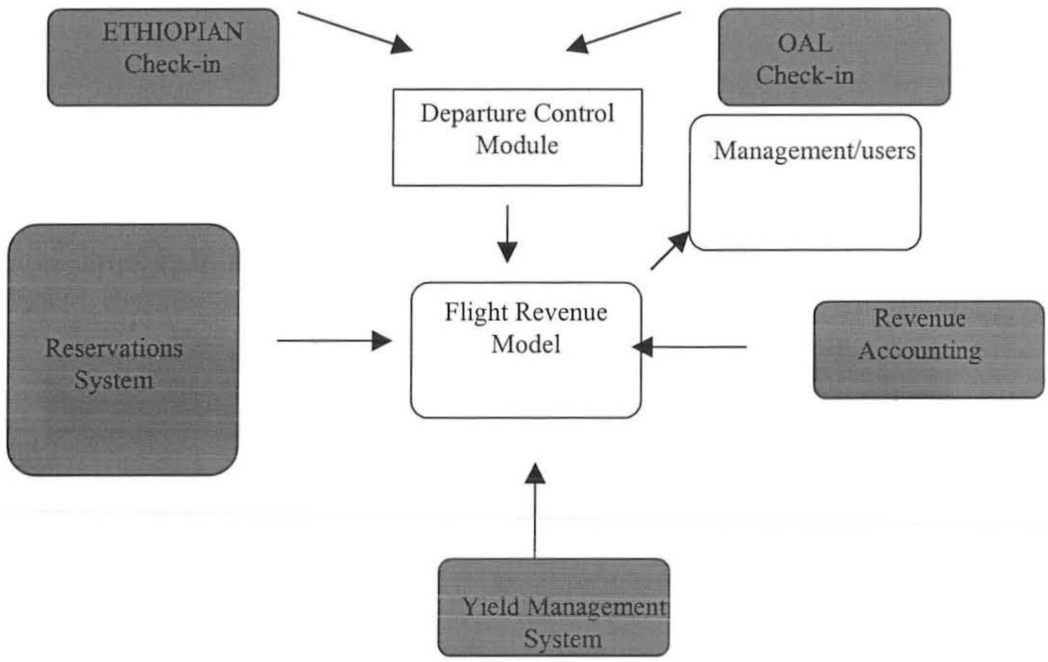


Figure 5.2: System Interfaces

Advance Booking Interface

A program will need to be written to download advance booking data from the yield management system. The application will aggregate the advance booking data to the flight segment, fare class level for each flight departure and output a flat file for subsequent transfer to the flight revenue model.

Schedule Interface

An interface program to the reservations system must be written to interrogate the reservation system and download, transaction by transaction, the necessary schedule information.

Revenue Accounting Interface

The current revenue accounting system can generate revenue information by fare class. The resulting output file (by flight number and day-of-week) can be downloaded to the flight revenue model on a scheduled basis (i.e. monthly, quarterly, etc.) using the same method identified for the Advanced Booking data. This data is required to retrain the model periodically to keep abreast of new trends.

Check-in Interfaces

An automatic link between the departure control system and the reservations system already exists. The post departure data can be downloaded to the reservation system and the same interface program which collects schedule information, described above, can collect the post departure data.

Finally, it is my belief that this research has some contribution to further research in the area. It has been able to successfully demonstrate that data mining applications can be an alternative approach to build information systems. Others can pursue similar research using different types of data mining applications or neural network models.

Some of the problems I encountered and the methodologies I used may help to guide others undertaking similar studies. I hope that the findings and experience noted throughout my research will stimulate others towards further exploring data mining applications.

B I B L I O G R A P H Y

Bhattacharya, G. (1994). **Information: Its Geneses, Capturing and Communicability.** Addis Ababa: AAU.

Bigus, Joseph P. (1996). **Data Mining With Neural Networks.** USA: McGraw-Hill Inc.

Bingham, John E. and Garth W.P, Davis. (1984). **A Handbook of Systems Analysis.** 2ed. Hong Kong: Macmillan Publishers Ltd.

Chisnall, P.M. (1991). **The Essence of Marketing Research**. London: Prentice Hall.

Chung, H. M., and Paul Gray. Data Mining. *Journal of Management Information Systems*. Vol. 16(1), Summer 1999.

Clark, David W. (1997). **An Introduction to Neural Networks.** n.c: n.p.

Connolly, T. (1995). **Revenue Optimization System Review.** USA: PROS Strategic Solutions, Inc.

Data Intelligence Group. (1995) **Data Mining Overview: An Introduction to Data Mining.** Cambridge, MA: n.p.

Eardley, A., Marshall, D. and Bob Ritchie. (1995). **Information Analysis.** London: Certified Accountants Educational Projects Ltd.

Edelstein, Herb. (1997). **Data Mining: Exploiting The Hidden Trends In Your Data**. USA: Miller Freeman Inc.

Edelstein, Herb. (1998). **Data Mining: Let's Get Practical.** USA: Miller Freeman, Inc.

Edelstein, Herb. (1997). Mining For Gold. *Information Week*. NY: CMP Media, Inc.,

Ethiopian Airlines-Controller (1975). *Management Policy & Procedure Manual*. Revised 1996, 1997. Addis Ababa: Ethiopian Airlines.

Ethiopian Airlines-Finance (1998). *Annual Report*. Oman: Fin Global Vision.

Ethiopian Airlines-Finance (1999). *Financial Report: Green Book*. Addis Ababa: Ethiopian Airlines.

Ethiopian Airlines-HRM (1992). *Manual of Organization*. Addis Ababa: Ethiopian Airlines.

Ethiopian Airlines-Information Systems (1999). *Overall Current Status Of Computerized Information Systems*. Addis Ababa: Ethiopian Airlines.

Ethiopian Airlines-Information Systems (1999).*Yield Information System Software Requirements*. Addis Ababa: Ethiopian Airlines

Ethiopian Airlines-Marketing (2000). *Flight Performance Report*. Addis Ababa: Ethiopian Airlines.

Ethiopian Airlines-Marketing (1971). *Sales and Services Procedure Manual (SSPM)*. Addis Ababa: Ethiopian Airlines.

Ethiopian Airlines-Marketing (1988). *Brining Africa Together: The Story of an Airline*. Nairobi: Camerapix Publishers Int'l.

Ethiopian Airlines-Marketing (2000). *Selamta*. Nairobi: Camerapix Publishers International.

Ethiopian Airlines-Marketing (2000). *Worldwide Timetable Express*. Dubai: Printing Services.

Hormby, S. (1992). **Neural Networks for Yield Management.** n.c.: Advanced Software Technology For Air Transport.

Lavin, M.R. (1992). **Business Information: How to Find It, How to Use IT**. Second Ed. Phoneix, Arizona: Oryx Press.

Pham, D. T., and X. Liu. (1995). **Neural Networks for Identification, Prediction and Control.** 2ed. Great Britain: Springer-Verlag London Ltd.

Rowley, Jennifer. (1990). **The Basics of Systems Analysis & Design for Information Managers**. London: Clive Bingley.

Singer, Gilbert. (1996). **Object Technology Strategies and Tactics**. New York: SIGS Books and Multimedia.

Small, Robert D. (1997) Debunking Data Mining Myths. *Information Week*. NY: CMP Media.

Small, Robert D., and Edelstein, Herbert A. **Scalable Data Mining.** USA: Two Crows Corp.

Whitten, Jeffrey L., Lonnie D. Bentley, Kevin C. Dittman. (1998).**Systems Analysis and Design Methods**. 4ed, Boston, MA: Irwin/McGraw Hill.

n.a., n.d. **Intelligence Technologies**. n.c., n.p.

Http://www.angoss.com/about/datamining.html (Site accessed March 2000)

Http://www.data-miners.com/products/vendors.html (Site accessed February 2000)

Http://www.Kdnuggets.com/software/classification.htm (Site accessed March 2000)

Http://www.SPSS.com/datamine/networks.htm (Site accessed February 2000)

APPENDICES

APPENDIX 1

The Pricing Decision Problem

At the outset of this research the major problem identified was the unavailability of flight revenue information, and the objective was to develop an information system that would provide this information as quickly as possible. After working on the ground work, analyzing the existing system, and making preliminary designs of the system to be, I realized that I was embarking on a project rather than a research. I was not really discovering anything new, and quite valueless at that since the system already existed; a much more powerful and better one than I could ever attempt to develop.

The real problem was not the unavailability of flight revenue information but the unavailability of timely flight revenue information; and this alone did not constitute a strong enough theme for a research. After carefully and in detail analyzing the root cause for the delay the major reasons are found to be a combination of manpower shortage, bureaucratic procedures and inefficient processes. This problem could be solved either by duplicating, at each airport, the module of the revenue accounting system which prorates the ticket values, or by hiring additional accountants, or by improving procedures and processes from airport departure control to accounting.

However, neither these procedural solutions nor the one of developing a new system effectively address the human element. To implement these solutions it would require training people to use the system and more importantly developing the procedures and policies to ensure data is timely and accurately input – this would entail extra costs, extra work, a change in mentality and extra working staff at every airport of ETHIOPIAN's destinations.

From the results of the analysis of the current system, I had identified other problems especially in the area of pricing. The main one being the pricing decision making process. The problem here was the inability for pricing analysts and management to quickly react to the market and develop competitive prices in time. The objective, as a result, became to develop a system that would help in improving this decision process.

However, after extensive investigation into the pricing process and the factors influencing pricing decisions, I realized this was not a problem that could be solved at this time. The main reasons were that the approval process was too centralized with only the CEO having the authority to approve most type of prices and the Executive Officer for Sales and Marketing, his immediate subordinate, to approve the remaining. With such a process the value of my system would greatly diminish. Another, and a more serious problem, was the inconsistency of the process. There are no properly defined rules by which prices are accepted or rejected.

Interview with pricing staff revealed that the same factors were not considered all the time to establish prices and they had no historical data on decisions being made. They worked only through a general guideline and most of the actions or decisions they take are highly subjective. Lastly, but not least, there isn't sufficient historical data with which to develop a model or even trace a pattern for the pricing decision making process.

Although I decided to drop pursuing this problem any further, it would be useful, however, to note the 'pricing decision' problem for ETHIOPIAN management to research and find a solution for, or a very interesting one for another thesis, if and when, the required data become available.

Appendix 2a

Interview & Questionnaire Questions

1. a. What application/information systems are used to perform various functions in sales and marketing?
   b. How do you rate the information each of these systems provide, in general, and in terms of revenue information, in particular ?

2. What are the five most important information you require to perform your task efficiently and to make effective decisions?

3. Do you believe information on flight revenue to be critical ?

4. Do you currently receive information on revenue realized by flight?

   If so, is it accurate? Is it timely?
   If not, what problems or disadvantages do you perceive by not receiving this information at all, or not in time.

5. Do you currently receive information on the forecast of revenue to be realized by flight?
   If so, is it accurate? Is it timely?
   If not, what problems or disadvantages do you perceive by not receiving this information at all, or not in time.

6. a. If information on revenue was provided how would you rate its usefulness?
   b. What opportunities do you believe the above information would bring ? what decisions does it support?

7. a. What kind of data is available in the application systems to support flight revenue information, in general? To support flight revenue forecast information, in particular?
   b. How accurate are these data ?
   c. Is data captured automatically?
   d. How is booking, schedule, actual boarded passenger and flight revenue data communicated to users?
   e. Are systems linked through automatic interface?

8. a. Is historical booking data available?
   b. Is historical flight revenue data available?
   c. Is forecast on booking data available?

9. If forecast on booking data is available then what is the problem in forecasting revenue using this information ?

10. a. What factors do you consider to decide prices?
    b. Please rank these factors according to their importance with weight (in %)
    c. Do you have historical data on factors used and pricing decisions made?
    If yes, how far back is this data available?
    d. Do you evaluate your pricing decisions?
    e. Is there any information available or used to help in this evaluation or to determine the performance of the different prices you issue ?
    If yes, what kind?
    If not, would information on the forecast of flight revenue to be realized or timely information on flight revenue realized help?

11. a. How are ranges for fare classes set ?
    b. are there any problems or constraints imposed? If so, what kind ?
    c. can fare-classess be used to accurately estimate or forecast flight revenue ?

    If not, why not?

12. How would you describe the communication of information between departments.

13. In relation to other departments in marketing, how would you rank your department's role in the process that lead to flight revenue?

14. a. Are major users of application and information systems at Ethiopian Airlines networked?
    b. Does Ethiopian Airlines have a LAN ?
    c. Does Ethiopian Airlines have a WAN ?

15. Would you consider information on forecast of revenue to be realized by flight (if accurate) to be strategic?

16. Do you believe information on the forecast of flight revenue would provide opportunity to gain competitive advantage ?

17. How stable is the environment you operate in?
    If unstable, volatile or extremely volatile would you agree that timely (day after departure) information on revenue realized is essential to take quick actions or make sound decisions?
    What about information on forecast of revenue to be realized?

18. a. Would you describe your route schedule as complex?
    If yes, would timely information on flight revenue help in rationalizing your schedule or your fare structure?
    What about forecast information on revenue to be realized by flight?

    If yes, would timely information on flight revenue help in rationalizing your fare structure?
    What about forecast information on revenue to be realized by flight?

19. a. What are the flight performance indicators that you currently use to evaluate the performance of a flight?
    b. Are the indicators currently used reliable?
    c. What is currently your best indicator to estimate flight performance and revenue?
    d. Please rank the indicators according to their importance, and assign weights (%)
    e. How would forecast information on revenue to be realized by flight and timely (day after departure) information on flight revenue realized rank against the above indicators?

20. Is load factor a good measurement of flight revenue?

21. How is flight revenue currently forecasted or estimated? Is the forecast accurate? What is the error?

22. In the absence of timely flight revenue information, is the current report on flight performance adequate to evaluate revenue of flights?

23. Are assumptions/estimates of fare class average accurate?
    If No, would information on flight revenue help to minimize errors ?

24. How is the communication between pricing and revenue management?

25. a. Would flight revenue information/forecast help to make better inventory decisions?
    b. Better frequent flyer actions and decisions
    c. Better promotional packages
    d. In redirecting advertising efforts

25. What factors are used to measure sales performance?

26. Would information on the forecast of flight revenue to be realized assist to develop better schedules and deciding when making major and minor schedule changes?

28. a. What are the criteria currently used to decide whether to cancel a flight or a stop or to combine two flights into one ?
    b. Would flight revenue information (forecast) help to make better combining, over flying, etc. decisions?
    If yes, to what extent:

29. a. Which of the following business units/functions play the primary role in the revenue process?
    b. Do you believe that if these business units had timely (day after departure) information on flight revenue realized or information on the forecast of revenue to be realized by flight that their performance would improve?
    If yes, how crucial would this piece of information be to improve the performance of these business units?

30. a. Is historical flight revenue information directly available in the financial system?
    If yes, how long does it take to get flight revenue information after the date of flight departure?
    b. If not, can it be made available indirectly?
    If so, how readily can it be made available ?

31. Can the financial system provide forecast information on revenue to be generated by flight?

    a. If yes, how far ahead of flight departure can the system forecast ? a day , a week, a month, a year ?
    b. If not, can it be made available indirectly?
    If so, how readily can it be made available ?

32. What historical data that can support flight revenue information and to what level of detail is available in the financial system ? in the reservations system? in the yield management system?

33. Is the staff in your business unit qualified?

34. What is the characteristic of your market segments?

35. Does the airline have a good understanding of what data mining is ?
    Have you done any development using data mining ?
    If no, how enthusiastic would you be to see a system developed using data mining?

36. a. Is the flight revenue information from the financial system of any use this late?
    b. What is the acceptable delay to make good use of information on flight revenue realized ?
    c. What is the acceptable error margin in terms of accuracy of the information on forecast of revenue by flight and of revenue realized by flight ?

37. If forecast on revenue by flight or if timely (day after departure) information on revenue realized by flight was to be provided in what medium would you like the output ?

38. a. What reports regarding flight performance do you currently receive?
    b. Are there any problems with these reports ?

APPENDIX 2b
INTERVIEW & QUESTIONNAIRE SAMPLE RESPONSE

INTERVIEWEE: *Wro Rahel Assefa*

POSITION: *Div. Mgr. Plnng. & Route Mgmt.*
DATE: *March 13, 2000*
TIME: *09:30 A.M.*
PLACE: *Ethiopian Airlines Head Office Building*
SUBJECT: *Flight Revenue Information*

TIME ENDED: 10:30

1. a. What important information do you require to perform your task and to make effective decisions ?
   - *Revenue*
   - *Cost*
   - *Market Share*
   - *Sales*
   - *Performance*

   b. Do you believe information on flight revenue to be critical ? *Yes*

   Is so, how would you rank it ?

The most critical	one of the most critical	average	less critical than most
☐	☑	☐	☐

2. a. Do you currently receive information on revenue realized by flight? Yes ☐ No ☑

   If so, is it accurate Yes ☐ No ☐

   is it timely Yes ☐ No ☐

   If not, what problems or disadvantages do you perceive by not receiving this information at all, or not in time.

   - *Difficult to evaluate the revenue performance of each flight on a daily basis*
   - *Inability to identify weak and strong flights to rationalize future schedule*

b. Do you currently receive information on the forecast of revenue to be realized by flight? *No*

If so, is it accurate Yes ☐ No ☐

is it timely Yes ☐ No ☐

If not, what problems or disadvantages do you perceive by not receiving this information at all, or not in time.

- *Inability to react quickly to changing market situations*
- *Unable to make advance evaluation on flight performance and take the necessary remedial actions*

c. If information on revenue was provided how would you rate its usefulness

in terms of the following:

	V.High	High	Med	Low	Useless
Information on revenue realized by flight provided the day following flight departure:	☑	☐	☐	☐	☐
Information on revenue to be realized by flight forecasted 1 month ahead of flight departure:	☐	☑	☐	☐	☐
Information on revenue realized by flight 3 months after flight departure:	☐	☐	☐	☑	☐

d. What opportunities do you believe the above information would bring ?

- *Advance planning*
- *Sound decisions*
- *Immediate performance feedback*
- *Proactive actions*
- *Optimize marketing mix*

3. What reports regarding flight performance do you currently receive?
   - *Delayed (after 3-4 months) information on revenue at the system level*
   - *Daily flight load report*

b. Are there any problems with these reports ?

- Inaccuracy ☐
- Insufficiency ☑
- Difficult to read ☐
- Difficult to use ☐
- Not generated on time ☑
- Other ☑ *No revenue information at flight level*

4. What kind of decisions does forecast information on flight revenue support?

   Same as 2d.

5. How would you describe the communication of information between departments?

Very effective	effective	not effective	poor	very poor
☐	☐	☑	☐	☐

6. In relation to other departments in marketing, how would you rank your role in the process that lead to flight revenue?

Critical	V.important	Important	Not so important	No role
☐	☑	☐	☐	☐

7. Would you consider information on forecast of revenue to be realized by flight (if accurate) to be strategic? Yes ☑ No ☐

If yes, how strategic	Extremely	Very	Somewhat	Not that much
	☑	☐	☐	☐

8. Do you believe information on the forecast of flight revenue would provide opportunity to gain competitive advantage? Yes ☐ No ☐

9. How stable is the environment you operate in?

Extremely volatile	volatile	unstable	OK	Stable
☐	☑	☐	☐	☐

If unstable, volatile or extremely volatile, would you agree that timely (day after departure) information on revenue realized by flight is essential to take quick actions or make sound decisions ?

Strongly agree	Agree	No opinion	Disagree	Strongly disagree
☑	☐	☐	☐	☐

What about information on forecast of revenue to be realized ?

Strongly agree	Agree	No opinion	Disagree	Strongly disagree
☑	☐	☐	☐	☐

10. a. What are the flight performance indicators that you currently use to evaluate the performance of a flight?

- *Yield*
- *Availed Seat Kilometre*
- *Revenue Seat Kilometre*
- *Load factor*
- *Lost per seat kilometre*

b. Are the indicators currently used reliable? Sufficient ? accurate ?

- *Not reliable*
- *Not sufficient*
- *Somewhat accurate*

c. What is currently your best indicator to estimate flight performance and revenue?

- *Yield*
- *Load factor*

d. Please rank the indicators according to their importance, and assign weights (%)
- *Yield 45%*
- *Load factor 45%*
- *Others 10%*

e. How would forecast information on revenue to be realized by flight and timely (day after departure) information on flight revenue realized rank against the above indicators? *More or less the same.*

Revenue information 45%
Load factor 45%
Others 10%

11. Is load factor a good measurement of flight revenue? Yes ☐ No ☑

12. In the absence of timely flight revenue information, is the current report on flight performance adequate to evaluate revenue of flights? Yes ☐ No ☑

13. a. Is the flight revenue information from the financial system of any use this late?

Yes ☑ No ☐ If yes, how useful: Very ☐ Somewhat ☐ Not much ☑

b. What is the acceptable delay to make good use of information on flight revenue realized ? One month ☐ One week ☐ One Day ☑ One Hour ☐ after flight departure.

c. What is the acceptable error margin in terms of accuracy of the information on forecast of revenue by flight and of revenue realized by flight ?

	+(-) 5%	+(-) 10%	+(-) 15%	+(-) 20%	+(-) 25%	+(-) 30%
forecast :	☐	☑	☐	☐	☐	☐
realized :	☑	☐	☐	☐	☐	☐

14. If forecast on revenue by flight or if timely (day after departure) information on revenue realized by flight was to be provided in what medium would you like the output ?

Hard copy report ☐ Diskette ☐ On-line ☑

Signature: Rafael Asse [signature]

DIVISION M... ...NING & ROUTE MANAGEMENT

APPENDIX 2C

Interview and Questionnaire Response Rate

Not all questions were given to all respondents. Questions were asked as applicable according to the function and level of the respondent. As a result, four groups of respondents were identified. The summary of number of candidates asked and responded by group type and question is shown herebelow.

Group I :	Areas Sales Managers
Group II :	Divisional and Regional Managers
Group III:	Department Managers and Senior Marketing Staff
Group IV:	Executive Officers
Type A :	Number of candidates who were 'asked' a particular question
Type R :	Number of candidates who 'responded' to a particular question

Ques.	Type	Group				Total	Resp.
		I	II	III	IV		Rate
1	A	15	0	7	0	22	73%
	R	9	0	7	0	16	
2	A	14	6	7	0	27	81%
	R	10	5	7	0	22	
3	A	15	6	7	1	29	86%
	R	11	6	7	1	25	
4	A	15	6	7	1	29	86%
	R	11	6	7	1	25	
5	A	15	6	7	1	29	86%
	R	11	6	7	1	25	
6	A	15	6	6	1	28	86%
	R	11	6	6	1	24	
7	A	0	2	1	0	3	100%
	R	0	2	1	0	3	
8	A	0	2	1	0	3	100%
	R	0	2	1	0	3	
9	A	0	1	1	0	2	100%
	R	0	1	1	0	2	
10	A	0	0	2	0	2	100%
	R	0	0	2	0	2	
11	A	0	0	2	0	2	50%
	R	0	0	1	0	1	
12	A	15	6	7	0	28	86%
	R	11	6	7	0	24	
13	A	15	6	6	0	27	85%
	R	11	6	6	0	23	
14	A	0	3	0	0	3	100%
	R	0	3	0	0	3	
15	A	15	6	7	1	29	86%
	R	11	6	7	1	25	
16	A	15	6	0	1	22	82%
	R	11	6	0	1	18	
17	A	15	6	2	1	24	96%
	R	15	6	2	0	23	
18	A	0	2	4	0	6	83%
	R	0	2	3	0	5	
19	A	15	6	0	1	22	82%
	R	11	6	0	1	18	
20	A	15	6	0	1	22	73%
	R	9	6	0	1	16	
21	A	0	1	2	0	3	100%
	R	0	1	2	0	3	

Ques.	Type	Group				Total	Resp.
		I	II	III	IV		Rate
22	A	0	5	2	2	9	89%
	R	0	5	2	1	8	
23	A	0	1	3	0	4	100%
	R	0	1	3	0	4	
24	A	0	0	2	0	2	100%
	R	0	0	2	0	2	
25	A	0	1	5	0	6	100%
	R	0	1	5	0	6	
26	A	15	3	3	1	22	55%
	R	8	3	0	1	12	
27	A	0	1	2	0	3	100%
	R	0	1	2	0	3	
28	A	0	0	3	0	3	100%
	R	0	0	3	0	3	
29	A	0	1	0	1	2	50%
	R	0	1	0	0	1	
30	A	0	2	1	0	3	100%
	R	0	2	1	0	3	
31	A	0	1	0	0	1	100%
	R	0	1	0	0	1	
32	A	0	2	1	0	3	100%
	R	0	2	1	0	3	
33	A	0	0	7	0	7	100%
	R	0	0	7	0	7	
34	A	15	0	0	0	15	67%
	R	10	0	0	0	10	
35	A	0	1	1	1	3	100%
	R	0	1	1	1	3	
36	A	0	7	0	1	8	88%
	R	0	6	0	1	7	
37	A	15	7	7	1	30	87%
	R	11	7	7	1	26	
38	A	15	6	3	1	25	80%
	R	11	6	2	1	20	
Total	Asked					508	
Qs	Resp.					405	79.7%
Total	Asked	15	9	8	2	34	88%
Q'N+I'vs	Resp.	11	9	8	1	30	

APPENDIX 2D

List of Interview and Questionnaire Respondents

Ethiopian Airlines:

Executive Officers

Ato Mekonnen Abebe	A/Executive Officer Marketing
Ato Mesfin Tassew	Chief Information Officer

Divisional and Regional Managers

Ato Abdulahi Ibrahim	Regional Manager Gulf, M.E. & Asia
Ato Alemayehu Tesfaye	Division Manager Computer Networks & Comms.
Ato Assefa Aitenfesu	Regional Manager Africa
Ato Damte Demeke	A/Regional Manager Europe & Americas
Ato Kemeredin Bedru	Division Manager Financial Information Systems
Wro. Makeda Yohannes	Division Manager Promotions & Customer Services
Wro. Martha Tilahun	A/Division Manager Market Development
Wro. Rahel Assefa	Division Manager Planning and Route Management

Department Managers and Senior Marketing Officers

Ato Belew Gugssa	Senior Marketing Officer – Pricing
Wrt. Bethlehem Tsegaye	Department Manager Sales Promotion
Ato Destaw Birke	Senior Marketing Officer Scheduling
Ato Fisseha Teklu	Department Manager Revenue Management
Ato Hailemeleko Mamo	Department Manager Customer Loyalty
Ato Henock Woubishet	Department Manager Marketing Systems Support
Wrt. Kidist Tibebu	Department Manager Ground Handling
Ato Paulos Legesse	Department Manager Tariffs & Industry Affairs
Ato Zewdu Hailemariam	Senior Marketing Officer – Market Development

Area & Sales Managers

Wro. Almaz Demissie	Area Manager China
Ato Amare T/Tsadika	Area Manager Pakistan
Ato Ayenew Alemneh	Area Manager India
Ato Bahiru Kefene	Area Manager Djibouti
Ato Busera Awel	Area Manager UAE
Ato Eskindir Alemu	Area Manager Mozambique
Ato Kiros Girmay	Area Manager Germany
Ato Mesfin Tessema	Area Manager Chad
Wro. Shewaye Asmelash	Area Manager Tanzania
Ato Solomon Dawit	Area Manager Israel
Ato Taye Mulat	Area Manager Mali

Others:

Lufthansa

Ato Yohannes Zerea	Sales Manager Ethiopia

APPENDIX 3

MECHANICS OF THE RESERVATIONS SYSTEM

Below is an illustration of how the linear nesting structure will operate in the reservations system under various booking request scenarios:

Scenario 1 – The following seat protection level is desired exclusively in each fare class:

Class	Exclusive Protection
S	15
B	20
H	25
K	40

The above protection levels would translate to different Authorization (AU) levels depending upon the type of nesting structure as follows:

(Assume these are initial AU settings and the seat available (SA) counts reflect NO bookings on hand).

	AU	SA
S	100	100
B	85	85
H	65	65
K	40	40

Scenario 2 – On the first reservation transaction, 10 sets are sold from 'H' class:

	AU	SA	
S	100	90	
B	85	75	
H	65	55	-10 seats decremented from H class
K	40	40	

Under linear nesting, the availability is decremented only from the class in which the seats were sold AND from all parent classes (i.e. 'B' & 'S').

Scenario 3 – On the second transaction, 18 seats are sold from 'K' class:

	AU	SA	
S	100	72	
B	85	57	
H	65	37	
K	40	22	- 18 seats decremented

Given linear nesting, 18 seats are taken from the class in which the bookings were made (i.e. 'K') and from all parent classes ('H', 'B' & 'S').

Scenario 4 – On the third transaction, 20 seats are sold from 'H' class.

	AU	SA
S	100	52
B	85	37
H	65	17
K	40	17

The availability in a hybrid nested environment would have rejected this transaction.

Given linear nesting, 20 seats are decremented up ward from 'H' class. This leaves 'H' class with 17 seats available. However, 'K' class had 22 seats available before this transaction. Since the rule of linear nesting is that the seat available count in the child class can never be greater than the seat available count in the parent class, the child class' seat available count must be decremented to equal the lowest seat available count of its parent(s).

In Scenario 3, and before Scenario 4, Seats Available (SA) in 'K' class amounted to 22. After the transaction of Scenario 4, the SA in 'K' is forced down to become 17, since the SA in 'K' (22 seats) exceeds that in 'M' (which is now 17).

Scenario 5 – On the fourth transaction, 1 seat is canceled in 'H' class and returned to inventory.

	AU	SA
S	100	53
B	85	38
H	65	18
K	40	18

Under linear nesting, 1 seat is returned to 'H' class and to each of the parent classes. Because 'K' class was artificially forced to an SA of 17 in Scenario Four, it can now increase up to 18 (which is the lowest SA count among the count among the three parents of 'K').

The mechanics of the reservation system will highly influence the relationship of the fare classes and the values of the fields that we are going to use to create the model for the prototype Flight Revenue Information System.

Appendix 4
Training parameter codes of Tables 4.8 - 4.14

A = Partition sample selection type for learn and test data sets.

B = Partition type for learn and test data sets.

C = Partition sample size out of the total records for learn and test data sets.

D = Weighting type.

E = Predictive model type.

F = Dependent variable.

G = Independent variables.

H = No. of records used for training out of learn data sets.

I = Record selection type.

J = Percentage of records used to test out of the learn data sets.

K = No. of iteration b/n overfit tests during learning.

L = No. of checks with worse results to cause termination during learning.

Training Paramaters Code.	Table No 4.8 – 4.11	Table No 4.12 – 4.14
M	No. of neurons in the hidden layer during learning.	Records correctly predicted in training (%).
N	No. of iterations during learning.	Time elapsed to complete training.
O	Error Mean Square (RMS) used during learning.	No. of actual iterations used during training.
P	Activation function type.	Records correctly predicted during validation.
Q	Records correctly predicted in training (%).	-
R	Time elapsed to complete training.	-
S	No. of actual iterations used during training.	-
T	Records correctly predicted during validation.	-

DECLARATION

The thesis is my original work and has not been presented for a degree in any other university.

(Signed)

Gobena Mikael

May 19, 2000

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

(Signed)

Ato Tesfaye Biru

May 19, 2000