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

PREDICTIVE MODELING USING DATA MINING TECHNIQUES IN SUPPORT OF INSURANCE RISK ASSESSMENT

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

Tesfaye Hintsay Atsmo

June 2002

Dedicated to my beloved father,

Hintsay Atsmo

ACKNOWLEDGEMENT

I would like to thank my advisors Ato Ermias Abebe, for his devotion and all rounded support during my work, and Ato Million Meshesha, for his keenness and constructive comments on my work. I am also grateful to all SISA staff for their contribution and support during my stay at the school.

My special thanks goes to the management and staffs of Nyala Insurance SC (NISCO), for their support and for giving me access to the data, and especially Ato Asratu Aemro and Ato Tamene, who are staff members of NISCO, for their understanding and constant support during my work.

I am grateful to my brother Ato Haddush Hintsay and his wife W/ro Wegayehu Asrat for always being there for me, and for supporting me all the way. I couldn't have been able to make this work with out them.

My special thanks are also due to my friends and classmates for their lenience and unreserved help. Finally, I am grateful to those who are not mentioned in name but who helped me much.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AI	Artificial Intelligence
ANNs	Artificial Neural Networks
CSS	California Scientific Software
CV	Commercial Vehicle
EIC	Ethiopian Insurance Corporation
IT	Information Technology
KDD	Knowledge Discovery in Databases
MIRS	Motor Insurance Renewal System
MIS	Management Information System
NISCO	Nyala Insurance SC.
OLAP	Online Analytical Processing
P&C	Property and Casualty
PV	Private Vehicle
RDMS	Relational Database Management System
SISA	School of Information Studies for Africa
UPA	Underwriting Profitability Analysis

Abstract

One of the important tasks that we have to face in a real world application is the task of classifying particular situation or events as belonging to a certain class. Risk assessment in insurance policies is one example that can be viewed as classification problem. In order to solve these kinds of problems we must build an accurate classifier system or model. Data mining techniques are powerful tools in addressing such problems.

This research describes the development of predictive model, which determines the risk exposure of motor insurance policies. Decision tree and neural network were used in developing the model. Since rejections of policy renewal are rare at Nyala Insurance SC. (NISCO), where the research was conducted as a case study, policies were classified into one of the three possible groups (Low, Medium, or High risk) on the basis of annual assessment made by NISCO. Six variables were extracted from the 25 variables used in this study. 940 facts (90% of the working dataset) were used to build both decision tree and neural network models. The remaining 116 (10 %) of the dataset were used to validate the performance of the models. The decision tree model, selected based on the meaningfulness of the rules extracted from it, correctly classified 95.69% of the validation set, and the classification accuracy for low, medium and high risk policies are 98.15%, 94.12%, and 92.86% respectively. The neural network model correctly classified 92.24 % of the validation set, high-risk groups are correctly classified, and low and medium-risk groups are classified with accuracy of 98.15% and 76.47% respectively. Some possible explanations for the relatively low performance of the neural network with medium policies are given. In addition, an interesting pattern was found between the two models that some policies misclassified by decision tree were correctly

classified by neural networks, and vice versa. This is a good indication that the hybrid of the two models may result in better performance.

CHAPTER ONE

INTRODUCTION

1.1. Background

The development of society and economic systems since prehistoric times has been paralleled by a growth in man's dependence upon systems. These systems are highly complex and highly vulnerable. Looked at in terms of the necessities of life, such as heat, light, and shelter, we are dependent on other men and on systems for providing them, which are vulnerable to instant disruption, whether from strike or property damage. It is against this background of system's vulnerability that the need for organized risk management has become self-evident (Bannister & Bawcutt, 1981). The management of risk is the universal problem, a problem for each and every one of us in the day-to-day life at work and play. The study of risk and the management of future uncertainty is therefore one of the most important topic for each of us.

In the business environment, there are two kinds of risks: dynamic and static (Dickson & Stein, 1999). Dynamic risks, often called speculative risks, arise from unexpected changes in the economic productivity of a given capital investment. They arise from market, management, and political sources and are ambivalent in nature: they can result in profit as well as loss (Bannister & Bawcutt, 1981). Static risks, often called pure risks, arise independently of the movements in the economy. They arise from loss of, or damage to physical assets, loss of possession of assets by fraud or criminal violence, loss of income

resulting from damages to property of others, etc. Static risks, unlike dynamic risks, can lead to losses only and are more subject to scientific control (Bannister & Bawcutt, 1981).

To try to eliminate risk in business enterprise is futile. Peter Drucker, as quoted by Bannister & Bawcutt (1981), comments that risk is inherent to the commitment of present resources to future expectation. He goes on to say, " the main goal of management science must be to enable business to take the right risks. Indeed, it must be to enable business to take greater risks - by providing knowledge and understanding of alternative risks and alternative expectations, by identifying the resources and efforts needed to achieve desired results - by mobilizing energies for contribution; and by measuring results against expectations, thereby providing means for early correction of wrong or inadequate decisions."

Risk management may be defined as a process of the identification, measurement and economic control of risks that threaten the assets and earnings of a business or other enterprise (Kaye, 2001). One of the main stages in the process of risk management is risk financing. The most widely used form of risk financing is, of course, insurance. The insurance market provides valuable service for the businessman in enabling him to deal with the financing needs of a large range of unscheduled and undesirable events. The ability to replace such possible losses by a certain, and much smaller, expenditure in the form of a premium has always appeared to most businessmen as a valuable risk trade off facility.

The insurance industry has historically been a growing industry. It plays an important role in insuring the economic well being of one country, but it does not have a high profile and therefore many people have little idea of the full role it plays. The insurance services provided to industry, and individuals, has far-reaching benefits both for those who insure and for the

country as a whole. It provides a form of peace of mind, or security, which is a vital importance in the industry and commerce. It also contributes to a general reduction in the economic waste, which follows from loss (Dickson & Stein, 1999).

The two basic functions in insurance are underwriting and rating, which are closely related to each other. Underwriting deals with the evaluation and selection of risks, and rating deals with the pricing system applicable to the risks accepted. The evaluation of the risks of the asset being insured is primarily based on the information submitted on proposal documentation which all owners of the assets are obliged to hand in. These documents include a number of prescribed characteristics of the asset as well as the owner.

In some types of insurance, such as property and casualty (P&C) insurance, major underwriting decisions, on both new and renewal acceptance, are made in the field. An on-site inspection can give a better idea of the solidity of a property than a sole analysis of the characteristics in the proposal form. On-site inspections enable the underwriter to make sure that the proposal submitted reflect the actual situation, on the one hand, and to obtain additional prospective information about the risk exposure of the property on the other. However, given the large number of policies, which have to be evaluated, and the limited number of employees there is not enough time left for an annual on-site inspection of each property. A model that can identify policies with excessive risk exposure can, therefore, be a useful tool to schedule the timing, intensity, and extent of a particular policy's examination and to focus on potential remedial action.

Traditionally, actuaries develop risk models by segmenting large population of policies into risk groups, each with its own distinct risk characteristics. Premiums are then determined for

each policy in a risk group based on the risk characteristics of the group, such as mean claim rate and mean claim severity, as well as on the cost structure of the company, its marketing strategy, competitive factors, etc (Apte et al., 1998). However, the basic tent in the industry is that no rating system can be perfect and competition therefore compels insurance companies to continually refine both the delineations they make among the risk groups and the premium they charge. The analytical methods employed by actuaries are based as much on statistical analysis as they are on experience, expert knowledge, and human insight (Dockrill et al., 2001). Thus, it is widely recognized that any risk model one develops is likely to overestimate the true levels of risk of some groups of policies and underestimates the risks of others.

Predicting trends and identifying good customers have always been difficult, owing to the need for extensive yet finely tuned analysis of data gleaned from years of customer interaction. In the insurance industry, there is a tremendous wealth of untapped potential in the voluminous raw data associated with policy and claims information. These historical data provide the greatest source of information on risk exposure and is the starting place for insurance risk modeling (Apte et al., 1998).

The incapability of human being to interpret and digest the accumulated data and make use of them for decision-making has created a need for development of new tools and techniques for automated and intelligent huge database analysis. As a result, the discipline of knowledge discovery or data mining in databases, which deals with the study of such tools and techniques, has evolved into an important and active area of research (Raghavan, Deogun & Sever, 1998).

Data mining is an automated process employed to analyze patterns in data and extract information (Trybula, 1997). Data mining is an extension of the Relational Database Management System (RDMS) that helps companies perform highly complicated tasks that are often beyond the capacity of human mind. The greatest difference between RDMS and Data mining is that while RDMS requires users to provide queries and extract information from the database, data mining automatically produces results from the patterns that it detects in the information.

Central to data mining is the process of modeling the data set. There are widely used generic modeling techniques, some of which are Neural networks, Agent Networks, Genetic Algorithm, Decision trees, and hybrid models (Bigus, 1996).

Data mining's extraction of meaning from huge databases is exactly what companies are looking for to increase profits through describing past trends and predicting future trends. It is currently used extensively in several financial institutions, as it is very effective at providing information used for customer profiling and behavior. Banks, in particular, are using corporate data to tailor their interaction with individual customers (Bigus, 1996). Insurance companies are also capitalizing on data mining technology. Many insurers are using data mining to help cross-sell existing customers on additional products or to identify new customers, and others are applying data mining techniques to reduce portfolio risk, and to identify policies that were issued based on incorrect or fraudulent information (Galfond, 1997).

Today the use of data mining technology as a support in business decision-making is growing fast. To the best of my knowledge, however, the potential applicability of data mining

technology in the Ethiopian business context, and in particular in the insurance business is not yet adapted. In the form of case study, an attempt has been made by Askale (2001) to assess the application of data mining technology in support of loan disbursement activity at Dashen Bank, one of the private commercial banks in Ethiopia. The other research, which is undertaken by Gobena (2000), on the possible application of data mining techniques that would help in forecasting flight revenue information in the airline industry (specifically Ethiopian Airlines) is another endeavor, which has been made so far. The results of both previous research works made in these areas have been encouraging. It is, therefore, with this understanding that this experimental research has been undertaken to develop a predictive model using data mining technology for the purpose of insurance risk assessment activities.

The research was conducted in a form of case study in the Nyala Insurance SC. (NISCO). The management's friendly and responsiveness to the researcher's enquiries to get access to and collect the required data for this study encouraged the researcher to conduct this research at NISCO. Nyala Insurance SC. was established in July 1995, and currently claims to be market leader among the seven private insurance companies in its profitability, market penetration, underwriting prudence and low claim ratio. There are about 18 classes of business that NISCO acts upon as a risk financing mechanism. One of these is the motor class. A typical motor policy is an amalgam of several different elements of cover and presents special problems for the insurer since each element impacts on a separate basic account: own vehicle, third party property damage, or third party personal injury (see chapter two for more detail). Moreover, it is the experience of the insurance company that motor class is characterized by relatively high claim frequency and cost. For these reasons, the researcher opted to experiment on motor insurance data.

1.2. Statement of the problem

Insurance underwriting has to do with the selection of subjects for insurance in such a manner that general company objectives are met. The profitability of underwriting depends on accurate risk assessment. The primary goal of the underwriting management is to protect the company against high claim cost due to predictable risk of policies. An increasing claim cost can lead to very serious problem for the company and, more often than not, adequate supervision at an early stage could help avert these problems. Underwriters identify the risks associated with a class of business, evaluate the degree of risk involved in terms of severity and frequency-and then cost the potential claims. Since many factors affect losses and expenses, the underwriting task is complex and uncertain. Bad underwriting has resulted in the failure of many insurance companies (Dockrill, et al. 2001).

Insurance underwriting relies heavily upon the "law of large numbers" (Dickson & Stein, 1999). In large homogeneous population it is possible to estimate the normal frequency of common events such as death and accidents. Losses can be predicted with reasonable accuracy, and this accuracy increases as the size of the group expands. However, there is often a tendency towards adverse selection, which the underwriter must try to prevent. Adverse selection occurs when those most likely to suffer losses are covered in greater proportion than others. Underwriters try to prevent adverse selection by analyzing the moral and physical hazards that surround the risk. A moral hazard exists when the applicant may either want an out right loss to occur or may have a tendency to be less than careful with property. Physical hazards are conditions surrounding property or persons that increase the danger of loss. Either of these hazards may exist on potential customers as well as the existing ones, and hence should be assessed for new as well as renewal businesses.

Underwriters place a great importance upon the renewal of policies. It is considerably more cost effective to renew a policy than take on a new one and, in addition, there is a comfort in having known the risk for a year or more (Kirby & Williams, 1998). To prevent adverse selection, underwriters try avoiding the offer of renewal to be an automatic process. The following are features that will give rise to special consideration at policy renewal (Ibid., 1998):

- Aspects of the risk that have changed since the previous renewal date. This would include high claims frequency or claims value, convictions notified, and new aspects such as a new driver to be added (in motor insurance).
- Aspects of the risk that have been present for at least one year and which have either attracted special terms in the past or place the risk in a non preferred category.

Every insurer will have risks in each of these categories either as a result of a positive underwriting decision taken at the outset, or because of midterm changes or by virtue of the operation of the policy.

However, an increasing size of the customer profile coupled with the small number of employees, made it difficult for the insurers, in particular local insurers, to regularly audit renewal acceptance. This is particularly the case in P&C insurance such as motor insurance, where the assessment of risk heavily depends on on-site inspection of the asset. On-site inspection of motor insurance, for instance, involves through and exhaustive survey of the car to be insured. According to surveyors at NISCO, on-site inspection of motor insurance may involve four classes: walk around inspection, under hood or tilt cab inspection, under body

inspection, and inside cabin inspection. To undertake these inspections, it takes 2 - 3 hours on average depending on the size and complexity of the car.

One alternative solution to this problem is to classify policyholders based on their degree of risk exposure using the portfolio of policy and claims data so that the insurer can easily identify those who demand priority attention from those that need little or no attention at the time of renewal. In this respect, data mining techniques could be effective tools in designing such a classification problem. With data mining, one can build a model that predicts the risk level of a policy by discovering highly complex relationships among the underwriting and claims data.

In this research an attempt has been made to develop a computer-based early warning system or model. This model enables the insurer to use massive historical policy and claims data in the course of investigating the risk exposure of customers at the time of policy renewal.

The primary goal of underwriting system in policy renewal business is not to predict policy rejection, but rather to identify conditions, which could lead to an unacceptable risk and, hence, require further inspection and perhaps closer monitoring (Dockrill et al., 2001). It is under this assertion that the model was built. The use of this model to predict such key future events can also improve the efficiency and profitability of the business.

The data mining technique, which will be used in developing the model, is neural network. The development of neural networks has been one of the most exciting developments from the IT community, which has found application in business (Stergiou, 1996). Neural networks have been applied to an increasing number of real world problems of considerable

complexity. Their most important advantage is in solving problems that do not have an algorithmic solution or for which an algorithmic solution is too complex to be found (Ibid., 1996). These problems include pattern recognition and predicting (which requires the recognition of trends in data).

1.3. Objectives of The Study

1.3.1. General Objective

The general objective of the study was to examine the potentials of data mining techniques in developing a predictive model in support of insurance risk assessment.

1.3.2. Specific Objectives

To accomplish the above stated general objective, the following specific objectives were carried out.

- Identify the features of the insurance risk assessment in particular to motor insurance, and classification mechanism in relation to the data mining techniques;
- Review literature on data mining technology with more emphasis to the neural network and decision tree techniques and their application in insurance business;
- Identify data sources and collect required data from NISCO;
- Explore and select one among the various data mining software that support neural network and decision tree techniques to experiment with insurance data;
- Build and train models and test their performance;

- Build an early warning system prototype using the model selected as a working model.
- Report the result and forward recommendation.

1.4. Research Methodology

To undertake the research described here, the following methodology was used.

1.4.1. Literature Review

The researcher has conducted literature review to assess data mining technology (both concepts and techniques) and research work in this field. Various books, journals, and articles and papers from the Internet have been consulted to understand the practice of risk assessment, in particular insurance risk assessment, and the potential applicability of data mining technology on the insurance industry.

1.4.2. Data Collection

The potential source of data used to undertake this research is mainly the policy and claims database used by each service unit (branch) of NISCO. The manual formats used to collect information about the vehicle and the owner both at the time of underwriting and claims request were also used in support of preprocessing the collected data.

By the time data collection was conducted, the insurance company had fifteen service units, two of which serves only for sister companies. These two service units were excluded, due to

the aggregation of the records under few policies, which made it difficult to extract the value of some important attributes. Thus, the researcher has conducted sampling on the remaining thirteen service units. The sampling technique used is purposive (judgmental) that all service units (four in number) located in Addis Ababa were included in the sample. The reason for purposive sampling was mainly due to the problem to data access from the branches outside Addis Ababa and the short time span for research undertaking. The inclusion of all service units found in Addis Ababa in the sample is because of the small number of records (on average 333) in each unit. The initial dataset included policies, renewed or newly accepted, whose expiry date is in the 2001 accounting period.

1.4.3. Data Preparation

The collected data has been preprocessed and cleansed in to a form suitable for the particular data mining software used in the study. The data preparation involved handling noisy data and unknown values, as well as accounting for missing data fields, deriving new fields from the existing ones, and summarization of data. Then initial relevant features based on the goal of the study have been identified in consultation with domain experts.

1.4.4. Training and Building Models

For reason of accessibility, *See5* and *BrainMaker* software were used in this study. See5 software, a decision tree classifier, is mainly used in this study for attribute selection that could be used as an input for the neural network. The BrainMaker software, neural network tool, was used to train and build neural network models. Feature selection and model building have been made iteratively using different decision tree and network parameter settings until

an acceptable result was obtained. Both tools have the facility to partition the dataset randomly into training and testing sets. Of the 1044 records, which are 90% of the total working dataset (1333), 940 (90%) of the 1044 facts were used for training and the remaining 104 (10%) were used for testing purpose.

1.4.5. Performance Evaluation

In feature selection phase, a working decision tree model was selected based on the consistency of the rule set extracted from the tree. For the purpose of performance evaluation of the neural network models, 10 percent (116) of the working dataset was randomly selected and put aside. After the training and model building stage was completed, 12 models were selected on the basis of their performance using the test dataset. Then, the selected models were evaluated for their predictive performance using this 116 dataset. Finally, statistical results are presented.

1.4.6. Prototype Development

An attempt has been made to develop an early warning system prototype by integrating the working neural network model with a database application. MS Access 2000 is used to design the database. The integration of the model with the database application was made using Visual C++ version 6.0 programming language.

1.5. Scope and Limitation

The insurance industry deals with risks of various classes of business. The underwriting problem discussed under the "Statement of the problem" section of this chapter exists in most of these classes of business. The scope of this research, however, is to examine the potentials of data mining techniques in developing predictive model in support of motor insurance underwriting.

The following are some of the limitations noticed during this research undertaking. First the research was originally planned to conduct a case study at the Ethiopian Insurance Corporation (EIC). However, for reasons that are hidden from the researcher, access to the data source was found difficult. A couple of time was spent during this phase, before changing to Nyala Insurance SC. (NISCO). Finding resources such as literature related to the study, alternative tools both for the neural networks and decision trees modeling was another limitations mainly due to price and ease of access.

1.6. Research Contribution

In this research an attempt was made to find out the applicability of data mining technology in the insurance industry. The result of the study shall be used as an input for the development of full-fledged data mining application using neural networks in supporting insurance underwriting activity. Although the study was aimed at addressing insurance problems in particular, the output of the study maybe used as a source of methodological approach for studies dealing with the application of data mining technology on similar problem areas.

1.7. Thesis Organization

This research report is organized into six chapters. The first chapter briefly discusses background to the problem area and data mining technology, and states the problem, objective of the study, research methodology, scope and limitation, and application of results of the research. Chapters two and three review background materials necessary to understand the basic concepts and results of this research. The concepts of the Motor insurance risk assessment are reviewed in chapter two, followed by a review of data mining technology in general and its application under chapter three. Chapter four explains the methods, neural networks and decision trees, used in this study. Chapter five presents the experimentation phase of the study. Results of training and testing of the neural network models were also discussed here, and finally, describes a prototype that demonstrates how data mining in general, and neural networks in particular significantly improves motor underwriting. Finally, Chapter six provides conclusion, and offers recommendations for future work.

CHAPTER TWO

INSURANCE RISK ASSESSMENT

2.1. Introduction

Insurance is a service industry, it is there to serve the needs of its customers and these needs do change. The underlying service is that of providing a risk transfer mechanism, but the nature of the risks for which this maybe necessary will alter as time passes. New products, processes, and industrial systems all bring new form of risk for which customers, be they corporate or private, will need protection. The management in the insurance industry that deals with the evaluation and selection of risks is the underwriting management.

Underwriting management involves the planning, development and control of a portfolio of policies (Dockrill, M. et al., 2001). To maintain a healthy portfolio, underwriting has to do with the selection of subjects for insurance in such a manner that general company objectives are met.

There are a number of subjects or classes of business that an insurance company underwrites. One of these is the motor class. This chapter reviews the motor insurance underwriting processes: motor policy classification, classification of risks, and risk assessment and improvement tasks. It also pointed out the problem associated with motor policy renewal, taking the practice of NISCO as a reference. Finally, how data mining could support the problem sought in motor insurance renewal is suggested.

2.2. Classification of Policies

The Nyala Insurance SC. classifies motor policies according to the indemnity, which they provide, and those ordinarily in use. These are described as follows.

2.2.1. Comprehensive Cover

A Comprehensive cover is the widest form of cover available, although it cannot protect against every conceivable risk. It provides protection against a wide range of contingencies including not only an indemnity in respect of the insured's legal liability for death of or bodily injury to or damage caused to the property of third parties arising out of the insured vehicle but also will indemnify the insured in respect of all damage/loss to the vehicle and its accessories caused by any accidental, external and physical means as a result of Collision, Overturning, Fire, Self-Ignition, Lightning, Explosion, Burglary, and House Breaking, Theft And Malicious acts. The policy exclusion among other things are wear and tear and/or depreciation of any motor vehicle or any parts of such motors, mechanical fracture and/or mechanical or electrical break down or failure of any part of ~~any~~ motor vehicle.

4.2.2. Third Party Cover

A Third party cover indemnifies the insured against liability at (Ethiopian) law for damages in respect of death of or bodily injury to third parties and damage to the property of such persons in the event of an accident caused by or through or in connection with any motor car described in the schedule of the policy.

2.2.3. Third Party Fire and Theft

In addition to the protection given by the third party insurance, this type of policy covers loss or damage to the insurer's own car as a result of fire, theft, or attempt theft.

2.3. Risk Assessment

Profitable underwriting depends on accurate risk assessment (Dockrill, et al, 2001). Underwriting management identify the risks associated with a class of business, evaluate the degree of risk involved in terms of severity and frequency, and other external factors.

The risk assessment process requires relevant and detailed statistical data, which is available from both internal and external sources. Internally, and most relevant, is the data extracted from the insurer's own policy and claims records. Externally, insurers have long recognized the value of exchanging claims and related data.

It should be noted that merely gathering information, however relevant, serves little purpose. Underwriters must organize the data in such a way that they are enabled to make the necessary judgments. The data must also be available in a manageable format.

2.3.1. Classification and Categorization of Risk

As a starting point, underwriters assess those features and characteristics, which they consider are necessary to assess the comparability of the range of risks within a particular class of

business. This process, subject to ongoing review in the light of claims experience, enables the classification of risk into groupings assessed to present a similar degree of hazard. For example, in motor insurance, different makes of comparable vehicles will be placed in the same classification and group for the purpose of rating each part of the policy (Dockrill, M. et al. 2001).

Insurers draw on their statistics and on their experience to issue special policies for various types of vehicles on the roads (Kirby & Williams, 1998). There are various categories of motor risks and a distinction is made in accordance with the use and type of the vehicles. The main classifications of motor policies at NISCO are as follows.

2.3.1.1. Private Vehicle (PV)

A motor vehicle used solely for private (social, domestic, pleasure, professional, purposes or business calls of the insured) purposes are classified as "Private Vehicles" and are insured under the "Private Motor Vehicles Policy." The term "Private Purposes" does not include use for hiring, racing, pace making, speed testing, and the carriage of goods in connection with any trade or business or use for any purpose in connection with the motor trade.

2.3.1.2. Commercial Vehicles (CV)

A wide range of vehicles that carry goods and passengers are classified under this heading and different rates of premium are applied depending on their use and type. These vehicles are sub-classified as follows.

Goods carrying Vehicles:- These are vehicles primarily constructed for carrying goods, and in the rate chart they are divided as:

Own goods: for carrying solely insured's goods, and

General cartage: for carrying goods for hire and further are subdivided as Light vehicles (up to 2350 cc.) and Heavy vehicles (over 2350 cc.).

Tankers: - Vehicles provided with tanks for transport of goods of liquid nature for the insured or others use.

Buses: -These are passenger-carrying vehicles including Omnibuses and microbuses with a seating capacity exceeding twelve including the driver's seat. They are further classified into Public and Own Service.

Taxis: -Vehicles carrying fare-paying passengers usually in a limited number not exceeding twelve seats including the driver.

Car Hire: -Cars which are hired to an organization or an individual, having less than 12 (twelve) seats including the driver's seat used for carrying passengers with or without a professional driver are categorized under this classification.

Motor Cycles: - These are two wheeled or three wheeled vehicles (including Motor Scooters and Mono Pedals), which in accordance with their use may also be private (with or without side car) or commercial (usually three wheeled with goods use).

Special Type of Vehicles: - As the name indicates this is a term given to vehicles, which are of unusual construction or adaptation and which does not fall under the above stated categories. This section includes vehicles such as Mobile Crane, Earth moving Equipment, ambulance, Agricultural vehicles, Road Rollers, Dumpers, Forestry vehicles, Fire Brigade vehicles etc.

The establishment of classifications is only the first step in the process. Within the same class of risk there is potential for a wide differential and underwriters identify different aspects, which they consider necessary to present a full and coherent picture. For Motor insurance, this includes the proposer's driving experience, claims history, occupation, annual mileage and so on.

2.3.2. Acceptance and Renewal Criteria

Acceptance criteria are all the factors, which can affect the incidence of damage, its spread and containment. These differ from one class of business to another and between different trades in each class of business.

The insurer must identify the terms and conditions under which they will accept a new policy or renew an existing one. The requirements must be capable of precise description and must be readily understood by all staff that makes such policy decisions (Dockrill, M. et al., 2001). At NISCO, underwriting guidelines are issued by management for both commercial and private risks.

Opinions about what constitutes an acceptable or superior risk, or one which merits corrective or remedial underwriting, vary from insurer to insurer and even between colleagues. Insurers set parameters of acceptability based on claims history, criminal records, trades or occupations and all other factors identified as pertinent. Such guidelines can affect not only the profile of the business underwritten but also of the business retained. Hence, It is essential that the underwriting management regularly audits both new and renewal acceptances.

2.3.3. Risk Improvement and Survey Criteria

The insurer, having specified the types of risks, which staffs are able to underwrite and the features of risks, which are acceptable and unacceptable, must insure that effective systems are in place to check that the information submitted on proposal documentation, is accurate and complete (Dockrill, M. et al., 2001). One possible mechanism for validating such information is the survey. The insurer must decide on the criteria for the surveys, including the size of the risk that will need to be surveyed for each trade or classification. Obviously, a survey is a snapshot of a risk; so it may also be important to decide how frequently re-surveys will take place.

The survey represents an opportunity for the insurer to work with the policyholder to improve the risk. As well as stipulating what must be done to bring the risk to a state, which the insurer considers acceptable (the risk improvement), the surveyor is able to discuss features of the operation of the business, which can be varied to the mutual benefit of the insured and insurer. Experienced surveyors can advise on range of risk enhancements including disaster recovery planning, health and safety, and security marking (Dockrill et al., 2001).

The survey criteria of the insurer are regularly reviewed and updated as the awareness of types of risk, which can be particularly vulnerable, increases. The survey criteria of an insurer dictate how much of the new business underwritten will be surveyed and how frequently newly acquired business will be re-surveyed. A prudent insurer analyses why they have retained their risks. Questions such as whether adverse features have been overlooked, producing inappropriately low premiums; or whether risks have been incorrectly classified, producing similar effects, will receive attention. It is also important to understand why business is not retained, particularly if it is of a good quality.

The result of a survey should not only be used to ensure that any risk improvements are acted upon, but also to acknowledge where a risk is particularly well managed and whether some reduction in rates may be appropriate to retain the business.

Another important aspect of risk improvement is incentive. One well-known incentive is the 'No Claim Discount'. This discount from the normal premium is awarded on a sliding scale, increasing each year up to 25%, if the policyholder does not make an "at fault" claim. People do not intend to have motor accidents but insurer believes that, in addition to discouraging small claims, No Claim Discount may encourage policyholders to be more careful.

2.3.4. Liaison Between Underwriting and Claims Management

Underwriting management inevitably involves several other organizational functions other than risk survey and improvement that supports the risk assessment activity. One of these is the relationship with the claims management.

According to the domain experts, liaison between the underwriting and claims management functions should be mandatory, not optional. This is true both for the setting of underwriting policies and during implementation, monitoring and revision.

The claims management function can constitute substantially to the development and refinement of underwriting policies, using experience of actual claims circumstances. The special nature of the claims function's involvement stems from its first-hand experience of underwriting policy implementation at customer level.

Claims handlers can give earning warning of deteriorating claims experience under individual policies. For example, successive request for claims in one year could be purely fortuitous or could be evidence of, at the very least, lack of care; this could result in non-renewal or penal renewal terms and/or increased premiums. The claims handler is likely to notice this first and the timing can be critical. It is imperative that the claims handler notifies the underwriter if renewal is imminent.

2.4. Renewal Procedures and Considerations

The bulk of a well-established insurance company's portfolio is made up of renewal business, hence it must insure that all policies are renewed on time and must take all appropriate measures to secure renewals. This involves sorting out renewable policies covering different risks and expiring in a given month. Most property and pecuniary insurance are written for one year. About six weeks before the expiry date each year provision is made for renewal of the insurance for a further year. These renewal policies shall be given to the supervisor for the

purpose of examining each risk separately based on claims experience and other contributing factors affecting the risk.

If an insurer decides to offer renewal, it will prepare a renewal notice to set out a brief summary of cover and premium required and issue this four to five weeks before the renewal date. The notice will set out any terms for payment. Renewal will then be processed if renewal instruction is received from the insured.

The renewal process involves several tasks. The insurer needs to apply a sequence of questions to each case at renewal; some answers can be obtained from the insurer's database by programming it to default if the particular facts appear on the file. There is a fundamental duty upon the insured at each renewal to reveal to the insurer any change of risk since inception, last renewal or last endorsement. The insurer has the options to accept or reject renewal at this point and to amend the policy premium or other terms to reflect the change. There are several reasons for an insurer refusing to accept (Wildman et al., 1998):

- It already has enough cases on the books of risks in the proposer's property, where this is hazardous, or

- It has decided, after suffering poor underwriting results and using its statistical records to identify the areas where the loss record was bad, to restrict acceptances on certain classes of businesses, certain geographical areas, or all cases having any heavy characteristics.

The next stage is to check the details of the particular case for any factors likely to influence the final decision. One of these factors is loss experience. The insurer identifies from loss

records the causes and effects of the accidents. In addition, on-site inspection is made by surveyors to:

i) Check if any cause is still present at the property (related to safety);

ii) Check if any factor, which contributed to the size of the claim, is still present;

iii) Correct, before going on risk, of any remaining causes in (i) or (ii).

The insurer sorts out the material facts to:

- Give a broad idea of the acceptance level or class into which the risk will fall if and when accepted;
- Identify facts which cannot be altered and which prevent the insurer taking any part of the risk;
- Identify facts which can be altered but which prevent the insurer taking any part of the risk until rectified;
- Identify facts which make the risk heavier but which can be accepted at an enhanced premium.

The primary goal of these underwriting activities in policy renewal business is not simply to predict refusal, but rather to identify conditions, which could lead to an unacceptable risk and, hence, require further inspection and perhaps closer monitoring. Insurers are in the insurance business to make money. If they waited until every thing was made low risk, another insurer would be likely to step in and take the business.

2.5. Risk Assessment Practice at NISCO

The activity of policy renewal business requires professionals in various fields, including actuaries and engineers. The number and quality of these professionals determines the success or failure of the renewal business undertaking. This is clearly proved by the practice at NISCO.

The underwriting procedure of NISCO states that the survey department at the main office mainly conducts pre-acceptance risk assessment of commercial vehicles. For private vehicles, service unit surveyors conduct the pre-acceptance risk assessment. In branches where a surveyor for this purpose is not available experienced senior staff, using the pre-risk assessment indicators provided by the Survey Department, shall carryout the assessment of the vehicle. In addition, the procedure says an already insured vehicle of a client, which is over 15 years of age, is subject to assessment every two years to this effect. However, in practice, risk assessment is undertaken mainly for those whose renewal instruction is received after the expiry date of the policy. In this case, renewal is effective from the date the instruction was received provided the existence and road-worthiness of the vehicle is ascertained.

The risk assessment conducted by the surveyors involves various stages (NISCO. 1999). This includes Walk around inspection, under hood or tilt cab inspection, under body inspection, and inside cabin check. To undertake these inspections, it took 2 - 3 hours on average depending on the size and complexity of the vehicle. In addition, the survey department undertakes the following tasks.

- Preparation of guidelines on handling of damaged vehicles

- Identify vehicle parts and current market price
- Recommend on loss prevention
- Training on engineering matters, and free consultancy on safety aspects and engineering matters, and so on.

The small number of surveyors coupled with the overload of tasks to be undertaken, has brought a problem in policy renewal. A report made by the survey department of the NISCO in an annual workshop of the company states that without rechecking the physical condition and even the existence of the vehicles, the service units consider vehicles for renewal (NISCO, 1999).

In order to meet these challenges, reengineering the existing policy renewal process is a necessary and the present research will contribute to this effort. The new process, suggested in this research, involves prioritizing customers on the basis of their risk exposure so that the insurer can identify those who need more attention at the time of renewal. To this end, a computer-based early warning system (or model) that can predict the level of risk exposure of a policy is appropriate solution towards this underwriting problem.

The system receives a sample of past cases, some of which are proved OK (Low risky), while some others turned out to be risky (medium or high). Each case is described by a set of attributes that detects its profile. The system may be used to verify when an insurance renewal should be given more attention with respect to the risk level associated with it before renewal is approved. For each new case, the company applies this model in order to predict the expected level of risk and get a recommendation on the treatment of the case. In this respect,

data mining techniques could be effective tools in designing such systems. Using data mining techniques, an insurance company can

- Underwrite much faster, in particular renewal acceptance;
- Reduce the cost of underwriting;
- Ensure that all the underwriters use a unified commonly accepted model of assessment and explicitly the underwriting factors that are associated with various levels of risk.
- Get objective and documented estimates on the expected level of risk associated with a policy renewal.

CHAPTER THREE

DATA MINING AND KNOWLEDGE DISCOVERY

3.1. Introduction

In earlier times, due to lack of existing automated information systems able to store data and to analyze them, companies suffered. However, recent advances in communication technologies, on the one hand, and computer hardware and software technologies, on the other hand, have made it all the more easy for organizations to collect, store and manipulate massive amounts of data. Over the last three decades, increasingly large amounts of critical business data have been stored electronically. It is estimated that the amount of information in the world doubles every 20 months (Lin & Cercone, 1997), and this volume will continue to increase in the future. Despite the growing volume of data collected by businesses, few have been able to fully capitalize on its value. This is due to the difficult task of fully analyzing these data and discerning the underlying patterns that can emerge.

Indeed, data reflects activities and facts about the organizations. Businesses, for example, use data to gain competitive advantage, increase efficiency, and provide more valuable services to customers. Yet, all these facts and opportunities are buried and concealed within mountains of data, unless tapped, analyzed and converted into useful information and knowledge for decision-making.

For many years, statistical theory and practice has been a traditional method to study and analyze data. This traditional method of turning data into knowledge in most application areas, such as marketing, finance, retail, insurance, science, etc., relies on manual analysis and interpretation. Moreover, it requires one or more analysts who becomes intimately familiar with the data and serves as an interface between the data and the users and products. This form of manual probing of a dataset is slow, expensive, and highly subjective. In fact, as data volume grows dramatically, this type of manual data analysis is becoming completely impractical in many domains. Hence, the need to scale up human analysis capabilities to handling the large number of bytes that we can collect is both economical and scientific (Fayyad, Piatetsky-Shapiro & Smyth, 1996).

Currently, an important new trend in information technologies is to identify and extract knowledge from data collected in information systems. As this knowledge is captured, it can be key to gaining a competitive advantage in an industry. This then creates an important need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data. These theories and tools are the subject of the emerging field of knowledge discovery in databases (KDD) (Fayyad, Piatetsky-Shapiro & Smyth, 1996).

The KDD field is concerned with the development of methods and techniques for making sense of data. The basic problem addressed by the KDD process is one of mapping low-level data, which are typically too voluminous to understand and digest easily, into other forms that might be more compact (for example, a short report), more abstract (for example, a descriptive approximation or model of the process that generated the data), or more useful (for example, a predictive model for estimating the value of future cases) (Fayyad, Piatetsky-

Shapiro & Smyth, 1996). At the core of the process is the application of a specific data mining methods for pattern discovery and extraction. The value of data mining is to proactively seek out the trends within an industry and to provide this understanding to organizations that maintain substantial amounts of information.

Data mining is the process of extracting out valid and yet previously unknown information from large databases and using it to make critical business decisions. It brings together the wealth of knowledge and research in statistics and machine learning for the task of discovering new snippets of knowledge in very large databases. This emerging technology can be put as one of the evolutionary steps in digital information processing.

Data mining, if done right, can offer an organization a way to optimize its processing of its business data. In this day and age, new data mining companies are springing up to the challenge of providing this service. Though data mining is improving the interaction between a business organization and its customers, there are many data mining companies that are trying to vertically integrate to offer the best services to broad markets (Piatetsky-Shapiro, 1999). This is done by focusing on a particular industry and trying to understand the types of information collected by companies in that sector.

This chapter discusses briefly the historical background of data mining and the processes associated with knowledge discovery in databases (KDD). Some of the types of data mining technologies are also described to a certain extent. Finally, applications of data mining in solving real world problems, with more emphasis to the financial sectors are reviewed.

3.2. A Brief History of Data Mining

Data mining and knowledge discovery have been attracting a significant amount of research, industry, and media attention of late (Fayyad, Piatetsky-Shapiro & Smyth, 1996). Though data mining is the evolution of a field with a long history, the term itself was only introduced relatively recently, in the 1990s. Statisticians, data analysts, and Management Information Systems (MIS) communities have mostly used it. It has also gained popularity in the database field. According to Piatetsky-Shapiro, cited in (Fayyad, Piatetsky-Shapiro & Smyth, 1996), the phrase 'Knowledge Discovery in Databases' (KDD) was first coined at the first KDD workshop in 1989 to emphasize that knowledge is the end product of a data-driven discovery. It has been popularized in the AI and machine learning fields (Fayyad, Piatetsky-Shapiro & Smyth, 1996).

There is some confusion about the terms Data Mining and KDD. Often these two terms are used interchangeably. Many authors agree in that KDD refers to the overall process of discovering useful knowledge from data and data mining refers to a particular step in this process. According to Fayyad, Piatetsky-Shapiro & Smyth (1996), KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data whereas Data mining is the application of specific algorithms for extracting patterns from data.

Data mining is one step at the core of the knowledge discovery process, dealing with the extraction of patterns and relationships from large amount of data. It is a promising interdisciplinary area of research shared by several fields such as database systems, machine learning, intelligent information systems, statistics, data warehousing, and knowledge

acquisition in expert systems (Lin & Cercone, 1997). It currently relies heavily on known techniques from statistics, AI, and machine learning, the three roots of data mining. These techniques are used together to study data and find previously hidden trends or patterns within.

The longest of these three is classical statistics. Knowledge discovery from data is fundamentally a statistical endeavor. Statistics provides a language and framework for quantifying the uncertainty that results when one tries to infer general patterns from a particular sample of an overall population. The concept of classical statistics such as regression analysis, standard distribution, standard deviation, standard variance, cluster analysis, and confidence intervals, all of which are used primarily to study data and data relationships, are the foundation of most technologies on which data mining is founded upon. Even in today's data mining tools and knowledge discovery techniques, classical statistical analysis still plays a significant role. The last few years have seen an increasing use of techniques in data mining that draw upon or are based on statistics; namely, in feature selection, data dependency involving two variables for constructing data dependency networks, classification of objects based on descriptions, discretization of continuous values, data summarization, predicting missing values, etc. (Lin & Cercone, 1997). The motivation behind this trend can be explained by the fact that statistical techniques for data analysis are well developed and in some cases, we do not have any other means to apply.

The second longest family line for data mining is artificial intelligence, AI. This discipline, which is built upon heuristics as opposed to statistics, attempts to apply human thought-like processing to statistical problems. AI is manly used to create new ways in addressing and solving very complex and math-driven problems. Researches in many aspects of intelligence

aim to understand human intelligence at all levels, including reasoning, perception, language, development, learning, and social levels; and to build useful artifacts based on intelligence (Rick & Knight, 1991).

The third family line of data mining is machine learning, which is the study of computational methods for improving performance by mechanizing the acquisition of knowledge from experience (Langley & Simon, 1995). Machine learning attempts to let computer programs learn about the data they study, such that programs make different decisions based on the characteristics of the studied data, using statistics for fundamental concepts, and adding more advanced AI heuristics and algorithms to achieve its goals. In the early work on machine learning, a number of theoretical and foundational issues of interest to data mining (for example, learning from examples, formation of concepts from instances, discovering regular patterns, noisy, and incomplete data, and uncertainty management, etc.) have been investigated (Lin & Cercone, 1997). As such, data mining, in many ways, is fundamentally the adaptation of machine learning techniques to business applications.

3.2.1. Generations of Data Mining Technology

Research and development from tools to solutions on data mining gave rise to three generations (Piatetsky-Shapiro, 1999). The first generation of what is now called data-mining systems appeared in the 1980s and consisted primarily of research-driven tools focused on single tasks. These tasks included building a classifier using a decision-tree or a neural network tool, finding clusters in data, or data visualization (Piatetsky-Shapiro, 1999). These tools addressed a generic data-analysis problem, and their intended user had to be technically sophisticated to understand and interpret the results. Furthermore, using more than one of

these tools on the same dataset was very complicated and required significant data and metadata transformations, which was not an easy task to achieve even for expert users.

The second-generation of data-mining systems called suites, which were developed around 1995, were driven largely by the realization that the knowledge discovery process requires multiple types of data analysis and most of the effort is spent in data cleaning and preprocessing. Some of the suites in this generation are SPSS's Clementine, Silicon Graphics' Mineset, and IBM's Intelligent Miner. They let the user perform several discovery tasks, such as classification, clustering, and visualization, and support data transformation and visualization.

Even though they empower data analysts, the second-generation systems were not without problems. Business users cannot use these systems directly because they require significant knowledge of statistical theory to be used properly (Piatetsky-Shapiro, 1999). As a result, the third generation systems came about resulting in vertical data-mining applications and solutions in the 1990s. These tools were primarily oriented towards solving specific business problems by sifting through piles of information stored in large databases to discover hidden patterns. The end results were pushed to front-end applications such as a decision support system. The interfaces were oriented to the business user and hid all the data mining complexity. Examples of such systems include HNC software's Falcon for credit card fraud detection, IBM's Advanced Scout for basketball game analysis, and NASD Regulation's Advanced-Detection system (Piatetsky-Shapiro, 1999).

3.3. Data Mining and Data Warehousing

Frequently, the data to be mined is first extracted from an enterprise data warehouse into a data-mining database. Data warehousing is a popular business trend of collecting and cleaning transactional data to make them available for online analysis and decision support (Fayyad, Piatetsky-Shapiro & Smyth, 1996). There is some real benefit if the data to be mined is already part of a data warehouse. The problems of cleansing data for a data warehouse and for data mining are very similar. If the data has already been cleansed for a data warehouse, then it most likely will not need further cleaning in order to be mined. Furthermore, the use of data warehouse will have addressed many of the problems in data consolidation (Two Crows Corporation, 1999). But a data warehouse is not a requirement for data mining. Setting up a large data warehouse that consolidates data from multiple sources, resolves data integrity problems, and loads the data into a query database can be an enormous task, sometimes taking years and costing millions of dollars. Alternatively, the data miner could mine data from one or more operational or transactional databases by simply extracting it into a data-mining database.

3.4. Data Mining and Online Analytical Processing (OLAP)

One of the most common questions from data processing professionals is about the difference between Data Mining and On-line Analytical Processing (OLAP). OLAP is a popular approach for analysis of data warehouses. It is part of the spectrum of decision support tools. OLAP tools focus on providing multidimensional data analysis, beyond the traditional query

and report tools, which describe what is in a database (Fayyad, Piatetsky-Shapiro & Smyth, 1996). OLAP is used to answer why certain things are true and verifies a hypothesis about a relationship with a series of queries against the data. For example, an analyst might want to determine the factors that lead to loan defaults. He or she might initially hypothesize that people with low income are bad credit risks and analyze the database with OLAP to verify or disprove this assumption. If that hypothesis were not borne out by the data, the analyst might try another feature, like debt, occupation, etc. or a combination of them as the best predictor of bad credit risks. In other words, the OLAP analyst generates a series of hypothetical patterns and relationships and uses queries against the database to verify them or disprove them. OLAP analysis is a deductive process.

The problem with OLAP comes when the number of variables being analyzed is in the dozens or even hundreds. It becomes much more difficult and time-consuming to find a good hypothesis, and analyze the database with OLAP to verify or disprove it. But data mining can do it. Another important point where data mining differs from OLAP is that rather than verifying hypothetical patterns, it uses the data itself to uncover such patterns. It is essentially an inductive process. For example, if the analyst were to use a data mining tool to identify the risk factors for loan default, the tool might discover that people with high debt and low income were bad credit risks, but it might go further and also discover a pattern that the analyst did not think to try, such as that age is also a determinant of risk. However, OLAP is complementary in the early stages of knowledge discovery process because it can help the data miner explore the data by, for instance, focusing attention on important variables, identifying exceptions, or finding interactions (Two Crows Corporation, 1999). This is important because the better the miner understand the data, the more effective the knowledge discovery process will be.

3.5. The Knowledge Discovery Process

For many people data mining seems just applying software but it is more than this. In fact, it is a process that involves a series of steps to preprocess the data prior to mining and post processing steps to evaluate and interpret the modeling results. The overall process of building and implementing a data mining solution is referred to as KDD.

Starting with the definition of the business problem, the KDD process is an iterative process requiring quite an important input from the user (Levin & Zahavi, 1999). The following summarizes the steps of the knowledge discovery process.

As in many problem-solving researches, the data-mining project should start with a clear definition of the business problem involved and the objective function. This may direct the KDD process and the data mining modeling involved.

Once the business problem and the objective of the project are clearly defined, the next step is to select the target dataset for analysis. This requires figuring out what data are needed, which data are most important and integrating the information. One needs to extract the target data to analyze in a way that is consistent with the business problem involved and the objective of the project.

Data Preparation and preprocessing of the selected dataset is the next step, which is often the most time consuming task of the KDD process, especially if data is drawn directly from the company's operational database rather than from data warehouse (Levin & Zahavi, 1999).

This step involves various data processing tasks such as overlaying of data from other sources, consolidating and amalgamating records, summarizing fields, checking for data integrity, detecting irregularities and illegal fields, filling in for missing values, trimming outliers, cleansing noise, etc. For data mining purposes, one also needs to understand the data, identify key predictors, trace non-linear relationships between data elements, point out important interactions, etc. Though tedious and somewhat boring, data preparation and preprocessing is definitely a critical step in the KDD process with significant impact on the quality of the modeling results.

The data-mining step then follows. It is only at this point that one invokes data mining models and tools to interrogate the data and convert it into a knowledge for decision making. This model-building step involves selecting data mining tools, transforming the data if the tool requires it, generating samples for training and testing the model, and finally using the tools to build and select a model. A brief review of the most common data mining technologies is provided in the next section.

Finally, the resulting model is evaluated and results are interpreted to make sure the model is any good, and convert the model results into useful knowledge for decision-making. In addition, as model that works today may not work tomorrow, it is necessary to monitor the behavior of the model to ensure it is meeting performance standards.

3.6. Data Mining Technologies

Data mining technologies may be categorized into three major application groups: Predictive modeling, descriptive modeling, and link analysis (Levin & Zahavi, 1999). In each of these

applications, data mining differs in the approach taken to solve problems. Each application is usually geared in solving a particular type of problem. That is, a specific algorithm will be favored over others depending in what the problem posed by the data miner is. Each of these applications will be explained in brief in the following sections.

3.6.1. Predictive Modeling

Prediction is arguably the strongest goal of data mining. In predictive modeling one identifies patterns found in the data to predict future values. Predictive modeling consists of several types of models: Classification models, regression models, and AI models.

Classification and regression are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends (Han & Kamber, 2001). Classification methods create classes by examining already classified cases and inductively finding the pattern (or rule) typical to each class. Data mining uses machine – learning methods using decision trees to classify objects based on a dependent variable.

Regression models are the leading predictive models (Levin & Zahavi, 1999). The most common regression models are linear regression models, for modeling continuous response, and logistic regression models, for modeling discrete choice response. The difference between classification and regression is the type of output that is predicted. Where as classification predicts class membership, regression models continuous valued functions. For example, a classification model may be built to categorize bank loan applications as either safe or risky, while a regression model may be built to predict the expenditure of potential customers on a certain product given their income and occupation (Han & Kamber, 2001). Regression is used

over classification where the predicted output can take on many possible values. Moreover, a regression problem can be turned into a classification problem by establishing categories in which certain set of values classify to every category (Brand & Gerritsen, 1998).

Another types of predictive model are AI-based models. The leading models in this category are neural networks models. Neural network model, biologically inspired model, is made up of a collection of processing units called neurons, connected by means of branches, each characterized by a weight representing the strength of the connection between the neurons. The weights of the branches connecting the nodes are determined through a training process by repeatedly showing the neural network examples of past cases for which the actual output is known, thereby inducing the system to adjust the strength of the weights between neurons. One important feature of neural networks that have become of a particular interest to data mining is that they offer a means for efficiently modeling large and complex problems in which there are several hundreds of independent variables that have many interactions. Neural networks may be used to predict discrete continuous choice.

The variety of applications that use predictive modeling techniques is plentiful. Classification and regression analysis is used to predict customer behavior, to signal potentially fraudulent transactions, to predict store profitability, and to identify candidates for medical procedures to name a few of the applications in which this kind of algorithm can be used (Brand & Gerritsen, 1998).

3.6.2. Descriptive Modeling

Description involves using some variables or fields in the database and focuses on finding human-interpretable patterns describing the data (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). While predictive models are mainly supervised learning models, descriptive models belong to the realm of unsupervised learning models. They interrogate the database to identify patterns and relationships in the data. One example of such models is clustering (or segmentation) algorithm.

Clustering is a common descriptive task where one seeks to identify a finite set of categories or clusters to describe the data. It involves partitioning data according to natural classes present in it, assigning data points that are "more similar" to the same "cluster". In clustering, no data are tagged before being fed to a function. Instead, the input to a clustering function is a collection of untagged records. The goal then of clustering, through a pre-established criterion, will be to sift through the data to produce a segmentation of the input records. Different clustering functions will hence yield different sets of sorted data. It is up to the miner to determine what meaning, if any, to attach to the resulting clusters (Berry & Linoff, 2000). Perhaps, the most common of all automatic clustering algorithms is K-means algorithm, which assigns observations to one of K classes to minimize the within-size-cluster-sums-of-squares (Levin & Zahavi, 1999). Another class of models are the self-organizing neural networks models. Examples of clustering applications in the context of knowledge discovery include discovering homogeneous subpopulations for customers in marketing databases. Marketing managers like the technique for identifying customer populations they may want to target, such as with a special promotion.

Visualization method is another powerful form of descriptive data mining (Berry & Linoff, 2000). It is a means for presenting data, both at the input and the output stages. Visualization techniques may help to discover relationships between features at the input stages, and explain the data mining results present them to the decision makers at the out put stage.

3.6.3. Link Analysis

Link analysis is concerned with finding rules between data elements. Two most common rules are affinity grouping (or association rule) and sequencing rules (Levin & Zahavi, 1999). The task of these rules is to determine which things go together.

In association tasks, a problem is solved by issuing a query against a database and finding out the affinities between existing variables. For example, these affinities can be expressed by stating "85% of customers who purchased items A, B, C also ended up purchasing D & E." As such, the goal of using associations is to find common relationship amongst the items, or variables, existing in a collection of records. Selling as many products to as many customers increases the potential revenue of a business and is an important component in understanding consumer preferences. One of the ways in which this goal may be realized is by understanding what products or services customers tend to be purchased at the same time, or later as follow-up purchases. As such, determining consumer purchasing behavioral trends is a very common application of data mining, and association and sequencing techniques can perform this kind of analysis (Berry & Linoff, 2000).

Sequential patterns are another form of a commonly used application in data mining. They are association rules with time dimension (Levin & Zahavi, 1999). A sequential pattern function

tries to detect frequently occurring patterns in the data. Sequential pattern mining functions can be used to detect the set of customers associated with some frequent buying pattern. As an example, if customers increase their purchase of a particular product A by 20% and also increase their consumption of another product B by 25%, then product C is found to also increase by 10%. Whether product C's increase in purchases was a direct result of the increase in sales of products A & B collectively is uncertain; what is certain is that this relationship holds and it is up to the data miner to examine this trend further by issuing further queries to determine whether or not a particular relationship is a casual one. In this manner, sequential patterns are used in detecting frequency of events occurring and their relationship to other events. Thus, association and sequencing tools analyze data to discover rules that identify patterns of behavior. Using an association or sequencing algorithm is frequently called market basket analysis (Berry & Linoff, 2000). Business managers or analysts can use such market basket analysis to plan discounting products, product placement, and timing of promotional sales. The industries in which association and sequencing data mining tools are mostly used include the retail, health care, financial, and insurance industries.

3.7. Application of Data mining Technology

In the last few years, KDD and Data Mining tools have been used mainly in experimental and research environments. Today, sophisticated tools, which aim at the mainstream business user, are rapidly emerging. Business questions that were previously impossible, impractical or unprofitable to address due to lack of processing capabilities can now be answered using data mining solutions. New tools hit the market nearly every month. The Meta Group estimates that the market size for Data Mining market will grow from $50 million in 1996 to 800 million by 2000 (Goebel & Gruenwald, 1999).

Good data mining application areas require knowledge-based decisions; have accessible, sufficient and relevant data; have a changing environment; have sub-optimal current methods; will not be outdated by imminent new technology; and provide a high payoff for the correct decisions (Piatetsky-Shapiro, 1999). Data mining offers value across a broad spectrum of industries that fit these requirements-including banking and credit, Customer relationship management, healthcare and human resources, insurance, marketing, retail, telecommunications, and manufacturing. A few of these applications are given below.

Telecommunications and credit card companies are two of the leaders in applying data mining to detect fraudulent use of their services. For example, HNC Falcon and Nestor PRISM systems are used for monitoring credit card fraud, watching over millions of accounts (Fayyad, Piatetsky-Shapiro & Smyth, 1996). In the telecommunications area, the Telecommunication Alarm-Sequence analyzer (TASA) was built in cooperation with a manufacturer of telecommunications equipment and three telephone networks. The system uses a novel framework for locating frequently occurring alarm episodes from the alarm stream and presenting them as rules.

Companies active in the financial market also use data mining to determine market and industry characteristics as well as to predict individual company and stock performance (Two Crows Corporation, 1999). In marketing, for instance, the primary application is database-marketing systems, which analyze customer databases to identify different customer groups and forecast their behavior. Estimates showed that over half of all retailers in U.S. are using or planning to use database marketing, and those who do use it have good results; for example, American Express reports a 10 to 15 percent increase in credit card use (Fayyad, Piatetsky-

Shapiro & Smyth, 1996). On investment, the LBS Capital Management deployed a system that uses expert systems, neural nets, and genetic algorithms to manage portfolios totaling $600 million; since it start in 1993, the system has outperformed the broad stock market (Fayyad, Piatetsky-Shapiro & Smyth, 1996).

Moreover, several studies had been carried out in applying neural network in the area of finance. In this field, neural networks have been seen to achieve certain amount of success. Some of these are summarized below.

One area of application was in company's bankruptcy prediction. In a study conducted by Odom and Sharda (1990), neural network has performed very well in predicting which companies are to enter into bankruptcy. The researchers have created a neural network bankruptcy forecasting system, which was applied to a sample of 129 firms from Moody's industrial manuals. Consequently, the neural network achieved a correct classification rate of 87 percent.

Another area where neural networks were found to be successful is bond-rating system. According to Dutta and Shekkar (1988), a neural network model developed using bond issues of 40 companies managed to achieve an 88 percent success rate. Similar research, conducted by Surkan and Singleton (1990), showed that bond rating of the 18 Bell Telephone companies diversified by AT&T for the years 1982-1988 has a prediction accuracy of 88 percent using neural network.

In a commodity market study, a neural network was modelled to recognize and direct a buy/sell pattern for the live cattle commodities futures market (Collard, 1993). The result is overwhelming with a 52% profit in only 178 trading days.

The aforementioned examples are certainly not exhaustive as data mining techniques, in particular neural networks, are being used in various investment related field.

3.7.1. Using Data Mining in the Insurance Industry

As in other sector of economy, the insurance industry has experienced many changes in information technology over the years. Advances in hardware, software, and networks have offered benefits, such as reduced cost and time of data processing and increased potential for profits, as well as new challenges particularly in the areas of increased competition. Insurance industry can make better use of modern data mining technologies to develop more accurate and better performing models that are generated in less time than with previously known techniques. By generating better, extensively tested models, insurance firms can more accurately address issues such as moral hazard in underwriting and the adverse selection in marketing. The researcher has tried to find documents on researches, which have been made so far on the application of data mining technology in support of various activities within the industry. The following are which the researcher came across from the Internet.

One notable research, which has been made on the data mining application in the insurance industry, is that of the IBM research group which results in the development of an application named UPA (Underwriting profitability Analysis) (Apte et al., 1999). The application embodies a new approach to mining Property & Casualty (P&C) insurance policy and claims

In a commodity market study, a neural network was modelled to recognize and direct a buy/sell pattern for the live cattle commodities futures market (Collard, 1993). The result is overwhelming with a 52% profit in only 178 trading days.

The aforementioned examples are certainly not exhaustive as data mining techniques, in particular neural networks, are being used in various investment related field.

3.7.1. Using Data Mining in the Insurance Industry

As in other sector of economy, the insurance industry has experienced many changes in information technology over the years. Advances in hardware, software, and networks have offered benefits, such as reduced cost and time of data processing and increased potential for profits, as well as new challenges particularly in the areas of increased competition. Insurance industry can make better use of modern data mining technologies to develop more accurate and better performing models that are generated in less time than with previously known techniques. By generating better, extensively tested models, insurance firms can more accurately address issues such as moral hazard in underwriting and the adverse selection in marketing. The researcher has tried to find documents on researches, which have been made so far on the application of data mining technology in support of various activities within the industry. The following are which the researcher came across from the Internet.

One notable research, which has been made on the data mining application in the insurance industry, is that of the IBM research group which results in the development of an application named UPA (Underwriting profitability Analysis) (Apte et al., 1999). The application embodies a new approach to mining Property & Casualty (P&C) insurance policy and claims

data for the purpose of constructing predictive models for insurance risk. UPA is rule-based system that utilizes probabilistic approach to discover risk characterization rules by analyzing large and noisy insurance data sets. The benefit assessment of the results suggests that this methodology provides significant value to the P & C insurance risk management process (Apte et al., 1999).

Williams G. J. and Huang Z. (1996) have also undertaken a research on the application of KDD for insurance risk assessment. The research was conducted using real world dataset, and decision tree techniques were used to identify significant risk areas with in an insurance portfolio.

The following are some of the business practices, outlined by SAS institute (2001), that data mining can help insurance firms.

Establishing rates: - An important problem in actuarial science concerns rate setting or the pricing of each policy. The goal is to set rates that reflect the risk level of the policyholder by establishing the "break-even" rate (or premium) for the policy. The lower the risk, the lower the rate. Although many risk factors that affect rates are obvious, subtle and non-intuitive relationships can exist among variables that are difficult if not impossible to identify without applying more sophisticated analysis. Modern data mining models can more accurately predict risk, therefore insurance companies can set rates more accurately, which in turn results in lower costs and greater profits.

Acquiring new customers and retaining policyholders: - another important business problem that is related to ratemaking is the acquisition of new customers. Traditional approaches

involve attempts to increase the customer base by simply expanding the efforts of the sales department. In contrast to traditional sales approach, data mining strategies enable analysts to define the marketing focus. Analysts in the insurance industry can utilize advanced data mining techniques that combine segmentations to group the high lifetime-value customers and produce predictive models to identify those in this group who are likely to respond to marketing campaign. Using a data mining techniques called association analysis, insurance firms can more accurately select which policies and services to offer to which customers. With this techniques insurance companies can segment the customer database to create customer profiles, conduct rate and claim analyses on a single customer segment for a single product, perform sequential market basket analysis on customer segments.

Data base segmentation and more advanced modeling techniques enable analysts to more accurately choose whom to target for retention campaigns. Current policyholders that are likely to switch can be identified through predictive modeling. By including nonlinear terms and more interaction terms, neural network models can generate more accurate data on the probability of policyholders switching. Additionally, decision tree models may provide more accurate identification by dividing (segmenting) the policyholders into more homogeneous groups.

Developing new product lines: - Markets change over time, and so do the products sought by customers. To adjust to market changes, insurance companies need to know what types of new policies will be profitable. Depending upon the goal of the firm, the new products can be then prioritized based on expected profit, expected number of new customers, and/or the expected speed of acceptance. Data mining technology can be used to identify customer

groups, to model their behaviors, to calculate expected profits fro marketing campaigns, and then to compare potential strategies for action.

Detecting fraudulent claims: - obviously fraudulent claims are an ever-present problem for insurance firms, and techniques for identifying and mitigating fraud are critical for the long-term success of insurance firms. In searching for fraudulent claims, analysts look for unusual associations, anomalies, or outlying patterns in the data. Specific analytical techniques adopt at finding such subtleties are market basket analysis, cluster analysis, and predictive modeling. Quite often, successful fraud detection analysis such as those from data mining project can provide a very high return on investment.

Providing Reinsurance: - Data mining technology is commonly used for segmentation clarity. In the case of reinsurance, a group of paid claims would be used to model the expected claims experience of another group of policies. With more granular segmentation, analysts can expect higher levels of confidence in the model's outcome. The selection of policies for reinsurance can be based upon the model of experienced risk.

CHAPTER FOUR

METHODS

4.1. Introduction

People have always sought better, easier and faster ways to do things, either because of a pressing need or purely for the satisfaction of discovery. Recall that many technological developments have been motivated either by military needs or the drive for automation and increased productivity in business. Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access such as more recently generating technologies that allow users to navigate through their data in real time. Artificial Neural Networks (ANNs) are one of the key technologies used in data mining.

An Artificial neural network is a system loosely modeled on the human brain. ANNs, like people, learn by example. A trained neural network can be thought of as an expert in the category of information it has been given to analyze. Neural networks are best at identifying patterns or trends in data. They have broad applicability to the real world problems.

This chapter elaborates more on the neural network and its structure, with more emphasis to the multilayer Back propagation neural networks. Backpropagation Neural networks are the most commonly used types of neural networks, which are also used in this study. It also reviews a bit about decision trees and its importance in feature (or attribute) selection. In this

study, Decision tree was used as a feature selection method for the input to the neural network.

4.2. Overview of Neural Networks

The human brain, with its 15 billion neurons, performs fantastically at such tasks as vision, speech, information retrieval, and complex spatial and temporal pattern recognition in the presence of noisy and distorted data (Jones, 1996). These are some of the things researchers are trying to train machines to do. Neural networks are a natural extension of exploring the limits of computing, in terms of methodology and theory. Artificial neural networks may be represented by different data structures, but they are each designed to make use of some of the organizational principles, felt to be, used by the brain (Cheng & McClain, 1998).

Neural networks are an attempt to simulate within specialized hardware or sophisticated software, the multiple layers of simple processing elements called neurons (Stergiou, 1996). Each neuron is linked to certain of its neighbors with varying coefficients of connectivity that represent the strengths of these connections. Learning is accomplished by adjusting these strengths to cause the overall network to output appropriate results.

What really sets ANNs apart and cause them to receive so much attention in both academia and business is that a properly designed neural network includes a self-training mechanism which allows it to analyze almost incomprehensible amounts of data, test for and discover relationships or connections among the data, and use these discoveries with its programmed formulas to make predictions about future trends or events. Moreover, the artificial neural network is then able to compare its predicted results to actual results and actually learn

overtime, much like a person does, by adjusting its formulas to reduce future discrepancies between predicted and actual results.

The ability of artificial neural networks to simulate learning in this manner and to search for and discover relationships and correlations within data makes them very adaptable for use in many problem domains. Moreover, their ability both to factor in an astonishing amount and variety of data and to learn to forecast more accurately overtime as the model gains experience is unique and invaluable for use in financial institutions, including insurance.

Often, neural networks are considered to be "black boxes" because of their non- linear behavior and are usually more complicated than other techniques. Training a neural network is a further challenge requiring setting numerous parameters. Furthermore, the output of a neural network is not as easily understood by the user as the output seen by a decision tree tool. Despite this, neural networks are proving their worth everyday in a wide variety of business applications (see chapter three for examples), and saving their users time and money in the process.

4.2.1. Neural Network Structure

Neural networks consist of a number of processing units, called neurons, analogous in some respects to the neurons found in the human brain. All of the processing of a neural network is carried out by this set of neurons (or units). Each neuron is a separate computation device, doing its own relatively simple job. A unit's job is simply to receive input from other units and, as a function of the inputs it receives, to compute an output value, which it sends to other

units. The system is inherently parallel in that many units can carry out their computations at the same time. Figure 4.1 shows a model of a processing unit.

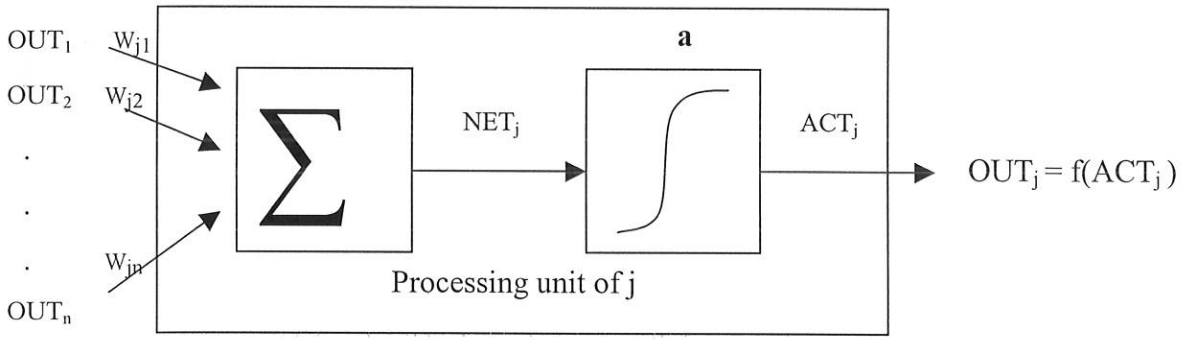

Figure 4.1: A processing unit

The units in a neural network are arranged in layers, usually classified as input, output, and hidden. Input unit receive inputs from sources external to the system under study. The output units send signals out of the system. The hidden units are those whose only inputs and outputs are within the system. They are not visible outside the system.

A network has one input and one-output layers but may have any number of hidden layers or none at all. Multi-layer networks maybe formed by simply cascading a group of single layers: the output to one layer provides the input to the subsequent layer. A layer consists of a set of weights and the subsequent units that sum the signals they carry (Berry & Lonoff, 2000). Figure 4.2 illustrates the basic structure of a one-hidden-layer neural network with l inputs, m-hidden units, and n output units.

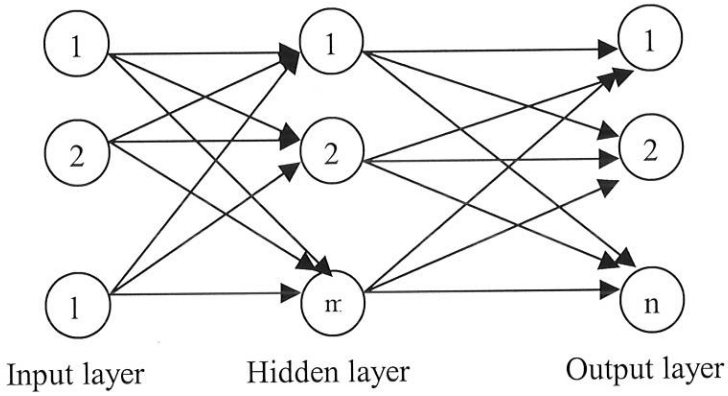

Figure 4.2: A one-hidden-layer neural network

The number of processing units in a layer will depend on the problem, which has to be solved. The choice of the number of input- and output units for a specific problem is quite straightforward. The choice of the number of units in the hidden layer(s) is, however, more difficult. There are only rules of thumb to help a researcher with this choice.

For most applications, only one hidden layer is needed. If there is no good reason to have more than one hidden layer, then you should stick to one. The training time of a network increases rapidly with the number of layers (Bigus, 1996).

4.2.2. State of activation

The level of activation of the units taken collectively represents the state of the system. It is convenient to look on the processing carried out by the system as the evolution of the system state. Activation of any particular unit induces or hinders the activation of units to which it is connected according to whether the interconnection is excitatory or inhibitory. According to Khanna, cited in (Kramer, 1994), the notion of activation per se may be viewed in two

different ways. First, the activation value of a unit indicates its degree of confidence that its associated feature is present or absent, as opposed to merely providing a yes/no answer regarding the presence or absence of a feature. Alternatively, the activation value of a unit might suggest the quantity of a feature that is present. The activation value is passed through a function to produce an output value.

4.2.3. The Output function: ($OUT_j = f(ACT_j)$)

Units interact by transmitting signals to other units. The strength of their signals, and therefore the degree to which they affect other units, is determined by their degree of activation. The output value can be seen as passing through a set of unidirectional connections to other units in the system. Associated with each unit, there is an output function *f*, which maps the current state of activation to an output signal. Sometimes, the output level is exactly equal to the activation level of the unit. In other cases, the output function is some sort of threshold function so that a unit has no effect on another unit unless its activation exceeds a certain value. Alternatively, the output function might be a stochastic function in which the output of the unit depends in a probabilistic fashion on its activation values.

4.2.4. Neural Network Topology

One important aspect used in classification of neural networks is their topology. The arrangement of neural processing units and their interconnection can have a profound impact on the processing capabilities of the networks (Bigus, 1996). Units are connected to one another. It is this pattern of connectivity that constitutes what the system knows and that determines how it will respond to any arbitrary input. The total pattern of connectivity can be

specified by defining the weights for each of the connections in the system. The *weight* or strength $\mathbf{w}_{ji}$ of a connection determines the amount of effect that unit i has on unit j.

Depending on the pattern of connectivity, two types of networks can be distinguished: *feedforward networks* and *recurrent networks.* Feedforward networks are 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 (Bigus, 1996). In this type of network, the data flows through the network in one direction, and the answer is based solely on the current set of inputs. *Feedforward networks* have no feedback connections, that is, they have no connections through weights extending from the outputs of a layer to the inputs of the same or previous layers. Feedforward networks have no memory; their output is solely determined by the current inputs and the values of the weights. *Recurrent networks* do contain feedback connections. In some configurations, according to Wasserman cited in (Kramer, 1994), recurrent networks recirculate previous outputs back to inputs; hence, their output is determined both by their current input and their previous outputs.

4.2.5. Rule of propagation $\left(NET_j = \sum w_{ji} OUT_i\right)$

All inputs to a unit are multiplied by their associated weights and summed to get the net input to that unit. The net input to a unit, along with its current activation value determine its new activation value.

4.2.6. Activation function $(ACT_j = a(NET_j))$

The activation function ***a*** combines the inputs impinging on a particular unit with the current state of the unit to produce a new state of activation. Whenever the activation value is assumed to take on continuous values, it is common to assume that *a* is a kind of sigmoid (i.e., S-shaped) function. In that case, an individual unit can saturate and reach a minimum or maximum activation value. With a logistic activation function, the limits of the output neuron are 0 and 1. With a hyperbolic tangent function, the limits are -1 and 1. According to Klimasauskas, cited in (Kramer, 1994), if the problem involves learning about 'average' behavior, logistic activation functions work best, but if the problem involves learning about 'deviations' from the average, hyperbolic tangent works best.

4.2.7. Training Neural Networks

In order to be immediately useful, a neural network must be trained before actually being used. A network is trained so that application of a set of inputs produces the desired (or at least consistent) set of outputs. Each such input (or output) set is referred to as a vector. Training is accomplished by sequentially applying input vectors, while adjusting network weights according to a predetermined procedure. Training algorithms can be categorized as *supervised* and *unsupervised* training.

Supervised training requires the pairing of each input vector with a target vector representing the desired output; together these are called a training pair. Usually a network is trained over a number of such training pairs. An input vector is applied; the output of the network is

calculated and compared to the corresponding target vector, and the difference (error) is fed back through the network and weights are changed according to an algorithm that tends to minimize the error. The vectors of the training set are applied sequentially, and errors are calculated and weights adjusted for each vector, until the error for the entire training set is at an acceptably low level.

In *unsupervised training* or clustering, the training set consists solely of input vectors. The training algorithm modifies network weights to produce output vectors that are consistent; that is, both the application of one of the training vectors or the application of a vector that is sufficiently similar to it will produce the same pattern of outputs. The training process, therefore, extracts the statistical properties of the training set and groups similar vectors into classes.

4.2.7.1. The Back-propagation Learning Algorithm

Back-propagation, the most successful of the current neural network algorithms, provides a systematic means for (supervised) training of multi-layer feed forward networks. A back-propagation network starts out with a random set of weights. The network adjusts its weights each time it sees an input-output pair. Each pair requires two stages: a forward pass and a backward pass. The *forward pass* involves presenting a sample input to the network and letting the activation of the units flow until they reach the output layer. The logistic function is normally used as activation function. There are, however, many functions that might be used; the back-propagation algorithm requires only that the function be everywhere differentiable. The logistic function satisfies this requirement.

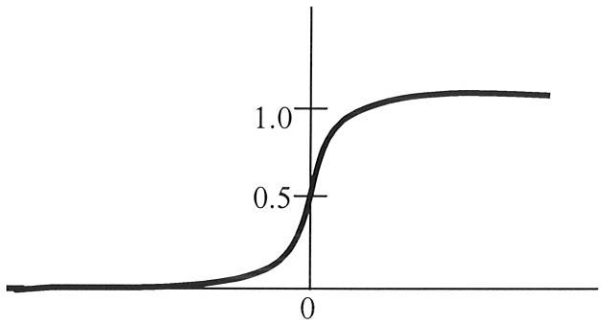

Figure 4.3: The sigmoid activation function of the back-propagation unit.

In the following, a logistic activation function is assumed. For simplicity, it is also assumed that the output level is equal to the activation level of the unit.

During the *backward pass*, the network's actual output vector (from the forward pass) is compared with the target output vector and error estimates are computed for the output units. The weights of the connections between the (last) hidden layer and the output layer o can be adjusted in order to reduce those errors. This adjustment is accomplished using the error signal (i.e., the target minus the actual output) multiplied by the derivative of the activation function, which is equal to $OUT_j(1-OUT_j)$ for a logistic activation function:

$$\delta_{jo} = OUT_{jo}(1-OUT_{jo})(TARGET_{jo} - OUT_{jo}) \quad \text{.....(1)}$$

Thereafter, the delta value for unit j in the output layer o, δ_{jo}, is multiplied by the output value from the source unit i in (hidden) layer q for the weight in question, OUT_{iq}. This product is in turn multiplied by the learning rate coefficient η and the result is added to the weight from unit i in layer q to output unit j:

$$\Delta w_{jio} = \eta \delta_{jo} OUT_{iq} \quad \text{.....(2)}$$

The learning-rate coefficient determines the average size of the weight changes (the step size).

Since hidden layers have no target vector, the training process described above cannot be used for adjusting the weights between subsequent hidden layers (if more than one hidden layer exists), and for adjusting the weights between the input layer and the (first) hidden layer. Instead, back-propagation trains the hidden layers by propagating the output error back through the network layer-by-layer, adjusting weights at each layer. These weights now operate in reverse, passing the delta value from the output layer back to the hidden layer(s). Each of these weights is multiplied by the delta value of the unit to which it connects in the subsequent layer. The delta value needed for unit j in layer p is produced by summing all such products for the subsequent layer q and multiplying by the derivative of the activation function:

$$\delta_{jp} = OUT_{jp}(1 - OUT_{jp})(\sum_i \delta_{iq} w_{ijq}) \quad \text{.....(3)}$$

For each unit in a given hidden layer, the deltas must be calculated, and all weights associated with that layer must be adjusted. This is repeated, moving back towards the input layer by layer, until all weights are adjusted.

Back-propagation networks are not without real problems, however, with the most serious being the slow speed of training (Rich & Knight, 1991). Even simple tasks require extensive training periods. Also, simple back-propagation does not scale up very well. The number of training examples required is super-linear in the size of the network.

4.3. Decision Trees

Data Mining uses machine-learning methods using decision trees to classify objects based on the dependent variable. There are two main types of decision trees (Two Craws Corporation, 1999). Decision trees, which are used to predict categorical variables, are called classification trees because they place instances in categories or classes. Decision trees used to predict continuous variables are called regression trees. Classification trees label records and assign them to the proper class. Classification trees can also provide the confidence that the classification is correct. In this case, the classification tree reports the class probability, which is the confidence that a record is in a given class. Regression trees, on the other hand, estimate the value of a target variable that takes on numeric values.

When a tree model is applied to data, each record flows through the tree along a path determined by a series of tests until the record reaches a leaf or terminal node of the tree. There it is given a class label based on the class of the records that reached that node in the training set or, in the case of regression trees, assigned a value based on the mean (or some other mathematical function) of the values that reached that leaf node in the training set. Figure 4.4 shows a simple classification decision tree, where the data has two possible classes: high risk and low risk.

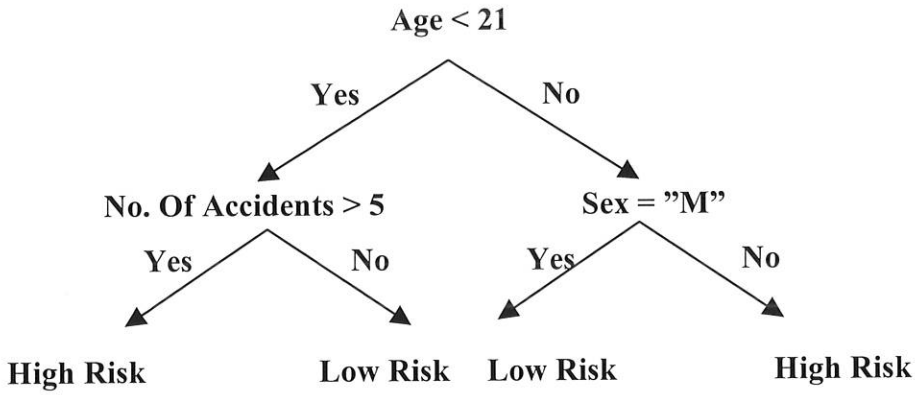

Figure 4.4: A simple Decision tree

Decision tree models are commonly used in data mining to examine the data and induce the tree and its rules that will be used to make prediction (Two Crows Corporation, 1999). Various decision tree algorithms such as CHAID (Chi-squared Automatic Interaction Detection), C4.5/5.0, CART (Classification and Regression Trees) and any with less familiar acronyms, produce trees that differ from one another in the number of splits allowed at each level of the tree, how those splits are chosen when the tree is built, and how the tree growth is limited to prevent over-fitting (Berry & Linoff, 2000). Today's data mining software tools allow the user to choose among several splitting criteria and pruning rules, and to control parameters such as minimum node size and maximum tree depth allowing one to approximate any of these algorithms.

4.3.1. Building Decision Trees

Decision trees are built through an iterative process of splitting the data up into partitions-and then splitting it up some more. This process is called Recursive partitioning (Berry & Linoff, 2000). The process starts with a training set consisting of pre-classified records. The goal is to build a tree that distinguishes among the classes. That is, the tree can be used to assign a class

to the target field of a new record based on the values of the other fields or independent variables.

Basically, all decision trees share the same structure. Starting from a root node (the whole population), tree classifiers employ a systematic approach to grow a tree in to branches and leaves. The first task is to decide which of the independent fields makes the best splitter. The best split is defined as one that does the best job of separating the records into groups where a single class predominates (Berry & Linoff, 2000). The measure used to evaluate a potential splitter is the reduction in diversity (or entropy). Most algorithms use a selection mechanism driven by the idea of maximizing the information gain (or, equivalently, minimizing the system entropy) (Costa, 2000). Lets suppose that we have a training set S at a particular node and, among others, an attribute A with values $\{a_1, a_2...a_n\}$. We can divide S into disjoint subsets $\{s_1, s_2... s_n\}$ where each s_i contains objects from S that have value a_i in attribute A. Suppose further that there are p positive examples and n negative ones in the node under analysis. The gain of using A as a branching attribute can be defined by:

$$gain(A) = I(p,n) - E(A) \quad \text{.....(4)}$$

Where:

$$I(p,n) = -\frac{p}{p+n}\ln\frac{p}{p+n} - \frac{n}{p+n}\ln\frac{n}{p+n} \quad \text{.....(5)}$$

Of course, this value is constant for a particular set S and defines the expected information content (E(A))of that set.

$$E(A) = \sum_{i=1}^{n}\frac{p_i + n_i}{p+n} I(p_i,n_i) \quad \text{.....(6)}$$

E (A), the expect information content, results from using A as the partitioning attribute.

To choose the best splitter at a node, the decision tree algorithm considers each input field. In essence, each field is sorted. Then, every possible split is tried. The diversity measure is calculated for the two partitions, and the best split is the one with the largest decreases in diversity (or equivalently with maximum information gain). This is repeated for all the fields. The winner is chosen as a splitter for that node.

Other two common diversity functions, which have simple formulas when there are only two outcomes, are:

i) $Min(P_1, P_2)$
ii) $2P_1(1-P_1)$, *Gini index*
where: P_1 *is the probability of class one.*

Tree building algorithms make their best split first, at the root node where there is a large population of records. Each subsequent split has a smaller and less representative population with which to work. To wards the end, idiosyncrasies of the training records at a particular node display patterns that are peculiar only to those records. These patterns are meaningless and harmful for prediction. By this depth in the tree, the model is over-fitting the training set. A well-known approach to this problem is pruning techniques. Pruning is the process of removing leaves and branches to improve the performance of the decision tree (Berry & Linoff, 2000). Pruning methods allow the initial decision tree to grow quite deep and then find ways to prune off the branches that fail to generalize.

4.3.2. Trees and Rules

Decision tree methods are often chosen for their ability to generate understandable rules. It is certainly true that for any particular classified record, it is easy to simply trace the path from

the root to the leaf where that record landed in order to generate the rule that led to the classification, and most decision tree tools have this capability. Many software products can output a tree as a list of rules in different format, including SQL code, pseudocode, or pseudo-English. However, since every split in a decision tree is a test on a single variable, decision trees can never discover rules that involve a relationship between variables. It is up to the miner to add derived variables to express relationships that are likely to be important.

4.3.3. Decision Trees and Attribute Selection

In order to effectively develop a model from the data, heuristics are needed to guide the machine learning process, which searches for an optimal model (Kononenko & Hong, 1997). An important issue in guiding the search is the quality of attributes.

Attribute may be relevant or irrelevant for the task at hand. When there are a large number of attributes, even some relevant attributes may be redundant in the presence of other attributes. Relevant attributes may contain useful information directly applicable to the given task by itself, or the information may be hidden among a subset of attributes.

An important problem is the selection of a reasonable subset of the available attributes so that the selected subset can adequately explain (or model) the target. This is especially true for neural network. Ideally, the neural network should be run on each possible set of variables to determine the perfect subset. The decision tree algorithm would help identify a handful of the most important variables (Berry & linoff, 2000). According to the authors, it is not uncommon for decision trees to be used for no other purpose than prioritizing the independent variables. The location and the frequency with which the variables appear in the tree can be of

use in gaining some insights into which variables appears to be important (Williams and Huang, 1996).

Attribute selection is an important phase in data mining researches. The importance of reducing the number of attributes from hundreds to within a few dozen, not only speed up the learning process, but also prevents most of the learning algorithms from getting fooled into generating an inferior model by the presence of many irrelevant or redundant attributes (Kononenko & Hong, 1997). This is mainly because most practical learning algorithms are necessarily heuristic in nature and they often are misled by the presence of many non-essential attributes.

CHAPTER FIVE

EXPERIMENTATION

In this chapter the researcher describes the source of data as well as the techniques that have been used in preprocessing and model building. Test results are then presented and discussed. In all of the work described, in terms of preprocessing and modeling, the approach used in this research can be applied to any task involving data mining.

5.1. Data Collection

The research was originally aimed to investigate the application of data mining in the Ethiopian Insurance Corporation (EIC). EIC, a government owned company, was primarily selected for it is the biggest insurance company in the country, with large number of customers, which implies huge database on both policy and claims data. But, data access was very difficult for the researcher, though a great effort is made. An alternative source of data then looked for, and Nyala Insurance SC. (NISCO) was found to be a good choice.

5.1.1. The Data

A precondition to any data mining is data itself. A good source of data for data mining purpose is identified to be the corporate data warehouse (Berry & Linoff, 2000). The reason for this is that the data is stored in common format with consistent definitions for fields and

their values. Unfortunately a corporate data warehouse is not available at NISCO. However, the company uses an automated system to store both policy and claims data at every service unit (or branch) located in different parts of the country. In addition, it uses manual format to collect information on the vehicle to be insured and the owner both at the time of underwriting and claims processing. These formats together with the database system provided the required information for the experimentation.

The original database is stored and maintained separately from the KDD process. Control and access to this original database is restricted due to the confidential nature of the material of interest. The employees of the company then perform the data selection, with input on data requirements from the researcher.

As mentioned in the methodology section of chapter one, for the purpose of data collection, four service units were chosen. In addition, from among the three types of cover that are available for motor insurance, comprehensive cover was chosen for this research. This is because comprehensive cover constitutes major proportion of the total motor policy existing in the company. Moreover, the company entertains relatively large number of motor claims from comprehensive cover.

The other point that had to be addressed during data collection was the time range considered. Nyala insurance started its operation in 1995. Due to few operational period of the company, the size of the customers in each of the sample service units is small. Hence policies, renewed or newly accepted, whose expiry date is in the 2001 accounting period were included in the dataset.

The initial source dataset was extracted from the motor vehicle insurance portfolio database of each service unit in the sample. Due to the little experience of the staffs in the Information Technology section of the NISCO in file Import/Export facilities of the system, coupled with limited access of the researcher to the system, the source data was obtained in printout form. Next, the real and symbolic data fields (attributes) within each of the policy and claims databases had to be combined into a single database that could be used for neural network training. In the course of doing this task, keying in the printed data back into a computer database consumed quite considerable amount of time. The number of records collected from the four service units is summarized in table 5.1 given below.

Service Unit	**Number of records collected**
Bole	332
Kera	466
Beklobet	293
Golla	241
Total	1332

Table 5.1: Distribution of collected data with respect to sample service units.

As discussed in the next sections, the maximum network size we had for the data was that with 71 inputs, 71 hidden and 3 output neurons. According to California Scientific Software (CSS), vendors of the BrainMaker software used in this study, the number of training facts for a network of this size should preferably be between 290 and 1450. And the collected data, i.e., 1332, was already within this range, and above the average number of facts recommended. The preprocessing of these data is described below.

5.2. Data Preprocessing

The database being used for the purpose of this research work has a number of limitations. The limitations include missing values in various fields and encoding problem. At the time of this work only a limited subset of all possible features (or attributes) were available in sufficient numbers to allow investigation. One of the aspects of modeling that is typical of real world tasks of this type is limitation in the data, whether in terms of noise or number (Edwards, P. J. et al., 1999). In the event when there is a comprehensive automated data collection process, perhaps an optimal model may be produced.

As mentioned in section 3.5 of chapter three, while data mining is a key stage in the knowledge discovery process, the other stages of the process often require considerable effort. One of these stages is data preprocessing. The purpose of the preprocessing stage is to cleanse the data as much as possible and to put it into a form that is suitable for use in later stages. Starting from the data extracted from the source database maintained by NISCO, a number of transformations had to be performed before a suitable working dataset was built.

5.2.1. Deciding on the right attributes

The policy and claims databases at NISCO consist of more than 120 attributes, some of which are collected from the manual format and others are defined for several purposes. The first task was to remove from the database those fields or attributes, which were irrelevant to the task at hand. This process was complicated by the fact that some obviously irrelevant attributes, such as BRANCH, where underwriting takes place, could contain surprising information. On the other hand, leaving irrelevant attributes in the dataset could lead to

abnormal results. The final list of input variables was compiled through an iterative process of consulting with experts on underwriting (mainly from NISCO) for ideas of what might be important and keeping the ones that proved to be important for this study. The initial set of attributes include:

- Branch
- Age Of Policyholder
- Sex
- License Grade
- Category
- Activity
- Class of business (Commercial or Private)
- Sub class of business (based on the purpose of use of vehicle)
- Number of years since policy inception
- Sum insured of vehicle
- Premium (averaged)
- Make of vehicle
- Model of vehicle
- Body type of vehicle
- Horsepower (or CC) of vehicle
- Seat Capacity
- Fuel type
- Age of vehicle
- Previous Claim Frequency
- Previous Claim Cost

- Claim Frequency (by the year 2001)
- Claim Cost (by the year 2001)
- Average claim to premium ratio
- Number of Years of no Accident
- Familiarity of garage
- Availability of spare parts
- Ease of repair
- Repair cost

All of these variables take on values that either falls into clearly defined categories (Branch, Sex, License Grade, Category, Activity, Business class, Sub Class, Make, Body Type, Fuel Type, Garage Familiarity, Spare Part availability, Ease of Repair, Repair Cost) or numeric values. Some of these variables were used as independent variables while others were used to derive other important variables.

5.2.2. Preparing Data for Analysis

The next step is preparing the collected data for analysis. This process involves summarization, data encoding, handling missing values, deriving new fields, and finally preparing the data into a form that is acceptable to the neural network (Berry & Linoff, 2000).

Data summarization is very important if there exists only few examples at the finest level of detail. In the collected data for this study, however, the distribution of facts was found to be reasonable, except that of the field MAKE with more than 49 possible categories. Regrouping of these categories on the basis of their manufacturers were suggested, but the experience of

surveyors at NISCO shows that the risk exposure of vehicles varies from one make to another as well.

The next stage was handling inconsistent data encoding. The fields that had data encoding problems were "MAKE", "MODEL" and "BODYTYPE." The difference in data encoding for these fields is attributed mainly to typing error. Correction is made on the value representation of these attributes by taking the standard naming used by manufacturers.

Another important point that is considered in preparing data for analysis was handling missing values. For several reasons, such as values unknown to the employee, giving less importance to the field, skipping while entering data into the database, and so on, most of the fields were suffering from missing values. The first step taken to handle this problem was to consult the manual formats prepared by the company to collect preliminary information on policyholders and the vehicle to be insured. Missing values that were skipped at the time of data entry were collected from the formats manually.

Apart from this, there are suggested ways for handling such missing values. One of these is that made by Two Crows Corporation (1999). For continuous variables, such as Age of vehicle, Average premium, vehicle value (or sum insured) in this study, it is suggested that missing values be replaced with the mean value for that field. Instead of taking the mean value of the entire field, however, records were grouped on the basis of some criteria, such as use of vehicle, make of vehicle, etc., used by the company in premium calculation and vehicle value estimation. Then the average of each group were considered in substituting the missing values. For some of the continuous variables it was difficult to use this method. The researcher as well as the domain expert did not find it logical to assign average value of fields,

like claim frequency and claim cost, to missing data for this may bring fallacy on the relationship between these two fields. For example, using average claim cost for missing value for a record with zero claim frequency is meaningless. Therefore the following approach was applied. In the case of policyholders whose policy year is above one, previous values were considered, but in the case of new customers, the whole record had to be deleted.

The other types of fields are the categorical fields. These fields can be grouped into ordinal, whose values can be meaningfully ordered, and nominal, whose values are unordered. According to the method suggested by Two Crows Corporation (1999) for handling missing data, missing values had to be replaced by the median for ordinal variables and the modal value for nominal values. Variables whose missing values were handled in the above way include those that represent the vehicles' make, body type, familiarity of garage, availability of spare parts, ease of repair, and repair cost.

Finally, for reasons of instances of many missing values, fields representing age of policyholder, license grade, model and Seat Capacity of vehicle had to be completely discarded. The number of records in the final working dataset, after removal of records during the preprocessing stage, is 1160.

The other important step in preprocessing is deriving other fields from the existing ones. Adding fields that represent the relationships in the data are likely to be important in increasing the chance of the knowledge discovery process yield useful result (Berry & Linoff, 2000). In consultation with the domain experts at NISCO, the following fields that are considered essential in determining the risk level of a policy were derived from the existing fields:

$$VehicleYear = Number\ of\ years\ since\ Policy\ inception$$

$$Average\ Claim\ Cost = \frac{Total\ Claim\ Cost}{Vehcle\ Year} \quad \text{.....(7)}$$

$$Average\ claim\ Frequency = \frac{Total\ Claim\ Frequency}{Vehicle\ Year} \quad \text{.....(8)}$$

$$Average\ Premium = \frac{Total\ Premium\ Collected}{Vehicle\ Year} \quad \text{.....(9)}$$

$$CTPR = \frac{Average\ Claim\ Cost}{Average\ Premium} \quad \text{.....(10)}$$

Other derived variables include Age of Vehicle, Total Frequency, Total Claim cost, and number of years of no accident. Table 5.2 shows final list of variables that have been used in this study.

The final task at this stage was preparing the data into a form that is acceptable to the neural network. Neural networks accept values in the range of 0 to 1 or –1 to 1. Fortunately, the *BrainMaker* software that was used in this study has the facility to automatically transform values into a form that can be understood by the neural network, i.e., in the range of 0 to 1. In this respect, the values in the numeric fields are scaled down to the range of 0 to 1 using the maximum and minimum values within the field. In the case of categorical fields, each unique value is considered as an input field with possible values of 0 or 1 depending on the presence (1) or absence (0) of the value in the field. The maximum possible size of a network using the independent attributes given in table 5.2 (25) is that with 71 inputs, 71 hidden and 3 output neurons.

No.	Attribute Name	Description	Attribute Type	Remark
1.	ID	Record Identifier		
2.	BRANCH	Branch	Categorical	
3.	CATEGORY	Category	Categorical	
4.	ACTIVITY	Activity	Categorical	
5.	BUSCLASS	Class of business (Commercial or Private)	Categorical: Private (P), Commercial (C)	
6.	SUBCLASS	Sub class of business (based on the purpose of use of vehicle)	Categorical: on the basis of their uses.	
7.	VEHYEAR	Number of years since policy inception	Continuous	
8.	VEHVALUE	Sum insured of vehicle	Continuous	
9.	AVGPRM	Premium (averaged)	Continuous	
10	MAKE	Make of vehicle	Categorical	
11	BODYTYPE	Body type of vehicle	Categorical	
12	POWER	Horsepower (or CC) of vehicle	Continuous	
13	FUELTYPE	Fuel type	Categorical	
14	AGEOFVEH	Age of vehicle	Continuous	Derived
15	PREVCFREQ	Previous Claim Frequency	Continuous	
16	PREVCC	Previous Claim Cost	Continuous	
17	CFREQ	Claim Frequency (by the year 2001)	Continuous	
18	CCOST	Claim Cost (by the year 2001)	Continuous	
19	TOTALCF	Total claim frequency	Continuous	Derived
20	TOTALCC	Total claim cost	Continuous	Derived
21	CTPR	Average claim to premium ratio per year	Continuous	Derived
22	NYNA	Number of Years of no Accident	Continuous	
23	GARAGE	Familiarity of garage	Categorical: Popular, Moderate, not Known	
24	SPARE	Availability of spare parts	Categorical: Dealer, Others, Both	
25	REPAIR	Ease of repair	Categorical: easy, Difficult	
26	RCOST	Repair cost	Categorical: Low, High	
27	RISKLEVEL	Level of risk	Categorical	Dependent

Table 5.2: Summary of attributes used

5.3. Defining the target Classes

A data mining training set is not complete until the data has pre-classified so that the data mining algorithms know what they are looking for (Berry & Linoff, 2000). Traditionally, underwriters classify renewal policies into acceptable and unacceptable policies based on the past claims experience and other contributing factors affecting the risk (Dockrill, et al., 2001). Since rejection of policies is very rare, and because of lack of other objective classification criteria, we had to resort to somewhat subjective criteria.

In this study, the classifications are based on general assessment made on the policy and claims data, such as age, sex, and activity involved of policyholder and make, Cubic capacity (CC), age, purpose of vehicle from underwriting data, claim frequency, claim cost, number of years of no accident of policy from claims data, and the annual assessment made by surveyors about the vehicles' familiarity of garage, availability of spare parts, and so on. These factors are rated as to how they affect the risk exposure of a policy, and are given points. Each policy in the data set has been classified into one of the three risk groups (Low, Medium, High) on the basis of its weak and strong points on these factors. The translation of the assessment into the classification of a policy into a specific risk group was made based on the number and the weight of a policy's weak and strong points given in the assessment reports.

Risk level	Number of records
Low	629
Medium	305
High	226
Total	1160

Table 5.3: Distribution of records with respect to risk level.

The classification approach used here has certain advantages. First, the employees of the NISCO can use all information about a policy that is available, plus any possible foreknowledge that they have about motor policy. Second, the standard amount of time for an evaluation of one single vehicle by NISCO surveyors is on average 2-3 hours. Such an investment of time is more than what one can expect to get otherwise, especially as the customer portfolio gets larger.

A disadvantage of this approach is that the assessment made by the domain experts is so general that might not fully describe the actual situation of a policy. Moreover, some important factors might not be known to NISCO. These problems are inevitable because of lack of workable objective criteria.

5.4. Feature Selection using Decision Trees

As mentioned in chapter four, section 4.3.3, feature (or attribute) selection an important problem in order to effectively develop a model from the data. This is especially true for neural networks. The neural network literature does not provide us with an efficient procedure to determine the "best" subset of variables for a certain application. Ideally, the neural network should be run on each possible set of variables to determine the perfect subset. The decision tree algorithm would help identify a subset of the most important variables (Berry & Linoff, 2000). Thus, Decision tree was used in this study to select the inputs for the neural network. To this end, an implementation of See5 software, the C5.0 version for windows based PC's, was used in developing the decision tree model.

5.4.1. Data organization for Decision Tree

The dataset prepared for this study, 1160, is divided into two: 90% (1044) for model building and testing set and the remaining 10% (116) for validation set. The decision tree software used in this study, See5, incorporates a facility to extract a random sample from a dataset used for model building and testing, construct a classifier from the sample and, and then test the classifier on a disjoint collection of cases. In this case, 90% of the 1044 records were used for model building and the remaining 10% for testing the resulting model.

These records are stored in two See5 application files, Data file and cases file. The Data file stores the records for model building and testing, and the cases file stores the validation dataset, which is set aside for later use in measuring the performance of the model. Both files have the same format.

Another important file during tree construction using See5 is the names file. It describes the name and possible values of the attributes and the classes.

5.4.2. Decision Tree Model Building

Taken as a whole, a decision tree is a classifier. Any previously unseen record can be fed into the tree. At each node it will be sent either left or right according to some test. Eventually, it will reach a leaf node and be given the label associated with that leaf. Here, we were more interested in generating rules to explain risk exposure of policies and to come to an understanding of the most important factors affecting the risk level of a policy than in simply classifying particular policies as risky or not, or predicting which future policies would be

risky. As a consequence, the decision tree that made the best predictions was not the one most useful for us. Instead, the one that generates sound rules was given priority in model selection.

With this understanding, numerous decision trees were constructed by taking a number of random 90-10 percent partitions of training and testing data, respectively, from the data file. For each decision tree, the corresponding rule sets were extracted.

Finally, on the basis of evaluation of the rule sets made by domain experts, a tree model whose rule set is found to be meaningful was selected as a working model for further stages of the study. The selected decision tree is depicted in figure 5.1. The extracted rules are annexed for reference.

```
Decision tree:

        CTPR > 1.2:
        :...NYNA <= 1: High (135/1)
        :   NYNA > 1:
        :   :...TCFREQ <= 1:
        :       :...VEHYEAR <= 3: Low (19)
        :       :   VEHYEAR > 3:
        :       :   :...AGEOFVEH <= 14: Low (8)
        :       :       AGEOFVEH > 14: Medium (4)
        :       TCFREQ > 1:
        :       :...NYNA <= 2: High (22)
        :           NYNA > 2:
        :           :...TCFREQ <= 2: Low (9)
        :               TCFREQ > 2: High (2)
        CTPR <= 1.2:
        :...AGEOFVEH <= 15:
            :...TCFREQ <= 0: Low (322)
            :   TCFREQ > 0:
            :   :...NYNA <= 1:
            :       :...NYNA <= 0: High (15)
            :       :   NYNA > 0: Medium (67/1)
            :       NYNA > 1:
            :       :...TCFREQ <= 1: Low (79)
            :           TCFREQ > 1:
            :           :...VEHYEAR <= 4: Medium (26)
            :               VEHYEAR > 4: Low (6)
            AGEOFVEH > 15:
            :...CC <= 1360:
                :...TCFREQ <= 0: Low (55)
                :   TCFREQ > 0:
                :   :...NYNA <= 1: Medium (11)
                :       NYNA > 1: Low (7)
                CC > 1360:
                :...TCFREQ <= 0: Medium (89)
                    TCFREQ > 0:
                    :...NYNA <= 0: High (4)
                        NYNA > 0:
                        :...VEHYEAR > 3: Medium (35)
                            VEHYEAR <= 3:
                            :...NYNA <= 1: Medium (18/1)
                                NYNA > 1: Low (7)
```

Figure 5.1: Decision tree constructed by See5.

5.4.3. Feature Selection Result

Although 25 independent variables were used in the input data, the decision tree selected only six variables. Table 5.4 lists the selected variables and the frequencies with which the variables appear in the tree. The listing is according to how the variable floats to the top of the tree. Decision tree building algorithms put the variable that does the best job of splitting at the root node of the tree (Berry & linoff, 2000). Such information can be of use in selecting variables that appear to be important for the input to the neural network model.

Variable	Frequency
CTPR	1
NYNA	6
AGEOFVEH	2
TCFREQ	4
CC	1
VEHYEAR	3

Table 5.4: Variable frequencies in the tree model

5.4.4. Classification Results

Though classification was not the primary goal at this stage, we did not find it natural to skip this step. A number of authors wrote about the classification and predictive power of decision trees. Berry and Linoff (2000) advise to use decision trees when the goal of study is to assign each record to one of the few broad categories.

The classification performance of the selected decision tree is measured using the validation dataset. The results of the model for the validation dataset are given in table 5.5. The Score column in the table shows the percentage of records that has been classified correctly.

	Predicted				
Actual	Low	Medium	High	**Total**	**Score**
Low	53	0	1	54	98.15%
Medium	0	32	2	34	94.12%
High	0	2	26	28	92.86%
Total	53	34	29	116	95.69%

Table 5.5: Confusion matrix for the selected tree model

One can see from the above table that the model performs quite well for each class. Only one policy (low risk) is misclassified as high-risk. Some of the reasons for the misclassification from one risk level to the upper or lower risk level of policies by the selected model might be the following. One possible reason is that the model is selected on the basis of the soundness of the rules that it generates rather than its classification performance. The other reason is that the group of medium risk policies is heterogeneous that they have certain peculiarities, which make them low or high risky.

5.5. Neural Networks Model Design and Building

As mentioned in the previous chapter, the combination of topology, learning rule, and learning algorithm define a neural network model. There are wide selections of neural network models. For data mining, perhaps the back-propagation network is the most popular

(Bigus, 1996). Thus, feed forward back-propagation neural network is used in this study to build risk model.

Once the data preparation is completed and the neural network model and architecture have been selected, the next step is to train the neural network. Training a neural network is a challenge requiring setting of numerous parameters. Taking into account the time available to undertake the study and a bewildering set of parameters to adjust, four networks have been designed and tested.

5.5.1. Network design

For all designs, the input of the network consists of the six numeric variables selected by the decision tree model. Hence, the input layer of the network consists of six units, one for each of the selected numeric attributes. The output layer contains three neurons: one for each risk group. Consequently, the output of the network consists of three scores between 0 and 1, one for each possible class. An advantage of the use of different output neurons for different classes is, that it facilitates the analysis of the effect of each input variable on each possible outcome (see section 5.5.4).

Two important questions regarding the network architecture are (Kramer, 1994): "how many hidden layers should we use?" and "how many neurons do we need in each hidden layer?" It is proven by several authors (see for instance Bigus (1996)) that a one-hidden-layer network with enough hidden neurons can approximate practically all vector functions. Therefore, a one-hidden-layer network is used in this study.

However, there is no exact formula to determine the ideal number of hidden neurons needed for a given application. The developers of *BrainMaker* software suggest on the minimum number of hidden neurons to be half of the sum of the input and output neurons. Thus, in this study, the researcher begins with a small number of hidden neurons in each of the network design, and adds neurons during the training process if the network is not learning. That is, six hidden neurons were used as a starting point, and add a neuron if the root mean square (RMS) error does not decrease over 100 training runs.

For the hidden and output neurons it is common to take a sigmoid (S-shaped) activation function (Bigus, 1996). In a comparison of the error surfaces of several activation functions, Kempka, cited in Kramer (1994) found sigmoid activation function to be the most desirable in neural networks that need to predict accurately. Since accurate prediction is an important quality of risk classification, a sigmoid activation function has been applied in this study.

In each of the four network designs, the absolute minimum and maximum of the data is used for normalization. Sometimes using the absolute minimum and maximum of the data causes problems, especially if the data contains outliers (Lawrence, 1994). The reason is that if the network uses the absolute minima and maxima that include the outliers, it will be difficult for it to tell the difference between facts with values close together. The network will understand extreme cases better, but it will not perform as well in the typical scenario. By adjusting the minimum and maximum values this bias can be reduced. However, the distribution of the data used in this study was fair with respect to each input variable except to that of the claim to premium ratio (CTPR). This is due to the fact that the number of policies subject to claim every policy year is very small that many policies have no claim cost. Hence we could not find it reasonable to make adjustment on the minima and maxima of the CTPR data.

Another important consideration while designing the network is setting the learning rate parameter. The data used for experimentation in this study are unevenly distributed; there are many more low-risk policies (629) than medium (305) and high-risk (226) policies. With a high learning rate, the network will tend to learn that all policies are low risk groups. A technique, which works well for many unevenly distributed data sets, is to set the learning rate to a very low value (CSS, 1998). The network will then train very slowly, because the steps towards the answer are very small, and the advantage here will be that it does not move too quickly in the direction of the majority solution. In order to be able to find the patterns for the minority data, low learning rates are used in this study. The rate varies in the range of 0.25 to 0.4 in step of 0.5. An overview of the network designs is given in table 5.6.

Design	**Learning Rate**	**Minimum number of Hidden Neurons**
A	0.25	6
B	0.30	6
C	0.35	6
D	0.40	6

Table 5.6: Summary of the trained neural networks

5.5.2. Data Organization for Model Building

One pitfall that a data miner should beware of in interpreting the model results is over-fitting (Levin & Zahavi, 1999). Over-fitting pertains to the phenomenon where one gets a very good fit on the data, which is used to build the model, but poor fit when the model is applied on a new set of observations. This problem could have serious implication in practice.

To avoid over-fitting, the available data is split into three parts. First, the working data set, with 1160 policies, is divided into two sets: The first, 90% (1044), is for a training set used for determining the values of the weights, and a testing set used for deciding when to stop training. The second dataset, the remaining 10% (116), is for validation set. While ninety percent (940) of the first dataset is used for training purpose, the remaining ten percent (104) is used for testing. In all cases, sampling is made on random basis, and care has been taken in that data is available for all possible outcomes (or classes) so the model can learn about all cases.

The performance on the test set is monitored. As long as this performance improves, training continues. When it ceases to improve, training is halted. Since the performance of the network may be influenced by the composition of the training and the testing set, three random 90-10 percent partitions have been generated. To estimate the expected performance of the network in the future, the validation set is used. This data set is strictly set apart and is neither used in training nor to determine the termination of the training process. The optimal network for this application will be the network with the best performance on the validation set.

Finally, the PC platform for this project was a (R) Pentium 4 CPU processor of 1.70GHz with 256MB RAM. This high-speed computer made the training of the neural networks fast.

5.5.3. Training and testing results

For each of the four network designs and three random data set partitions, the network is trained over 5000 runs. Numerous models for each of the four designs were developed by adjusting the network parameters like the testing tolerance from 0.4 down to 0.1, and number

of hidden neurons interactively for better network performance. After the training had stopped, the development over the runs of the root mean square (RMS) error on the training set and the number of errors on the testing set were analyzed. *BrainMaker* has the capability of storing these statistics on separate files. If a network still showed improvement over the last runs, training would be continued for another 1000 runs.

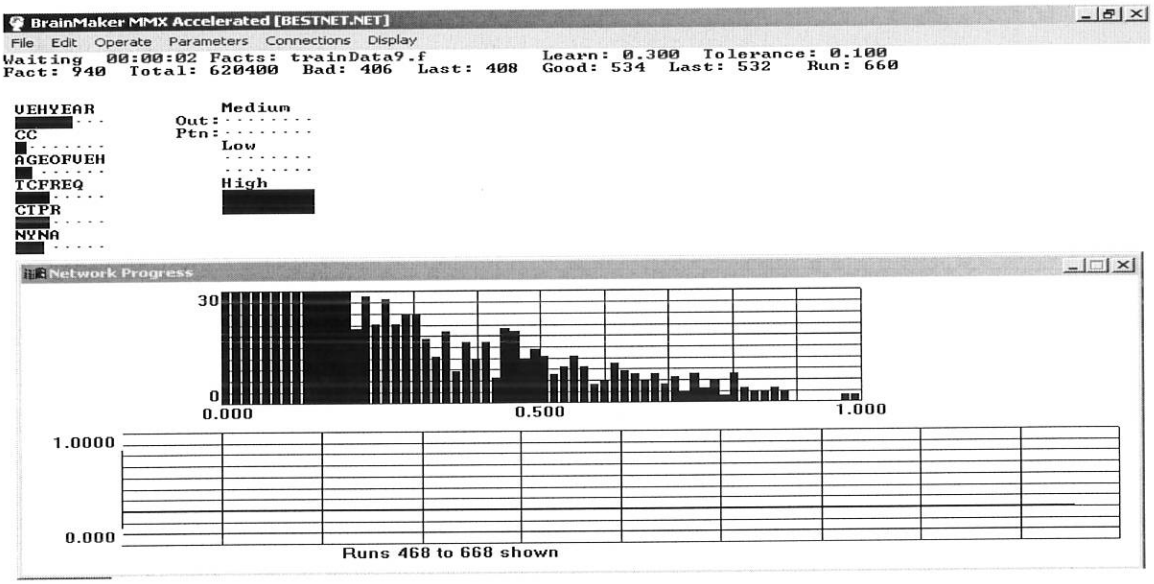

Figure 5.2: BrainMaker Neural Network while in training

After the completion of the training process, for each of the 12 (3x4) partitions, the networks from the runs with the minimum number of errors on the testing set were selected. The final choice from the 12 networks is based on the performance of these networks on the validation dataset. The percentages of policies correctly classified by the different networks for the validation set are given in table 5.7.

Partition	Network Design			
	A	B	C	D
1	89.66	87.1	91.35	87.1
2	92.24	91.34	92.24	91.34
3	81.05	81.9	84.48	79.31

Table 5.7: Percentage of correctly classified policies

From the above table one can see that the overall performance of the network models towards this classification problem is encouraging. Moreover, most of these models have good classification accuracy for high-risk policies. The majority of policies that are misclassified are those in medium risk level. The reason for this misclassification is as discussed earlier for the decision tree model.

Two networks, 2A and 2C, have better performance with maximum score of 92.24%. Of these networks, network 2C is preferred for the following reasons. First, the network correctly classifies the high-risk policies whereas network 2A misclassifies 2 high-risk policies. Second, the misclassification of low risk policies as high risk plus high-risk policies as low-risk is small compared to that of network 2A (this is 0 for 2C and 2 for 2A). The classification tables for the networks 2C and 2A are given in tables 5.7 and 5.8 respectively. The Score column in the table shows the percentage of records that has been classified correctly.

Actual	Predicted			Total	Score
	Low	Medium	High		
Low	53	1	0	54	98.15%
Medium	7	26	1	34	76.47%
High	0	0	28	28	100%
Total	60	27	29	116	92.24%

Table 5.8: Confusion matrix for the selected Network (Network 2C)

Actual	Predicted			Total	Score
	Low	Medium	High		
Low	51	2	1	54	94.44%
Medium	3	30	1	34	88.24%
High	1	1	26	28	92.86%
Total	55	33	28	116	92.24%

Table 5.9: Confusion matrix for Network 2A

The selected network, 2C, has 10 hidden units, and is obtained after training the network with training tolerance of 0.1 and testing tolerance of 0.3, for 650 runs. The overall performance of the network is satisfactory. The number of medium risk policies that has been classified correctly is, however, small. One possible reason for the misclassification of medium policies might be that the group of medium-risk policies is very heterogeneous. A number of policies might be classified as medium risk although they have certain peculiarities, which make them less risky or more risky. The other possible reason is that, either the employees of the company are not sure whether a policy is low or high risk, and use the medium class as a 'don't know " class. Therefore, a policy is classified, as medium while it is actually low or

high risk. In this case it is not strange that the model shows slightly low performance on medium-risk policies.

The performance of the model on the most important class, the high-risk level, is quite good. All high-risk policies get a high priority, and most policies, which the model classified as high risk, are truly so. These are useful properties for a system, which will be used as a tool to support the supervision by the underwriting management. In this case the insurer can concentrate on the most important cases.

5.5.4. Interpretation of the network weights

Although it is difficult to explain why a neural network reaches a particular conclusion, something can be said about the relative importance of the different input variables for the different output units. According to Yoon et al., cited in Kramer (1994), this can be done by calculating the strength of the relationship between each input and each output neuron, which is measured by the following statistic:

$$RS_{ji} = \frac{\sum_{k=0}^{n} \left(W_{ki} \times W_{jk} \right)}{\sum_{i=0}^{m} \left| \sum_{k=0}^{n} \left(W_{ki} \times W_{jk} \right) \right|}$$

where : RS_{ji} is the relative strength beween the i^{th} input and the j^{th} output units;

W_{ki} is the weight between the k^{th} hidden unit and the i^{th} input unit; and

W_{jk} is the weight between the k^{th} hidden unit and the j^{th} output unit

In multivariate analysis, this approach is applied frequently to determine the proportion of variation of one variable in relation to that of all variables. This statistic measures the strength of the relationship between the i^{th} input and the j^{th} output neuron to the total strength of all input neurons with respect to the j^{th} output neuron. The result is the percentage of weights between all input neurons and a specific output neuron. Table 5.8 presents the relative strengths for the current network.

Input Variable	**Output pattern**		
	Low	Medium	High
AGEOFVEH	-0.0370	0.0398	0.0314
CC	-0.0341	0.0309	0.0013
CTPR	0.0307	-0.1613	0.1996
NYNA	0.0551	0.0360	-0.0830
TCFREQ	-0.1487	0.0440	0.0979
VEHYEAR	0.0264	0.0219	-0.0820

Table 5.10: Relative strengths between the input and the output variables.

The results given in table 5.10 shows more or less the reality in insurance risk assessment. According to the underwriters at NISCO, Age of vehicle and CC are two of the important factors used in insurance premium calculation. Their experience shows that as the age and CC of the vehicle increases, then its risk exposure also becomes higher. Thus, as expected, the network shows positive correlation between these variables and high-risk level. Total claim frequency (TCFREQ) is also positively related with high-risk, and negatively related to low-risk level. That is, higher number of frequency will result in higher score for high risk, and lower score for low-risk, all other things being equal. Similarly, the average claim to premium ratio (CTPR) shows higher score for high-risk, and lower score for low-risk, though not

negative. It is known that this ratio is a key driver to profitability. As this ratio increases, the profit decreases, and hence high risk is associated.

Whereas, the other two variables, number of years of no accident (NYNA) and number of years vehicle is insured (VEHYEAR), are negatively related to high-risk, and positively related to low-risk level. The reason for the relationship between NYNA and risk level of a policy is immediate. The relationship between VEHYEAR and risk level shown on the above table might be attributed to the fact that as VEHYEAR increases, the policyholder becomes more loyal to the company.

5.6. Comparison Between Decision Tree and Neural Networks

The scores of the neural network for low- and high-risk policies are very high. Especially the score for high-risk policies is very convincing. The overall score of the neural network model is, however, slightly lower than that of the decision tree model (92.24 versus 95.69 per cent). This may reflect the fact that neural networks can benefit from more extensive data preparation as well as a greater degree of experimentation with model parameters.

The neural network fails to recognize medium-risk policies as good as that of low and High-risk policies. Apart from the possible explanations given earlier, an additional explanation for the low performance of the neural network may be that, given the highest overall score, the network was selected on the basis of its performance on high-risk policies, and on the (lowest) number of misclassifications from low to high-risk and vice versa. The network performed best on both criteria, whereas other networks performed somewhat better on medium-risk policies.

Although there is a consensus between the tree and network models, they disagree on the classification of several policies. Some policies are misclassified by one of the models though not by the other. A combination of both models could, therefore, lead to an even better performance than the performance of the separate models. The misclassification of a policy by the decision tree model may be compensated for by the correct classification of the neural network model, and a misclassification by the neural network model may be compensated for by the correct classification of the decision tree model. Due to time limitation, however, testing the combined models was not done.

5.7. The Renewal System: A Prototype

Data mining can be used for much more than decision support applications (Bigus, 1996). The trained model can be used to process transactions and perform classification and predictions on data in the real time.

In this study an attempt was made to develop an operational application prototype named Motor Insurance Renewal System (MIRS) that uses the neural network model developed in the above process. The model is deployed as part of the application where it is used to score customers propensity to risk exposure prior to renewal request of policy is accepted. MIRS contains MS access database, the neural network, and the visual C++ program. The structure of MIRS is given in figure 5.3.

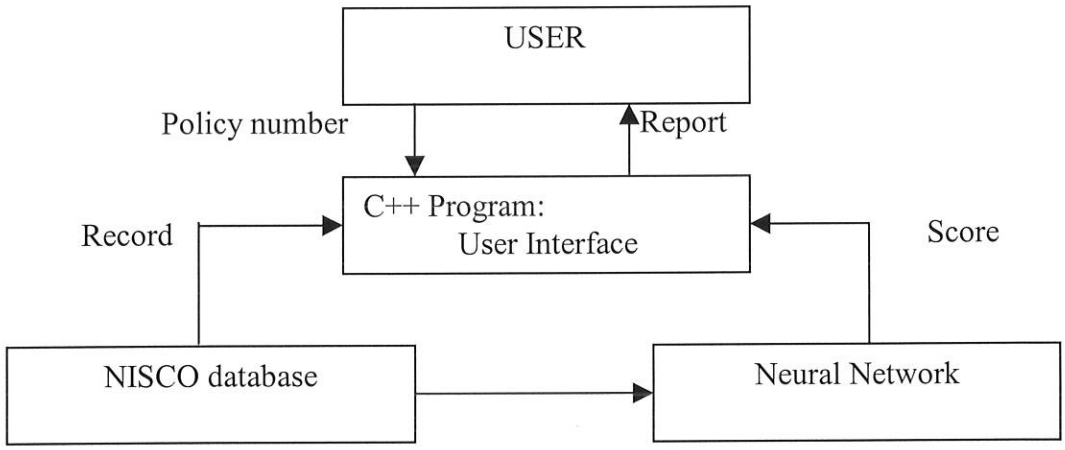

Figure 5.3: The structure of MIRS.

The database holds the policy and claims data in two separate tables linked with a common field, policy number. It also holds a table that stores the value of the variables used in neural network model building. The neural network calculates scores based on the six variables extracted from the database.

The C++ program can be seen as the main program. It functions as an interface for the user. After the user has started the program, the program first asks the customer policy number, which is the unique identifier, to be evaluated. When this number is given, the program searches for the record with this policy number from the database, and extracts the required attributes, which are then passed through the neural network model for prediction. The program then reads the output values from the output file of the neural network, and post-processing functions such as scaling and conversion into categorical values is performed. This post-processed information is then reported to the user as a result. The report is both written down into a file and onto the screen. The user, employee of the company, can then use this report so that he can decide on further action.

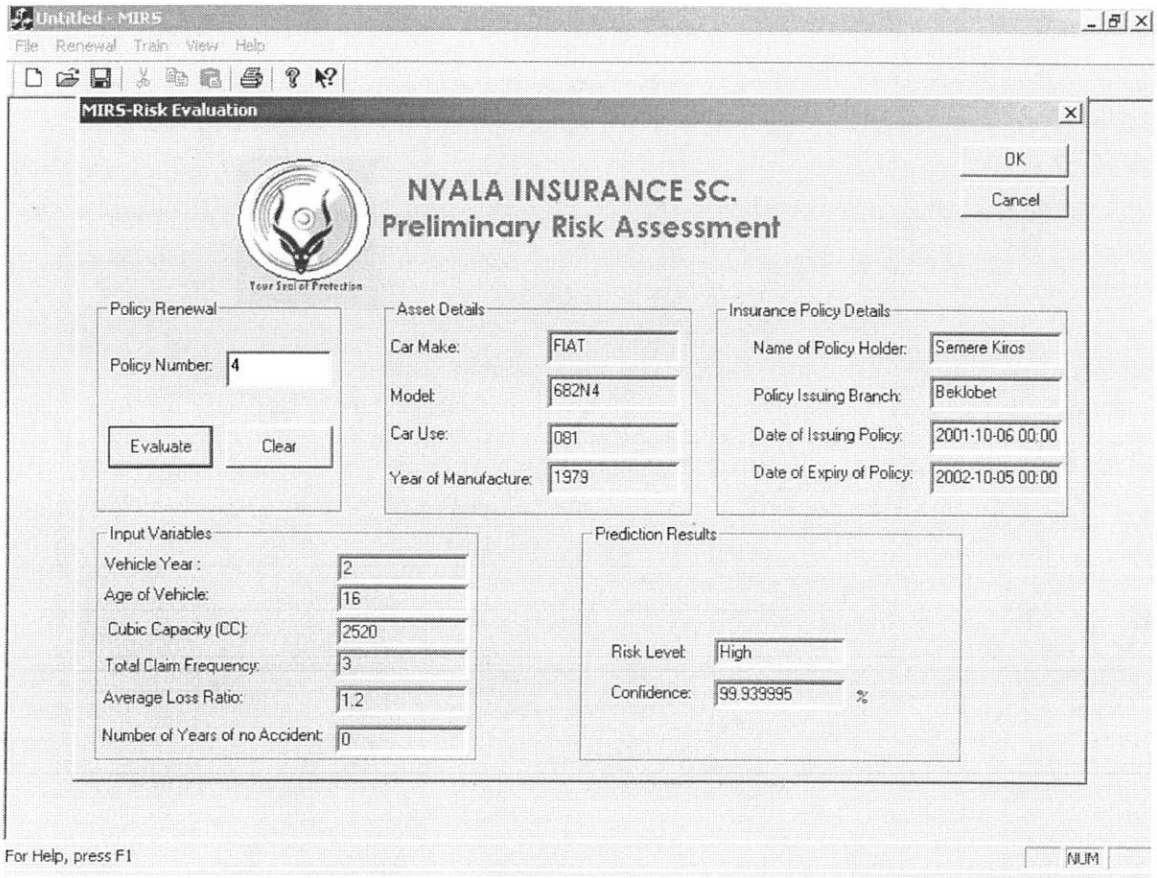

Figure 5.4: Dialog box for the preliminary risk assessment report of MIRS.

5.7.1. Maintaining The Neural Network Model

Once deployed, the predictive accuracy of the neural network must be monitored. A neural network that is trained and then deployed for a long period is like having an employee who is trained to perform a specific task and then does nothing to update his or her skills (Bigus, 1996). One possible way for network maintenance is to deploy the network while it is still in training mode so that it can learn from experience after it is deployed. This is called online learning (Bigus, 1996). But the neural network used in this study, the back propagation neural network, does not support online learning.

The other alternative way to customize the neural network model is that one could periodically need to retrain the neural network using the latest data that captures the latest trends and give feedback to the network on its performance. If the network makes an incorrect prediction or decision, make a new training example by taking the input data and adding the known correct answer as the desired output. In this way, one can modify the behavior of the network. An attempt is also made on MIRS to support an ease access to the network tool, Brainmaker software, when retraining of the model is needed.

CHAPTER SIX

CONCLUSION AND RECOMMENDATIONS

6.1. Conclusion

Data mining, extracting meaningful patterns and rules from large quantities of data, is clearly useful in any field where there are large quantities of data and some thing worth learning. In this respect, the insurance industry is a potential area for data mining. It is filled with lots of data and professionals who already make sense of all the data to manage risk.

In this research, an attempt was made to assess the potential applicability of data mining technology in support of risk assessment activity in the insurance industry. This experimental research, which employed the commonly used methodological approach in data mining researches, made use of two predictive modeling techniques, decision tree and neural networks, to address the problem.

Defining the target class was one of the Challenging tasks in this research. The initial data collected from the NISCO was not pre-classified. So before moving to the successive stages of the study, the data had to be classified in such a way that the insurer can easily prioritize policies according to their risk exposure for assessment during policy renewal. The classification approach applied in this study differs from the one used by insurance companies. The approach used is based on the idea that it is more useful to have a tool that focuses on identifying conditions that can lead to a rejection, than one that focuses on

predicting policy rejection itself. A problem with the classification approach used in this study is that it is not error-free. That is, there is no guarantee that the classification based on the analysis of the general assessment report is correct. The classifications are based on a subjective analysis of the assessment reports, and the assessment reports are the result of a subjective analysis of annual reports. Therefore, the classifications based on an analysis of the assessment reports should be considered as initial classifications, which can be refined (or improved) iteratively.

Thus, even though no objective classification criterion is available, a model to classify policies according to their risk exposure can still be built. Two basic tasks made in model building are feature selection using decision tree and predictive model building using neural network. Various experiments were made iteratively by making adjustments on the modeling parameters in both tasks to come up with meaningful results.

In the feature selection phase, more emphasis was given to the soundness of the rule sets extracted from the resulting decision trees. Accordingly, the better decision tree selected as a working model generates meaningful rules that would assign new policy records to the classes. Though, the domain experts selected about 25 attributes that they believed to be important factors in predicting risk exposure of a policy, the decision tree extracts only six of them as important classifiers of the policies. Moreover, the classification accuracy of the decision tree was so convincing, that among the 116 validation data, 95.69% of them were correctly classified. The misclassification of the remaining records is mainly attributed to the problem associated with the definition of the target variable.

The six significant variables of the decision tree were used as input to the neural network in developing predictive model. The best neural network model, which is selected as a working model among the numerous models generated during the training phase, was able to correctly classify 92.24 per cent of the 116 policies in the validation data set. All high-risk policies were correctly classified. Furthermore, no policy with a high-risk exposure was misclassified as low risk, and no policy with a low risk exposure was misclassified as high risk. Conditional on the assumption that the actual risk exposures with which the outcomes of the neural network model have been compared are correct, it can be concluded that the model is quite able to identify most policies with an excessive risk exposure.

Compared to the result of decision tree, the neural network performs slightly low. This performance difference by no means shows the weakness of the predictive capability of neural networks. Whereas, the model-building phase revealed that neural networks can benefit from extensive work on data preparation, and experimenting iteratively by altering the various parameters.

In addition, the comparison of the results of the decision tree and neural network models showed an interesting pattern. Policies that are misclassified by one model are correctly classified by the other. This might be an indication that the combination of the models could result in a better classification performance.

The last phase of this study was prototype development. This was the result of good performance of the working model. Here an attempt was made to show that research on data mining could result in a new tool that support in addressing the problem at hand. The prototype, named MIRS (Motor Insurance Renewal System) is developed under the

assumption that work has been done on designing the required database. MIRS would not take over the full evaluation process. Rather, the output of MIRS would be used to establish priorities for the assessment round, which would then be performed by employees of the company. In this way, the verification round can be done quicker, and potential risky customers will be treated first in the assessment round. A thorough analysis by human experts will, however, still be necessary.

To conclude, results from the study have shown that the problem in insurance risk assessment, in particular assessment made during policy renewal, could be leveraged using data mining techniques.

6.2. Recommendations

This research work is conducted mainly for academic purpose. However, it is the researcher's belief that the findings of the research will help initiate financial sectors to work on the application of data mining technology to gain competitive advantage in their field. Moreover, the research work can contribute a lot towards a comprehensive study in this area in the future.

In the course of doing this study and on the basis of the findings of the research work, the researcher has come up with a sort of tasks that need more consideration in future work.

At the time of this work, only a limited number of all possible attributes were available with their values in the database of the company. Moreover, inconsistency in filling out the required information, dealing more on the financial information and giving less importance to

others like age and sex of driver, which perhaps are the most important risk factors, were noticed. During data preprocessing the researcher was forced to exclude such important factors. Since data is the most important component in data mining research, proper handling and concern to information is strongly recommended. Furthermore, the researcher would like to suggest future work including the excluded variables so that better models could be developed.

Another future task that need more concern is the definition of the target variable. Even though the result of the study shows better performance in classifying policies according to their risk exposure, still more refinement of the initial classifications is needed until the performance of the model is satisfactory.

In this research an attempt was made to make use of hybrid techniques. The result of the two models from the decision tree and neural network was so encouraging. Moreover, the consensus between them and the ability to classify correctly those policies that are misclassified by the other model is another interesting area that needs more effort to be made in the future to find out how the combined model will change the performance of either model.

Finally, the deployment of the working model that results from the data-mining phase is the end product of many researches in the field of KDD. As mentioned in the prototype development section, the neural network model used in this study, back propagation neural network, has no online learning capability. For problems that change through time, like insurance risk, online learning is a good choice. This allows the neural network learn from experience after it is deployed. Hence the researcher recommends on using alternative neural

network models, for example Adaptive resonance neural networks and Probabilistic neural network, as suggested by Bigus (1996) that can be trained online.

REFERENCES

Apte, C. et al. (1998). Insurance Risk Modeling Using Data Mining Technology. Research Report available at URL: http://www.ibm.com

Askale Worku (2001). *Data Mining Application in Support of Loan Disbursement Activity At Dashen Bank SC.* A Thesis Submitted in Partial Fulfillment of the requirement for the Degree of M.Sc. I.S. Addis Ababa University: Addis Ababa

Bannister, J. E. and Bawcutt, P. A. (1981). *Practical Risk Management.* London: WITHERBY & CO. LTD.

Berry, M. J. A. and Linoff, G. (2000). *Mastering Data Mining: The art and Science of Customer Relationship Management.* New York: John Wiley & Sons, inc.

Bigus, J.P. (1996). *Data Mining With Neural networks in Solving Business Problems-From Application to Development to Decision Support.* New York: McGraw-Hill.

Brand, E. and Gerritsen, R. (1998). Classification and Regression. Available at URL: http://www.dbmsmag.com

Cheng, W. and McClain, B. W. (1998). Artificial Neural Networks make their Mark as a powerful tool for Investors. Available at URL: http://members.tripod.com/kalle_t/AI_networks_invest.html.

Collard, J. E. (1993). *Commodity Trading with a Three Year Old.* In Trippi, R. R. and Turban, E. eds. Neural Networks in Finance and Investing, Pages 411-420.

CSS (1998). *Neural Networks Simulation Software User's Guide and Reference Manual.*, Nevada City, CA: California Scientific Software (CSS).

Deogun, J.S. et al. (1998). Feature Selection and Effective Classifiers. *Journal of American Society for Information Science*, 49 (5).

Dickson, G. and Stein, W.M. (1999). *Risk and Insurance*. Study course 510. London: CII Publishing Division.

Dockrill, M. et al. (2001). *Underwriting Management*. Study Course 815. London: CII publishing Division.

Dutta, S. and Shekkar, S. (1988). Bond Rating: A Non-conservative Application of Neural Networks. *Proceedings of the IEEE International Conference on Neural Networks*. San Diego, Pages 443-450.

Edwards, P. J. et al. (1999). The Application of Neural Networks to the Papermaking Industry. *IEEE Transactions on Neural Networks*. Vol. 10, No. 6.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI magazine.* At URL: http://citeseer.nj.nec.com/fayyad96from.html

Galfond, Glenn (1997). Data Mining can Unearth A Competitive Edge. Available at URL: http://rmisweb.com/rmisartc/nu/100697.htm.

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

Goebel, M. and Gruenwald, L. (1999). A Survey of Data Mining and Knowledge Discovery Software Tools. URL: http:// www.acm.org/sigkdd/explorations/issue1-1/survey.pdf

Jones, L. W. (1996). *Neural Networks*. Available at URL: http://ei.es.vt.edu/~history/NEURLNET.HTML.

Kaye, D. (2001). *Risk Management*. . Study course 655. London: CII Publishing Division.

Kirby, P. and Williams, G. (1998). *Motor Insurance*. Study Course 765. London: CII Publishing Division.

Kononenko, I. And Hong, S. J. (1997). Attribute Selection for Modeling. At URL: http:// www.research.ibm.com/dar/papers/ pdf/fgcshong_with_cover.pdf

Kramer, B. (1994). The evaluation of Dutch Non-life Insurance Companies: A comparison of an Ordered Logit and a neural Network Model. At URL: http://www.ub.rug.nl/eldoc/som/95A20/95A20.pdf

Langley, P and Simon, H.A. (1995). Applications of Machine Learning and Rule Induction. URL: http:// citeseer.nj.nec.com/langley95applications.html -

Lawrence, J. (1994). *Introduction to Neural Networks, Design, Theory and Application.* Nevada City: California Scientific Software Press.

Levin, N. and Zahavi, J. (1999). Data Mining. At URL: http://www.urbanscience.com/Data_Mining.pdf

NISCO (1999). Experience and Recommendation on the Service rendered by TC. A report prepared for the 7th NISCO workshop. Addis Ababa.

Odom, M. & Sharda, R.A. (1990). A neural Network Model for Bankruptcy Prediction. *Proceedings of the IEEE International Conference on Neural Networks*. San Diego, Pages 147-162.

Piatetsky-Shapiro, G. (1999). The data Mining Industry Coming of Age. *IEEE Intelligent Systems*. URL: www.kdnuggets.com/gpspubs/ ieee-intelligent-dec-1999-x6032.pdf

Raghavan, V.V; Deogun, J.S. and Sever, H. (1998). Knowledge Discovery and Data Mining: Introduction. *Journal of American Society for Information Science*, Vol. 49 (5).

Rick, E. and Knight, K. (1991). *Artificial Intelligence.* 2nd ed. New York: McGraw Hill, Inc.

SAS Institute Inc. (2001). Data Mining In the Insurance Industry: Solving Business problems using SAS® Enterprise Miner™ Software. http://www.sas.com.

Stergiou, C. (1996). What is a Neural Network? Available at URL: http://www.doc.ic.uk/~nd/surprise_96/journal/vol1/cs11/article1.html.

Surkan, A. J. and Singleton, J. C. (1990). *Neural Networks for Bond Rating Improved by Multiple Hidden Layers. Proceedings of the IEEE International Conference on Neural Networks*. San Diego, Pages 443-450.

Trybula, W.J. (1997). Data Mining and knowledge Discovery. *Annual Review of Information Science and Technology*, vol. 32.

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

Wildman, P.; Wright, J. D and McNamara, M. (998). *Principles of Property and Pecuniary Insurances*. Course 745. London: CII Publishing Division.

Williams (Jr.), C.A.; Smith, M.L. and Young, P.C. (1995*). Risk Management and Insurance*.7th ed. New York: McGraw-Hill.

APPENDICES

Appendix I

Rule set extracted from the working decision tree.

```
See5 [Release 1.16]      Wed Jun 19 19:20:31 2002

    Options:
        Rule-based classifiers
        Tests on discrete attribute groups
        Use 90% of data for training
        Pruning confidence level 30%

Read 940 cases (27 attributes) from trainData.data

Rules:

Rule 1: (322, lift 1.8)
        AGEOFVEH <= 15
        TCFREQ <= 0
        ->  class Low  [0.997]

Rule 2: (238, lift 1.8)
        AGEOFVEH <= 15
        TCFREQ <= 1
        NYNA > 1
        ->  class Low  [0.996]

Rule 3: (103, lift 1.8)
        CC <= 1360
        TCFREQ <= 0
        ->  class Low  [0.990]

Rule 4: (64, lift 1.8)
        CC <= 1360
        CTPR <= 1.2
        NYNA > 1
        ->  class Low  [0.985]

Rule 5: (37, lift 1.8)
        VEHYEAR > 4
        AGEOFVEH <= 15
        CTPR <= 1.2
        ->  class Low  [0.974]


Rule 6: (67/2, lift 1.8)
        VEHYEAR <= 3
        TCFREQ > 0
        NYNA > 1
        ->  class Low  [0.957]
```

```
Rule 7: (9, lift 1.7)
        TCFREQ > 1
        TCFREQ <= 2
        CTPR > 1.2
        NYNA > 2
        ->  class Low  [0.909]

Rule 8: (89, lift 3.7)
        CC > 1360
        AGEOFVEH > 15
        TCFREQ <= 0
        ->  class Medium  [0.989]

Rule 9: (31, lift 3.7)
        VEHYEAR <= 4
        TCFREQ > 1
        CTPR <= 1.2
        NYNA > 1
        ->  class Medium  [0.970]

Rule 10: (4, lift 3.1)
        VEHYEAR > 3
        AGEOFVEH > 14
        TCFREQ <= 1
        CTPR > 1.2
        ->  class Medium  [0.833]

Rule 11: (275/120, lift 2.1)
        TCFREQ > 0
        CTPR <= 1.2
        ->  class Medium  [0.563]

Rule 12: (52, lift 5.2)
        NYNA <= 0
        ->  class High  [0.981]

Rule 13: (199/41, lift 4.2)
        CTPR > 1.2
        ->  class High  [0.791]

Default class: Low
```

Appendix II

Source Code used to evaluate the risk level of a policy given its policy number

```
void Creport2::OnEvaluate()
{

FILE *BrRTS,*OP,*OP1;
CString IDNumber, Prediction, SConfidence;
bool found=FALSE, found2=FALSE;
CString CharID,CharVEHYEAR, CharAGEOFVEH, CharCC, CharTCFREQ, CharCTPR,
CharNYNA;
CString CharMAKE,CharMODEL, CharBUSINESS, CharMANYEAR, CharDATEFROM,
CharDATETO, CharBRANCH, CharNAME;
CNNData MIRSrst;
CPolicyData1 MIRSrst2;
int i,InputUnits=6;
float Input,Low=0, Medium=0, High=0, float Confidence=0;

CWnd* pWnd1 = GetDlgItem(IDC_EDIT1);
CWnd* pWnd3 = GetDlgItem(IDC_EDIT3);
CWnd* pWnd4 = GetDlgItem(IDC_EDIT4);
CWnd* pWnd5 = GetDlgItem(IDC_EDIT5);
CWnd* pWnd6 = GetDlgItem(IDC_EDIT6);
CWnd* pWnd7 = GetDlgItem(IDC_EDIT7);
CWnd* pWnd8 = GetDlgItem(IDC_EDIT8);
CWnd* pWnd9 = GetDlgItem(IDC_EDIT9);
CWnd* pWnd10 = GetDlgItem(IDC_EDIT10);
CWnd* pWnd11 = GetDlgItem(IDC_EDIT11);
CWnd* pWnd12 = GetDlgItem(IDC_EDIT12);
CWnd* pWnd13 = GetDlgItem(IDC_EDIT13);
CWnd* pWnd14 = GetDlgItem(IDC_EDIT14);
```

```
CWnd* pWnd15 = GetDlgItem(IDC_EDIT15);
CWnd* pWnd16 = GetDlgItem(IDC_EDIT16);
CWnd* pWnd17 = GetDlgItem(IDC_EDIT17);
CWnd* pWnd18 = GetDlgItem(IDC_EDIT18);


pWnd1->GetWindowText(IDNumber);

MIRSrst.Open();
MIRSrst.MoveFirst();
while(!MIRSrst.IsEOF()&&(!found))
{
        MIRSrst.GetFieldValue("ID",CharID);
        if(IDNumber.Compare( CharID ) == 0)
                found=TRUE;
        MIRSrst.MoveNext();
}

if(!found)
MessageBox("Policy not found!!!!!!");
else {
MIRSrst.MovePrev();
        MIRSrst.GetFieldValue("VEHYEAR",CharVEHYEAR);
        MIRSrst.GetFieldValue("AGEOFVEH",CharAGEOFVEH);
        MIRSrst.GetFieldValue("CC",CharCC);
        MIRSrst.GetFieldValue("TCFREQ",CharTCFREQ);
        MIRSrst.GetFieldValue("CTPR",CharCTPR);
        MIRSrst.GetFieldValue("NYNA",CharNYNA);

        MIRSrst2.Open();
        MIRSrst2.MoveFirst();
        while(!MIRSrst2.IsEOF()&&(!found2))
        {
                MIRSrst2.GetFieldValue("ID",CharID);
```

```
        if(IDNumber.Compare( CharID ) == 0)

                found2=TRUE;

        MIRSrst2.MoveNext();

}

if(!found2)
MessageBox("Policy not found!!!!!!");
else {
MIRSrst2.MovePrev();
MIRSrst2.GetFieldValue("MAKE",CharMAKE);
MIRSrst2.GetFieldValue("MODEL",CharMODEL);
MIRSrst2.GetFieldValue("CLASS",CharBUSINESS);
MIRSrst2.GetFieldValue("DATEOFMANUFACTURE",CharMANYEAR);
MIRSrst2.GetFieldValue("POLICYPERIODFR",CharDATEFROM);
MIRSrst2.GetFieldValue("POLICYPERIODTO",CharDATETO);
MIRSrst2.GetFieldValue("BRANCH",CharBRANCH);
MIRSrst2.GetFieldValue("NAME",CharNAME);


pWnd3->SetWindowText(CharMAKE);
pWnd4->SetWindowText(CharMODEL);
pWnd5->SetWindowText(CharMANYEAR);
pWnd6->SetWindowText(CharBUSINESS);
pWnd7->SetWindowText(CharNAME);
pWnd8->SetWindowText(CharBRANCH);
pWnd9->SetWindowText(CharDATEFROM);
pWnd10->SetWindowText(CharDATETO);
pWnd11->SetWindowText(CharVEHYEAR);
pWnd12->SetWindowText(CharAGEOFVEH);
pWnd13->SetWindowText(CharCC);
pWnd14->SetWindowText(CharTCFREQ);
```

```
        pWnd15->SetWindowText(CharCTPR);
        pWnd16->SetWindowText(CharNYNA);

}
if ((BrRTS = fopen ("BrainRTS.in","w"))== NULL)
        {
                MessageBox("Can not open input File");
                return ;
        }
fprintf(BrRTS,"facts run\n");

fprintf(BrRTS,"%s\t%s\t%s\t%s\t%s\t%s\t",CharVEHYEAR, CharCC, CharAGEOFVEH,
CharTCFREQ, CharCTPR, CharNYNA);
fclose(BrRTS);


//The following two statements are to make the output file of the brainMaker empty.
OP=fopen("BrainRTS.out","w");
fclose(OP);

//The following statement calls  the BrainMaker Software and predicts the output for the
//input given in the input file "BrainRTS.in using the working NN model "BrainRTS.net"

spawnl(P_WAIT,"Brainmak.exe","BrainRTS.net","-b",NULL);
OP=fopen("BrainRTS.out","r");

for (i=1;i<=InputUnits;i++)
fscanf(OP,"%f",&Input);

fscanf(OP,"%f%f%f",&Medium,&Low,&High);

if ((Low>Medium)&&(Low>High))
{
        Prediction="Low";
```

```
        Confidence=(Low/(Low+Medium+High))*100;
}
if ((Medium>Low)&&(Medium>High))
{
        Prediction="Medium";
        Confidence=(Medium/(Low+Medium+High))*100;
}
if ((High>Medium)&&(High>Low))
{
        Prediction="High";
        Confidence=(High/(Low+Medium+High))*100;
}
OP1=fopen("Test.out","w");
fprintf(OP1,"%f", Confidence);
fclose(OP1);
OP1=fopen("Test.out","r");
fscanf(OP1,"%s", SConfidence);
fclose(OP1);

pWnd17->SetWindowText(Prediction);
pWnd18->SetWindowText(SConfidence);
fclose(OP);
}
}
```

Declaration

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

Tesfaye Hintsay

June 2001

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

Ato Ermias Abebe

Ato Million Meshesha