ADDIS ABABA UNIVERSITY
FACULTY OF INFORMATICS
DEPARTMENT OF INFORMATION SCIENCE

Predictive Data Mining Technique in Insurance:
(The Case of Ethiopian Insurance Corporation)

A thesis submitted to Graduate Studies of Addis Ababa University in partial fulfillment of the requirements for the Degree of Masters in Information Science.

By: Mesfin Fikre Woldmariam

June 2005

ADDIS ABABA UNIVERSITY
FACULTY OF INFORMATICS
PHOTOGRAPHIC LAB

ADDIS ABABA UNIVERSITY
LIBRARIES
P O BOX 1176
ADDIS ABABA ETHIOPIA

ADDIS ABABA UNIVERSITY
FACULTY OF INFORMATICS
DEPARTMENT OF INFORMATION SCIENCE

Predictive Data Mining Technique in Insurance:
(The Case of Ethiopian Insurance Corporation)

A thesis submitted to Graduate Studies of Addis Ababa University in partial fulfillment of the requirements for the Degree of Masters in Information Science.

By: Mesfin Fikre Woldmariam

June 2005

ADDIS ABABA UNIVERSITY FACULTY OF INFORMATICS BIBLIOGRAPHIC LAB

ADDIS ABABA UNIVERSITY LIBRARIES P.O.BOX 1176 ADDIS ABABA ETHIOPIA

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

ADDIS ABABA UNIVERSITY
FACULTY OF INFORMATICS
BIBLIOGRAPHIC LAB

PREDICTIVE DATA MINING TECHNIQUE IN INSURANCE: THE CASE OF ETHIOPIAN INSURANCE CORPORATION

BY
MESFIN FIKRE

Name and Signature of Members of the Examining Board

Ato Getachew Jemaneh, Chairman, Examining Board

Dr. B.L. Desai, Advisor

Ato Tesfaye Biru, Examiner

Chairman, Faculty	Signature	Date
Chairman, Graduate Council	Signature	Date

The soil is refreshed when sown with successive changes of seed, and so are our minds if cultivated by different subjects

The Letters

Pliny the Younger

Dedication

I dedicate this thesis to my parents

ACKNOWLEDGMENT

First and foremost my heart felt gratitude goes to my advisor Dr.B.L.Desai for his valuable support. I might not come to the finalization of this work by this time had his timely support wouldn't have been added.

My special thank goes to Ato Mesfin Tegaye, Supervisor of General Accident claim, Ato Girma head of records section and Ato Lulie staff member of Ethiopian Insurance Corporation for thier expertise and friendly support in discussing some doubts during data collection and attribute selection time.

In addition to this I would like to thank w/o Fetlework Fikre, w/o Haregewoyne Tesfayie, Daniel Mamo, Mestawot Girma, Herute Mamo, Rahel Mamo and many others for their concern and encouragement.

Finally, I am grateful to all SISA communities for their social and academic relationships.

Table of Content

Dedication.....i
Acknowledgement.....ii
Table of Content.....iii
List of Tables.....vi
List of Figures.....vii
Abbreviation of terms.....viii
Abstract.....ix

CHAPTER ONE: BACKGRAUND 1
1.1 Introduction.....1
1.2 Statement of the Problem.....6
1.3 Objectives of the Study.....9
1.3.1 General Objective.....9
1.3.2 Specific Objectives.....9
1.4 Method of the Study.....10
1.4.1 Literature Review.....10
1.4.2 Data Collection and Processing..... 10
1.4.3 Model building.....11
1.4.4 Model evaluation.....12
1.5 Application of the Result..... 13
1.6 Scope and Limitation of the Study.....14
1.7 Organization of the Thesis.....15
CHAPTER TWO: DATA MINING AND KNOWLODEGE DISCOVERY 16
2.1 Introduction.....16
2.2 Data Mining Processes.....16
2.3 Data Mining Functionalities.....18

2.3.1 Classification.....19
2.3.2 Regression.....19
2.3.3 Affinity Grouping.....20
2.3.4 Deviation Detection.....20
2.4 Data Mining Techniques and Algorithms.....21
2.4.1 Clustering.....21
2.4.2 Neural Network.....22
2.4.3 Decision Tree.....23
2.4.3.1 Decision Tree Building.....24
2.4.3.2 Appropriate Problems for Decision Tree Learning.....25
2.5 Applications of Data Mining.....26
2.5.1 Using Data Mining in the Insurance Industry.....27

CHAPTER THREE: INSURANCE IN ETHIOPIAN INSURANCE CORPORATION. ..31
3.1 Organizational Profile.....33
3.2 Major Insurance Covers in EIC.....34
3.3 Personal Accident Policy.....35
3.3.1 Scope of cover.....36
3.3.2 Policy (Benefits) classification.....36
3.4 Group personal Accident policy.....38
3.5 General Exceptions in the Cover.....40
3.6 Underwriting Practice in EIC.....41
CHAPTERFOUR: DATACOLLECTION AND PREPARATION.....45
4.1 Introduction.....45
4.2 Data Collection.....46
4.3 Data Preparation.....49.
4.3.1 Data Transformation.....50
4.3.2 Attribute Selection.....53
4.3.3 Derived Attributes.....54

CHAPTER FIVE: EXPERMENTATION AND ISCUSSIONS.....57

5.1 Introduction..... ... 57
5.2 Building Model..... 59
5.3 Issues in Model Selection..... 68

CHAPTER SIX: CONCLUSIONS AND RECOMMENDATIONS 72

6.1 Conclusions..... 72
6.2 Recommendations..... 74

REFERENCES..... 75

APPENDIX..... ... 78

Portion of deployed Decision Tree..... 78
Generic Rules generated for the Tree.....81

List of Tables

Table 1: Identified and collected attributes.....48
Table 2: Attributes used in the experimentation.....56
Table 3: Confusion matrix for experiment 1..... 61
Table 4: Balanced partitioning based on "PAcc_Within" attribute..... 65
Table 5: Confusion matrix for experiment 2.....67

List of Figures

Figure 1: KnowledgeSTUDIO data partitioning dialogue.....59
Figure 2: Decision tree predictive model for experiment 1.....60
Figure 3: Decision tree for experiment 1.....62
Figure 4: Balancing data partitioning dialogue.....64
Figure 5: Decision tree predictive model for experiment 2.....66
Figure 6: Decision tree for experiment 2.....66
Figure 5: Decision tree predictive model for experiment 2.....66
Figure 7: Manually expanded decision tree.....71

Abbreviation of Terms

DM	Data Mining
DT	Decision Tree
EEPCO	Ethiopian Electric Power Corporation
EIC	Ethiopian Insurance Corporation
ETC	Ethiopian Telecommunication Corporation

ABSTRACT

One of the important tasks that we face in 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 of the many areas, which uses classification as problem solving approach.

In order to solve problems whose solution could be categorical, we must build classification models. Data mining techniques have powerful tools for building models thereby addressing the problems.

This research study addresses the issues, techniques, and feasibility of building and deploying predictive model (s), which determines the risk exposure of individuals, i.e. the study was based on claims data of Personal accident at Ethiopian Insurance Corporation.

Ethiopian Insurance Corporation classifies claims into two classes as Small claim and Big claim. To meet the objectives of the research, 1,600-dataset records, each records having 23 attributes had been collected. But after the data is preprocessed, the total dataset records used for this study were reduced to 1543. And among the 23 attributes, 6 of them were selected by discussing with the insurance experts for final model building.

A close examination about the distribution of the data reveals that the data has an imbalanced distribution, which affects the accuracy of the model in favor of the dominant class, in this case the "Small" class. Thus, in order to solve this problem, dataset balancing based on "PAcc_Within" was taken and the result has shown that the accuracy was improved by far.

Beside this, the researcher has found that the frequency occurrence (class distribution) of the values of an attribute has a great impact on the accuracy of the model.

Lastly, analyzing the economic impact of the models and a separate accuracy measures for each class of risk category is used in order to compare the models and select one for deployment. Accordingly a model, which was built using knowlodgeSEEKER algorithm and balanced data partitioning, based on "PAcc_Within" attribute was selected as a final model. This model has an overall accuracy of 96.61% and a classification error of 0.29% and 27% for Small and Big risk claim classes respectively.

CHAPTER ONE

BACKGROUND

1.1 Introduction

Insurance Industry is a service-giving industry providing risk transfer from the insured to the insurer called Insurance Company. With the passage of time the industry is undergoing profound changes as a result of globalization, deregulation, and technology. In essence, it is moving from a product oriented to a customer oriented service provider. This is to mean that as with the advancement with technology and passage of time Companies are trying to provide customers exactly with what they are looking for than providing one product to all customers. This concept is called customer oriented.

Deregulation allowed non-insurance companies, such as banks, retailers, to enter the insurance market and compete against the established firms. At the same time, deregulation also allowed insurance companies to enter different financial markets, such as investment banking, requiring the development and marketing of new products and identification of potential customer groups.

Globalization accentuated with this change, insurance companies are now facing competitors from many parts of the world, many of which move aggressively into new markets trying to attract customers by allowing low rates and improved services.

Technology further added to the momentum of change, mostly due to different aspects of computerizations. The result again is an increased focus on the customer, his needs, and the tailoring of insurance policies to fit those needs on a profitable basis.

Given the current state of the insurance landscape, carriers competing for customers, dealing with rising claim rates and high loss ratios, as it could be seen in this study, the need have been greater for insurers to establish and maintain a complete view of their current and prospective policyholders.

A key part of this complete view is the specific claim risk that each customers poses to the company. Evaluating claims propensity gives insurance providers an objective way to adjust premium rates based on relevant, customer-specific information. Accurately predicting claims propensity throughout an organization can dramatically improve loss ratios and enables providers to charge lower premiums to low-risk clients (principle of individualization) as presented by SAS Institute (2003).

To improve loss ratios through more effective pricing and marketing, insurance providers need a solution that gathers relevant claims data from every corner of the enterprise as well as third-party data sources to provide both claims propensity predictions and claim size predictions

This trend toward an individualization of services and policies is supported and implemented by increasingly sophisticated business software. Relevant data storage and retrieval of customer policy and business data in a database and data warehouse and the use of these data in data mining systems is becoming a must for today's insurance firms.

With advancement in information technology, the amount of data stored in insurance databases is increasing rapidly (Vesanto 2002). These huge databases contain a wealth of data and constitute a potential goldmine of valuable business information. As new and evolving loss exposures emerge in the ever-changing insurance environment, the form and structure of insurance databases change. Finding the valuable information hidden in those databases with high-dimensional vector and identifying appropriate models is a difficult task.

Insurance Corporations although are service giving, they have to make some amount of profit for existence, expansion (product diversification) and for growth. But their profitability depends on the accuracy of risk assessment and ability of predicting the future. Insurance underwriting experts use their own subjective judgments in assessing the risk exposure of clients thereby affecting the profitability of the organization. Their primary goal is to protect the company against high claim cost due to predictable risk of policies.

Clients of Ethiopian Insurance Corporation can be either new customer or existing customers. From these customer categories, underwriters argue that renewing the policies for existing customers is considerably more cost effective than taking a new customer. But the renewal process is time consuming for the fact that they are required to manually analyze many risk factors for many clients. Thus underwriters of the Corporation demand the process to be an automatic process.

In addition to this an increasing size of the customer profile added with the small number of the company's employee, made it difficult for the company to regularly audit renewal acceptance manually.

One alternative solution to these problems is to classify policyholders based on their degree of risk exposure using data mining techniques; which enables the Corporation to predict the risk level of a policyholder by discovering the relationship between attributes from past database.

But what is data mining?

According to (Thearling 2003), data mining is the extraction of hidden predictive information from large databases; it is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make *proactive, knowledge-driven* decisions by answering business questions that traditionally were time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations, thus overlooking datasets which have valuable business information.

But by using today's data mining technology, it was possible to analyze massive dataset records for an information easily rather than throwing away. Penzias (1999) states that "Data mining will become more important, and companies will throw away nothing about their customers because it will be so valuable. If you are not doing this, you are out of business." In business there are clear benefits for investing in data mining.

As data mining provides results that improve business processes and thereby increase profits, it is becoming a competitive weapon in business industries.

Given a database of sufficient size and quality like insurance databases, data mining technology can generate new business opportunities by providing the following capabilities:

- Automated production of trends and behaviors, like insurance fraud;
- Automated discovery of previously unknown patterns, like patterns between work category and level of risk;
- To analyze massive databases in short minutes. Faster processing means that users can automatically experiment with more models' to understand complex data. Analyzing larger databases, in turn, yield improved predication.

In other words, in larger databases, where the depth and breadth is too high, one can easily understand how difficult it will be if processing is done manually. Because more columns, means that the analyst must often limit the number of variables he/she examines due to time constraints. Yet variables that are discarded because they seem unimportant may carry information about unknown patterns. But by using data mining tools, users can explore the full depth of database, without pre-selecting subsets of variables and decrease the associated lose of information.

Therefore, the advancement in the capability of technology added to its reduction in cost lets business organizations to change their data in to asset, following a profitability equation of:

Data ⟶ Information ⟶ Knowledge ⟶ Better Decisions.

But what can business do to ensure that their use of data follow this path? Data mining technologies should be applied throughout the process to guarantee this success, (Mangold 1998). How data mining works?

How exactly is data mining able to tell important things that one didn't know or what is going to happen next? The technique that is used to perform these feats in data mining is called *molding.* Modeling is simply the act of building a model in one situation where one knows the answer and then applies it to another situation that he/she don't.

Therefore, this study is was concerned with problem of realizing predictive models based on knowledge of associations in the context of insurance dataset. Like: work environment – risk of death.

1.2 STATEMENT OF THE PROBLEM

From a simple preliminary observation about the performance of the Corporation for one year (2002/2003) and brief discussion with the insurance experts, the Corporation has earned a growth revenue of 83.8% from non-life insurance coverage, 11.1% from life, and 5.1% from other businesses. On the other hand it has paid claims of 83.6% for non-life insurance cover, 13.5% for life and 2.9% for other businesses. From this figures one can easily understand the fact that non-life insurance coverage is taking the largest proportion both as a source of revenue for the Corporation and also as the largest expenditure to the Corporation, which catches the researcher's attention for further analysis.

Non-life insurance covers has many different types of covers among which Personal Accident policies that take the line share both as a source of income for the Corporation and as the largest amount of expenditures to the Corporation (as seen above). As a result of this, the researcher has focused on the analysis of different claim data of Personal Accident policies in the Corporation. With respect to Personal Accident claims data analysis, the Corporation is classifying claims into either "Small" or "Big" claim risk category based on the amount of claims they pose to the Corporation. But with regards to predicting about the future uncertainties (expected claim to be paid), the Corporation is not using its past dataset efficiently and effectively.

This means the Corporation isn't using its past data for future anticipation of the different risk levels and do not categorize claims according to reason of request (benefits sought) which could has helped it in identifying the type of benefits sought most and the type of benefits which are not sought most (frequently). This in turn could have helped the Corporation to adopt the "right " marketing strategies, which tailors, to individual clients' behavior called principle of "individualization of service". As a result of this, the Corporation has got problem of reserving an approximate amount of fund as a contingency for its claim request from clients working in different work category thereby affecting its budgetary plan and operation.

But it is possible to get decisive information about the future by studying the different risk factors of clients and prepared accordingly well in advance, which is one of the strategies in today's competitive business environment. This anticipation is possible through data mining technologies, which helps one to have the complete picture of its clients' behavior. But because of the many dimensional risk variables kept by the Corporation added with the manual data handling and processing approach, the Corporation couldn't be able to operate efficiently and effectively.

- Underwriters in the Corporation use their own personal judgments (based on past experience) in forecasting out standing claims, which clients are expected to come with based on the clients' work category or occupation. This subjective method of risk assessment in estimating the outstanding claim amount and setting premium rate is important because the experts utilize their current knowledge. However relying only on experts' judgment and current knowledge is not sufficient for predicting out standing claims for the fact that there are many risk variables that can be used to analyze outstanding claims. But as the number of variables increase it becomes difficult to understand and/or analyze hidden relationships between variables. From this, it is evident that there is a need for building (developing) predictive model that assists the workers of the Corporation in estimating the claims expected from each clients' based on their work category.

- The Corporation is classifying claims based on amounts they claim for irrespective of their work category. But classifying clients based on their claim amount and work category is possible by analyzing the characteristics and relationship of claim amounts to work category of customers, reason of claims like death, permanent total disablement, total temporarily disablement and medical expenses of Personal Accident using data mining techniques. Analyzing such variables will help to find what category of workers claim for which type of the above benefits, the level of risk associated with the above benefit types and etc. It was believed that such analysis could be realized by automatic means.

Lastly, as the assessment of the collected data, shows at about the termination of the insurance policy cover, most of the customers of the Corporation have been seen claiming for repayment of medical expenses which might be suspected as a sort of fraud when they don't want to renew it and some times clients claim for an insurance coverage as soon as they have got the cover which is suspicious to be fraud.

Therefore, the underwriters of the Corporation want to study whether the clients have some health problems prior to getting the insurance coverage or not. This information helps them to accept or reject the contract from such clients.

1.3 OBJECTIVE OF THE STUDY

1.3.1 General objective

The general objective of this study is exploring the potential usefulness of data mining techniques in large dataset organizations, particularly, Ethiopian Insurance Corporation in assisting the workers of the Corporation to facilitate the policy renewal process of Personal Accident Policies

1.3.2 Specific objective

- To explore the specific areas where data mining technology can be applied in an insurance industry;
- To identify the nature of Risk and Risk assessment practices at the Corporation;
- To analyze the relationship between different risk factors like work category, level of risk, benefits sought and timing of the insurance claim;
- To build a predictive model (classification model) that helps underwriters to predict the risk category of clients by studying claims of past dataset records so that underwriters can reserve the estimated fund for different claims that may come to the Corporation.
- To study the most frequent type of benefits (death, total permanent disablement) sought by clients and the relationship with class of risk and work category.

1.4 METHOD OF THE STUDY

In order to meet the above objectives and elaborate the usefulness predictive data mining technology, the researcher has conducted the following basic activities:

1.4.1 Literature Review

Extensive literature review has been conducted on the role of data mining technologies in transforming raw data into information in general and the different application areas of data mining in insurance. The literatures reviewed are journal articles, Internet resources, books, and insurance related thesis works etc.

While reviewing these literatures, the researcher hasn't included a theory of the domain that allows mechanistic mathematical definitions of the model. Because the aim of this study is to build a model, the connection between inputs and outcome using observed associations between attributes of risk factors.

1.4.2 Data collection and processing

Personal Accident data has been collected from Ethiopian Insurance Corporation by using document analysis, and discussions with insurance experts in the department (section). The reason behind is that almost all the data is in hard copy, which requires the researcher's personal analysis and observation of the documents and digitizing the data has been also conducted. This was done for the fact that in order to apply data mining technologies every data has to digitized in to its equivalent electronic format. Then after the data is preprocessed for further analysis by the tool called knowlodgeSTUDIO. The preprocessing includes the following activities:
Dealing with outliers, data transformation, dealing with missing values and aggregating the required attributes has been done.

Discussion with experts was made in order to see the technical gap about data analysis technologies-like data mining, in order to know the important attributes for decision making, and the criteria used to classify claim records as "Small" and "Big". This is because understanding the available technologies and its capabilities is a stepping-stone for efficient utilization of the raw data.

1.4.3 Model building

In order to build the predictive model, many experiments have been done for different reasons. As the name implies building a model is different from building an application. In model building the algorithm is expected to learn the different patterns in the dataset and so that the learned knowledge by the algorithm could be applied on the new dataset. In order to analyze the dataset and build the above type of model, the different available tools are considered like SAS Enterprise Miner and Knowledge STUDIO software tool(s). But because of the researcher's familiarity with knowlodgeSTUDIO, its free availability of its trial version and for the fact that this tool includes the commonly used data mining techniques like decision tree, neural network, and clustering it was selected for this study. In this study the researcher has adopted Decision Tree technique for its visual expandability by users if they demanded it and it produces easily understandable rules for end user of the model. These rules can be transformed in to a software code for deployment.

Beside this Microsoft excel 2000, has been used to store the initial data so that the software tool can import and change the dataset into its own format for further processing.

1.4.4 Evaluation of the model

After the data is processed the predictive model has been built and tested for its accuracy. The built model was evaluated for its performance accuracy. Irrespective of its limitations *average accuracy* measure is automatically generated for each model in the experimentations for each predictive model.

But as this measure is not an effective approach for the data under study, which is characterized as highly imbalanced dataset distribution of two classes'-Small and Big risk classes, separate accuracy measure was calculated manually for each class under each experiments by taking the figures from confusion matrix. And the economic impact of the model was considered in order to compare the models for deployment.

1.5 APPLICATION OF THE RESULT

In this study an attempt has been made to explore the application of data mining technology, particularly claim prediction modeling in insurance company. The output of this research can be used as:

- A starting point for further research in this area and other related areas, and come up with a full-fledged predictive model to be used by the Company (these areas are discussed at the end of chapter two);

- It brings the available data mining technology and their potentials in filtering information out of raw data to the attention of the officials of the Company so that the gap of technical know how will be get minimized;

- By considering the decision tree generated and the rules produced, the company can give more attentions to some attributes which determines the characteristic nature of the customers and their risks under consideration;

- Evaluating claims propensity gives insurance providers an objective way to adjust premium rates based on relevant, customer-specific information and risk factors;

- Finally, this study can give hands on experience for the researcher for further research in the future.

1.6 Scope and Limitation of the study

In his way of conducting this research study, the researcher has limited his scope as follows:

- Rather than attempting to build predictive for many areas the study is limited to focus on only classification of claims by studying their relationship with work category.

- The study also addresses the different issues to be considered in building predictive models like accuracy measures, economic considerations, and the role of imbalanced data distribution on the accuracy of the models than discussing the mathematical logic behind predictive models.

The major limitations of this study are many, among which time and the nature of data takes the line share in a sense that:

- As the name indicates this study is concerning on the data from insurance industry where there is a high market competition for customers and where information about customers is confidential. This has an effect of not knowing the rate of premium charges the Corporation is using. In order to solve this problem the researcher has taken some measures like: looking for data attributes, which do not require knowing the details of premium charges for every individual in the data set. As a result, it was found that taking dataset of clients being insured by their employing organization is found to be promising for the fact that premium determination is in group base. This limits the distribution of the data set type and the different risk factors thereby affecting the thesis work.

- In addition to this, the data in the Corporation is manual and many of the fields are found in different forms, which makes data collection to be difficult and time consuming, as a result the data of one branch office of the Corporation is taken, which again limits the variations in the dataset distribution thereby affecting the research work.
- Lastly, because this research study was based on data collected on clients from two corporations, namely Ethiopian Telecommunication Corporation and Ethiopian Power Corporation, its application is limited to clients of these Corporations.

1.7 Organization of the thesis

The overall organization of this study is presented into six chapters as follows: **Chapter one** provides the background to the study problem, by discussing the introductory parts of the subsequent chapters so that the detail explanations and discussions to be made at respective chapters.

Chapter two discusses the different Data Mining Processes, DM Functionalities, DM techniques and algorithms like (Automatic clustering, Neural network, and Decision tree), and finally the chapter ends by discussing some of the application areas of data mining in insurance industry.

Chapter three provides an overview of risk assessment practice at Ethiopian Insurance Corporation and, policy renewal practices, and type of risk covers.

Chapter four discusses the data collection and preprocessing like handling missing values, outliers, data transformation and selecting "best" attributes for the experimentation phase for what has been discussed in the previous chapters (theme of the thesis)-model building.

Chapter five this chapter discusses the different experiments that have been done to build the predictive model. It provides some discussions on the tool was used, the algorithms and accuracy measuring approaches are discussed on top of the issues of class imbalance. Lastly;

Chapter six this chapter ends the thesis by providing the result of the experimentation and discussions in the form of summery called conclusions and recommends sighting potential areas to be done in the future, if any.

CHAPTER TWO

DATA MINING AND KNOWLEDGE DISCOVERY

2.1 Introduction

In this section the researcher has made an attempt to overview the data mining process, data mining functionalities, data mining techniques and algorithms, and their potential applications in general and particularly in insurance industry.

2.2 Data Mining Process

Data mining (DM) is the process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules (Berry and Linoff, 2000). A typical data mining process includes data acquisition, data integration, data exploration, model building, and model validation. Both expert opinion and data mining techniques play an important role at each step of this information discovery process. It combines techniques from Machine Learning, Pattern Recognition, Statistics, Database theory, and Visualization to extract concepts, concept interrelations, and interesting patterns automatically from large corporate databases. In order to attain these and other related knowledge discovery, passing the basic and fundamental steps is a prerequisite activity like:

Data acquisition.

The first step is to select the types of data to be used. Although a target data set has been created for discovery in some applications. DM can be performed on a subset of variables or data samples in a larger database called training set while holding back some of the datasets for latter validation of the model.

Data preprocessing

Once the target data is selected, the data is then preprocessed for cleaning, scrubbing, and transforming to improve the effectiveness of discovery. During this preprocessing step, researchers remove the noise or outliers if necessary and decide on strategies for dealing with missing data fields and accounting for time sequence information or known changes. Then data is transformed to reduce the number of variables by converting one type of data to another (e.g., numeric ones into categorical values) or deriving new attributes.

Building model

The third step of DM refers to a series of activities such as deciding on the type of DM operations; selecting the DM technique; choosing the DM algorithm, and mining the data. First, the type of DM operation must be chosen. The DM operations can be classified as classification, regression; segmentation, link analysis, and deviation detection (*see section 2.3 for details*).

Based on the operations chosen for the application, an appropriate data-mining technique is then selected. Once a data-mining technique is chosen, the next step is to select a particular algorithm within the DM technique chosen (*see section 2.4 details*). Choosing a data-mining algorithm includes a method to search for patterns in the data, such as deciding which models and parameters may be appropriate and matching a particular data-mining technique with the overall objective of data mining. After an appropriate algorithm is selected, the data is finally mined using the algorithm to extract novel patterns hidden in databases.

Result Interpretation and model evaluation.

The fourth step of the DM process is the interpretation and evaluation of discovered patterns. This task includes filtering the information to be presented by removing redundant or irrelevant patterns, visualizing graphically or logically the useful ones, and translating them into understandable terms by users.

In the interpretation of results, the researcher determines and resolves potential conflicts with previously known knowledge or decides to redo any of the previous steps. The extracted knowledge is also evaluated in terms of its usefulness to a decision maker and to a business goal. Then extracted knowledge is subsequently used to support human decision making such as prediction and to explain observed phenomena.

(Guo) stated that the four-step process of knowledge discovery should not be interpreted as linear, but as an interactive, iterative process through which discovery evolves.

2.3 Data Mining Functionalities

Once the dataset was prepared for mining, the next step is to define the scope and the subject of study. This is referred as choosing DM functionality (DM tasks). According to (Guo) the most common types of DM functionalities are: classification, regression, affinity grouping, and deviation detection. Classification and regression are useful for prediction, whereas affinity grouping and deviation detection are for description of patterns in the data. A DM application typically requires the combination of two or more DM operations.

2.3.1 Classification

The goal of classification is to develop a model that maps a data item into one of several predefined classes. Once developed, the model is used to classify a new instance into one of the classes.

Examples include the classification of customers based on the degree of risk they bring to the Corporation in association with their work category which alerts the insurance underwriters to adjust their premium rates.

By analyzing the amount of claims that the Corporation has accepted in the past, it can predict the future amount based on the work category of clients.

Beside these, through predictive classification, the Corporation could be able to assess the possibilities of fraud by analyzing the date the insurance cover commences and the date the customer came with a claim.

2.3.2 Regression

This operation builds a model that maps data items into a real-valued prediction variable.
Models have traditionally been developed using statistical methods such as linear and logistic regression. Both classification and regression are used for prediction. The distinction between these two models is that the output variable of classification is categorical, whereas that of regression is continuous valued (Berry & linoff 2000).

In this study classification model was discussed for predicting the classification of insurance customers based on the amount of claims they posses.

2.3.3 Affinity Grouping or Association Rules

Affinity grouping is used to establish correlation between attributes of database records. Its typical application is market-basket analysis, where the technique is applied to analyze point-of-sales transaction data to identify product affinities. Another application could be finding the relationships between work category and the amount of claims they pose to an insurance firm, which is the major study of this thesis.

2.3.4 Deviation Detection (Outlier Analysis)

This operation focuses on discovering interesting deviations. Deviation Detection is usually performed after a database is segmented to determine whether the deviations represent noisy data or unusual casualty.

Most DM methods discard outliers as noise or exceptions. However, in some applications such as fraud detection, deviation detection is often the source of true discovery and more interesting than the more regularly occurring once since deviations represent anomaly from some known expectation or norm..

Once the business problem, objective, and data mining tasks were identified the next step is looking for an algorithm and technique, which meets the required business objective.

2.4 Data Mining Techniques and Algorithms

The most prominent and the heart of DM process to accomplish many of the above-indicated tasks is called Predictive Modeling. In this technique the business issue under consideration (example assessing customer risk) is investigated using historical data, which allow an identification of the relevant business risk classes (as Small claim risk, or Big claim risk).

These characteristics involve the various properties of customer data and their relationships: Any new observation (client) will be assigned to a business class of the above type on the basis of risk characteristics. The results are statistically tractable with quantifiable reliability i.e. it is the essence of predictive modeling to develop the model on known data and to apply the model on new datasets.

Predictive models (classification) can be built by using neural network, decision tree analysis, and/or automatic clustering.

2.4.1 Clustering

Clustering techniques are employed to segment a database into clusters, each of which shares common and interesting properties. The purpose of segmenting a database is often to summarize the contents of the target database by considering the common characteristics shared in a cluster whereby it increases the similarities between objects within the same cluster and increasing the difference between different clusters. K-means algorithm to clustering is the common algorithm used by different DM software tools.

Clustering is recommended when:

The data analyst suspects that there are natural groupings that may represent groups of customers or products that have a lot in common with each other. These may turn out to be naturally occurring customer segments for which customized marketing approaches are justified.

More generally, clustering is often useful when there are many competing patterns in the data making it hard to spot any single pattern.

When combined with other techniques like affinity grouping, creating clusters of similar records decreases the complexity within clusters and so that other DM techniques are more likely to succeed with their task (Berry & Linoff 2000)

2.4.2 Neural Networks

Neural networks constitute the most widely used technique in data mining. They imitate the way the human brain learns and use rules inferred from data patterns to construct hidden layers of logic for analysis. Neural network methods can be used to develop classification, regression, link analysis, and segmentation models. A neural net technique represents its model in the form of nodes arranged in layers with weighted links between the nodes. Neural networks actually represent complex mathematical equations, in lots of summations, exponential functions, and many parameters.

One important characteristic of neural networks is that they are opaque (Guo), which means there is not much explanation of how the results come about and what rules are used i.e. they are the most widely known and the least understood of the major data mining technique (Berry & Linoff 2000).

Many machine learning and data mining researchers argue that neural networks do not allow well when there are many hundreds of input features. Large hundreds of features make it more difficult for the network to find patterns and can result in long training phases that never converge to a good solution. As a result of this for insurance organizations with a great depth of statistical information, neural networks are ideal because they can't identify and analyze changes in patterns, situations, or tactics for much number of numeric and continuous valued variables.

Therefore, in order to use neural nets effectively, they need to be combined with decision tree.

(Berry & Linoff 2000) suggested that neural networks can work well with decision tree methods for the fact that decision trees are good at choosing the most important variables and these can then be used for training a network, in his experimentation (Tesfaye 2002) has used decision trees for variable selection and feed the result to the net. And he found that the neural network has performed well irrespective of the fact that there are some objects, which the decision tree has classified correctly while the net has misclassified. He stated that the hybrid of the two models (technique) might improve the overall performance. This indicates how decision trees can optimize the performance of neural network through attribute selection.

2.4.3 Decision Trees

Induction techniques develop a classification model from a set of records - the training set of examples. The training set may be a sample database, or an entire data warehouse. Each record in the training set belongs to one of many predefined classes, and an induction technique induces a general concept description that best represents the examples to develop a classification model. The induced model consists of patterns that distinguish each class. Once trained, a developed model can be used to predict the class of unclassified records automatically.

Induction techniques represent a model in the form of either decision trees or decision rules. These representations are easier to understand, and their implementation is more efficient than those of neural network

There are two types of decision trees, namely classification trees and regression trees. Classification trees label records and assign them to the appropriate class by providing the confidence that the classification is correct. In such cases, the tree reports the class probability, which is the confidence that a record belongs to a given class. On the other hand regression trees estimate the value of a target variable that takes on numeric values (Two Crows Corporation 1999).

When a tree model is applied to data, each record flows through the tree along a path determined by a serious 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 the node in the training set.

2.4.3.1 Decision Tree Building

The first step in building a decision tree is to decide which of the independent attributes make the best split for the tree. The best split is defined as one that does the best job of separating the records into groups where a single class predominates. The measure used to evaluate the splitter is reduction in diversity or entropy. Most algorithms use the idea of information gain (Costa, 2000).

In choosing the best splitter at a node, the decision tree algorithms calculate the information gain for each input field and decide the best splitter. This process repeats until the tree is fully grown.

Tree building algorithms make their best split first at the root node where there is a large population of record. Each subsequent split has a smaller and less representative of the whole dataset with which to work.

With increasing the depth of the tree the model over fits the training set. An approach to this problem is called Pruning. Pruning is the process of removing leaves and branches to improve the performance the decision tree (Berry and Linoff 2000).

Decision tree techniques are chosen for their ability of generating easily understandable rules. It is certainly true that for any classified record it is easy to simply trace the depth from the root to the leaf where that record landed in order to generate the rules that lead to the classification and most decision tree tools have this capability.

2.4.3.2 Appropriate Problems for Decision Tree Learning

Decision tree are best suited to problems with the following characteristics (Mitchell, 1997):

- When attribute-value pairs represent instances like-Small or Big;

- When the target function has discrete output values. In such situation decision tree assigns a classification to each class as Small or Big. The simplest case exists when there are only two possible classes (binary classification).

- When the problem at hand is a disjunctive description in nature.

- When the training data may contain outliers and missing values i.e. decision tree learning methods are robust to the above problems - both errors in classifications of the training examples and errors in the attribute values that describe the cases.

- Decision tree methods can be used even when some training examples have unknown values.

Although decision trees have the advantage of allowing human experts to easily understand the solution of a problem by representing the flow of the solution graphically they are not without limitation. Perhaps the weakest point of decision trees in modeling is their sensitivity to changes in data, hence also to noise (Zied et al).

Therefore, which algorithm to choose for business problems, involving classification is difficult. Growe. A (1999) stated that there is a controversy over the relative merits of doing classifications using artificial intelligence (AI) tools such as neural net works versus employing statistical methods like Decision Trees. Computer Science researchers have been arguing that the nets are performing more than decision trees. But Growe (1999) hasn't seen this argument in his experimental study. He keeps on citing the work of Couvreur and Couvreur (1997) as follows:

'For us, decision tree and neural networks are complementary tools, with considerable overlap not only in their fields of application but also in their theoretical foundations...when compared fairly, neural network and modern statistical methods like decision tree perform similarly, both in terms of quality of results and in terms of computational costs. In some applications neural networks will outperform their statistical counterparts, in others they will not.'

So which technique to use for this study? Irrespective of the above controversies, from the nature of the dataset of this research and the potentiality of decision tree the researcher has used decision tree technique in this study.

2.5 Applications of Data Mining

Now a days DM is increasingly popular because of the substantial contribution it can make to firms. It has been used in a wide variety of business activities as it offers value across a broad spectrum of industries, (Meheretwork 2004).

Two Crows Corporation (1999) mentioned that telecommunications and credit card companies are identified as the two leaders in applying DM technology to detect fraudulent use of their services. For example Telecommunication Companies are service-providing companies, which provide their services not in a pre-paid manner as s result of this, they often suffer from insolvent customers who has used the provided services without paying their dues. Ethiopian Telecommunication Corporation is one of these companies, which is loosing considerable amount of money now a day's (Gashaw (2004). With regards to this, he has made an experimental research in predicting customer insolvency at Ethiopian Telecommunication Corporation. And he found that DM has a promising potential application in telecommunication industries.

Further more, they mentioned that DM is found to be fruitful in medical applications where it can be used to predict the effectiveness of surgical procedures, medical tests. In this regard, Shegaw

(2002) has built a model that can support in the prediction of chilled mortality at Butajira hospital, indicating the potentials of DM in medical areas.

From business organizations point of view, credit companies can leverage their vast warehouses of customer transaction data to identify customers most likely to be interested in a new credit product or fraudulent credit card use, determine credit card spending by customer groups. With the objective of building a model that can support loan-decision making process at Dashen Bank S.C, Askale (2001) and Meheretwork (2004) have investigated the potential application of DM in banking sector.

2.5.1 Using Data Mining in the Insurance Industry

From insurance industries perspective, data mining technology can often improve existing actuarial models by finding additional important variables, and by identifying interactions between risk factors. The improvement can take different forms like for example by using some software tools like knowlodgeSTUDIO the models will let insurance people to understand important risk factors like "age", "work category", level of education and etc so that they could charge proportional premium rate. But manually (or without such aid), it is too difficult for underwriters to estimate which factors are more important.

DM can help insurance firms make crucial business decisions and turn the new found knowledge into actionable results in business practices such as product development, marketing, claim distribution analysis, asset liability management and solvency analysis.

The researcher has made some literature review to see what has been done so far in insurance industry and come up with the following:

Underwriting Profitability Analysis (UPAS) System: This system is developed in IBM research Group by (Apte et al, 1998). The application embodies a new approach to mining Property & Casualty (P&C) insurance policy. This system is a rule-based system that utilizes probabilistic

approach to discover risk characterization by analyzing large and noisy insurance data. The benefit assessment of the results suggested that this methodology provides significant value to the P&C insurance risk management process (Apte et al, 1999).

Motor Insurance Renewal System (MIRS): This system was built as a partial fulfillment for the degree of Masters by (Tesfaye 2002) at Addis Ababa University. He has built a prototype for Insurance Company. His system was built by using neural network and decision tree techniques. In his study he has demonstrated the potential application of DM technology in motor insurance.

According to (Guo) specifically, DM techniques can be applied in the following areas from insurance industries point of view.

Establishing Rates

An important problem in actuarial science is concerning premium 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 is the rate.

Modern DM models (tools) can help companies to predict risk and thereby enable to set rates.
One critical question in rate making is knowing what are the risk factors or variables that are important for predicting the likelihood of claims and the size of a claim? 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 analyses.

Modem data mining models such as decision trees and neural networks can more accurately predict risk than current actuarial models, therefore insurance companies can set rates more accurately, which in turn can result in more accurate pricing and hence a better competitive position. In this line SAS Institute (2003) has made a comparable study on rate setting. The study has used premium rates, costs and other risk factors of clients. By calculating all the costs associated with each client, it is possible to set approximately objective rates for each clients.

This study was conducted by using motor claim datasets and SAS Enterprise Miner software tool and the result is seen to be promising.

Customer Level Analysis

Successfully retaining customers requires analyzing data at the most appropriate level, the customer level, instead of across aggregated collections of customers. Using the Associated Discovery DM technique, insurance firms can more accurately select which policies and services to offer to which customers SAS institute (2001).

Traditionally, simply expanding the efforts of sales department can increase market shares.
In contrast to this, DM strategies enable analysts in insurance industry for market expansions in the form of customer retention.
Database 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.

Identifying the target group for retention campaigns may be improved by modeling behavior of policyholders. With these techniques insurance companies can segment customer database to create customer profile, conduct rate and claim analysis on a single customer segment for a single product etc.

Developing New Product Lines

Markets change over time, and so do products sought by customers (Tesfaye 2002). To adjust to the market demand, insurance companies need to know what types of new policies will be needed and will be profitable from the market analysis point of view.

Insurance firms can increase profitability by identifying the most lucrative customer segments and then prioritize marketing campaigns accordingly. Problems with profitability can occur if

firms do not offer the "right" policy or the "right" rate to the "right" customer segment at the "right" time (Berry & Linoff, 2000).

With DM operations such as association analysis, insurance firms can now utilize all of their available information to better develop new products and marketing campaigns.

Reinsurance

DM can be used to structure reinsurance more effectively than using traditional methods. Data mining technology is commonly used for segmentation. 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 and not just the generalization that it is a long tailed book of business.

Estimating Outstanding Claims Provision

The settlement of claims is often subject to delay, so an estimate of the claim severity is often used until the actual value of the settled claim is available. The estimate can depend on the following factors:

- Severity of the claim.
- Likely amount of time before settlement.
- Effects of financial variables such as inflation and interest rates.
- Effects of changing social mores.

DM operations such as Link Analysis and Deviation Detection can be used to improve the claim estimation.

The estimate of the claims provision generated from a predictive model is based on the assumption that the future will be much like the past. If the model is not updated over time, the assumption becomes that the future will be much like the distant past. However, as more data become available, the predictive DM model can be updated, and the assumption becomes that the future will be much like the recent past.

Data mining technology enables insurance analysts to compare old and new models and to assess them based on their performance. When the newly updated model outperforms the old model, it is time to switch to the new model. Given the new technologies, analysts can now monitor predictive models and update as required.

An important general difference in the focus between existing actuarial techniques and DM is that DM is more oriented towards applications than towards describing the basic nature of the underlying phenomena. Instead, the focus is on producing a solution that can improve the predictions for future premiums. DM is very effective in determining how the premiums related to multidimensional risk factors such as customers' type of work and the level of risk he/she will bring to the insurance Corporation.

Detecting Fraudulent Claims

Fraudulent claims are an ever-presented problem for insurance firms, and techniques for identifying fraud are critical for the long-term success of insurance firms. In searching for fraudulent claims, analysts look for unusual associations, outlying in the data etc.

As far as the knowledge of the researcher is concerned, there are a few studies have been done so far in Ethiopian context. Tesfaye (2002) has done on Motor insurance risk assessment. Still his work doesn't cove any of the above application areas directly. But indirectly since all the above areas are interdependent (one area can be an input for the other area) building a model for one application area can be used to generate an input for the other. For example if one builds a rate establishing predictive model, the same model can be used for customer level analysis, new product development analysis, reinsurance, and estimating the outstanding claims. This interdependence is due to the commonly used risk factors, i.e. the risk factors used for rate estimation is used for outstanding claim estimation, reinsurance etc.

As a result of this, in this research study an attempt has been made to classify claim risks in two categories by studying the association between claim amounts with work category. The model built in the study has the ability to: predict the existence of fraud, assist in premium rates setting based on the claim requested, and also estimate out standing claims by studying the relationship between amount of claims and work categories. Therefore, this study addresses the potential application of DM technology in different application areas in insurance Corporations.

In order to see these potential applications, Ethiopian Insurance Corporation (EIC) has been selected for the reasons indicated in chapter four of this study. Therefore, it is natural to discuss about EIC's risk assessment practices and the type of insurance covers before collecting data for model building and experimentation as a result of this the next chapter is about EIC's risk assessment practices.

CHAPTER 3

INSURANCE IN ETHIOPIAN INSURANCE CORPORATION

3.1 Organizational Profile

Ethiopian Insurance Corporation (EIC) was established in 1976. The Corporation came in to existence with paid up capital of 11 million Ethiopian birr aiming with the objectives of engaging in all classes of insurance business in Ethiopia, and ensure the insurance services reach the broad mass of the people. Earlier, EIC was operating the business for about nineteen years under protected monopolistic system as state owned-sole insurer. But after the demise of the Marxist regime in mid-1991 a fundamental change has taken place and there was a shift in political, economic and social orientation from totalitarianism to that of liberalism. As a result, EIC was re-established as public enterprise in (1994) with a paid up capital of 61 million birr.

Currently, the Corporation has 6 main branches, 21 sub branches and 4 contact offices through out the country. Out of these main and sub branches, 6 main branches, 3 sub branches, and 1 contact office are located in the capital city of the Country i.e. Addis Ababa.

With regards to staff development, as of June 2004, the Corporation has 1135 employees with different educational background. Both the management and staff are highly skilled professionals with outstanding academic achievements as well as professional insurance qualifications and many years of practical experience in the insurance industry.

From organizational structure and management perspective, the Corporation is administered by the Board of Management, which reports to the Public Financial Institutions Agency. In a new derive for excellence of service, the Corporation recently embarked upon a comprehensive program of re-engineering its business and restructuring its organization. As a result of re-organization, the new Top Management Team now comprises the Managing Director, plus four

Deputy Managing Directors respectively responsibility for: Operations, Human Resource Management and Property Administration, Finance and Investment, and Marketing.

3.2 Major Insurance Covers in EIC

Ethiopian Insurance Corporation has two major types of insurance covers namely: Non-life insurance and Life insurance covers. Looking the different classes of risk covers the Corporation is providing is believed to be a stepping stone for the subsequent discussions

Non-life Insurance

This policy accounts for the largest percentage of the total policies in a sense that they are taking the line share of both the revenue and expenditure sections of the financial statements of the Corporation. This cover includes the following basic types:

- Aviation (Cargo & Hull)
- Burglary & House breaking
- Bonds
- Consequential Loss (Business interruption)
- Fidelity Guarantee
- Fire & Allied Peril
- Householders Comprehensive
- Legal expense
- Personal / Group Personal Accident
- Rural
- Marine (Cargo & Hull)
- Motor
- Money
- Plate glass
- Professional Indemnity
- Public Liability
- Travel Coupon
- Workmen's Compensation
- Inland Carrier Liability
- Goods-in-transit
- Product Liability
- Horticultural-plantation insurance

Life Insurance Covers

Unlike that of non-life insurance coverage policies, life insurance policies take a lesser proportion in generating revenues for the Corporation and expenditures relatively. This class of cover includes the following:

- Endowment
- Anticipated Endowment Policy (with Profit)
- Children's Education Policy (with Profit)
- Endowment Annuity
- Medical
- Individual Medical Assurance
- Group Medical Assurance
- Term
- Individual Term life Assurance
- Regular Term life Assurance
- Modified Large group Term life Assurance
- Mortgage Protection Assurance
- Whole Life Assurance

As stated in the previous sections, this study focuses on claims dataset that the Corporation has handled on Personal and Group Personal Accident policies. Therefore, before collecting data on this policy for further study, it is natural to discuss some basic points on the policy.

3.3 Personal Accident Policy

Personal accident insurance is used to provide compensation in the event of death by accident, and to mitigate loss of income and any additional expenses incurred if an individual is disabled from attending to his usual work. Cover can also be arranged to provide for disablement due to sickness.

3.3.1 Scope of cover

The term personal Accident insurance offered by the Corporation refers to various forms of covers of which the most common are:

Accident only policies

These policies cover the consequence of accidental bodily injury due to specific circumstances like:

- Occupational accident
- Non-occupational accident
- Part time, hobbies or sport, Travel, and Criminal Violence

Accident and all sickness policies

In addition to the cover provided by accident only policy, disablement due to sickness or disease is also covered. The cover does not apply to illness contracted within 21 days of the commencement of the period of insurance. 21 days is referred to as incubation period for most illness. The exclusion prevents selection against the insurer by proposer who may have already contracted an illness. But anticipating the existence of health problem before 21 days is difficult for the underwriters thereby creating a room for fraud.

3.3.2 Policy (Benefits) classification

This section discusses the fundamental and valid causes of the injury and the reason of claim on which data for this research is collected i.e. the policy normally include a scale of benefits (reasons) to provide a sum payment of money following an accident that resulted in:

- Death
- Total loss of sight in one or both eyes
- Total loss of one or both limits
- Permanent total disablement (PTD)

It is natural that when a certain injury occurs, the claim may not be payable until 12 or 24 months after an accident has occurred and must be certified by a medical doctor. Beside this, it may take at least that period of time to determine weather the disablement is both total and permanent. Immediately after an accident, it may be difficult to make a definite prognosis and the insured person will probably qualify for weekly benefits under the "Temporary total disablement" heading. If the injury does not show any progress or improvement he/she may fall under a title called "Permanent total disablement".

Permanent total disablement can be defined in a number of ways according to the extent of cover, which insurers wish, to provide. The most common definition is:

Permanent total disablement from gainful employment of any and every kind: in this case it means that the injured client was unable to get back to his usual task as the result of the accident. He/She was said to be permanently disabled from any task. Or (PTD) Permanent total disablement from the insured's usual occupation and any other occupation from which the insured person is fitted by knowledge and training. This means that also the injured was incapable to work on his usual task; he might be capable to engage in other task. Here with such type of variations in cover the premium charge also varies.

♦ Temporary total disablement (TTD)

This applies to the insured person if he/she is unable to engage himself/herself from his/her usual occupation due to accident or sickness. This disablement is of a total type in a sense within certain period of time the injured was incapable to do any thing. Compensation in the form of weekly benefit is usually payable for a maximum of 52 weeks. An excess of seven days may be applied to payment of benefit for accident and sickness.

♦ Medial expenses

These are expenses, which are necessarily incurred for treatment following an accident. They include cost of medical, surgical, prescription made by qualified doctor or hospital and ambulance changes. The affected clients can get a refund for such type of medical expenses up on presentation of payment receipt to the Corporation.

♦ Legal expenses

Some insurers provide this additional benefit. These are expenses necessary to pursue actions against negligent third parties following bodily injury.

♦ Family expenses

Many insurers now offer polices which provide personal accident and /or sickness cover for the insured and all members of his family. Demand for its type of cover has increased as individuals have become more aware of the medical costs following an accident or sickness to housewife or a child. The normal exclusion also applies to family policy.

3.4 Group personal Accident policy

This policy type is almost similar to the above but except that the benefit is paid on condition that the accident is group based. Under such types of policies, one person, firm or association affects this type of policy to insure the members of the group. There is a condition in the policy that no insured person shall have an independent right of action against the insurer.

The basic difference between Personal and General Personal Accident is that in the case of Group Personal Accident the participating member may contribute the whole or part of the premium through the collecting agent, or non-contributory as with employees who offer the benefit of group scheme as an attraction of his services.

Benefits of the policy

Often the normal full scale of benefits is provided, but sometimes, for example, for younger people, the death benefits are scaled down considerably since they are unlikely to have dependents. On the other hand permanent partial or permanent total disablement benefits are raised.

Insurers are reluctant to pay full wages for temporary total disablement without due regard to the compensation to which the injured person may be entitled from the state. Full wages payment from the insurers in addition to any compensation from the state may easily act as a deterrent against resumption of work as soon as possible. For this reason insurers will not accept proposal form under which the employee may obtain more than the balance between such compensation and his/her normal full wages.

Both the insurer and the insured persons benefit from the issue of a group or collective policy in a sense that the insurer's expense is kept to be a minimum and the premium rate for the group is lower too.

Also there is other difference, there is one common proposal form, one renewal notice form, and one set of record. Usually, too, there is no selection against the insurer for all members of the group are required to insure. As a result the insured will benefit from reduced premium while the insured person benefit from wider cover. Specifically the benefits can have a range of different types like:
If the accident results in a death or loss of limbs or legs the annual salary or the agreed capital sum of money will be paid. However if the accident results in Permanent total disablement or Temporary total disablement, the sum of money equal to one half of weekly earnings will be paid.

.

3.5 General Exceptions in the Cover.

As the activity of insurance corporations is accepting the risks of second parties for insurance cover, the coverage is always subjected to some preconditions agreed upon the proposal form. For either of the above policy the cover would normally be excluded in respect of:

- The insured person being under the influence of or being affected by alcohol
- The insured person willfully exposing himself to needless peril (except in as attempt to save human life) or the insured person committing or attempting to commit suicide, self inflicted injury or disease is also excluded.
- The insured person suffering from any physical defect or infirmity which existed prior to an accident
- The insured person engaging in aviation, otherwise than air travel which shall mean mounting on or dismounting from and traveling in any fully licensed passenger aircraft or as a passenger and not as member of the crew.
- Child birth, pregnancy, venereal disease or AIDS
- Sickness occurring within 21 days of the commencement of sickness cover. This is to exclude ailments, which the insured has contacted prior to the insurance contract effected and is designed to prevent selection against the insurer.

3.6 Underwriting Practice in EIC

The Corporation and the insuring (firm) i.e. employer of the clients will make an agreement based on the terms and conditions specified in the proposal form. Proposal form is a document containing fundamental points for agreement between parties. Some of the questions included in a Personal Accident Proposal form are occupation, age, weight/height, and accident within the previous 5 years.

♦ ***Occupation***: The insurer requires to know whether the pro-poser is engaged in manual work, self-employed or an employee. If he has more than one occupation details of each will be needed. This is required basically for premium determination and accordingly the Corporation categorizes occupations into four main classes as:

Class I: No accident/ health risk professional, administrative and clerical staff

Class II: Slight accident/ health risk skilled occupation with moderate level of manual work, semi-skilled occupation with small amount of manual work

Class III: Moderate occupation/ health risk skilled occupation with predominant manual work, semi-skilled occupation with moderate level of manual work

Class IV: Appreciable accident/ health risk physically strenuous work, semi-skilled occupation with predominant manual work. (This class contains basically Contract and Daily Laborers).

From this research study point of view, it is occupation category, which plays significant role in determining premium charge. Accordingly, the work category or occupation takes the most important attribute.

♦ ***Age (birthday):*** Generally speaking the risk of accident and illness increase with age. Elderly persons are more likely to contract illness than young people and are prone to accident by

stumbling or falling. Fractures are more easily sustained because bones become more reckless, and this can affect the accident risk. They are also more likely to suffer from infections diseases. But the Corporation will not take in to consideration about the age of clients coming from public organizations. This is because the employing organization themselves follow the age of every clients for their own purpose i.e. to calculate the retirement date. But age is important for clients coming privately (buying insurance cover for themselves).

♦ ***Weight/ Height:*** The weight and height will be compared with a table of the average of sex and age. Further investigation will be made if the proposer is 20% or more over or under the average.

♦ ***Accidents within previous five years:*** The existence of previous accidents may be indicative of potentially increased claim frequency, which causes a loss to the Corporation. Such transfer of risk has to be prevented by the Corporation by taking some measures. Example because of the nature of a particular activity or occupation an injury may claim for an insurance coverage frequently or he/she might leave for a long absence, which again passes the burden to the Corporation.

Beside the effect of high claim frequency stated above, if the client has been affected severely by previous case, he/she has more probability to bring a larger amount of claim for which the Corporation couldn't have been requested. Therefore, accessing past medical history of the clients is an important factor for the Corporation.

As can be seen above, if the clients are coming from public organizations, being insured by their employers, the Corporation uses the above work categories to determine premium rates. For example clients under Class I have relatively small risk exposures compared to clients under Class IV. Accordingly, in issuing the agreement policy, the Corporation calculates the cumulative 5 years salary of clients. This cumulative is important for the Corporation to limit the amount of claim to be requested by clients. With in the policy time period, clients can't request a sum total of more than their 5-year salary within the insurance period. For example, a client may present 10 or more claims to the Corporation within the insurance period and get a cover, as far as the sum of these claims is not more than the sum of his/her 5-year salary. Therefore, the

Corporation has to continuously update claims data files for clients and let them be informed how much they are left with. But this is difficult when individual file updating is done manually for many clients.

Beside this, the Corporation alerts its clients 2 months before the insurance period is about to expire. Accordingly, clients can accept the renewal as it is or with some changes like change in work category which requires the Corporation to revise the previous premium rate. Again this is a difficult task for the Corporation.

When an accident occurs, the injured client will fill a form called claim notification form in triple copies and submit to the Corporation. Then the underwriter examines the previous claims paid to this client to see if the sum total is more than the client's 5-year salary (as stated above). If the previous amount is not more than the required amount, the Corporation starts to pay the claim and if not the requested claim will be rejected.

Such procedure has a lot of shortcomings like:

1. Problem of underestimating out standing claims

The Corporation may not have enough funds to settle every claim, because many "Big" claims may be presented to the Corporation unexpectedly in a period, which affects the Corporation's financial status. But, if the Corporation has identified the source and amount of "Big" and "Small" claims, it can reserve an approximate amount so that the above problem may be reduced.

Moreover, classifying the clients according to their amount of claim level they bring (big or small), and work category will facilitate the claim settlement process, estimating the expected amount to be claimed from clients of each work category, and to revise the premium rate accordingly.

2. Problem for existence of fraud

The other problem is that some clients are seen attempting to request an insurance coverage as soon as they get the insurance cover. Studying this behavior for clients is very important for the Corporation so that they could avoid such suspicious to fraud. This is to say that they could attempt to transfer previous risks to the Corporation for which it is not liable. The same holds true while they are at about the termination of the contract. They claim for some medical expenses when they want to terminate the contract and do not want to continue for the next time.

These and other related problems are believed to be resolved by using data mining technologies (basically by classification). This could be done by classifying the clients according to their level of risk exposure (work category), the type of benefits they are looking for the type of clients who are claiming at the start and end of the contract and etc.

Once this classification is done different marketing strategies can be implemented like: market differentiation (segmentations) retaining clients who are "best" for the Corporation, charging different premium rate according to risk level, work category etc.

So in order to approach the above problems, the relevant dataset has to be collected and analyzed. Accordingly, the next chapter deals about data collection and preprocessing.

CHAPTER FOUR

DATA COLLECTION AND PREPARATION

4.1 Introduction

In the previous chapter of this study the researcher has made an attempt to give an overview of the different risk covers and somewhat in a detailed manner about the type of cover practice of EIC for Personal Accident Policy. Accordingly claims data for the above "benefits listed" is collected from the above said Corporation to meet the objectives and problems stated in this study.

As this chapter discusses the basic and fundamental stage of the DM processes that have been presented earlier, it was argued that 60% or more of the DM effort goes into preparing the data, which is the process of extracting the data from its original source, adding calculated columns, performing joins, generating new attributes, and cleaning the data. But in this experimental study more than 60% of the effort has gone for data preparation. This has happened for the fact that the data at the Corporation was manual, different fields are at different forms which requires the researcher to digitize the data and collect into one common form.

Before data is prepared for mining the source and type of data to be collected, approaches of preparing the data for mining, selecting attributes for building the predictive model are discussed in this chapter.

4.2 Data Collection

Like any other DM functionalities the most important requirement is the availability of appropriate data both in its size and its quality. In the context of insurance policies and claims, information regarding the policyholder, risks covered, cumulative risk behavior, and variety of additional factors that may be relevant for the modeling purpose is collected from the Corporation. North Western Branch Office of Ethiopian Insurance Corporation has been selected randomly as a source of datasets. This branch is one of the largest branches found in Addis Ababa.

Personal Accident policy clients of the Corporation are basically employees of public firms like: Ethiopian Telecommunication, Ethiopian Electric Power Corporation, Ethiopian Road Authority, and many more. But when we assess the ratio (size) claims coming to the Corporation most of them are from Ethiopian Telecommunication Corporation and Ethiopian Electric Power Corporation. Therefore, in order to meet the objective of this research, claim data records of Ethiopian Electric Power Corporation and Ethiopian Telecommunication Corporation is collected

Beside these from the nature of the activities these Corporations are engaged in, they have many employees that could be classified into different classes based on some attributes like: work category (A, C, D, and 2\24), which enables the researcher to study the relationship between level of claims and work categories.

For the purpose of premium rate determination, the Corporation classifies clients in four groups as indicated above. Workers under category "A" are employees who are permanent employees but have an insurance coverage while they are on duty. Any accidents occurred while they are not at work is not covered by the Corporation. Employees under category "C" are contract employees and have an insurance cover as long as they have got an injury with in the contract time period. "D" category stands for daily laborers who also have an insurance cover provided that the accident has occurred while the client is at work.

On the other hand, employees under category "2\24" are permanent employees who have got 24-hour insurance coverage. As the result of this, EIC can charge uniform amount of premiums for clients working under similar work category. This in turn makes the researcher not to opt for knowing the rate of premium charge of each client, which remains confidential to the Corporation.

As stated in earlier chapters, all the data at the Corporation is manual, which causes the data collection process to be time taking and tiresome. As a solution to these problems, the researcher has collected Personal Accident claims data from one branch office of the Corporation. The collected data size from this branch office accounts for about 1,600 records each has 23 attributes as indicated in table 1. This size of data is believed to be sufficient to train machine learning algorithms like decision trees.

With regards to the format, feasibility, and accuracy of the data is concerned, all the dataset was collected from the file cabinet of the branch for each records manually in one common form by contract workers. This was a time consuming and tiresome task ever seen in this study for the fact that some attributes are in different forms. During data collection the experts are checking for the consistency of the data while it was being gathered on one common format to avoid discrepancies among datasets, if any. This implies there is no question of data feasibility.

Attribute Name	Descriptions
Claim Number	A number to identify records uniquely
Insured birth date	In a dd/mm/yy order format
Work_Catagory	The category of work one is working
Branch number	Branch office number
Branch name	Name of the branch office
Beginning insurance date	The date the policy start to be valid
Beginning insurance month	The month the policy start to be valid
Beginning insurance year	The month the policy start to be valid
Ending insurance date	The last date the policy will expire
Ending insurance	The last month the policy will expire
Ending insurance	The last year the policy will expire
Accident date	The date the accident has occurred
Accident month	The month the accident has occurred
Accident year	The year the accident has occurred
Claim frequency	Number of claims per accident
Name of injured	The name of affected individual
Form of payment	Mode of payment
Salary per month	Monthly salary of the injured
Salary of 5 years	Cumulative salary of 5 years
Reason of payment	The reason why claim is requested
Number of rest days	Number of rest days given to the injured
BSG	Bandit, Shifta, Gorilla
Claims paid	Amount of money paid for claim requested
Sector	Business sector of insured organization

Table 1: identified and collected attributes

4.3 Data Preparation

The data collected is too dirty; in a sense that some columns like "insured birth date" has a missing value. Therefore, before applying the DM software tool directly on the dataset for analysis, it is essential to test the qualities of the attributes to the study, and to interrelate the heterogeneous data sets. As a result of this, preparing data input for a DM tool investigation usually consumes the bulk of the researcher's efforts and time. But investing a lot of effort and time in preparing (cleaning) the data to improve its quality is an important task for the fact that the quality of the data determines the successfulness of the mining process. Accordingly this section has taken more than 60% of the total time of the study.

Missing Values

Missing values of attributes are seen as an endemic in the datasets of this research. It is one kind of dirty data. These fields whose values aren't filled in may occur for different reasons like: empty values, nonexistence values, incomplete data, and/or data may not be collected.

In this research the researcher has found that the attribute "insured birth date", "claim frequency", and "number of rest days " columns have almost missing values from many of dataset records collected. There are different approaches to handle missing values for attributes like ignoring the column, filling the values with average values, etc. By consulting with the experts about the relevance of the attributes for decision-making, the researcher has used "ignoring columns (attributes)" approach for the above attributes.

Handling Outliers

It is uncommon for some values in a column to be quite rare or far out-of-bounds from others in the same column. From the attribute values collected in this research, "claims paid", and both "salary" data columns have been identified as having outliers. There are many approaches to handle these data values like: do nothing, and ignore the record, etc. Although the algorithm used

in this study is robust in the presence of outliers, neither of the above approaches is used for the reason that the ultimate goal is different from the nature of data. Stemming from the business objective of the Corporation, classification of claim records based on the amount they claimed for and considering the values in the "claims paid" column, which has a continuous value, this attribute has been transformed into two categorical value, Small/Big, thereby solving the above problem **(see section 4.4.1)** below.

With regards to the "salary" data columns, it is not important for business problems under consideration and also it has an outlier; this column has been ignored from further analysis.

4.3.1 Data Transformation

In preparing data into a min-able form for the required DM task, the data types of the attributes need to be in line with the capabilities of the learning algorithm i.e. If the attribute that is tested is a nominal one, the number of children produced is usually the number of possible values of the attribute i.e. the higher the number of values of the attribute, the more the number of branches to be produced from that tree. Therefore, the values (cardinality) of every attributes have to be limited to a certain amount.

For example the value of the attribute "claims paid" has many values in a continuous valued form, having an infinite number of values, which could end up by many branches in the decision tree. Thus many branches mean, it is difficult for one to understand the out come of the decision tree.

Therefore, columns with the value of numeric attributes has to be restricted to small number of categorical values, unless otherwise, they will end up decision trees that are messy and difficult to understand, this is because the tests on any single numeric attribute are not located together but is scattered along the path.

In this research, the researcher has identified "claims paid" attribute to have many number of values or cardinalities. According to the learning algorithms (schemes) used for this research, it is obvious that this attribute has much number of values in its column, which again can end up with many branches in our decision tree.

In order to solve this problem, many DM researchers recommended various approaches like:

Decision tree pruning after the algorithm has generated it. But for this research this technique is inapplicable for the fact that the values for these attributes can take so many values that tree pruning is time consuming.

The other alternative, to deal with numeric attributes like ("claims paid") thereby reducing the decision tree branch into a meaningful form is discretization or transformation (converting continuous numeric values into categorical one).

According to (Witten & Frank 2000), even through decision tree schemes can handle numerical attributes as they are, but better results in the form of faster work can be acquired, if the attributes are transformed.

Accordingly, the researcher has transformed the continuous numeric valued attribute, "claims paid", into two values as "Small" or "Big" by giving column heading called "Risk_category". This is to mean that from this time onwards, the "claim paid" attribute name has been replaced by "Risk_catagory" name and it has only two values (cardinality). Under "Risk_catagory" column heading, the values of the attribute was Small if the amount of claims paid is less than or equal to 2,500 birr and otherwise into Big.

Another attributes which need sort of transformation (aggregation) is (are) attributes associated with the timing, previously, each of these attributes has three columns i.e. the attributes "beginning of insurance date", beginning of insurance month", and "beginning of insurance year" are aggregated together and presented as a single attribute. A similar aggregation of values of

columns has been made for ending insurance date and accident date. As a result of this we can have a single column for beginning of insurance date, ending of insurance date and accident date.

The above aggregation of attributes have been made in order to calculate some other derived variables which are important for this research i.e. according to the business demand of the Corporation like studying the behaviors of clients with respect to the time they get an insurance cover and with respect to the time they are about to terminate the contract which the Corporation may suspect for the existence of fraud.

Attribute (Column) reduction

On the other hand some attributes (columns) like, "Branch number", "Branch name", Policy number", and "Sector" of the insured organization i.e. (ETC, EEPCO) have only one value in their respective columns. These unary-valued columns do not contain any information that helps to distinguish between different records thereby missing their usefulness for further analysis.

Accordingly, the researcher has made a close examination of these columns (above columns) and made a discussion with the underwriter experts concerning their information content and relevance. They agreed that these attributes don't have relevant information content in a sense that if the value of an attribute is unique where lies the predictive power of models?. Therefore, the researcher has excluded them form further analysis in this research study.

Furthermore, some columns like "form of payment" which has two cardinality as (cash payment or check payment), but the values for this attribute is almost-unary for many of the records understudy. On top of this, it doesn't play a significant role in decision making. So as a rule of the thumb such columns are also too informative that they need to be neglected for the DM task. From the ideas collected from the insurance experts about this field ("form of payment", this field has not have much information for decision making, as a result of this; again the researcher has excluded this attribute from further analysis.

The other attribute, which has a cardinality of two in the form of "Yes" and "No", is the "BSG" attribute. While examining the values under this column all the records have the value of "yes" in their respective column (i.e. each clients have an insurance coverage for BSG), which turns to be again informative and need to be neglected from analysis.

Unlike the above attributes, which have unary values, there are other columns like ("claim number", and "name of injured") that take on a different value for every single row-or nearly every row and they are ignored. Generally these columns are also not informative, for the fact that they do not have predictive value (Witten & Frank 2000). In doing so, they prevent DM algorithms from finding patterns contained in multiple rows.

4.3.2 Attribute Selection

Once the attributes are identified and transformed (aggregated) as was seen in the previous section, the attributes have to be selected for the experimentation, and for model building purpose. For this considering the risk factors (variables) that are important for predicting the likelihood of claims and the size of a claim are important criteria.

In this study, a heuristic approach is followed to select relevant variables for the tasks under consideration i.e. in order to identify, which risk factors (attributes) are determinant and decisive for the problem under consideration, discussing about the attributes with underwriters is very important approach. This approach of attribute selection for further analysis is called heuristic approach.

Accordingly, the researcher has identified the following attribute variables to be very important for decision making thereby for model building in the next chapter. These are: "work_Category", "Reason for Payment", and "Risk_catagory" (after it was transformed from the initial attributes).

The underwriters in the Corporation have a great interest on "Work_Catagory" attribute in order to examine the association between work category and level of risk or claim. The Corporation has categorized its clients into 4 categories based on work categories as indicated in this chapter as (2\24, A, C, and D). Therefore, this attribute was selected in order to classify which customers are high risk and which are not and this helps to set appropriate priemum charges.

As discussed in chapter three, there are different benefits that the Corporation is giving to its clients like payments for death, total permanent disablement, total temporary disablement, and medical expenses and the combination of these. Therefore, including this attribute helps the Corporation to examine the relationship between levels of risk as (Small of Big) and reason of payment. In doing so the underwriters can be able to examine for which type of benefit its clients are claiming for.

The other attribute, which underwriters of the Corporation are highly interested is "Risk_catagory". This attribute is used as dependant attribute in order to classify clients as Big claim client or as Small claim client. This classification is very important for many reasons like: estimating the amount of outstanding claims, to examine client's risk exposure behavior and level of claims.

4.3.3 Derived Attributes

"PAcc_Within" attribute

On top of the above attributes, the underwriters need to know about the nature of claims of individuals (clients) with respect to the time period they get the insurance cover and with respect to the time period the insurance contract will terminate to avoid fraud. But through the above attribute variables it is difficult to analysis this timing effect.

This is to say that clients have been seen attempting to transfer previous accident (injury) to the Corporation unnecessarily i.e. the underwriters of the Corporation wants to study whether the clients have some health problems prior to getting the insurance coverage or not. This

information helps them to accept or reject the contract from such clients. This could be possible by attributes, which could be derived from the other attributes.

Therefore, in order to study this behavior of clients', the researcher has derived an attribute by subtracting the value(s) of the attribute "big insurance date" (after it was aggregated) in the form of (mm/dd/yy) from the value(s) of the attribute "Accident date" (after it was aggregated too) in the form of (mm/dd/yy). Then the result was given an attribute name called "PAcc_Within". This attribute has a cardinality of four as quarter 1(Q1), quarter 2 (2), quarter 3 (Q3), and quarter 4 (Q4).

The lower the value of this attribute like (Q1), the higher the probability for that client to have prior health problem thereby indicating the possibility for a fraud to exist and vis versa i.e. if clients came up with claims as soon as they get an insurance coverage from the Corporation this might be an indication that such clients might commit fraud. In order to study such behavioral pattern, the researcher has calculated this value for the whole dataset records (as indicated above) and gave the name of the attribute as "PAcc_Within".

On the other hand, if the value of the above attribute ("PAcc_Within") is large like Q4, it means that clients are claiming for medical expenses and other related expenses while they are about the end of the termination of their contractual agreement. Such clients have to be identified and special marketing strategies have to be designed for them.

Finally, based on the underwriters' business demand about the relevance of the attributes to the problem at hand, the researcher has selected the best and most determinant attributes to be used in the experimentation of this study with their possible values showing the cardinality of the attributes in table 2 below.

Work_Catagory	PAcc_Within	Reason_Payment	Risk_Catagory
A	Q1	DEA	Small
C	Q2	PTD	Big
D	Q3	TTD	
2\24	Q4	ME	
		TTD_ME	
Cardinality=4	Cardinality=4	Cardinality=5	Cardinality=2

Table 2: Attributes used in the experimentation.

Remark:

The values under the attribute "Work_catagory" are used for the purpose of premium determination. The Corporation charges the same premium rate for every individual under the same work category.

Meanings for the above attribute values:

A refers to Permanent employees who has got an insurance cover if an accident occurs while he/she is at work

D refers to daily laborer again these individuals get the cover if the accident occurs while they are on duty.

C refers to contract individuals and get the cover while they are at work

2\24 refers to Permanent employees having 24 hours of insurance cover

The values Q1, Q2, Q3, Q4, stands for quarters **1, 2, 3, 4**, for the attributes "PAcc_Within" As stated in the previous discussions these values are very important to anticipate the existence of fraud. **DEA** refers to payment for death

- **PTD** refers to payment for Permanent Total Disablement
- **TTD** refers to payment for Temporary Total Disablement
- **ME** refer to payment for Medical Expense
- **TTD_ME** refers to payment for both at the same time.

CHAPTER 5

EXPERMENTATION AND DISCUSSIONS

5.1 Introduction

Once data is preprocessed and transformed into its min-able form, the next task is to look for a DM technique and tool, which is robust to the nature of data and the DM task under consideration. Accordingly, the researcher has found that knowledgeSTUDIO version 4.1.1 software tool has the required techniques and measures, which are suitable for this research.

KnowledgeSTUDIO is new generation software that employs today's most advanced DM techniques (decision tree, neural network, clustering, and predictive models) packaged together in a familiar to Microsoft office-style working environment. It provides a number of ways to visually explore and express patterns in a data. It has the capacity to help with the whole DM process – from preparing data through to producing final graphs and reports as a result of this it was selected for this study. It has two algorithms for the decision tree technique embedded in the software. These are KnowlogdeSEEKER and HeatSEEKER.

KnowlodgeSEEKER is a flexible algorithm that is especially good for exploration purposes and manual tree expanding according to the preference of the expert. The algorithm brings the best attributes to the root node based on the statistical relationship between attributes to the dependant attribute. But if the need arises, an expert can expand beyond the one generated by the software by using the command called "split by force".

Unlike KnowlodgeSEEKER, HeatSEEKER is a fast algorithm that is especially good for automatically generating tree. Moreover it works best with fewer number of attribute variables.

The numbers of attributes used in this research are relatively small, which create a room for using either of the above algorithms. But as a result of shortage of time only one algorithm (knowlodgeSEEKER) was considered in the experimentation of this study. By using this algorithm, three (3) experimental models have been built and one whichever has more accuracy (as measured by the confusion matrix) has been taken as final Predictive Model for deployment.

With the selected algorithm (knowlodgeSEEKER), Adjusted-P-value Bonferroni Adjustment measure is used in all the experiments for the fact that it is the most statically sound approach given that it reduces the possibility of producing chance results and puts all variables on a common statistical footing.

5.2 Building Model

In order to build a model the processed data was analyzed using the above software tool. This section explains model building and comparison between them.

Experiment 1

In order to build a model, data has to be imported to the software and it has to be partitioned into two partitions as: Learning and Validation as indicated below using the data partitioning wizard of the tool.

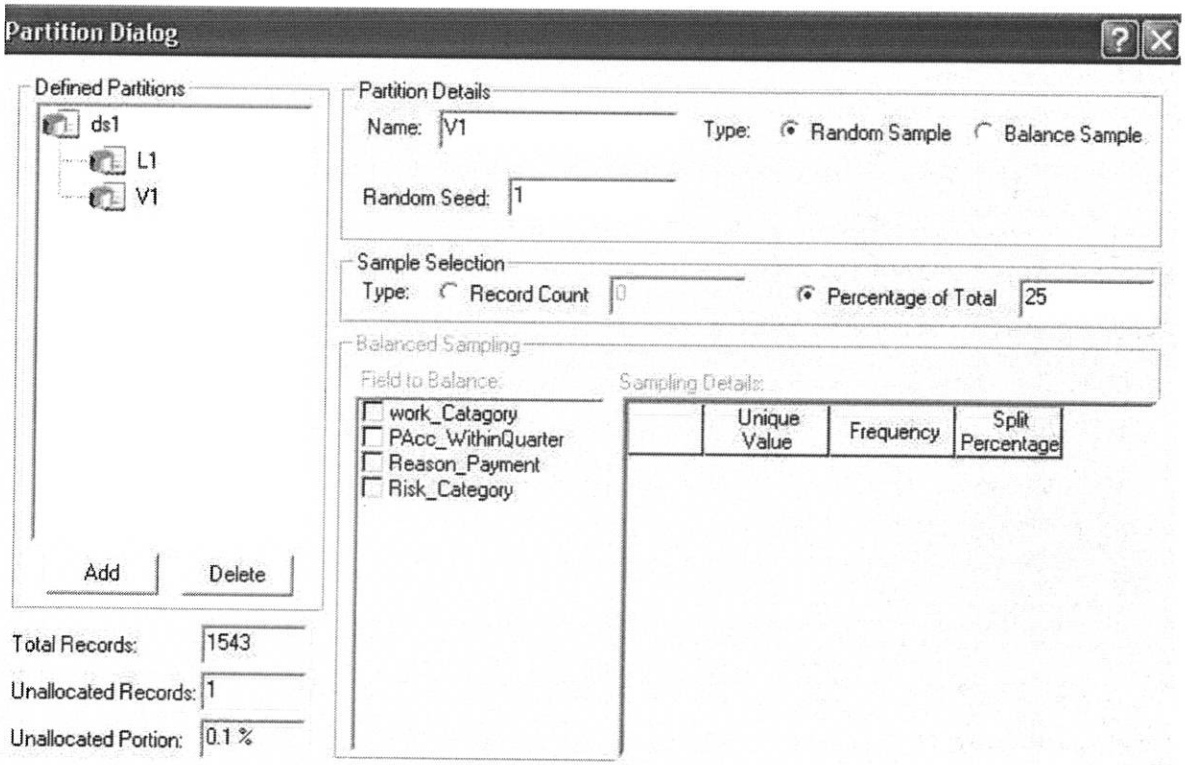


Figure 1: KnowledgeSTUDIO data partitioning Dialogue

According to Two Crows Corporation (1999), the ratio of data partitioning can range between 5% and 33% and choosing between these ranges is a matter of judgment. Accordingly, the 1543-dataset records are partitioned with a ratio of 75% (1157 records) for learning (as represented by L1) and 25% (385) records (as represented by V1) for validating the trained data. In order to partition the datasets, a randomized sampling method was used as can be seen on the above figure for this experiment (see how random sampling method is selected above). From the above figure there is one record, which is not allocated in neither of the partitions.

Once the dataset was partitioned, the next step was building the predictive model. In order to build the model, the researcher has used the "Learn" (L1) dataset, stated above to train the model and generate the predictive decision tree by using "Risk_category" as a dependant variable and considering the rest variables as independent variables.

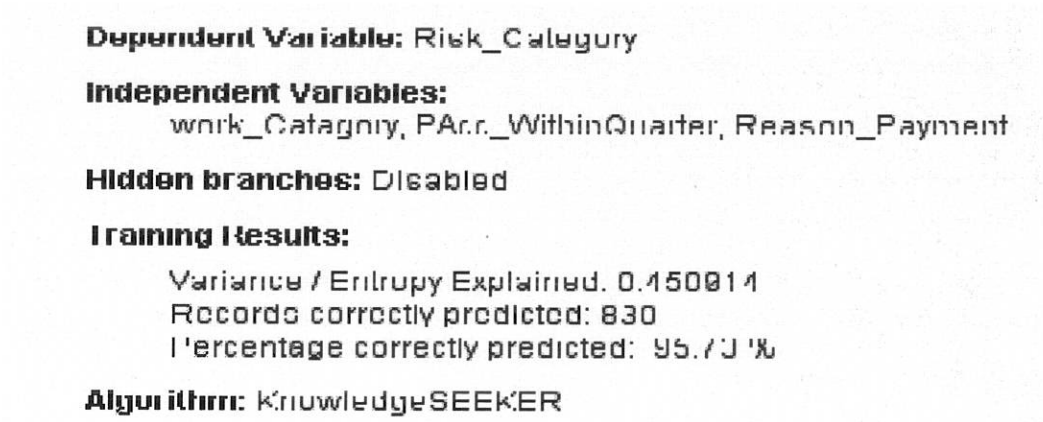


Figure 2: Decision Tree Predictive model for experiment 1

In order to see the performance accuracy of this model and apply the model else where, if it performs well, the model has to be validated (tested) on new dataset called validation dataset.
Validation is the process of assigning outcomes (scores) to records in a dataset for which the dependant variable values are already known. The predicted values are then compared to the known values, to determine how accurate the predictive model is.
Accordingly, when this model is brought to the other dataset (on 385 datasets) the model has a performance accuracy of 95.84% as measured by the confusion matrix indicated by table 3 below.

Confusion Matrix - Risk_Category

		Predicted	
		BIG	SMALL
Actual	BIG	21	14
	SMALL	2	348

Total records	385
Correctly predicted	369
Percentage	95.84%

Table 3: Confusion matrix for experiment 1

From the confusion matrix (indicated by table 3), it can be seen that the model has an average accuracy of 95.84% accuracy level, which can be calculated as follows:

Percentage= [(21+348)/385] x100%

- Where 21 refer to records, which are actually under Big risk class category and predicted under the same category by the model;
- 14 refer to records, which are actually under Big risk category but predicted to be under Small risk category;
- 2 refer to records, which are actually under Small risk category but predicted to be under Big category;
- 348 refer to records, which are actually under Small risk category and predicted to be at the same category.

In Order to visually see and analyze the relationship between the different risk factors like level of risk as "Small" and "Big", the following decision tree was generated automatically on the validation dataset records.

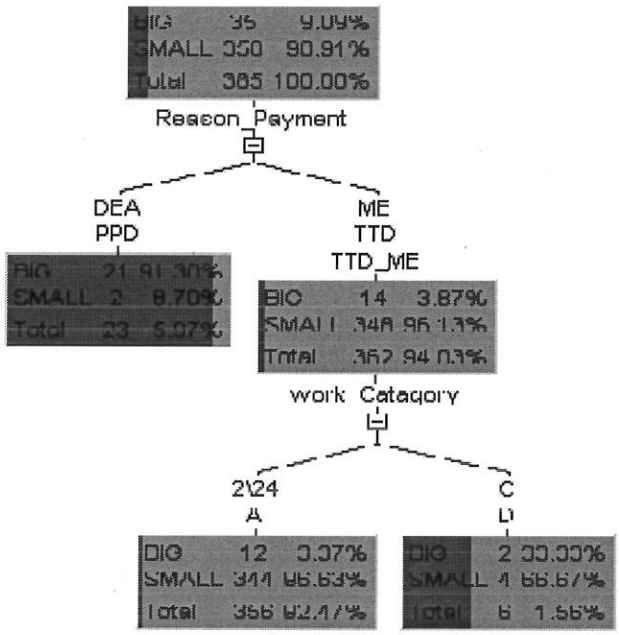


Figure 3: Decision tree for experiment 1

The above figure (figure 3) was generated by the model on the validation set data. As the tree indicates, most clients "Reason_Payment" is the most statistically important to predict the class of risk as "Small" or "Big". Further more the tree indicates that (94.03%) request claim payment for medical expenses (ME), total temporary disablement (TTD) and for both medical expense and total temporary (TTD_ME) while payment for death (DEA) and permanent total disablement (PPD) accounts for 5.97% of the total record, in this case validation dataset.

From payments made for death (DEA) and permanent total disablement (PPD), most of the claim request (91.30%) was categorized under "Big" risk category while only 5.97% of the dataset claim request falls under Small risk categories.

In addition to this, we can explore the relationship between levels of risk (as "Small" and/or "Big"), reason for payment and work categories by moving downward through the tree. From the lower leaf of the right tree, it can be seen that, "Work_Category" is the second most important attribute to predict the class of risk level. From this last leaf of the tree it was evidenced that, 92.47% of the claims are coming from clients working under "2\24" and "A" work category. This is as the result of the fact that there are more "A" and "2\24" type of clients than the other type. But what is important to see is that most of the "Big" claim request is from clients of "2\24" and "A" and the request is for death and permanent total disablement.

Evaluating the dataset distribution used in this study reveals that the class distribution of the dataset was imbalanced, (imbalanced datasets or class imbalance occurs when the different classes in the dependent variable do not have equal number of cases) i.e. the number of records in Small dataset records is by far more than that of Big claim dataset records, with the ratios of 1405:138, which indicates the existence of class imbalance.

Using separate accuracy measure (for each risk class) is one alternative to overcome class imbalance problem i.e. computing the accuracy or equivalently error rates for each class is important to judge about the performance of the model. Accordingly, the above model has an error rate of 40% [14/ (21+14)] for Big risk class category and 0.57% [2/ (2+348)] for Small risk class category.

Class imbalance causes models to be overwhelmed by the large classes and ignore the small ones. This is because the model attempts global quantities such as errors rates, not taking the data distribution in to consideration. As a result, examples from the majority classes are well classified whereas examples from the minority class tend to be misclassified.
(See error rates for experiment 1 above).

This is mainly because decision tree learning is highly influenced by the majority class in the tree building process. Again during the pruning process majority classes have higher chances of influencing newly generated nodes (Japkowicz et al).

All the above discussions support the assertion that the reason for the failure of the first experiment is the problem of class imbalance. As a result the following experiment attempt to balance the distribution of dataset records in the sample.

Experiment 2: (based on "PAcc_Within" attribute

This experiment was done in order to improve the shortcoming that was seen by experiment 1 above. As stated above class imbalance has a significant effect on the accuracy of the model. So in order to see the role of the frequency occurrence (distribution) of the value of the attribute variables in the dataset as a base for balanced partitioning, this experiment is done based on the attribute "PAcc_Within" using knowlodgeSEEKER algorithm. This attribute has a cardinality of 4 as a result 25% of each has been considered in the experiment.

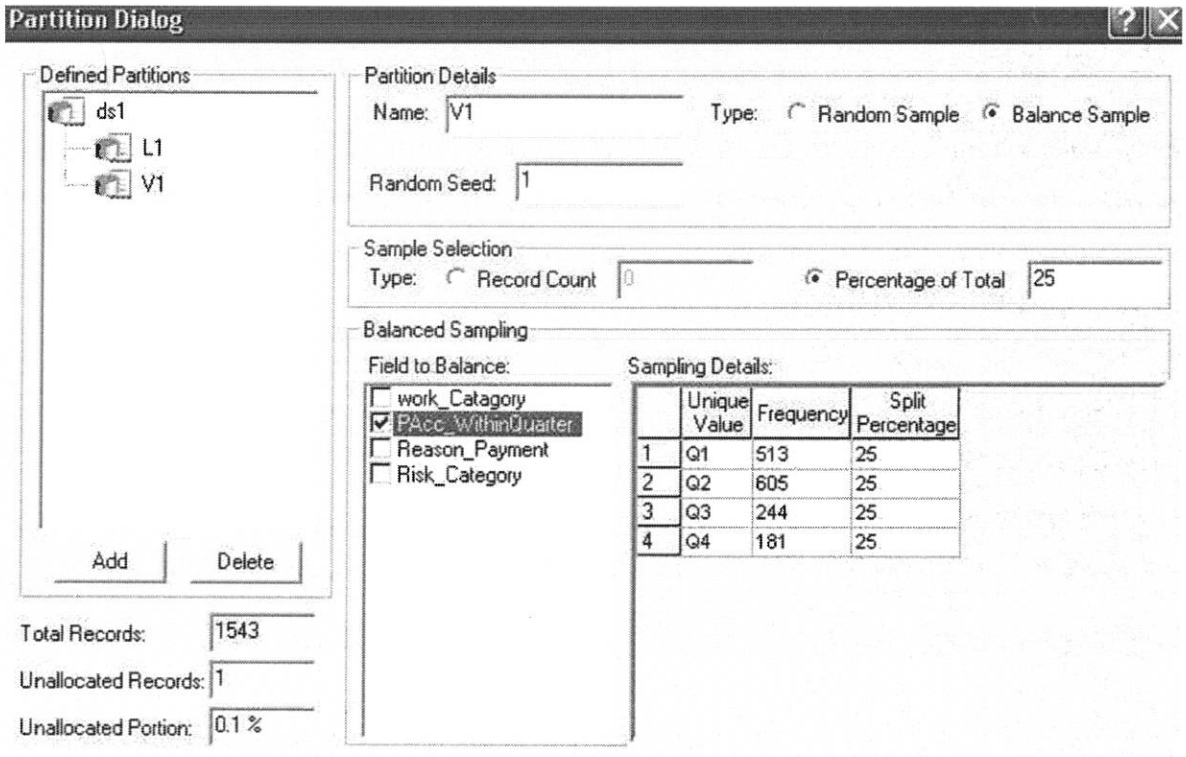


figure 4 knowlodgeSTUDIO data partitioning Dialogue

As indicated above, this experiment is based on balancing the distribution of the dataset (we can see how this was done above)

The purpose of this partitioning is to distribute dataset records proportionally through out both the learning and validating datasets. Accordingly, the validation dataset partition includes 25% of 513 (which is equivalent to 128) for value "Q1", 25% of 605 (which is equivalent to 151) for value "Q2", 25% of 244 (which is equivalent to 61) for value "Q3", and 25% of 181 (which is equal to 45) for value "Q4" as indicated in the last column of table 4 below.

Unique Value	*Frequency*	*Split Percentage*	*Exact amount of each*
Q1	*513*	*25%*	*128*
Q2	*605*	*25%*	*151*
Q3	*244*	*25%*	*61*
Q4	*181*	*25%*	*45*

Table 4: balanced partitioning based on value of "PAcc_Within" attribute

Using the above dataset composition for validation purpose, it is expected to balance the composition of the dataset distribution for learning dataset too, whereby minimizing the effect of class imbalance problem seen in experiment 1.

Once the dataset distribution is determined (with the above value composition), partitioning the dataset into two classes is a natural task and is done accordingly as indicated above.

Figure 5 below shows the decision tree predictive model generated.

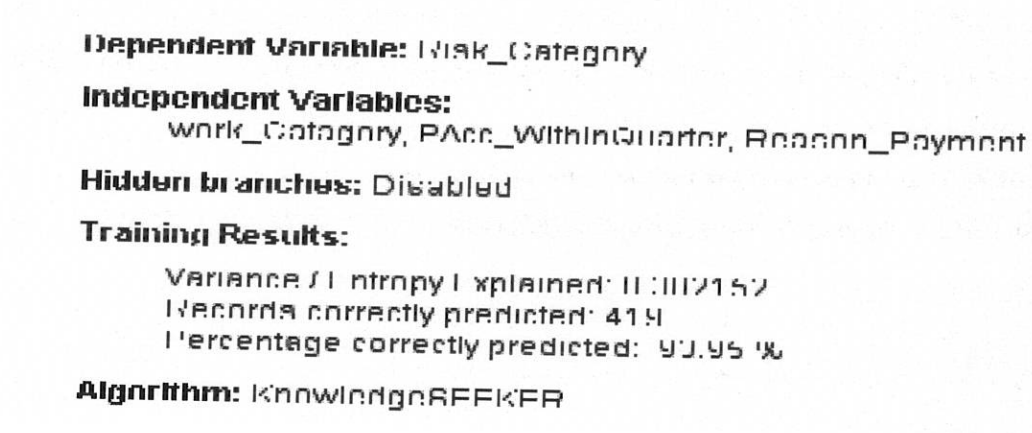


Figure 5: Decision tree predictive model for experiment 2

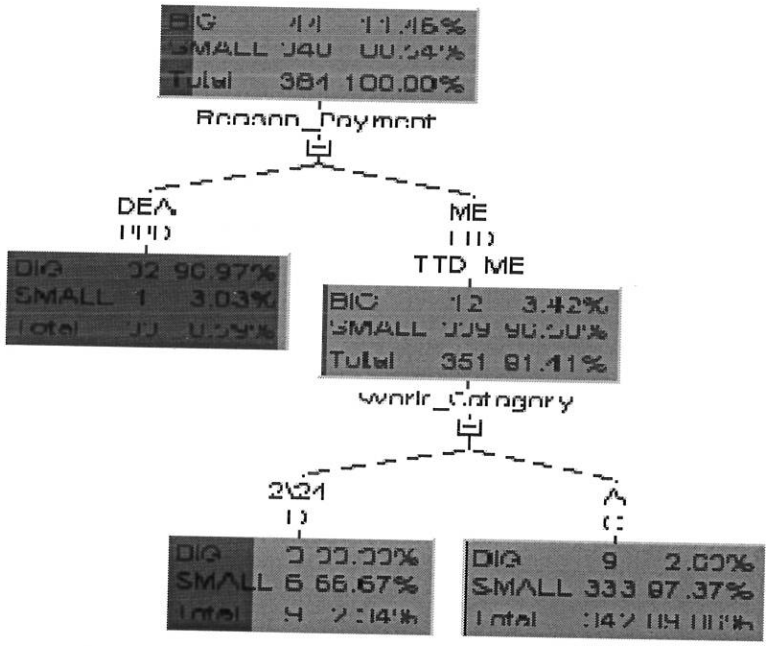


Figure 6: Decision tree for experiment 2

Comparing the decision tree of experiment 1 with that of experiment 2, it can be seen how balancing dataset affects the distribution of the minority class. In experiment 1 the dataset size of Big risk category is 35 and there are 350 records under Small risk category.

But in experiment 2, above, the dataset size for the minority class, Big risk category, has been increased from 35 to 44 thereby improving the average accuracy of model 1 from 95.84 % to 96.61% (for model 2). (See table 5 below).

Confusion Matrix - Risk_Category

		Predicted	
		BIG	SMALL
Actual	BIG	32	12
	SMALL	1	339

Total records	384
Correctly predicted	371
Percentage	96.61%

Table 5: Confusion matrix for experiment 2

From the confusion matrix indicated above, this experiment has a classification error of 27% for Big claim risk category while it has a rate of 0.29% for Small risk categories. In comparison to experiment 1, this model has more accuracy for predicting both Small claims and Big claim risk categories than model1.

From this experimental result, given equal size of the dataset size and the same partitioning ratio (75:25) in each models, one can easily confirm how class imbalance affects the accuracy of models. In short, balancing the partition based on "PAcc_Within" attribute has more encouraging result than randomized partitioning approach illustrated in experiment 1

So which model has to be selected for final deployment? Before, selecting a model for deployment, the business objective and problem of the Corporation has to be considered.

The following section discusses these issues.

5.3 Issues in Model Selection

Before selecting the intended predictive classification model some basic issues like dataset distribution, performance measure approaches to be used, and economic implications of the classification errors for the Corporation was considered.

Issue 1: Performance measures

In order to tackle problems with imbalanced datasets along with some other difficult problems performance measures other than the traditional ones are used in this study.

The performance measures mostly used for classification problems are the overall accuracy or equivalently the total error. This method works fine for the majority class problems (Small claim class), however highly imbalanced class distribution may require high accuracy for the minority as well as for the majority class, which is not guaranteed from the overall accuracy i.e. for problems with high imbalance in class distribution the average accuracy measure is inappropriate. In datasets with two classes, out of which one is very rare, even accuracy rate of 100% may not be satisfactory, because the error that stems from the minority case is always disproportionably large compared to the errors that stems from the majority cases.

Conversely, several prediction problems require high precision, achieved when the correctly classified cases in any given class is large compared to the incorrectly classified cases in the other class therefore additional performance measure that counts the accuracy rates separately for the majority and minority classes is required. For example, in experiment 1, the average accuracy of the model is 95.84%, which seams encouraging. But considering the error rates separately, this model is rather discouraging for the fact that it has an error rate of 40% for Big claim risk category. Therefore, it is recommended to compute a separate accuracy measure than simply considering the overall average accuracy, which is a misleading measure in the case of imbalanced dataset. Even the Corporation doesn't want such a big error for Big claim risk category clients.

Issue 2: Altering the class distribution in the training set

Altering the class distribution in the training set is another way used to deal with problem of imbalanced class distribution.

Classification algorithms should be trained using datasets reflecting the characteristics of all classes, the larger and smaller classes. Previous studies (weiss and Provost 2003) have shown that training with "natural" distribution does not always give the best model; therefore, new datasets should be constructed to aid the training process.

Especially when the minority cases are rare, a testing set "enriched" with minority examples is needed in order for the induction algorithm to create the models with strong predictive capabilities for both classes. Taking the variables of interest to partition the dataset in a balanced way could do this. In order to build a model with a better overall accuracy and separate accuracy, balancing the distribution of the class has shown great change. For example with the same dataset size of 384 (for validation dataset) the two experiments show different accuracy measures. From these experiments, experiment 2 has been accepted by underwriters of the Corporation, for its being more accurate than the other.

Issue 3: Economic impacts

Finally, a study on the economic impact of the classification result for the organization leads to an evaluation framework for all involved models. For this study apart from the accuracy classification results, a cost/ benefit analysis is made based on the business criteria in the problem.

It is argued that different models are indicated as "best" depending on the relative gain from correct classification in the minority class compared to the cost for incorrect classification in the majority class.

By discussing the impacts of the above models, underwriters of EIC gave more preference for the last model for many reasons like: experiment 1 has more errors for Big claim risk category and even it has lesser overall accuracy measure compared to experiment 2. Therefore, by taking the economic impact of the models in addition to accuracy measurement, the model built under experiment 2 has more advantage for final deployment for the fact that it has a lesser error rate for both classes and even has a higher average accuracy.

As can be seen from the decision tree generated under experiment 2, "Reason_Payment" is seen as the best statistically important attribute in order to predict risk levels ("Small" or "Big") and "Work_Category" was seen to be the second important attribute. But from the business problem of the Corporation point of view, "Work_Category" was more important than "Reason_Payment) this is because the Corporation can use the latter attribute to set premium rate than the former.

Therefore, by using the "Forced Split" command of the tool (software), "work_category" attribute has brought to the top of the decision tree followed by "Reason_Payment".
(See figure 7)

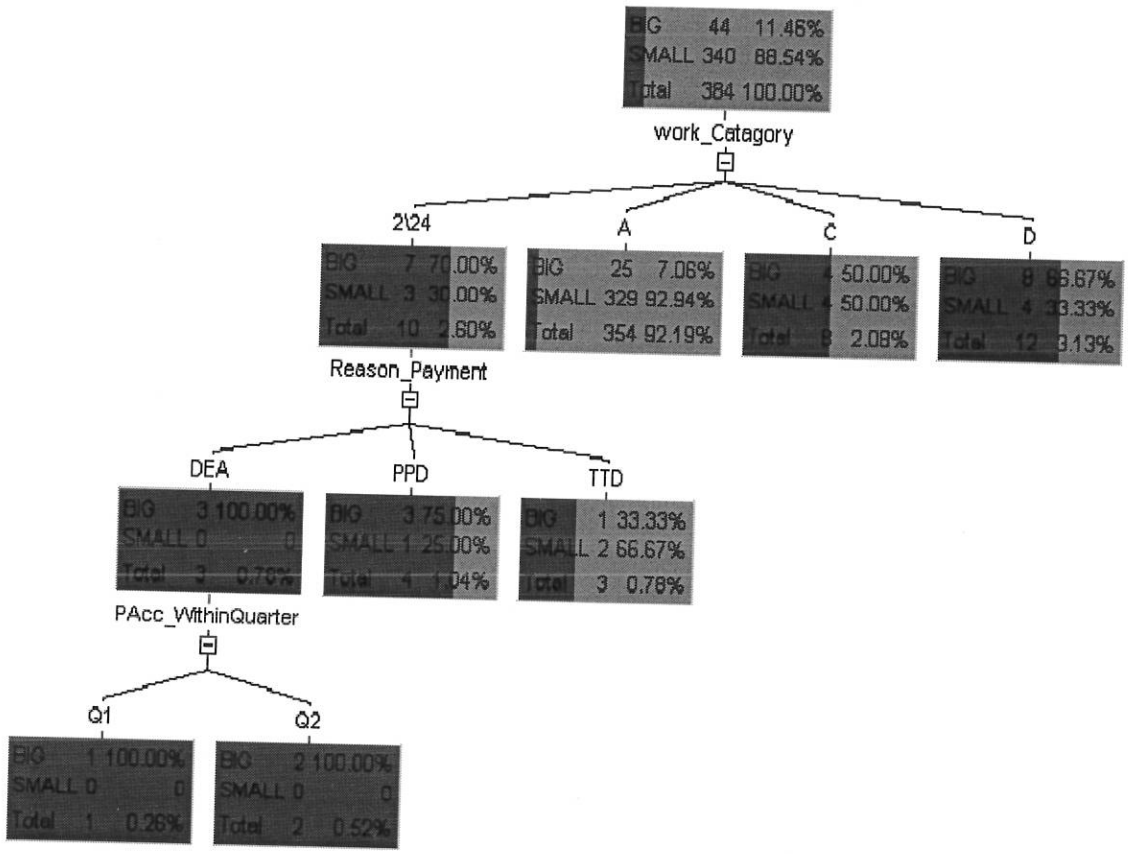


Figure 7 manually extended decision tree

The tree shows a manually extended form the model to be deployed. As can be seen, clients under work category of 2\24 claim for death, permanent total disablement, and temporary total disablement. When we see the death payments (benefits sought), almost all are claimed at first (Q1) and second (Q2) quarters from their entrance of contractual agreement with the Corporation. As indicated in the problem statement of this study, underwriters have to give more emphasis to these clients for the fact that they may attempt for making fraud.

Similarly, we can grow the nodes of the above tree and see the relationship between the risk factors. And this has been done but the relationship has been indicated as "If…Then" rules and attached as an appendix at the end of this document.

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusions

In his way of conducting this study the researcher has come up with the following concluding remarks.

1. In this experimental study, the researcher has identified clients with high possibility of risk exposures, thereby identifying, which customers are coming with "Big" or "Small" claims. According to the experimental result, 78.42% of the test data has a work category of "A" while work category "C" and "D" takes the smallest ratio. But examining the risk classes, indicates that although small by frequency size clients under "C" and "D" are more interesting in a sense that they account for 91.67% for Big risk class and 8.33% for Small both equally, showing clients under these work deserve some attentions. This model will help underwriter to adjust their premium rates with the logic of the higher the risk the higher has to be the rate. Another alternative for the Corporation is that it has to give more attentions to attract and retain workers under category "A" followed by workers under category "2\24" for the fact that most of these groups pose Small claim to the Corporation than the rest.

2. Another interesting view observed is that almost all payments made for "TTD_ME", "TTD", and "ME" are categorized under Small risk claim category while most of the payments made for "DEA" and "PTD" are categorized under the Big claim risk categories. Still this model gives an insight for underwriters for further analysis and takes some measure accordingly.

3. With regards to accuracy measure, the average accuracy measure is not a good performance measure of models for the fact that this approaches favor the dominant class, a class with many number of dataset records. In order to solve this problem an alternative approach has to be used like accuracy measurement, which considers each class separately. In building predictive models considering the frequency occurrence of the values of the attributes is another factor, which deserves attention. Accordingly in this research this effect has been seen and balancing the class distribution measure was taken as a solution. This balancing method based on attribute partitioning was proved to be promising.

4. On top of this, the predictive techniques also help underwriters to study the behavior of individuals with respect to the timing they claim for repayment of some expenses, which again lets them examine the existence of fraud in the Corporation, if any. This is to mean that with claim records in an insurance database, one can find that combinations of different factors are important predictors of fraud. One we have this knowledge, we can use this to show the potential for fraud in a new claim and investigate effectively.

5. Finally, the researcher has a strong belief that data mining technology (predictive models) can play a great significance role in today's insurance business by assisting insurance underwriters by letting them know which customers are posing Big claim risk and which of them pose Small risk to the Corporation thereby identifying profitable and non profitable clients.

6.2 Recommendations

While conducting this study, many issues have come to the attentions of the researcher. Most of which require a separate study by them selves. Some of them are:

1. The key to gaining a competitive advantage in the insurance industry is found in recognizing that customer databases, if properly managed, analyzed, and exploited, are a very important, and valuable corporate assets. Therefore, Insurance firms can unlock the intelligence contained in their customer databases through modern data mining technology that uses predictive modeling, data warehousing, data processing, and combinations of these to quickly answer crucial business questions with greater accuracy. But the Corporation has to build similar predictive models for other clients too and integrate with this model.

2. It has been seen how balanced data partitioning gets an improvement over a random sampling approaches. Therefore, the researcher strongly recommends similar type of experiments to be done on other attributes too.

3. This study has used knowlodgeSEEKER algorithm found in knowlodgeSTUDIO. It was discussed that the other algorithm (HeatSEEKER) can also be used if the numbers of attributes in the study are small relatively. As a result the researcher recommends a study to be conducted using this algorithm.

REFERENCES

Angoss Software Corporation. (2001). KnowledgeSTUDIO Data Mining User Guide. http://www.angoss.com

Apte,C.et.al (1999). Insurance Risk modeling using Data mining Technology. IBM Research report.

http://www.ibm.com

Askale Worku (2001). The application of data mining technique in supporting loan disbursement activity at Dashen Bank S.C. A thesis submitted in partial fulfillment of the degree M.Sc.I.S. Addis Ababa University. Addis Ababa.

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

Costa, E. (2000). Classification problems: an old question with new solutions: http://www.dcs.napier.ac.uk/coil.summerschool/docs/problem_costa.pdf

Gashaw Mulatu (2004). Application of data technology in support of customer insolvency prediction at Ethiopian Telecommunication Corporation. A thesis submitted in partial fulfillment of the degree M.Sc.I.S. Addis Ababa University. Addis Ababa

Girma Belew (2004). The role of data mining in the risk assessment of Customs: With special reference to Ethiopian Customs Authority. A thesis submitted in partial fulfillment of the degree M.Sc.I.S. Addis Ababa University. Addis Ababa

Growe.(1999). Comparing Algorithms and Clustering Data: Components of the data mining process. A thesis submitted to the department of computer science and Information Systems at Grand Vally State University)

http://www.qnetr2k.com/thesis/modelThesis.pdf

Guo.L, Applying Data Mining Techniques in Property Casualty Insurance (Ph.D Dissertation paper). University of Central Florida.
http://www.casact.org/pubs/forum/03wforum/03wf001.pdf

Japkowicz, N., and Stephen, S. The class imbalance problem: A systematic study. School of Information Technology and Engineering University of Ottawa. http://www.iro.umontreal.ca/~kegl/ift3390/H04/Presentations/Imbalanced/**systematic**StudyJapkowicz.ps

Lanna, A. S. Brazilian Customs Risk Management Model:
Selection of Import operations for Customs Examination.
http:// www.iadb.org/int/customsjpn/Practicas/Risk%20Management%20Model.pdf

Llyod-williams. Discovering the hidden secrets in your-data: the data mining approach to information. Information Research, Vol.3, number 2
http://www.informationr.net/ir/3-2/paper36.html

Mangold.J (1998). Turning straw into Gold: How data mining Transforms data from and expense into an asset. An article published in DM Review Magazine.
http://www.dmreview.com/toc.cfm?issueid=1372

Meretework Shawel (2004). Possible application of data mining technology in supporting risk assessment: The case of NIB International Bank S.C.
A thesis submitted in partial fulfillment of the degree M.Sc.I.S.
Addis Ababa University. Addis Ababa

Penzias (1997). Why use data mining?
http://www.vencore.com/cases.htm

SAS Institute Inc. (2001,2003). Data Mining in the Insurance Industry: Solving business Problems Using SAS Enterpriesminer Software.
http://www.sas.com

Shegaw Anagaw (2002). Application of data mining technology to predict child mortality patterns: The case of Butajira Rural Health Project. A thesis submitted in partial fulfillment of the degree M.Sc.I.S. Addis Ababa University. Addis Ababa

Vesanto.J (2002). Data exploration process based on the self-organizing map: Dissertation for the degree of Doctor of Science in Technology. http://www.cis.hut.fi/research/reports/biennial02-03/cis/-biennial-report-2002-2003-doctorial-dissertations.pdf

Tesfaye Hintsay (2002). Predictive Modeling Using Data Mining Techniques in support of Insurance Risk Assessment. (Unpublished M.Sc.I.S) Addis Ababa University. Addis Ababa

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

Thearling (2003). An introduction to Data Mining: Discovering Hidden Value in your Data Warehouse http://www.thearling.com/index.htm

Vinterbo (1999). Predictive models in medicine: Some methods for construction and adaptation http://www.dsg.bwh.harvard.edu/people/svinterb/pubs/vinterbo1999.pdf

Weiss, G. M. Mining with Rarity: A Unifying Framework. AT & T Laboratories. http://www.research.rutgers.edu/~gweiss/papers/sigkdd04.pdf

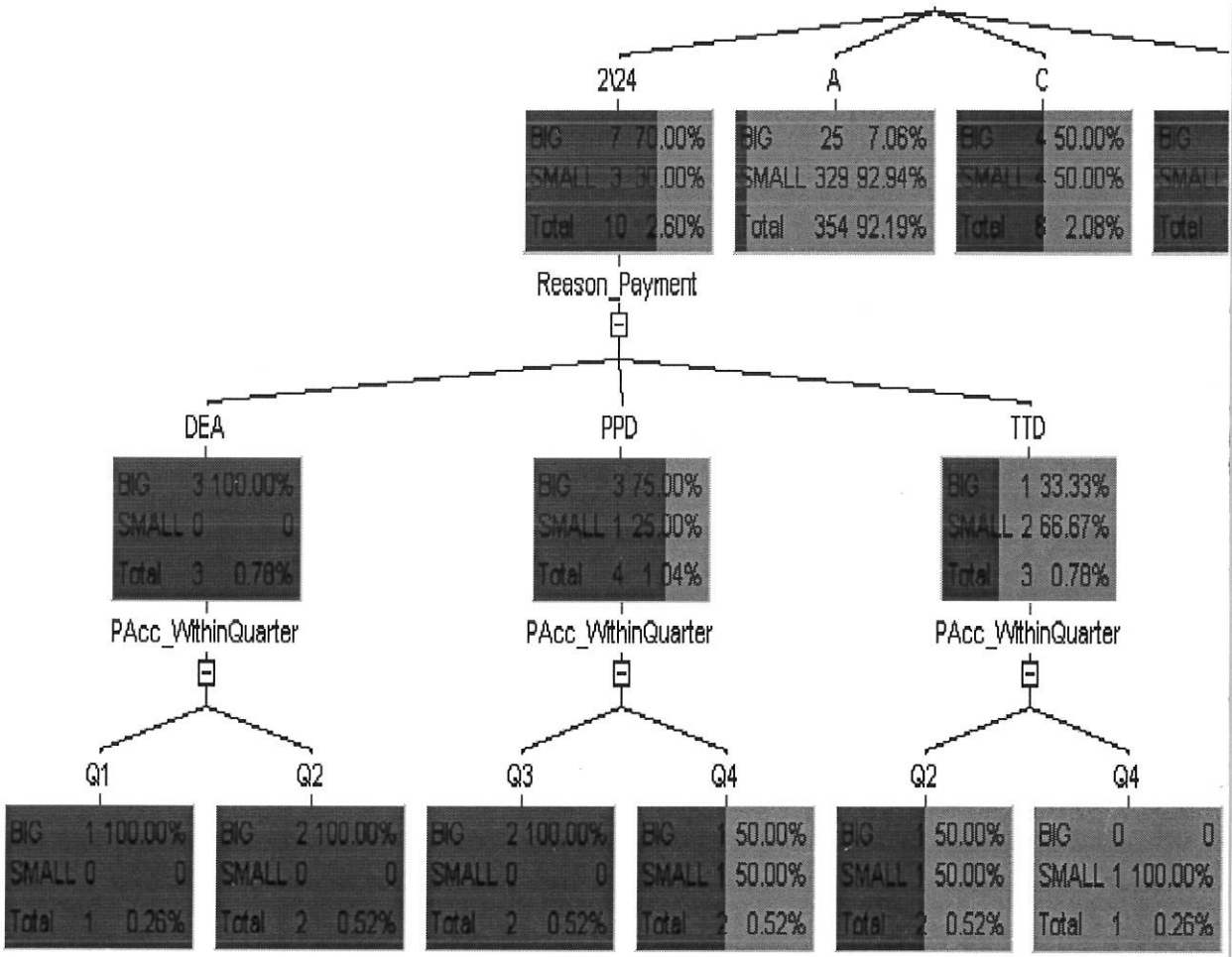
Witten, I. H., & Frank, E. (2000). Data Mining- Practical Machine Learning Tools with Java Implementation. Morgan Kaufman Publishers

Zied Elouedi1, Khaled Mellouli1, and Philippe Smets2 Classification with Belief Decision Trees. http://irida.ulb.ac.be/~PSMETS/BDT-AIMS

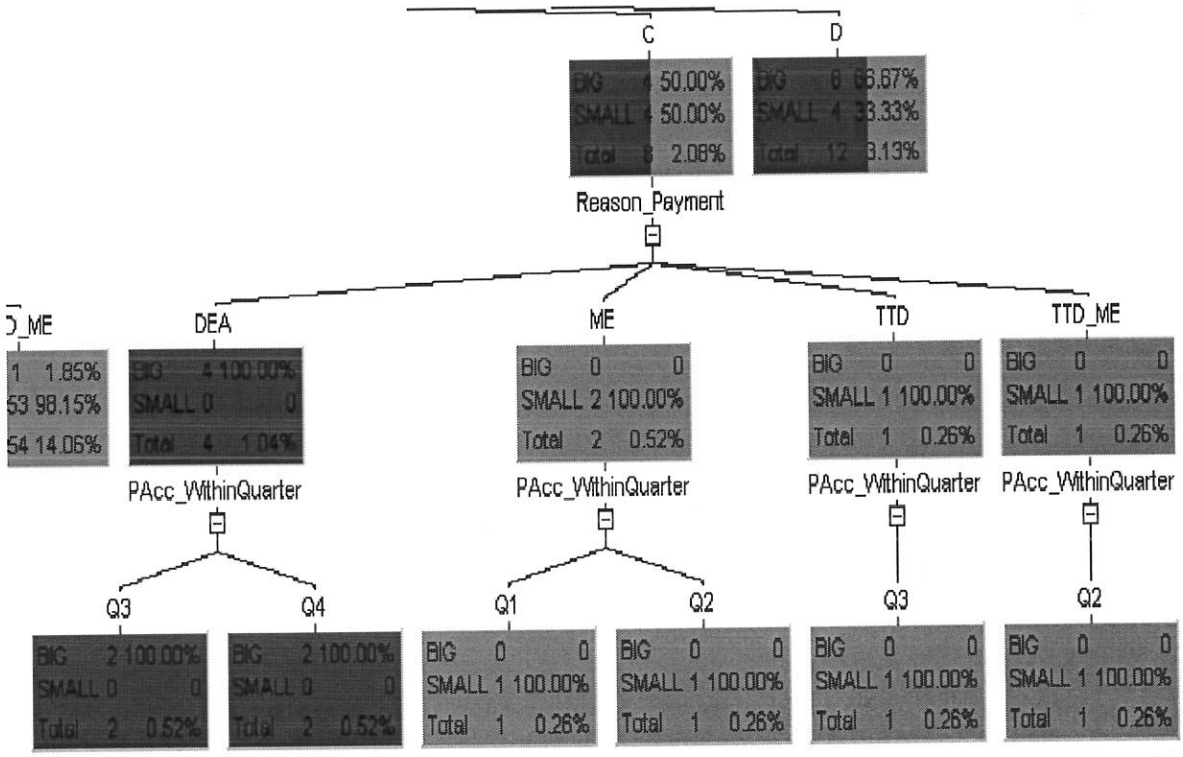
APPENDIX

Decision Tree for the selected Model

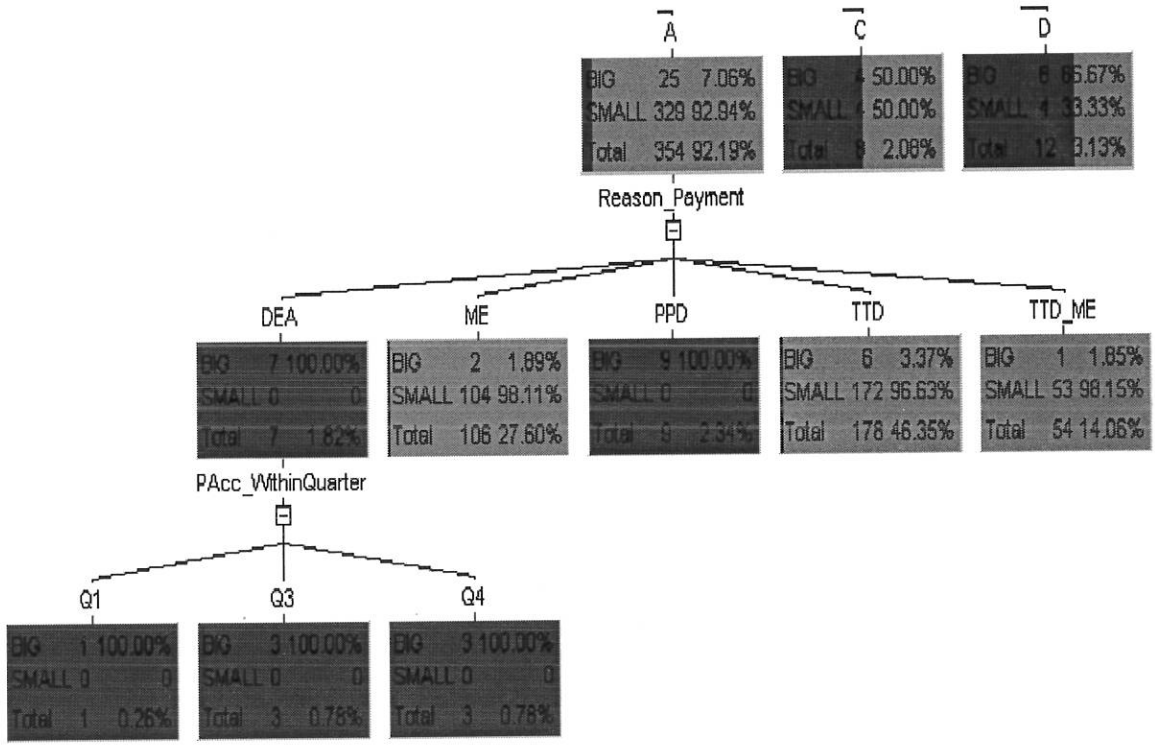
The root node and a grown node for 2\24 work category.



The grown tree for work category "C"



The following is some portion of growing category "A"



Generic Rules

```
RULE #1
 (Whole Tree)
     Risk_Category = BIG   0.114583333333
     Risk_Category = SMALL   0.885416666667

RULE #2    if
              work_Catagory = 2\24
           then
              Risk_Category = BIG   0.7
              Risk_Category = SMALL   0.3

RULE #3    if
                work_Catagory = A
           then
                Risk_Category = BIG   0.0706214689266
                Risk_Category = SMALL   0.929378531073

RULE #4    if
                work_Catagory = C
           then
                Risk_Category = BIG   0.5
                Risk_Category = SMALL   0.5

RULE #5 if
           work_Catagory = D
           then
                Risk_Category = BIG   0.666666666667
                Risk_Category = SMALL   0.333333333333

RULE #6    if
                work_Catagory = 2\24
                Reason_Payment = DEA
           then
                Risk_Category = BIG   1
                Risk_Category = SMALL   0

RULE #7    if
                work_Catagory = 2\24
                Reason_Payment = PPD
           then
                Risk_Category = BIG   0.75
                Risk_Category = SMALL   0.25
```

```
RULE #8   if
               work_Catagory = 2\24
               Reason_Payment = TTD
          then
               Risk_Category = BIG   0.333333333333
               Risk_Category = SMALL   0.666666666667

RULE #9   if
               work_Catagory = 2\24
               Reason_Payment = DEA
               PAcc_WithinQuarter = Q1
          then
               Risk_Category = BIG   1
               Risk_Category = SMALL   0

RULE #10  if
               work_Catagory = 2\24
               Reason_Payment = DEA
               PAcc_WithinQuarter = Q2
          then
               Risk_Category = BIG   1
               Risk_Category = SMALL   0

RULE #11  if
               work_Catagory = 2\24
               Reason_Payment = PPD
               PAcc_WithinQuarter = Q3
          then
               Risk_Category = BIG   1
               Risk_Category = SMALL   0

RULE #12  if
               work_Catagory = 2\24
               Reason_Payment = PPD
               PAcc_WithinQuarter = Q4
          then
               Risk_Category = BIG   0.5
               Risk_Category = SMALL   0.5

RULE #13  if
               work_Catagory = 2\24
               Reason_Payment = TTD
               PAcc_WithinQuarter = Q2
          then
               Risk_Category = BIG   0.5
```

```
                work_Catagory = A
                Reason_Payment = DEA
                PAcc_WithinQuarter = Q1
          then
                Risk_Category = BIG  1
                Risk_Category = SMALL  0

RULE #21  if
                work_Catagory = A
                Reason_Payment = DEA
                PAcc_WithinQuarter = Q3
          then
                Risk_Category = BIG  1
                Risk_Category = SMALL  0

RULE #22  if
                work_Catagory = A
                Reason_Payment = DEA
                PAcc_WithinQuarter = Q4
          then
                Risk_Category = BIG  1
                Risk_Category = SMALL  0

RULE #23  if
                work_Catagory = A
                Reason_Payment = PPD
                PAcc_WithinQuarter = Q2
          then
                Risk_Category = BIG  1
                Risk_Category = SMALL  0

RULE #24  if
                work_Catagory = A
                Reason_Payment = PPD
                PAcc_WithinQuarter = Q3
          then
                Risk_Category = BIG  1
                Risk_Category = SMALL  0

RULE #25  if
                work_Catagory = A
                Reason_Payment = PPD
                PAcc_WithinQuarter = Q4
          then
                Risk_Category = BIG  1
                Risk_Category = SMALL  0
```

```
RULE #26  if
               work_Catagory = A
               Reason_Payment = TTD
               PAcc_WithinQuarter = Q1
          then
               Risk_Category = BIG  0.0188679245283
               Risk_Category = SMALL  0.981132075472

RULE #27  if
               work_Catagory = A
               Reason_Payment = TTD
               PAcc_WithinQuarter = Q2
          then
               Risk_Category = BIG  0.0344827586207
               Risk_Category = SMALL  0.965517241379

RULE #28  if
               work_Catagory = A
               Reason_Payment = TTD
               PAcc_WithinQuarter = Q3
          then
               Risk_Category = BIG  0.0625
               Risk_Category = SMALL  0.9375

RULE #29  if
               work_Catagory = A
               Reason_Payment = TTD
               PAcc_WithinQuarter = Q4
          then
               Risk_Category = BIG  0.0208333333333
               Risk_Category = SMALL  0.979166666667

RULE #30  if
               work_Catagory = A
               Reason_Payment = TTD_ME
               PAcc_WithinQuarter = Q1
          then
               Risk_Category = BIG  0
               Risk_Category = SMALL  1

RULE #31  if
               work_Catagory = A
```

```
                Reason_Payment = TTD_ME
                PAcc_WithinQuarter = Q2
          then
                Risk_Category = BIG  0
                Risk_Category = SMALL  1

RULE #32 If
                work_Catagory = A
                Reason_Payment = TTD_ME
                PAcc_WithinQuarter = Q3
          then
                Risk_Category = BIG  0.0909090909091
                Risk_Category = SMALL  0.909090909091

RULE #33  if
                work_Catagory = A
                Reason_Payment = TTD_ME
                PAcc_WithinQuarter = Q4
          then
                Risk_Category = BIG  0
                Risk_Category = SMALL  1

RULE #34  if
                work_Catagory = C
                Reason_Payment = DEA
          then
                Risk_Category = BIG  1
                Risk_Category = SMALL  0

RULE #35  if
                work_Catagory = C
                Reason_Payment = ME
          then
                Risk_Category = BIG  0
                Risk_Category = SMALL  1

RULE #36  if
                work_Catagory = C
                Reason_Payment = TTD
          then
                Risk_Category = BIG  0
                Risk_Category = SMALL  1

RULE #37  if
                work_Catagory = C
                Reason_Payment = TTD_ME
```

```
            then
                 Risk_Category = BIG   0
                 Risk_Category = SMALL   1


RULE #38   if
                 work_Catagory = C
                 Reason_Payment = DEA
                 PAcc_WithinQuarter = Q3
            then
                 Risk_Category = BIG   1
                 Risk_Category = SMALL   0

RULE #39   if
                 work_Catagory = C
                 Reason_Payment = DEA
                 PAcc_WithinQuarter = Q4
            then
                 Risk_Category = BIG   1
                 Risk_Category = SMALL   0

RULE #40   if
                 work_Catagory = C
                 Reason_Payment = ME
                 PAcc_WithinQuarter = Q1
            then
                 Risk_Category = BIG   0
                 Risk_Category = SMALL   1

RULE #41   if
                 work_Catagory = C
                 Reason_Payment = ME
                 PAcc_WithinQuarter = Q2
            then
                 Risk_Category = BIG   0
                 Risk_Category = SMALL   1

RULE #42   if
                 work_Catagory = C
                 Reason_Payment = TTD
                 PAcc_WithinQuarter = Q3
            then
                 Risk_Category = BIG   0
                 Risk_Category = SMALL   1
```

```
RULE #43  if
                work_Catagory = C
                Reason_Payment = TTD_ME
                PAcc_WithinQuarter = Q2
          then
                Risk_Category = BIG  0
                Risk_Category = SMALL  1

RULE #44  if
                work_Catagory = D
                Reason_Payment = DEA
          then
                Risk_Category = BIG  1
                Risk_Category = SMALL  0

RULE #45  if
                work_Catagory = D
                Reason_Payment = ME
          then
                Risk_Category = BIG  0
                Risk_Category = SMALL  1

RULE #46  if
                work_Catagory = D
                Reason_Payment = PPD
          then
                Risk_Category = BIG  1
                Risk_Category = SMALL  0

RULE #47  if
                work_Catagory = D
                Reason_Payment = TTD
          then
                Risk_Category = BIG  0.4
                Risk_Category = SMALL  0.6

RULE #48  if
                work_Catagory = D
                Reason_Payment = DEA
                PAcc_WithinQuarter = Q4
          then
                Risk_Category = BIG  1
                Risk_Category = SMALL  0

RULE #49  if
                work_Catagory = D
```

```
            Reason_Payment = ME
            PAcc_WithinQuarter = Q4
        then
            Risk_Category = BIG  0
            Risk_Category = SMALL  1


RULE #50  if
            work_Catagory = D
            Reason_Payment = PPD
            PAcc_WithinQuarter = Q1
        then
            Risk_Category = BIG  1
            Risk_Category = SMALL  0

RULE #51  if
            work_Catagory = D
            Reason_Payment = PPD
            PAcc_WithinQuarter = Q3
        then
            Risk_Category = BIG  1
            Risk_Category = SMALL  0

RULE #52  if
            work_Catagory = D
            Reason_Payment = PPD
            PAcc_WithinQuarter = Q4
        then
            Risk_Category = BIG  1
            Risk_Category = SMALL  0

RULE #53  if
            work_Catagory = D
            Reason_Payment = TTD
            PAcc_WithinQuarter = Q1
        then
            Risk_Category = BIG  0
            Risk_Category = SMALL  1

RULE #54  if
            work_Catagory = D
            Reason_Payment = TTD
            PAcc_WithinQuarter = Q2
        then
            Risk_Category = BIG  0
```

```
                Risk_Category = SMALL  1

RULE #55  if
                work_Catagory = D
                Reason_Payment = TTD
                PAcc_WithinQuarter = Q3
          then
                Risk_Category = BIG  0.5
                Risk_Category = SMALL  0.5

RULE #56  if
                work_Catagory = D
                Reason_Payment = TTD
                PAcc_WithinQuarter = Q4
          then
                Risk_Category = BIG  1
                Risk_Category = SMALL  0
```

Declaration

I, the undersigned, declare that this thesis is my original work and has not been presented as a partial degree requirement for a degree in any other university and that all sources of materials used for the thesis have been duly acknowledged.

Mesfin Fikre Woldmariam

June 2005

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

---

Dr. B. L. Desai

June 2005

ADDIS ABABA UNIVERSITY
FACULTY OF INFORMATICS
BIBLIOGRAPHIC LAB