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

APPLICATION OF DATA MINING TECHNIQUES TO SUPPORT IMPORTERS AND EXPORTERS DELINQUENCY PREDICTION: THE CASE OF NATIONAL BANK OF ETHIOPIA

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

BY
YISHAK YILMA

January 2009

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

FACULTY OF INFORMATICS

INFORMATION SCIENCE DEPARTMENT

APPLICATION OF DATA MINING TECHNIQUES TO SUPPORT IMPORTERS AND EXPORTERS DELINQUENCY PREDICTION: THE CASE OF NATIONAL BANK OF ETHIOPIA

BY

YISHAK YILMA

Name and Signature of Members of the Examining Board

_____	_____	_____
Chair Person, Examining Board	Signature	Date
_____	_____	_____
Advisor	Signature	Date
_____	_____	_____
Chair Person, Faculty	Signature	Date
_____	_____	_____
Chair Person, Graduate Council	Signature	Date

DEDICATION

I would like to dedicate this paper to my GOD, who allows me to start and come to the end, and always supports me to pass all the challenges I went through. Thank You.

ACKNOWLEDGMENT

First and foremost I offer my sincerest gratitude to my advisor, Dr Manoj V.N.V, who has supported me throughout my thesis with his knowledge, patience and encouragement.

I am especially indebted to National Bank of Ethiopia ISD staffs, Ato Tesfaye Teshome, Ato Assefa Yeshanew, Ato Wondesson Tsegaw for their unreserved support until the end of my study. Alem W/gerima, also provided me the necessary information for the research work in addition to his good advices. Thank you.

I would like to thank my friends Amha Birru, Dagne Minda, Haileyesus Mulugeta, Meshesha Legese and my classmates for their friendship advice and support throughout my study.

I am also very grateful to Aklile Mitiku and Tsegaye G/Medhin for their valuable comment on the study.

My special thanks also go to my family for their love, advice and encouragement, and overwhelming and tremendous support in the pursuit of this study.

Table of Content

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN	Artificial Neural Networks
CAD	Cash Against Document
CART	Classification And Regression Trees
CHAID	Chi-squared Automatic Interaction Detection
CRISP–DM	Cross-Industry Standard Process for Data Mining
DSF	Decision-Support Functions
FEMoS	Foreign Exchange Monitoring System
F/C	Foreign Currency
GNU	General Public License
HS	Harmonized System
IBM	International Business Machine
IBO	International Banking Operations
IMF	International Monetary Fund
KDD	Knowledge Discovery in Databases
NBE	National Bank of Ethiopia
OLAP	On-Line Analytical Processing

ABSTRACT

The application of Data mining is becoming increasingly common in both the private and public sectors. Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce costs, risks, and increase sales.

Foreign currency is a scarce financial resource in Ethiopia. This scarcity calls for the consolidation and fostering of existing financial management systems to ensure optimum utilization of the available resource. The National Bank of Ethiopia (NBE) has the task of monitoring the settlement of importers and exporters foreign exchange commitments as per the existing directives. Currently in the NBE there are many importers and exporters who are delinquent and expected to settle their commitment.

Thus, it is the aim of this study to examine the potential applicability of data mining technology in building a predictive data mining model that helps to predict potentially delinquent or non delinquent importers and exporters in relation to their utilization of the foreign currency.

To conduct the study, the researcher adopted the Cross-Industry Standard Process for Data Mining (CRISP–DM) process model. Several predictive classification models were built both in decision tree and neural network techniques using WEKA software. The best performing model was chosen by comparing the models using standard evaluation criteria such as accuracy, precision, recall and interpretability.

According to the evaluation results, both techniques have shown a promising performance. However, the best models for both export and import transactions were obtained using decision tree techniques. The decision tree approach brings about 94.02% accuracy in the case of predicting export transactions and 98.03% for import transactions. Moreover, the models built by

the decision tree show better results in terms of precision, recall and interpretability for both transactions. Thus, compared to neural network, the decision tree approaches are more applicable in addressing the research problem.

Accordingly, some important rules are derived using the selected attributes such as, MethodofPayment, BaseOfShipment, Country_Region, AmtOfBirrIn_Range, Currency, Validity_period and EconomicSector that are relevant in business decision making.

In general, the results obtained from the study proved the potential applicability of data mining technology to predict importers and exporters into predefined classes (delinquent and non-delinquent) based on their transaction characteristics.

CHAPTER ONE

1 INTRODUCTION

1.1 Background

Over the past several years, there has been a rapid increase in the volume of information collected and stored, with some observers suggesting that the quantity of the world's data approximately doubles every year [12]. The increase in computer hardware technology in the past has led to large supplies of powerful and affordable computers, data collection equipment and storage devices. This technology has resulted in the processing and accumulation of excess amount of data without being analyzed and used to discover important knowledge from it. This gave rise to a data rich but information poor situation where there is a widening gap between the explosive growth of data and its types, and the ability to analyze and interpret it.

It was also recognized that information is at the heart of business operations and that decision-makers could make use of this information to gain valuable insight into the business. Database Management systems gave access to the data stored but this was only a small part of what could be gained from the data. Traditional on-line transaction processing systems are good at putting data into databases quickly, safely and efficiently but are not good at delivering meaningful analysis in return. Analyzing data can provide further knowledge about a business by going beyond the data explicitly stored to derive knowledge about the business. Hence there is a need for a new generation of automated and intelligent tools and techniques known as Data Mining, to look for patterns in data [7].

Data mining is the process of discovering, extracting and analyzing meaningful patterns, structure, models, and rules from large quantities of data [4]. The process is automatic or semi automatic, with interactive steps such as problem and data understanding, data selection, data preprocessing and cleaning, data transformation, incorporation of appropriate domain knowledge to select data mining task and algorithm, application of data mining algorithm(s), and knowledge interpretation and evaluation. It sits at the common frontiers of several fields including Data Base Management, Statistics, Artificial Intelligence, Machine Learning, Pattern Recognition, and Data Visualization. Today, data mining has a major impact in business, industry, and science.

There are different functionalities in data mining such as characterization and discrimination, association analysis, classification and prediction, cluster analysis, etc. These functionalities are used to specify the kind of patterns found in data mining tasks.

Data mining is used to build predictive and descriptive models. A predictive model is used to explicitly predict values while a descriptive model is applied to describe patterns existing in the data.

A number of researches have been done to show the possible applications of data mining techniques in different parts of the world including Ethiopia. Researchers tried to prove its applicability in many domain areas and organizations in Ethiopia. The first attempt was made by Gobena [1999] that was on the application of data mining technology and techniques in Ethiopian Airlines and this work was extended by Henok [2002] and Denekew [2003]. Askale [2001], Tesfaye [2002] and Meherework [2004] also conducted other research works on the application of data mining technology in banking and insurance industries for risk management purpose. Moreover Yoshep [2004], Fekadu [2004] and Abenet [2005] were assessed the potential

applicability of data mining techniques in the Ethiopian context with different areas and organizations. Therefore, this research work also attempts to explore the potential applicability of the data mining technology in National Bank of Ethiopia in order to reduce the number of delinquent customers with the utilization of foreign currency.

The National Bank of Ethiopia is the central bank of the Federal Democratic Republic of Ethiopia established by proclamation No. 83/1994 in order to "foster monetary stability, a sound financial system and such other credit and exchange conditions, as are conducive to the balanced growth of the economy of the country."

The Bank manages and administers the country's international reserves required for the payments of import as well as for meeting external debt servicing and other payment obligations. In its functions, the Bank endeavors to maintain, at all times sufficient international reserve fund, while considering the risk and return factors associated with management of these reserves.

The International Banking Operations Department is the key department to undertake part of the tasks entrusted to the Bank as promulgated by the establishing proclamation. This department has two wings; namely, the operation wing and the monitoring wing. The monitoring wing, which is the focus of this research, has the statistics, forex bureaux follow up and transfer, and foreign exchange follow up divisions.

The Foreign Exchange Follow-up division is entrusted with the task of monitoring the settlement of import and export foreign exchange commitments as per the existing directives and procedures of the Bank.

To this effect, the division

- Follows up the utilization of foreign exchange allowed for the purpose of imports.
- Follows up the repatriation of foreign currency to Ethiopia from exports made to other countries.
- Carries out ex-post verification of export and import documents issued by commercial banks.

Volume of Transaction

The volumes of transaction in some of the functions are tabulated as follows.

Table 1.1-1: Volume of transactions

Type of Transaction	# of Transactions/ Month
Import Permits	2725
Import Utilization	2800
Import Declaration	2800
Export Permit	1880
Export Utilization	300

1.2 Statement of the Problem

There is a scarce foreign currency in Ethiopia that needs optimization. As a result, proper management of foreign exchange is vital in order to avoid misappropriation and thereby maximize its use from limited supply. The National Bank of Ethiopia has a sole responsibility, given by the Government, to manage foreign currency reserves. The reserve of foreign currency is highly dependent on the utilization of the foreign currency by importers and the system of control on export transactions.

The NBE has importer and exporter monitoring database system that registers transactions of each importer and exporter in the country. The system registers permitted amount for each transactions when permit is given to an importer or exporter by different commercial banks acting on behalf of NBE. The importer or exporter is expected to inform the Bank of the amount of permit utilized by presenting legal documents from the company that they are dealing with. The amounts of used permit are then registered in the system as well.

Despite the existence of import and export transactions' monitoring database system within the NBE, the Bank has been facing with delinquency problems with respect to the utilization of foreign currency by importers and exporters. Under normal circumstances, an importer has to utilize the foreign currency properly for the intended items and finally present a legal document, which certifies the intended item is imported, to the NBE. If the importer fails to do so then it will be classified as delinquent. Similarly, an exporter has to submit the equivalent amount of foreign currency obtained from the items it has exported. Otherwise, it will be classified as delinquent. As a result of these delinquencies, the Bank faces difficulties to properly manage the Country's foreign currency. This leads to suboptimal utilization of the scarce foreign currency reserve.

Therefore, it is the aim of this research to develop a model that helps predict the delinquency behavior of importers and exporters with respect to the utilization of foreign currency by using appropriate data mining tools and techniques.

1.3 Justification of the Research

To date, the NBE has no practical mechanism to predict the behavior of importers or exporters on the utilization of foreign currency. As a result, there is a huge amount of obligation unsettled for so many years and there are very many customers who are delinquent and required to clear up their commitment.

In order to optimally utilize the scarce foreign currency with respect to import and export transactions, the NBE should therefore need to put in place appropriate mechanisms that enable it to oversee the behavioral tendency of importers and exporters. One mechanism can be, building a predictive data mining model that helps to predict the potentially delinquent or non-delinquent importer or exporter in relation to their utilization of the foreign currency. Because applying such a model aids the Bank to identify the delinquent client well in advance and with greater degree of confidence, it serves as a critical precautionary tool for the Bank's management.

This research is, therefore, an attempt to demonstrate potential mechanisms that could be applied by the NBE in its effort to controlling delinquency problems associated with foreign currency management.

1.4 Objective of the Research

The general and specific objectives of this research work are the following: -

1.4.1 General Objective

The general objective of the research is to explore the potential applicability of data mining technology in developing a model that can predict importers and exporters transactions as delinquent or non delinquent.

1.4.2 Specific Objectives

In order to achieve the above stated general objective, the following specific objectives are formulated.

- Conduct a thorough review of literature on the existing data mining techniques in general, and their application in identifying critical patterns.
- Select and extract the dataset required for analysis from the database.
- Prepare the data for analysis which includes handling inconsistent and noisy data, accounting for missing values, deriving other fields from existing ones, transformation and integration.
- Assess different data mining application tools supporting different data mining functions and techniques and that are more appropriate to the problem domain, and select the best tool.
- Build and train data mining models.
- Evaluate the accuracy of the models.
- Report results and make recommendations.

1.5 Research Methodology

1.5.1 Literature Review

The relevant literatures were intensively reviewed to assess data mining technology, both concepts and techniques, and researches in this field. Various books, journals, magazines, articles, and papers from the Internet pertaining to the subject matter of data mining and Knowledge Discovery Process in Databases (KDD) were thoroughly reviewed to understand the potential applicability of data mining technology in general and classification and prediction rules in particular with respect to the problem of the domain.

1.5.2 Business Understanding

Interviews, observations and document review were made to analyze the business problem, and have good background knowledge in interpreting results of the data mining process.

1.5.3 Data Source and Collection

The data source for this research was the foreign exchange monitoring database system of the National Bank of Ethiopia. The database records importer and exporter transactions. More than 500,000 data records of importer and exporter transactions were available in the foreign exchange monitoring database system. These records are found in a number of tables related to each other and each table has a number of attributes such as MethodOf Payment, Currency, HsCode, HsDescription, DateOfIssued, ValidityDate, ExchangeRate, AmountDeclared, AmountCancelled, CountryName, FobForeignCurrency, UnitOfMeasurment, QualityDesc, GradeDesc, Quantity, TypeofProduct, UnitPrice, PermitNo, etc. A thorough investigation and discussion with the domain experts were made on the selection of data attributes that related to the problem of the domain.

1.5.4 Data Preparation and Transformation

Data preprocessing is an important step that needs a considerable amount of time since the quality of data determines the quality of the output information mined from the respective dataset. In reality large databases are subject to noise, inconsistency and missing values that can hinder the quality and efficiency of the mining process. Hence, the collected data were preprocessed into a form that is suitable for data mining purpose. Pre-processing tasks like handling noisy data, unknown values, missing values, deriving new fields from the existing ones, and transformation of data were performed.

1.5.5 Model Building and Evaluation of Results

As the theme of the research was building a predictive classification model, decision tree and neural network algorithms were selected as data mining techniques for the purpose. The reasons for the selection of these two data mining techniques are explained in detail in the literature review and modeling chapters. The results obtained from the selected algorithms were tested and compared in order to determine the best algorithm for the problem domain. WEKA software was used as data mining tool for building the models.

After the models were built and trained, the next action was evaluating or assessing and interpreting the results of the models. This step would help in selecting the best model that finds an interesting pattern. While evaluating a model, its accuracy, the types of errors and the cost associated with it were considered. One of the various approaches employed to deal with such issues was confusion matrix. This matrix helps understand how well the model predicts and also equally aids to see exactly where things may have gone wrong. Finally, based on assessment and evaluation of such models the best model that explores an interesting pattern was selected. And this provides a good ground for result interpretation and recommendation.

1.6 Scope and Limitation of the Research

The scope of the research focuses on assessing the potential applicability of data mining technologies, specifically, predictive modeling, in supporting importers and exporters delinquency prediction at the NBE in relation to their use of foreign currency. To this end, the research analyses the problem from the perspective of two critical sectors that have significant influence on the foreign currency reserve. The analysis will concentrate on the delinquency classification problems vis-à-vis importers and exporters.

The major limitations encountered while undertaking this research were the difficulty in accessing important attributes such as exchange rate and absence of full details for some attributes such as commodity details. Apart from creating workload on data preprocessing task, this was a constraint on the dimensionality of the data collected.

1.7 Organization of the Thesis

The thesis is organized in to six chapters. The first chapter presents introduction of the research. In the second chapter, basic data mining concepts and applications are discussed as reviewed from literature and similar research works. The third chapter reviews the work flow of the existing system. The fourth chapter illustrates the step by step tasks executed for data understanding and preparation phase. The fifth chapter deals with the experimentation process. The final chapter concludes the research work and provides some useful recommendations based on the results of the experimentation.

CHAPTER TWO

2 LITERATURE REVIEW

2.1 Introduction

The modern world is a data-driven one. We are surrounded by data, numerical and otherwise, which must be analyzed and processed to convert it into information that informs, instructs, answers, or otherwise aids understanding and decision-making. This is the age of the Internet, intranets, data warehouses and data marts, and the fundamental paradigms of classical data analysis are ripe for change [13].

Progress in storage technology is allowing vast amounts of raw data to accumulate in both private and public databases. Insurers, banks, hotel chains, airlines, retailers, telecommunications and other enterprises are rapidly accumulating information from day to day transactions with their customers.

Nowadays each individual and organization – business, family or institution – can access a large quantity of data and information about itself and its environment. This data has the potential to predict the evolution of interesting variables or trends in the outside environment, but so far that potential has not been fully exploited.

According to Witten and Frank [28], there is a gap between the generation of data and our understanding of it. As the volume of data increases, the proportion of it that people understand decreases. There is always hidden, potentially useful information in all these data, which should be made explicit. The abundance of data and the need for powerful data analysis tools has been described as a data rich and information poor situation. The fast growing, large amount of data,

collected and stored in large and many databases, has far exceeded our human ability for comprehension without powerful tools.

According to Han and Kamber [9], having similar view stated that 'Our capabilities of both generating and collecting data have been increasing in the last several decades, contributing factors being the computerization of many businesses, scientific and governmental transactions, advances in data collection tools ranging from scanned text and image platforms to satellite remote sensing systems'. This explosive growth in stored data has generated an urgent need for new techniques and automated tools that can assist us in transforming the vast amount of data into useful information and knowledge.

Data might be one of the most valuable assets of any corporation but only if it knows how to reveal valuable knowledge hidden in raw data. Data mining allows extracting diamonds of knowledge from the historical data, and predicting outcomes of future situations. It helps optimize business decisions, increase the value of each customer and communication, and improve customer satisfaction.

2.2 Overview of Data Mining

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets [26]. These tools can include statistical models, mathematical algorithms, and machine learning methods (algorithms that improve their performance automatically through experience, such as neural networks or decision trees). Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction.

Data mining can be performed on data represented in quantitative, textual, or multimedia forms. Data mining applications can use a variety of parameters to examine the data. They include association (patterns where one event is connected to another event, such as purchasing a pen and purchasing paper), sequence or path analysis (patterns where one event leads to another event, such as the birth of a child and purchasing diapers), classification (identification of new patterns, such as coincidences between duct tape purchases and plastic sheeting purchases), clustering (finding and visually documenting groups of previously unknown facts, such as geographic location and brand preferences), and forecasting (discovering patterns from which one can make reasonable predictions regarding future activities, such as the prediction that people who join an athletic club may take exercise classes) [12].

As an application, compared to other data analysis applications, such as structured queries (used in many commercial databases) or statistical analysis software, data mining represents a difference of kind rather than degree. Many simpler analytical tools utilize a verification-based approach, where the user develops a hypothesis and then tests the data to prove or disprove the hypothesis.

In contrast, data mining utilizes a discovery approach, in which algorithms can be used to examine several multidimensional data relationships simultaneously, identifying those that are unique or frequently represented.

Data mining should be applicable to any kind of information repository. This includes relational database, data warehouses, transitional databases, advanced database systems, flat files, and the World Wide Web [2]. When data mining is applied to relational databases, one can go further searching for trends or data patterns. For example data mining systems may analyze customer

data to predict the credit risk of new customers based on their income, age, and pervious credit information.

2.3 Data Mining and Knowledge Discovery

Data mining is the science and technology of exploring data in order to discover previously unknown patterns. Data Mining is a part of the overall process of Knowledge Discovery in databases (KDD). The accessibility and abundance of information today makes data mining a matter of considerable importance and necessity [8].

Traditionally data collection is considered to be one of the most important stages in data analysis. An analyst (e.g., a statistician) used the available domain knowledge to select the variables to be collected. The number of variables selected was usually small and the collection of their values could be done manually (e.g., utilizing hand-written records or oral interviews). In the case of computer-aided analysis, the analyst had to enter the collected data into a statistical computer package or an electronic spreadsheet. Due to the high cost of data collection, people learned to make decisions based on limited information [18].

Unfortunately, as the amount of machine readable information increases, the ability to understand and make use of it does not keep pace with its growth. Data mining is a term coined to describe the process of sifting through large databases in search of interesting patterns and relationships. Practically, Data Mining provides tools by which large quantities of data can be automatically analyzed. Some of the researchers consider the term "Data Mining" as misleading and prefer the term "Knowledge Mining" as it provides a better analogy to gold mining [14].

Data mining derives its name from the similarities between searching for valuable information in a large database and mining rocks for a vein of valuable ore. Both imply either sifting through a large amount of material or ingeniously probing the material to exactly pinpoint where the values reside. It is, however, a misnomer, since mining for gold in rocks is usually called “gold mining” and not “rock mining”, thus by analogy, data mining should have been called “knowledge mining” instead. Nevertheless, data mining became the accepted customary term, and very rapidly a trend that even overshadowed more general terms such as knowledge discovery in databases (KDD) that describe a more complete process. Other similar terms referring to data mining are: data dredging, knowledge extraction and pattern discovery.

The Knowledge Discovery in Databases (KDD) process was defined by many, for instance Fayyad et al. [7] define it as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data". Friedman considers the KDD process as an automatic exploratory data analysis of large databases. Hand views it as a secondary data analysis of large databases. The term "Secondary" emphasizes the fact that the primary purpose of the database was not data analysis. Data Mining can be considered as a central step of the overall process of the Knowledge Discovery in Databases (KDD) process. Due to the centrality of data mining in the KDD process, there are some researchers and practitioners that use the term "data mining" as synonymous to the complete KDD process [19].

Several researchers have proposed different ways to divide the KDD process into phases. According to Oded & Lior [18], a hybridization of these proposals and suggests breaking the KDD process into the following eight phases.

(1) Developing an understanding of the application domain, the relevant prior knowledge and the goals of the end-user.

(2) Selecting a data set on which discovery is to be performed.

(3) Data Preprocessing: This stage includes operations for Dimension Reduction (such as Feature Selection and Sampling), Data Cleaning (such as Handling Missing Values, Removal of Noise or Outliers), and Data Transformation (such as Discretization of Numerical Attributes and Attribute Extraction)

(4) Choosing the appropriate Data Mining task such as: classification, regression, clustering and summarization.

(5) Choosing the Data Mining algorithm: This stage includes selecting the specific method to be used for searching patterns.

(6) Employing The Data mining Algorithm.

(7) Evaluating and interpreting the mined patterns.

(8) Deployment: Using the knowledge directly, incorporating the knowledge into another system for further action or simply documenting the discovered knowledge.

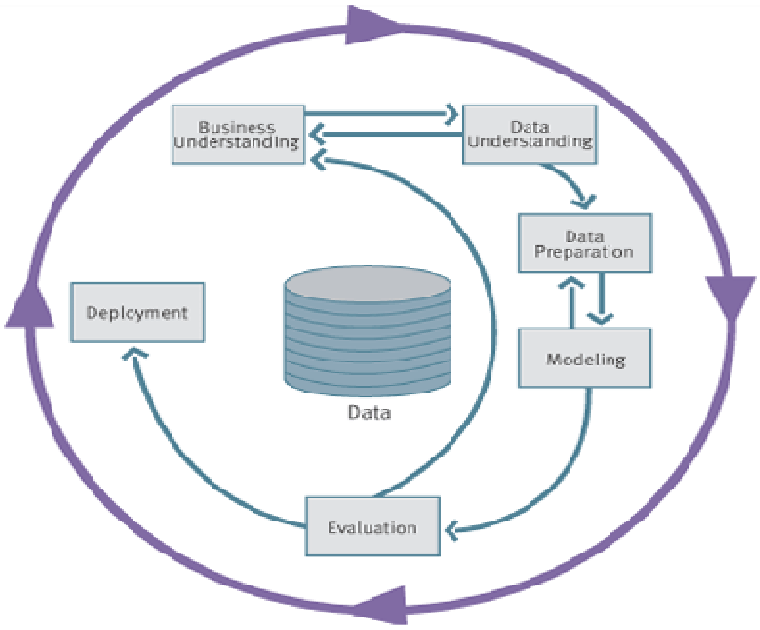
The KDD is an iterative process. Once the discovered knowledge is presented to the user, the evaluation measures can be enhanced, the mining can be further refined, new data can be selected or further transformed, or new data sources can be integrated, in order to get different, more appropriate results.

2.4 Models for Data Mining

In the data mining literature, various "general frameworks" have been proposed to serve as blueprints for how to organize the process of gathering data, analyzing data, disseminating results, implementing results, and monitoring improvements.

One such model, CRoss-Industry Standard Process for Data Mining (CRISP-DM) is suggested by a consortium including SPSS and others published in 1999 [21]. This was done, among other reasons, in response to a perceived need to have an industry, tool, and application neutral process that would provide a point of commonality for the introduction of new ideas and applications in an environment where the KDD process is lengthy, complex and has many variations in approach to the individual parts of that process. CRISP-DM has major steps of business understanding, data understanding, data preparation, modeling, evaluation and deployment.

Figure 2.4-1: CRISP–DM process model.



Note that the phase sequence is adaptive. That is, the next phase in the sequence often depends on the outcomes associated with the preceding phase. The arrows indicate the most significant dependencies between phases. The following points state the phases of CRISP-DM as stated by [21]. Further descriptions were also presented referring to other literatures.

2.4.1 Business Understanding

This initial phase focuses on understanding the research objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

2.4.2 Data Understanding

The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

2.4.3 Data Preparation

The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record and attribute selection as well as transformation and cleaning of data for modeling tools.

Explaining the need to pre-process data, Daniel [5] stated that, 'To be useful for data mining purposes, the databases need to undergo preprocessing, in the form of data cleaning and data transformation. Data mining often deals with data that hasn't been looked at for years, so that much of the data contains field values that have expired, are no longer relevant, or are simply missing. The objective is to minimize the 'garbage' that gets into our model so that we can minimize the amount of garbage that our models give out'.

According to Dorian [6], in his book 'Data Preparation for Data Mining', estimated that data preparation alone accounts for 60% of all the time and effort expanded in the entire data mining process.

Data Cleaning

The tasks of data cleaning attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

Missing Values: There are different methods to handle missing values such as:

- Ignore the record (usually done when the class label is missing)
- Fill in the missing value manually (time consuming and may not be feasible)
- Use global constant to fill in the missing values (Replace all missing values by the same constant)
- Use attribute mean to fill in the missing values (especially for numeric values)
- Use the most probable value to fill in the missing value

Noisy Data and Outliers: - As Han and Kamber [9] described noise as a random error or variance in a measured variable. Outliers are unusual data values that are not consistent with most observations. Commonly, outliers result from measurement errors, coding and recording errors, and, sometimes, are natural, abnormal values. Such non representative samples can seriously affect the model produced later".

There are two strategies for dealing with outliers and noisy values:

- Detect and eventually remove outliers as a part of the preprocessing phase, or
- Develop robust modeling methods that are insensitive to outliers.

2.4.4 Data Integration and Transformation

Data mining often requires data integration- the merging of data from multiple data stores. The data may also need to be transformed into forms appropriate for mining [9]. Data are transformed or consolidated in to forms appropriate for mining and involves many activities such as Smoothing (to remove noise from data), Aggregation (summary or aggregation operations are applied), Generalization (low level data are replaced by higher-level concepts), Normalization (attributes are scaled to fall within a specified ranges), Attribute Construction (feature construction-new attributes are constructed).

2.4.5 Modeling

A model is a high-level, global description of a data set. It takes a large sample perspective. It may be descriptive- summarizing the data in a convenient and concise way- or it may be inferential, allowing one to make some statement about the population from which the data were drawn or about likely future data values [10].

In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often necessary.

2.4.6 Evaluation

Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model and review the steps executed to construct the model to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue

that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

2.4.7 Deployment

Creation of the model is generally not the end of the research. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. It often involves applying 'live' models within an organization's decision making processes, for example in real-time personalization of Web pages or repeated scoring of marketing databases.

However, depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise. In many cases it is the customer, not the data analyst, who carries out the deployment steps. However, even if the analyst will not carry out the deployment effort it is important for the customer to understand up front what actions need to be carried out in order to actually make use of the created models [24].

2.5 Data Warehousing, OLAP and Data Mining

2.5.1 Data Warehousing for Data Mining

According to Kantardzic [13] one of the global definitions is that 'The data warehouse is a collection of integrated, subject-oriented databases designed to support the Decision-Support Functions (DSF), where each unit of data is relevant to some moment in time'.

Based on this definition, a data warehouse can be viewed as an organization's repository of data, set up to support strategic decision-making. The function of the data warehouse is to store the

historical data of an organization in an integrated manner that reflects the various facts of the organization and business. The data in a warehouse are never updated but used only to respond to queries from end users who are generally decision-makers. Typically, data warehouses are huge, storing billions of records. In many instances, an organization may have several local or departmental data warehouses often called data marts. A data mart is a data warehouse that has been designed to meet the needs of a specific group of users. It may be large or small, depending on the subject areas [11].

Although the existence of a data warehouse is not a prerequisite for data mining, in practice, the task of data mining, especially for some large companies, is made a lot easier by having access to a data warehouse.

Data mining represents one of the major applications for data warehousing, since the sole function of a data warehouse is to provide information to end users for decision support. Unlike other query tools and application systems, the data-mining process provides an end-user with the capacity to extract hidden, nontrivial information. Such information, although more difficult to extract can provide bigger business and scientific advantages and yield higher returns on 'data warehousing and data mining' investments.

2.5.2 OLAP vs Data Mining

One of the most common questions from data processing professionals is about the difference between data mining and OLAP (On-Line Analytical Processing). However, they are very different tools that can complement each other.

According to a description by Two Crows Corporation [27], OLAP is part of the spectrum of decision support tools. Traditional query and report tools describe what is in a database. OLAP goes further; it's used to answer why certain things are true. The user forms a hypothesis about a relationship and verifies it with a series of queries against the data. In other words, the OLAP analyst generates a series of hypothetical patterns and relationships and uses queries against the database to verify them or disprove them.

Data mining is different from OLAP because rather than verify hypothetical patterns, it uses the data itself to uncover such patterns. It is essentially an inductive process than OLAP which is a deductive process. However, OLAP is complementary in the early stages of the knowledge discovery process because it can help you explore your data, for instance by focusing attention on important variables.

The derivation of answers from data in OLAP is analogous to calculations in a spreadsheet; because they use simple and given-in-advance calculations, OLAP tools do not learn from data, nor do they create new knowledge. They are usually special-purpose visualization tools that can help end-users draw their own conclusions and decisions, based on graphically condensed data. OLAP tools are very useful for the data-mining process; they can be a part of it but they are not a substitute [13].

2.6 Data Mining Functionalities

According to Osmar [19], the kinds of patterns that can be discovered depend upon the data mining tasks employed. By and large, there are two types of data mining tasks: descriptive data mining tasks that describe the general properties of the existing data, and predictive data mining tasks that attempt to do predictions based on inference on available data.

The data mining functionalities and the variety of knowledge they discover are briefly presented as follows:

2.6.1 Characterization

Data characterization is a summarization of general features of objects in a target class, and produces what is called characteristic rules. The data relevant to a user-specified class are normally retrieved by a database query and run through a summarization module to extract the essence of the data at different levels of abstractions. For example, one may want to characterize the OurVideoStore customers who regularly rent more than 30 movies a year. With concept hierarchies on the attributes describing the target class, the attribute oriented induction method can be used, for example, to carry out data summarization. Note that with a data cube containing summarization of data, simple OLAP operations fit the purpose of data characterization [19].

2.6.2 Discrimination

Data discrimination produces what are called discriminate rules and is basically the comparison of the general features of objects between two classes referred to as the target class and the contrasting class. For example, one may want to compare the general characteristics of the customers who rented more than 30 movies in the last year with those whose rental account is lower than 5. The techniques used for data discrimination are very similar to the techniques used for data characterization with the exception that data discrimination results include comparative measures [19].

2.6.3 Association Analysis

Association analysis is the discovery of what are commonly called association rules. It studies the frequency of items occurring together in transactional databases, and based on a threshold called support, identifies the frequent item sets. Another threshold, confidence, which is the conditional probability than an item appears in a transaction when another item appears, is used to pinpoint association rules. Association analysis is commonly used for market basket analysis [19].

2.6.4 Classification

Classification analysis is the organization of data in given classes. Also known as supervised classification, the classification uses given class labels to order the objects in the data collection. Classification approaches normally use a training set where all objects are already associated with known class labels.

The classification algorithm learns from the training set and builds a model. The model is used to classify new objects. For example, after starting a credit policy, the OurVideoStore managers could analyze the customers' behaviors vis-à-vis their credit, and label accordingly the customers who received credits with three possible labels "safe", "risky" and "very risky". The classification analysis would generate a model that could be used to either accept or reject credit requests in the future [19].

2.6.5 Prediction

Prediction has attracted considerable attention given the potential implications of successful forecasting in a business context. There are two major types of predictions: one can either try to predict some unavailable data values or pending trends, or predict a class label for some data. The latter is tied to classification. Once a classification model is built based on a training set, the class

label of an object can be foreseen based on the attribute values of the object and the attribute values of the classes. Prediction is however more often referred to the forecast of missing numerical values, or increase/ decrease trends in time related data. The major idea is to use a large number of past values to consider probable future values [19].

2.6.6 Clustering

Similar to classification, clustering is the organization of data in classes. However, unlike classification, in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes. Clustering is also called unsupervised classification, because the classification is not dictated by given class labels. There are many clustering approaches all based on the principle of maximizing the similarity between objects in a same class (intra-class similarity) and minimizing the similarity between objects of different classes (inter-class similarity) [19].

2.6.7 Outlier Analysis

Outliers are data elements that cannot be grouped in a given class or cluster. Also known as exceptions or surprises, they are often very important to identify. While outliers can be considered noise and discarded in some applications, they can reveal important knowledge in other domains, and thus can be very significant and their analysis valuable [19].

2.6.8 Evolution and Deviation Analysis

Evolution and deviation analysis pertain to the study of time related data that changes in time. Evolution analysis models evolutionary trends in data, which consent to characterizing, comparing, classifying or clustering of time related data. Deviation analysis, on the other hand,

considers differences between measured values and expected values, and attempts to find the cause of the deviations from the anticipated values.

It is common that users do not have a clear idea of the kind of patterns they can discover or need to discover from the data at hand. It is therefore important to have a versatile and inclusive data mining system that allows the discovery of different kinds of knowledge and at different levels of abstraction. This also makes interactivity an important attribute of a data mining system [19].

2.7 Data Mining Techniques

2.7.1 Decision Tree

A Decision tree is a classifier expressed as a recursive partition of the instance space. A decision tree consists of nodes that form a Rooted Tree, meaning it is a Directed Tree with a node called root that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called internal node or test nodes. All other nodes are called leaves (also known as terminal nodes or decision nodes) [18].

In a decision tree, each internal node splits the instance space into two or more subspaces according to a certain discrete function of the input attributes values. In the simplest and most frequent case each test considers a single attribute, such that the instance space is partitioned according to the attribute's value. In the case of numeric attributes the condition refers to a range.

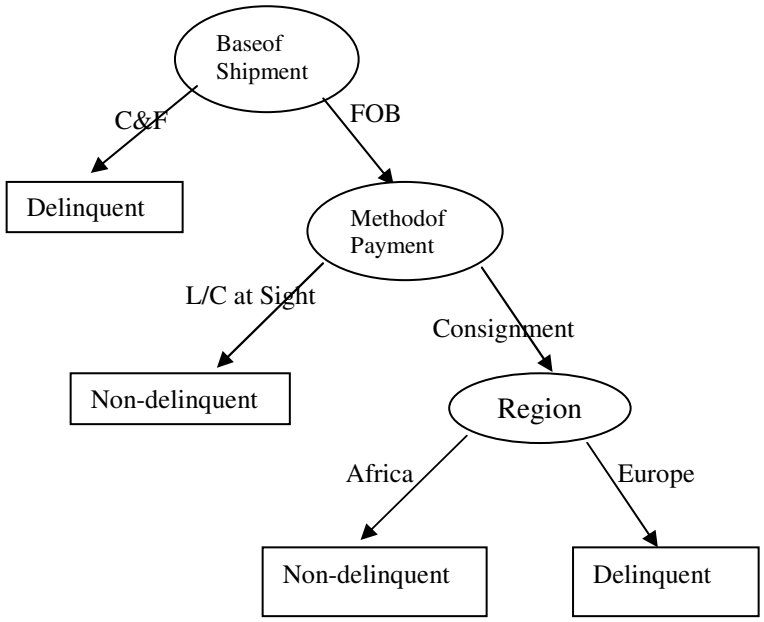
A decision tree is a predictive model that, as its name implies, can be viewed as a tree. Specifically each branch of the tree is a classification question and the leaves of the tree are partitions of the dataset with their classification [24].

Classification is similar as the process of finding a set of models (functions) that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. In line with this Han and Kamber [9] assert decision trees are one of the most commonly used algorithms used to perform classification.

Decision trees models are commonly used in data mining to examine the data and induce the tree and its rules that will be used to make predictions. A number of different algorithms may be used for building decision tree including CHAID (Chi-squared Automatic Interaction Detection), CART (Classification And Regression Trees), Quest, and C5.0 [26].

Figure 2.7-1 describes a decision tree to the classification problem (whether or not an export transaction is delinquent or non delinquent). Internal nodes are represented as circles whereas leaves are denoted as rectangles. The root node in this example is “BaseOfShipment”. Depending on the mode of shipment the root node split in to branches, each representing one of the possible answers. Each branch will lead either to another decision node or to the leaf node. Each node is labeled with the attribute it tests, and its branches are labeled with its corresponding values. By navigating through the decision tree, one can assign a value or class to a specific case by deciding which branch to take, starting at the root node and moving to each subsequent node until a leaf node reached.

Figure 2.7-1: Sample Decision Tree



Naturally, decision makers prefer a less complex decision tree, as it is considered more comprehensible. Furthermore, the tree complexity has a crucial effect on its accuracy performance. Usually large trees are obtained by over fitting the data and hence exhibit poor generalization ability. Nevertheless a large decision tree can be accurate if it was induced without over fitting the data. The tree complexity is explicitly controlled by the stopping criteria used and the pruning method employed.

2.7.1.1 Decision Tree Induction

The basic algorithm for decision tree induction is a greedy algorithm which constructs decision trees in a top-down recursive divide-and-conquer manner. The algorithm is a version of ID3, a well-known decision tree induction algorithm [9].

The basic strategy is as follows:

- The tree starts as a single node representing the training samples.
- If the samples are all of the same class, then the node becomes a leaf and is labeled with that class.
- Otherwise, the algorithm uses an entropy-based measure known as information gain as a heuristic for selecting the attribute that will best separate the samples into individual classes. This attribute becomes the “test" or “decision" attribute at the node. In this version of the algorithm, all attributes are categorical, i.e., discrete-valued. Continuous-valued attributes must be discretized.
- A branch is created for each known value of the test attribute, and the samples are partitioned accordingly.
- The algorithm uses the same process recursively to form a decision tree for the samples at each partition. Once an attribute has occurred at a node, it need not be considered in any of the node's descendents.
- The recursive partitioning stops only when any one of the following conditions is true:
    1. All samples for a given node belong to the same class, or
    2. There are no remaining attributes on which the samples may be further partitioned. In this case, majority voting is employed .This involves converting the given node into a leaf and labeling it with the class in majority among samples. Alternatively, the class distribution of the node samples may be stored; or
    3. There are no samples for the branch. In this case, a leaf is created with the majority class in samples.

2.7.1.2 Attribute Selection Measure.

The information gain measure is used to select the test attribute at each node in the tree. Such a measure is referred to as an attribute selection measure or a measure of the goodness of split. The attribute with the highest information gain (or greatest entropy reduction) is chosen as the test attribute for the current node. This attribute minimizes the information needed to classify the samples in the resulting partitions and reflects the least randomness or "impurity" in these partitions. Such an information-theoretic approach minimizes the expected number of tests needed to classify an object and guarantees that a simple (but not necessarily the simplest) tree is found [9].

Let S be a set consisting of s data samples. Suppose the class label attribute has m distinct values defining m distinct classes, Ci (for i = 1,…,m). Let si be the number of samples of S in class Ci. The expected information needed to classify a given sample is given by:

$$I(s1, \ldots, sm) = -\sum_{i=1}^{m} P_i \log_2(P_i) \quad \text{----------} \quad (2.1)$$

Where: $P_i = \frac{s_i}{s}$ is the probability that an arbitrary sample belongs to class C_i.

s_i – is the number of samples of S in class C_i.

To select the test attribute (i.e., the best attribute for splitting), the entropy and information gain need to be calculated for each attribute. Therefore, if an attribute A has v distinct values, { a1, a2, …, av }, then attribute A can be used to partition S in to v subsets, { S1, S2, …, Sv }, where Sj contains those samples in S that have value aj of A. If A were selected as the test attribute (i.e., the best attribute for splitting), then S1, S2, …, Sv would correspond to the branch grown from

the node containing the set S. The entropy, or expected information based on the partitioning into subsets by an attribute A, is given by:

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \ldots + s_{mj}}{s} . I(s_{ij}, \ldots, s_{mj}) \quad \ldots\ldots (2.2)$$

Where: s_{ij} - is the number of samples of class C_i in a subset S_j

$\frac{s_{1j} + \ldots + s_{mj}}{s}$ - acts as the weight of the j^{th} subset and is the ratio of number of samples in the subset to total samples in S

The smaller the entropy value, the greater will be the purity of the subset partitions. It should be noted that, for a given subset Sj,

$$I(s_{1j}, s_{2j} \ldots, s_{mj}) = -\sum_{i=1}^{m} p_{ij} \log_2 (p_{ij}) \quad \ldots\ldots (2.3)$$

Where: $p_{ij} = \frac{s_{ij}}{|S_j|}$ - is the probability that a sample in S_j belongs to class C_i

As a result, the encoding information that would be gained by branching on attribute A is:

$$\text{Gain (A)} = I(s_{1j}, s_{2j} \ldots, s_{mj}) - E(A) \quad \ldots\ldots (2.4)$$

In other words, Gain (A) is the expected reduction in entropy caused by knowing the valued of the attribute A.

The attribute with the highest information gain is considered as the most discriminating attribute of the set under consideration. So, an attribute that yields maximum information gain will be

chosen for data set partitioning. Then, a node is created and labeled with the chosen attribute, branches are formed for each value of the attribute, and the samples are partitioned accordingly. The same criteria will then be applied to each split sample. The iterative divide and conquer process executes until no further split is required.

2.7.1.3 Tree Pruning

When a decision tree is built, many of the branches will reflect anomalies in the training data due to noise or outliers. Tree pruning methods address this problem of over fitting the data. Such methods typically use statistical measures to remove the least reliable branches, generally resulting in faster classification and an improvement in the ability of the tree to correctly classify independent test data [9].

There are two common approaches to tree pruning.

- In the prepruning approach, a tree is “pruned" by halting its construction early (e.g., by deciding not to further split or partition the subset of training samples at a given node). Upon halting, the node becomes a leaf. The leaf may hold the most frequent class among the subset samples, or the probability distribution of those samples.

  When constructing a tree, measures such as statistical significance, information gain, etc., can be used to assess the goodness of a split. If partitioning the samples at a node would result in a split that falls below a prespecified threshold, then further partitioning of the given subset is halted. There are difficulties, however, in choosing an appropriate threshold. High thresholds could result in oversimplified trees, while low thresholds could result in very little simplification.

- The post pruning approach removes branches from a “fully grown" tree. A tree node is pruned by removing its branches.

Alternatively, prepruning and post pruning may be interleaved for a combined approach. Post pruning requires more computation than prepruning, yet generally leads to a more reliable tree.

2.7.1.4 Extracting Classification Rules from Decision Trees

The knowledge represented in decision trees can be extracted and represented in the form of classification IFTHEN rules. One rule is created for each path from the root to a leaf node. Each attribute-value pair along a given path forms a conjunction in the rule antecedent (“IF" part). The leaf node holds the class prediction, forming the rule consequent (“THEN" part). The IF-THEN rules may be easier for humans to understand, particularly if the given tree is very large [9].

A rule can be “pruned" by removing any condition in its antecedent that does not improve the estimated accuracy of the rule. For each class, rules within a class may then be ranked according to their estimated accuracy. Since it is possible that a given test sample will not satisfy any rule antecedent, a default rule assigning the majority class is typically added to the resulting rule set.

2.7.1.5 Advantages of Decision Trees

Decision trees make few passes through the data (no more than one pass for each level of the tree) and they work well with many predictor variables. As a consequence, models can be built very quickly, making them suitable for large data sets.

Decision trees handle non-numeric data very well. This ability to accept categorical data minimizes the amount of data transformations and the explosion of predictor variables inherent in neural nets.

Additionally the decision tree is appreciated for its cognitive nature. In other words, they allow human experts to easily understand the solution of a problem [29].

2.7.1.6 Limitations of Decision Trees

Decision trees are not at all without any limitations. Perhaps the weakest point of decision trees in modeling is their sensitivity to change in data, hence also to noise [29]. In contrast to the rather stable problems of learning of the classical kind, such as a feed-forward artificial neural network, the problem of learning decision trees are inherently ill-posed. Again when compared to the neural network they lack the ability to visualize the non linear relationship that may exist in the independent variables [1].

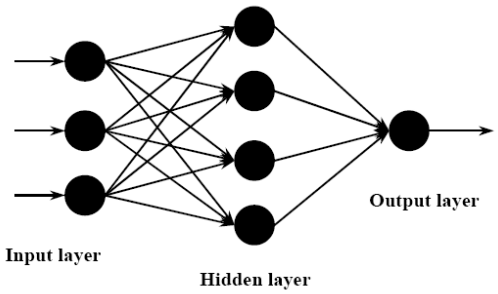
A common criticism of decision trees is that they choose a split using a "greedy" algorithm in which the decision on which variable to split doesn't take into account any effect the split might have on future splits. In other words, the split decision is made at the node "in the moment" and it is never revisited. In addition, all splits are made sequentially, so each split is dependent on its predecessor. Thus all future splits are dependent on the first split, which means the final solution could be very different if a different first split is made.

2.7.2 Neural Networks

Neural Networks are an expanding and interdisciplinary field bringing together mathematicians, physicists, neurobiologists, brain scientists, engineers, and computer scientists. Seldom has a field of study coalesced from so much individual expertise, bringing a tremendous momentum to neural network research and creating many challenges [24].

A neural network starts with an input layer, where each node corresponds to a predictor variable. These input nodes are connected to a number of nodes in a hidden layer. Each input node is connected to every node in the hidden layer. The nodes in the hidden layer may be connected to nodes in another hidden layer, or to an output layer. The output layer consists of one or more response variables [26].

Figure 2.7-2: Neural network structure



Neural networks involve long training times, and are therefore more suitable for applications where this is feasible. They require a number of parameters which are typically best determined empirically, such as the network topology or “structure". Neural networks have been criticized for their poor interpretability, since it is difficult for humans to interpret the symbolic meaning behind the learned weights. These features initially made neural networks less desirable for data mining [9].

Advantages of neural networks, however, include their high tolerance to noisy data as well as their ability to classify patterns on which they have not been trained. In addition, several algorithms have recently been developed for the extraction of rules from trained neural networks. These factors contribute towards the usefulness of neural networks for classification in data mining [26].

The goal of training the neural network is to estimate the connection weights so that the output of the neural net accurately predicts the test value for a given input set of values. The most common training method is back propagation.

2.7.2.1 Defining a Network topology

One important aspect used in classification of neural network is their topology. The arrangement of neural processing units and their interconnection can have a profound impact on the processing capabilities of the network. Unites are connected one another. It is this pattern of connectivity that constitutes what the system knows and that determines how it will respond to any arbitrary input. The total pattern of connectivity can be specified by the weight for each of the connections in the system.

Depending on the pattern of connectivity, two types of networks can be distinguished: feedforward networks and recurrent networks. Feedforward networks are the most common structures for the well-known back-propagation algorithm. Feedforward networks have no feedback connections, that is, they have no connections through weights extending from the output of a layer to the input of the same or previous layers. Recurrent networks do contain feedback connections. Recurrent networks recirculate previous output back to inputs; hence their output is determined by their current input and their previous outputs [3].

There are no clear rules as to the “best" number of hidden layer units. Network design is a trial by error process and may affect the accuracy of the resulting trained network. The initial values of the weights may also affect the resulting accuracy. Once a network has been trained and its accuracy is not considered acceptable, then it is common to repeat the training process with a different network topology or a different set of initial weights [26].

2.7.2.2 Learning Using Neural Nets

Artificial neural nets have been successfully used for recognizing objects from their feature patterns. For classification of patterns, the neural networks should be trained prior to the phase of recognition process. The process of training a neural net can be broadly classified into three typical categories, namely,

- Supervised learning
- Unsupervised learning
- Reinforcement learning.

2.7.2.2.1 Supervised Learning

The supervised learning process requires a trainer that submits both the input and the target pattern for the objects to get recognized. In supervised learning if the training examples comprise input vectors x and the desired output vectors y, then training is performed until the neural network "learns" to associate each input vector x to its corresponding and desired output vector y. Among the supervised learning algorithms, most common are the back-propagation training. An input vector is applied; the output of the network is calculated and compared to the corresponding target vector, and the difference (error) is feed back through the network and weights are changed according to an algorithm that tends to minimize the error. This process is continued until the error for the entire training set is an acceptably low level [3].

2.7.2.2.2 Unsupervised Learning

Unsupervised learning only input vectors x are supplied; the neural network learns some internal features of the whole set of all the input vectors presented to it.

The process of unsupervised learning is required in many recognition problems, where the target pattern is unknown. The unsupervised learning process attempts to generate a unique set of weights for one particular class of patterns. The objective of unsupervised learning process is to adjust the weights autonomously, until an equilibrium condition is reached when the weights do not change further. The process of unsupervised learning, thus, maps a class of objects to a class of weights. Generally, the weight adaptation process is described by a recursive functional relationship [3].

2.7.2.2.3 Reinforcement learning (reward-penalty learning)

This type of learning may be considered as an intermediate form of the above two types of learning. Here the learning machine does some action on the environment and gets a feedback response from the environment. The learning system grades its action good (rewarding) or bad (punishable) based on the environmental response and accordingly adjusts its parameters. Generally, parameter adjustment is continued until an equilibrium state occurs, following which there will be no more changes in its parameters [3].

2.7.2.3 Backpropagation

The Back-propagation training algorithm for training feed−forward networks was developed by Paul and later by Parker and Rummelhart and McClelland. This type of network configuration is the most common in use, due to its ease of training [27].

Backpropagation learns by iteratively processing a set of training samples, comparing the network's prediction for each sample with the actual known class label. For each training sample, the weights are modified so as to minimize the mean squared error between the network's prediction and the actual class. These modifications are made in the “backwards" direction, i.e.,

from the output layer, through each hidden layer down to the first hidden layer (hence the name backpropagation) [9].

Backpropagation training is simply a version of gradient descent, a type of algorithm that tries to reduce a target value (error, in the case of neural nets) at each step. The algorithm proceeds as follows [27].

Feed forward: The value of the output node is calculated based on the input node values and a set of initial weights. The values from the input nodes are combined in the hidden layers, and the values of those nodes are combined to calculate the output value.

Backpropagation: The error in the output is computed by finding the difference between the calculated output and the desired output (i.e., the actual values found in the training set). Next, the error from the output is assigned to the hidden layer nodes proportionally to their weights. This permits an error to be computed for every output node and hidden node in the network. Finally, the error at each of the hidden and output nodes is used by the algorithm to adjust the weight coming into that node to reduce the error.

This process is repeated for each row in the training set. Each pass through all rows in the training set is called an epoch. The training set will be used repeatedly, until the error no longer decreases. At that point the neural net is considered to be trained to find the pattern in the test set [27].

Because so many parameters may exist in the hidden layers, a neural net with enough hidden nodes will always eventually fit the training set if left to run long enough. But how well it will do on other data? To avoid an over fitted neural network which will only work well on the training

data, you must know when to stop training. Some implementations will evaluate the neural net against the test data periodically during training. As long as the error rate on the test set is decreasing, training will continue. If the error rate on the test data goes up, even though the error rate on the training data is still decreasing, then the neural net may be over fitting the data.

In summary, Neural networks are remarkable for their learning efficiency and tend to outperform other methods (like decision trees) when no highly relevant attributes exist, but many weakly relevant ones are present. Furthermore, Artificial Neural Network (ANN) can easily be adjusted as new examples accumulate.

However according to Leul [15], the drawbacks of applying neural networks to data mining include: difficulty in interpreting the model, difficulty in incorporating prior knowledge about the application domain in a neural network, and, also, long learning time, both in terms of CPU time, and of manually finding parameter settings that will enable successful learning. The rule extraction algorithm, described in [15], makes an effective use of the neural network structure, though the weights of the links between the nodes remain meaningless, and the rules are extracted in a deterministic (Boolean) form. The network is pruned by removing redundant links and units, but removal of entire attributes (Feature selection) is not considered.

2.8 Application of Data Mining

Data mining is used for a variety of purposes in both the private and public sectors. Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce costs, enhance research, risk assessment and increase sales. For example, the insurance and banking industries use data mining applications to detect fraud and assist in risk assessment (e.g., credit scoring). Using customer data collected over several years, companies can develop models that

predict whether a customer is a good credit risk for the different services the bank delivered, or whether an accident claim may be fraudulent and should be investigated more closely [11].

Data mining offers value across a broad spectrum of industries. Telecommunications and credit card companies are two of the leaders in applying data mining to detect fraudulent use of their services.

Medical applications are another fruitful area. The medical community sometimes uses data mining to help predict the effectiveness of a procedure or medicine. Pharmaceutical firms are mining large databases of chemical compounds and of genetic material to discover substances that might be candidates for development as agents for the treatments of disease [11].

Retailers can use information collected through affinity programs (e.g., shoppers' club cards, frequent flyer points, contests) to assess the effectiveness of product selection and placement decisions, coupon offers, and which products are often purchased together.

Companies such as telephone service providers and music clubs can use data mining to create a "churn analysis," to assess which customers are likely to remain as subscribers and which ones are likely to switch to a competitor.

2.8.1 Data Mining Application in Banking

Currently, huge electronic data repositories are being maintained by banks and other financial institutions across the globe. Valuable bits of information are embedded in these data repositories. The huge size of these data sources make it impossible for a human analyst to come up with interesting information that will help in the decision making process [16].

Data mining can contribute to solving business problems in banking and finance by finding patterns, causalities, and correlations in business information and market prices that are not immediately apparent to managers because the volume data is too large or is generated too quickly to screen by experts.

The banking industry in general has stared hard at its customer data to analyze customer behavior, and it has learned valuable lessons from other industries that use Data-mining. Although banks have employed statistical analysis tools with some success for several years, previously unseen patterns of customer behavior are now coming into clear focus with the aid of new Data-mining tools. Data-mining is the automated analysis of large data sets to find patterns and trends that might otherwise go undiscovered. By studying these patterns/trends, banking executives can predict with increasing precision how customers will react to various changes, for example, interest rate adjustments [16].

Many banks now days are making more and more use of data mining in their day-to-day activities, however many companies will not admit to their techniques , due to polices set because of the competitive market. But there are few cases where banks discuss their experience [17].

For example Bank of Montreal has reported having analyzed mortgage customer's transactions in checking, savings and other accounts for insight into who is at risk of defaulting. The bank found that a certain type of customer is in the habit of paying bills late but has the means to fulfill his or her obligations. By further analyzing the transactional behavior of customers across all their accounts, the bank can see which customers experience periodic cash flow crunches and which may truly be in danger of defaulting [16].

There are numerous areas in which data mining can be used in the banking industry, which include customer segmentation and profitability, credit scoring and approval, predicting payment default, marketing, detecting fraudulent transactions, cash management and forecasting operations, optimizing stock portfolios, and ranking investments. In addition, banks may use data mining to identify their most profitable credit card customers or high-risk loan applicants. There is, therefore, a need to build an analytical capability to address the above-stated issues and data mining attempts to provide the answer [16].

Marketing: One of the most widely used areas of data mining for the banking industry is marketing. The bank's marketing department can use data mining to analyze customer databases and develop statistically sound profiles of individual customer preferences for products and services. By offering only those products and services that customers really want, banks can save substantial money on promotions and offerings that would otherwise be unprofitable. Bank marketers, therefore, need to focus on their customers by learning more about them. Bank of America, for instance, uses database marketing to improve customer service and increase profits. By consolidating five years of customer history records, the bank was able to market and sell targeted services to customers [16].

'Cross-selling': is another marketing area where data mining can be extensively used. Here, a service provider makes it attractive for a customer to buy additional products or services with the same business. The more products and services a bank can provide for customers, the more likely the bank is to retain those customers. For example, several leading private and foreign banks in India (ICICI, HSBC, etc.) use data mining to find customers with demand deposit accounts who may be interested in a home equity loan. A model is built of the customers who already have home equity loans and this model is used to pinpoint other customers who may also be interested.

Another example is Bank of America, which has recently completed a project with IBM's data mining tools to search its database of corporate clients and try to figure out what products the clients may need next [16].

Risk Management: Data mining is widely used for risk management in the banking industry. Bank executives need to know whether the customers they are dealing with are reliable or not. Offering new customers credit cards, extending existing customers lines of credit, and approving loans can be risky decisions for banks if they do not know anything about their customers. Data mining, however, can be used to reduce the risk of banks that issue credit cards by determining those customers who are likely to default on their accounts. An example was reported in the press of a bank discovering that cardholders who withdrew money at casinos had higher rates of delinquency and bankruptcy. It is a common practice on the part of banks to analyze customers' transaction behaviors in their deposit accounts to determine their probability of default in their loan accounts. Data mining can also derive the credit behavior of individual borrowers with installment, mortgage and credit card loans, using characteristics such as credit history, length of employment and length of residency [16].

Fraud Detection: Another popular area where data mining can be used in the banking industry is in fraud detection. Being able to detect fraudulent actions is an increasing concern for many businesses; and with the help of data mining more fraudulent actions are being detected and reported. Two different approaches have been developed by financial institutions to detect fraud patterns. In the first approach, a bank taps the data warehouse of a third party (potentially containing transaction information from many companies) and uses data mining programs to identify fraud patterns. The bank can then cross-reference those patterns with its own database for

signs of internal trouble. In the second approach, fraud pattern identification is based strictly on the bank's own internal information. Most of the banks are using a 'hybrid' approach.

One system that has been successful in detecting fraud is Falcon's 'fraud assessment system'. It is used by nine of the top ten credit card issuing banks, where it examines the transactions of 80 per cent of cards held in the US. Mellon Bank also uses data mining for fraud detection and is able to better protect itself and its customers' funds from potential credit card fraud [16].

Customer Acquisition and Retention: Not only can data mining help the banking industry to gain new customers, it can also help retain existing customers. Customer acquisition and retention are very important concerns for any industry, especially the banking industry. Today, customers have so many opinions with regard to where they can choose to do their business. Executives in the banking industry, therefore, must be aware that if they are not giving each customer their full attention, the customer can simply find another bank that will. Data mining can also help in targeting 'new' customers for products and services and in discovering a customer's previous purchasing patterns so that the bank will be able to retain existing customers by offering incentives that are individually tailored to each customer's needs [16].

In general, Data Mining techniques can be of immense help to the banks and financial institutions in this area for better targeting and acquiring new customers, fraud detection in real time, providing segment based products for better targeting the customers, for risk management on different services delivered, analysis of the customers' purchase patterns over time for better retention and relationship, detection of emerging trends to take proactive stance in a highly competitive market adding a lot more value to existing products and services and launching of new product and service bundles[16].

CHAPTER THREE

3 OVERVIEW OF THE EXISTING SYSTEM

3.1 Organization Profile

The National Bank of Ethiopia is the central bank of the FDRE established by proclamation No. 83/1994 in order to "foster monetary stability, a sound financial system and such other credit and exchange conditions, as are conducive to the balanced growth of the economy of the country."

As stipulated in the Monetary and Banking Proclamation No. 83/1994, the primary duties and responsibilities of the National Bank of Ethiopia include conducting and implementation of monetary policy and exchange rate policy, regulation and supervision of banks and other financial institutions, issuance of currency, maintaining and managing of gold and foreign exchange reserves, provision of refinancing facilities to banks and other financial institutions, as well as banking service and provide financial advice to the Government.

The monetary policy, the Bank maintains monetary stability with the view of achieving macroeconomic stability and sustainable economic growth and development. The Bank implements exchange rate policy with the objective of enhancing the country's competitiveness in the global economy while fostering macro-economic stability.

The Bank manages and administers the country's international reserves required for the payments of imports as well as for meeting external debt servicing and other payment obligations. In its functions, the Bank endeavors to maintain, at all times, sufficient international reserve fund, while considering the risk and return factors associated with the management of these reserves.

In the monetary and foreign exchange policies design and implementation activities, the Bank collects statistical data from various sources, produces policy oriented research outputs, prepares periodic reports and disseminates them to various users, both domestically and internationally.

The powers and duties vested on the bank include:

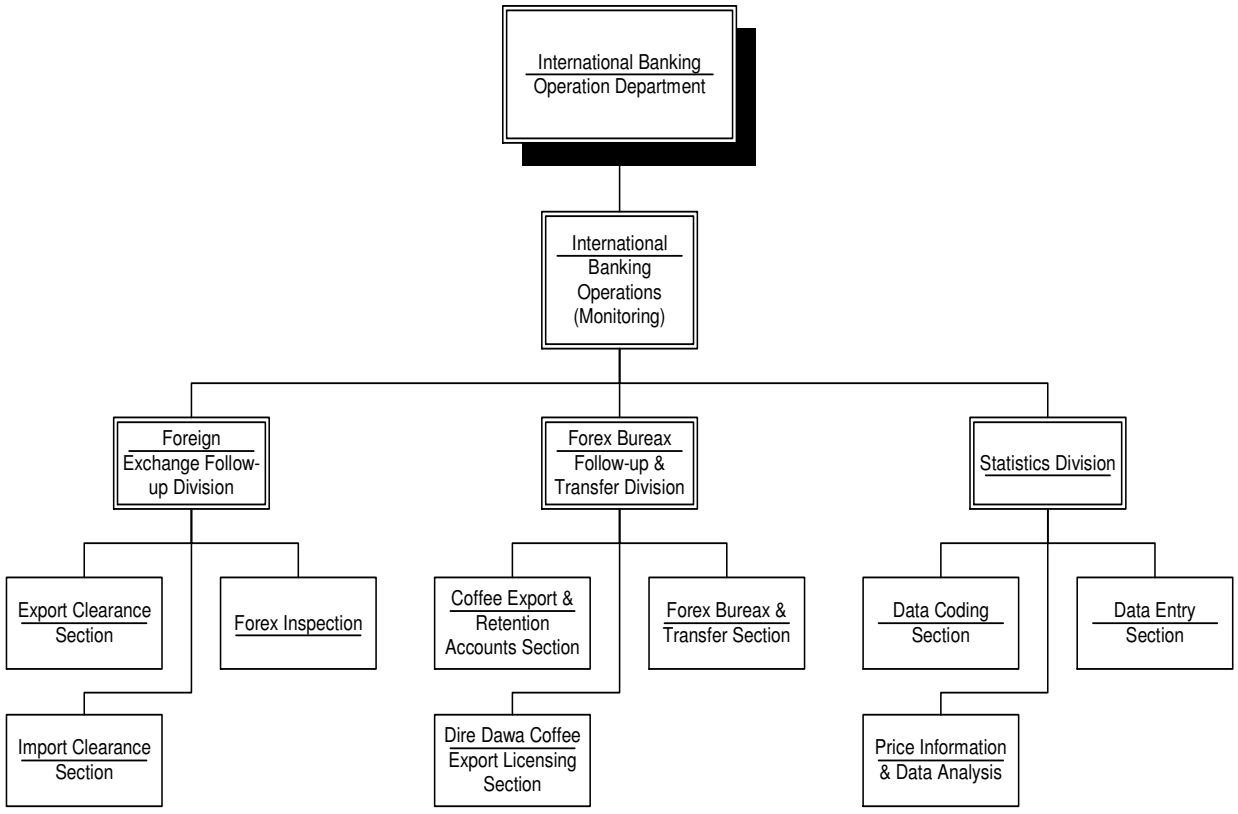
- Regulate the supply and availability of money & credit and applicable interest and other changes.
- Set limits on gold and foreign exchange assets which banks and other financial institutions authorized to deal in foreign exchange and hold in deposits.
- Set limits on the net foreign exchange position and on the terms and amount of external indebtedness of banks and other financial institutions.
- Make short and long-term refinancing facilities available to banks and other financial institutions.

In order to undertake these duties, the Bank has organized itself into various operational and support units. The Bank's Departments consist of Legal, Administration, Accounts, International Banking Operations, Information Systems, Treasury, Economic Research, Ethiopian Institute of Banking and Insurance and an Executive Assistant Office. Four Departments, including the Executive Assistant Office and Ethiopian Institute of Banking and Insurance, directly report to the Governor while the rest departments are answerable to the Vice-Governor.

3.2 The International Banking Department

The International Banking Operations Department is the key department to undertake part of the tasks entrusted to the Bank as promulgated by the establishing proclamation. This Department has two wings; namely, the Operation wing and the Monitoring wing. The Operations wing has the Reserve Management, Inter-Bank Foreign Exchange Banking, and External Debt follow-up Divisions established under it. The Monitoring wing, which is the focus of this research study, has the Statistics, Forex Bureaux Follow-up and Transfers, and Foreign Exchange Follow-up Divisions.

Figure 3.2-1: IBO Department Structure



3.2.1 The Foreign Exchange Follow-up Division

This division is entrusted with the task of monitoring the settlement of import and export foreign exchange commitments as per the existing directives and procedures of the Bank. To this effect, the division

- Follows-up the proper utilization of foreign exchange allowed for the purpose of imports,
- Follows-up the repatriation of foreign currency to Ethiopia from exports made to other countries
- Carries out ex-post verification of export, import, and invisible transaction documents issued by commercial banks.

The division also provides information to different arms of the Government like Inland Revenue Authority and tax offices.

3.2.1.1 Import Registration and Follow-up

The National Bank of Ethiopia has the responsibility of following up foreign exchange payments for imports made to the country.

Import Application

The import application form of Foreign Exchange Monitoring System (FEMoS) is used to register, amend, utilization ticket and cancellation ticket entry and declaration entry. The import application will allow the user to register many commodities with one permit. The end result of import application is to produce delinquent importers and various statistical reports. Permit will be delinquent if it has not been declared by the imported in a way of presenting the import

declaration to the clearance section of the bank within the time period specified for each method of payment.

Utilization and cancellation of ticket entry will be allowed for permits already registered and as a default, you can't utilize more than an issued amounts unless and otherwise the mode of payment setup is changed from the default during parameter setup by the bank. The system will check utilization ticket if it is above issue or not.

Declaration entry will be done after utilization ticket is entered but sometimes one may want to enter declaration before utilization. In this case one should setup the method of payment parameter such that declaration precedes utilization.

Importer registration

Every permit to be registered should have an importer and this importer should have a unique account number which will be supplied by the NBE. The details of the importer information will be maintained in the system. You can also make importer active or inactive so that you can enter permit for active importers only. Making as an inactive will not allow registering permit for that importer.

Import permit registration

The first task in import application is registration of import permits. All import permits issued by commercial banks and the NBE should be reported to the NBE's IBD Monitoring Wing in the first day of each week (for previous week's transactions). When the data is incomplete during registration of permits, the concerned party should be informed so that required data has to provide on time.

The process begins

1. when the customer bank delivers a set of import permits (of a WEEK) with associated documents to the receiving Clerk
2. The receiving Clerk receives the import permits and registers some attributes on the Import Permit Registration ledger. The attributes recorded in the ledger are import permit number, date of issue, name of importer, quantity and description of commodity, currency type, amount approved in F/C and in Birr, country of origin, and a remark, if any. The receiving Clerk then passes the permit and all required documents to the Coder.
3. The encoder then puts the required codes on the data input form designed for computer application and completes the form by filling certain fields by extracting them from the permit. Then after the encoder passes all documents together to the Section Head
4. The Section Head checks for missed permits and any discrepancies and forwards the permit to Data Entry Section
5. The data entry Clerk enters the data to the Import Permit Registration System, proofreads and passes all the documents to the File Room, again with a dispatch, to be filed.
6. The documents are then filed by the Filing Clerk

Utilization / Cancellation ticket registration

Utilization and cancellation ticket information shall be entered for the permit registered in the system. The system will accept ticket for the existing permit numbers only. All payments made by importers needs to be reported to the NBE. The system will not accept utilization ticket exceeding the permitted value unless to do so. The ticket entry is used to capture utilization for

import and export permits. If the importers don't utilize the issued permit then they have to cancel it by submitting required documents

Declaration registration

The import permit issued and utilized should be settled by entering declaration for the import permit issued. The importers are, therefore, required to submit customs declarations to ensure the entry of the permitted items to the country. It is based on this declaration data that delinquent importer list is generated by the system

1. The process begins when the importer requests to settle his/her outstanding import permits or required to do so. For this purpose, the importer submits major documents (customs declaration, Chamberized invoice, bill of lading/airway bill) and supplementary documents (certificate of origin, packing list, carrier invoice) to the counter Clerk at the import clearance section
2. The Filling Clerk search and retrieve the file of the Importer
3. The Clerk examines if there are any discrepancies in the permitted, utilized and declared records as to types of goods, quantity of goods, price of items, freight amount, etc.
4. If there is no discrepancy, the checker fills an import declaration input form that contains date, quantity, declaration number, amount in foreign currency and in birr
5. The checker, after signing on the input form, passes the declaration with the import control copy to the section head for counter checking
6. The section head signs and passes the documents to filling Clerk/Counter Clerk.

7. The Commitment filing Clerk/Counter Clerk after sorting the declarations and permits, dispatches the permit to the Data Entry section and submit the declaration to the importer
8. The filling clerk/Counter Clerk proofreads the data against the print out then dispatches the finalized permits to the main File Room and file active permits to Customer/Importers file.

Issuance of Import Clearance

Importers need to have a clearance from the NBE for the settlement of their commitments to obtain new import permit. This process describes the procedures that take over in issuing an import clearance.

1. The process begins when the Importer requests for import clearance
2. The counter Clerk/Filing Clerk search and find the customer file and checks if there is any unsettled commitment.
3. If there is no unsettled commitment, the Clearance Clerk prepares the clearance certificate and forward to the Import Clearance Section Head for verification.
4. The Import Clearance Section Head, after verification and sample checking passes the clearance to the Foreign Exchange Follow-up Division Head for signature.
5. The Foreign Exchange Follow-up Division Head after signing passes the document to Clerk
6. The Clerk issues the clearance to the Importer.

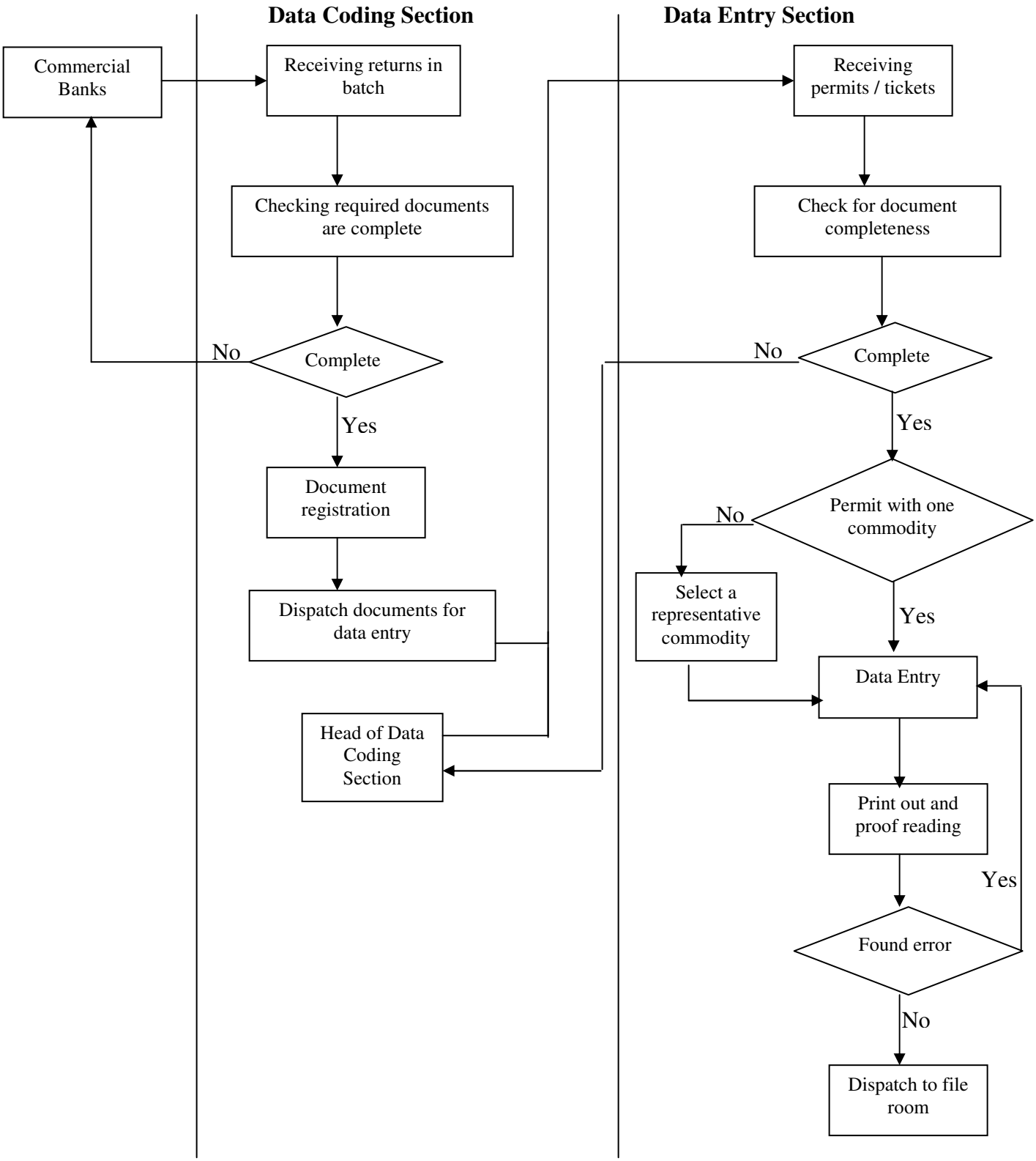
Method of Payment

There are various types of payments methods for imports. These include:

- Import on Letter of Credit basis
- Import on Cash Against Document at sight basis
- Import on Advance payment basis, and
- Import on Franco-valuta basis
- Import on grants and loans basis

The import transactions and the amount permitted for payment will be registered by the NBE and the necessary follow-up will be made in order to ensure the amount permitted is fully utilized and the items for which the payments are made are duly imported to the country unless the permit is cancelled.

Figure 3.2-2: Work flow for permit Registration Activity



3.2.1.2 Export Registration and Follow-up (EXP)

The NBE controls the export income of the country since this is the major determinant of the Balance of Payments of the Country. Every export from Ethiopia will be registered and the necessary follow-up will be done in order to ensure the repatriation of the foreign currency back to Ethiopia.

The methods of payment for export transactions are:

- Export on Letter of Credit Basis
- Export on Advance Payment Basis
- Export on Consignment Basis
- Cash Against Document (CAD)

Export Application

The export application form of FEMoS is used to register and amend utilization tickets and cancellation tickets entry and custom information entry. The export application will allow the user to register many commodities with one permit. The end result of export application is to produce delinquent exporters and various statistical reports. Permit will be delinquent if the exporter fails to repatriate the amount of foreign currency within the time period specified for each method of payment.

Utilization and cancellation ticket entry will be allowed for permits already registered and can't utilize more than issued amounts unless and otherwise the mode of payment setup is changed from the default during parameter setup by the bank. The system will check utilization ticket if it is above issue or not.

Exporter registration

Every export permit to be registered should have exporter and this exporter should have a unique account number which will be supplied by the NBE. The details of the exporter information will be maintained in the system module. You can also make exporter active or inactive so that you can enter permit for active exporters only. Making as an inactive will not allow registering permit for that exporter.

Export permit registration

The first task in export application is registration of export permits. Permits issued by different banks will be received and captured to the system daily. When the data is incomplete during registration of permits, the concerned party should be informed so that required data has to provide on time. The following are some of the information's captured by the system. These are

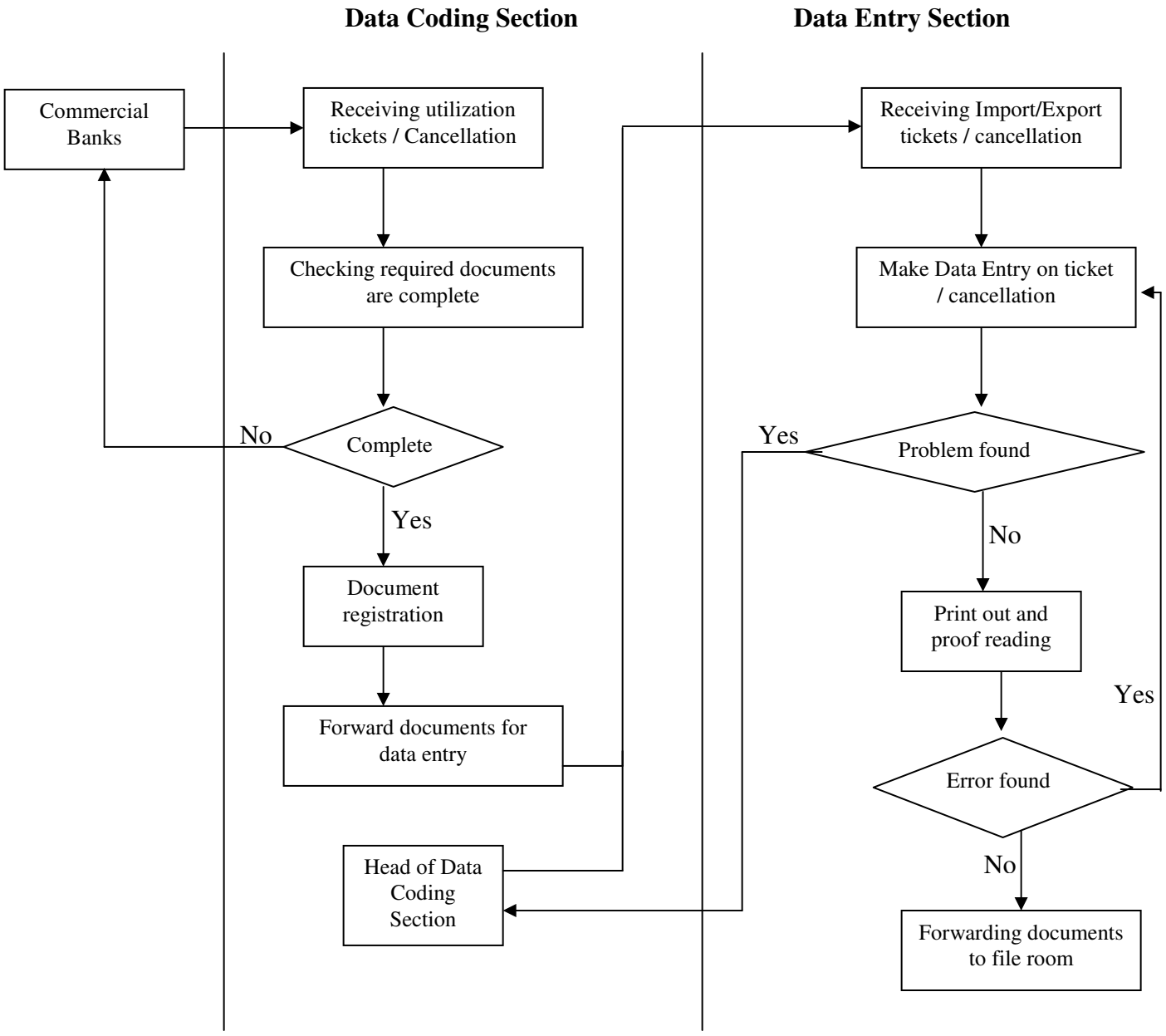
- General export permit information such as permit number, Date of issue, buyer country, method of payment, currency, etc
- Commodity details like commodity HS Code, quantity of commodity, quality/grade of commodity, etc

Utilization / Cancellation ticket registration

Utilization and cancellation ticket information shall be entered for the export permit registered in the system. The system will accept ticket for the existing permit numbers only. The system will not accept utilization ticket exceeding the permitted value unless to do so. The ticket entry is used to capture utilization for import / export permits. This process describes the procedures of receiving the export receipts/tickets.

1. The process begins when the customer bank delivers export tickets to the receiving Clerk along with other visible and invisible documents.
2. The receiving Clerk receives the tickets and passes the documents to data coders.
3. The Coder then puts the required codes such as Unit Code, Country Code on the export tickets and passes the documents to data entry section.
4. The IMF system data entry Clerk enters the data into the IMF system, sorts the export utilization tickets and dispatches them to Export Clearance Section.
5. Export Clearance Section passes entry on export ledger manually and dispatches the documents to the coder.
6. The coder then captures some of the attributes into the data input form, and pass the tickets with the input form again to the data entry section
7. The data entry Clerk enters the data to the export tickets database and passes all the documents to the File Room with a dispatch. The documents are then filed in the respective files

Figure 3.2-3: Activity work flow for ticket utilization and cancellation registration



CHAPTER FOUR

4 DATA COLLECTION AND DATA PREPARATION

4.1 Introduction

The landmark research work by Fayyad, et al [7] outlined the basic steps of the data mining process. Brachman and Anand [24], complemented this process by focusing on the human-process interaction. This was superseded by the practical and comprehensive Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology [21].

This chapter being the fundamental part of this research work describes the methods and procedures followed to accomplish the research objectives following the CRISP-DM standard. It begins with summaries of the business/domain understanding of the research target followed by data understanding, data preparation, modeling, and evaluation methods used. It also shows the method of data mining tool evaluation and selection used in the research. The details of modeling and evaluation are described under modeling chapter.

4.2 Business/Domain Understanding

In a research process, a good understanding of domain/business area is critical for the success and effectiveness of the overall research work. This is usually done through a study and closer look at the target domain through a survey. As stated in Peter, et al [21], in this phase, the major objectives and requirements should be clearly described and translated in to data mining problem definition.

As stated in the aforementioned discussions, the reserve of foreign currency is highly dependent on utilization of the foreign currency by importers and the proper control of transaction of goods

by the exporters respectively. Thus, the Bank develops a database that is used to registers and monitor transactions of the importers and exporters in the country. This information system identifies the delinquency status of importers and exporters at a given point in time. But the system does not allow the Bank to predict whether the importer or exporter would be delinquent or not at any point in time using their previous performance.

Having thoroughly understood the business problem through continuous interactions with the Bank, the researcher, considers that one mechanism that the NBE can employ to ensure precautionary knowledge on delinquency is to build a predictive data mining model

In an attempt to understand the business problem, the researcher has applied a number of approaches. These comprise the following:

- Frequent visit to the Bank and intensive discussion with experts in the domain,
- Attached for some period of time with the operators in order to gain a thorough understanding of the operational processes,
- In the course of interview and learning from users (experts and other relevant personnel), the researcher:
    - Has tried to look into their ideas, suggestions, and alternatives.
    - Has had a smooth relationship in explaining the benefits of the outcome of the research work and the importance of their contribution.

4.2.1 Data Mining Tool Selection

The selection of data mining software with the required capabilities that performs the required tasks of data mining was one of the challenges in the process. Nonetheless, the WEKA software has been chosen for the construction of the research model. The selection of this software was based on the review of relevant features in the available literature and has included the consideration of important checklist.

In line with [26] and [9] approach, the checklist for selecting the data mining tool includes:

- The data mining tasks that the tool is intended for (Decision tree and Neural Network)
- The Algorithms supported (J48 and MultilayerPerceptron)
- The system architecture on which the tool runs
- The operating system on which the tool runs (Windows, etc)
- Data preparation is by far the most time consuming aspect of data mining. Everything a tool can do to ease this process will greatly expedite model development. Some of the functions that a tool may provide include:
    - Data cleanup, such as handling missing data or identifying integrality violations;
    - Data description, such as row and value counts or distribution of values;
    - Data transformations, such as adding new columns, performing calculations;
    - Data sampling for model building or for the creation of training and validation data set;
    - Selecting predictors from the space of variables;
- The ability to consolidate data from multiple sources;
- Scalability, the maximum number of records the tool can handle;
- Visualization capabilities;

WEKA is an open source data mining tool that is developed by the University of Waikato, New Zealand. WEKA, a machine-learning algorithm in Java, constitutes several machine learning algorithms for solving real-world data mining problems. It is written in Java and runs on almost any platform. The algorithms can either be applied directly to a dataset or called from one's own Java code. WEKA is open source software issued under the GNU General Public License. WEKA explorer supports almost all data mining tasks. WEKA has also user-friendly interface. In addition to the learning schemes, WEKA also comprises several tools that can be used for datasets pre-processing and visualization [20].

4.3 Data Understanding

Using the right data for data mining task is one of the primary keys for its success. As stated in the earlier chapters, the source of data for this research is the FEMOs database of the NBE. This database registers permitted amount for each transactions when permit is given to an importer or exporter by different commercial banks acting on behalf of the NBE.

The Bank has made this database operational since 2003 G.C. There are three main forms for each category where most of the data is extracted. These are:

- The general import registration form which contains information that describes general information about the import permit, information about the supplier, information about the total permit amount in foreign currency, other information like economic sector of the importer, etc.
- The Commodity details form that includes information like commodity HS Code, name of manufacturer, quantity of commodity, etc.

  All of the tables or forms are linked using the permit number.

Likewise, the same is true for the export transaction in such a way that

- The general export registration form contains information that describes general information about the export permit, information about the buyer, information about the total permit amount in foreign currency, other information like economic sector of the exporter, etc.
- Commodity details like commodity HS Code, quantity of commodity, quality/grade of commodity, etc.
- The utilization and cancellation ticket form contains information like transaction date, transaction type, permit values, etc. This form is used for both import and export transactions. The exporter tables are linked using export permit number as unique key.

4.3.1 Initial Data Collection

The first task in the collection of data was to look through the attributes of import and export transactions. This was done by building a query and extracting the attributes from different tables for both types of transactions. The extracted attributes were then exported to excel spreadsheet and studied in detail with the help of the domain expert. This brought about the first selection of the attributes.

After a thorough investigation and discussion with the domain experts, the researcher decided to use the year 2007 data as it was deemed sufficient enough as a sample to build the predictive model for both type of transactions. It has been observed that the 2007 data have better quality compared to the previous years. One reason for the poor quality of the previous years' data was that some were not familiar to the information system and that had necessitated the need to make frequent corrections on the records. However, through time, the frequency of data correction has

reduced dramatically, hence also the preference for the 2007 data as the research sample. The initial attributes obtained are described in Table 4.3-1 and Table 4.3-2.

Table 4.3-1: Summary of the attributes for import transaction

Attribute	Data Type	Description
AmountDeclared	Currency	Shows the amount of money declared out of Permitted.
AmountPermitted	Currency	Shows the amount of money Permitted.
Country	Text	The country where the products imported.
Currencycode	Text	The currency type used for the Payment.
Date of Issue:	Date	The issue date of the permit.
Economic Sector	Text	It shows whether the business is private or non-private.
Exchange rate	Number	Exchange rates of different Currencies.
GradeDesc	Text	Grade description for some Items.
HsCode	Text	Code given to each imported Items.
HsDescription	Text	Descriptions for imported Items.
ImporterName	Text	Name Given to the importer.
MethodOf Payment	Text	Method of Payment of the importers.
Permit Number:	Text	Uniquely identifies the import transaction.
QualityDesc	Text	Quality of description for some Items.
Quantity	Number	Amount of Items imported.
Type of Business	Text	It shows the specific business sector.
UnitOfMeasurment	Text	Unit of Measurements for Each item.
UnitPrice	Currency	Price of Each Item.
ValidityDate	Date	Maturity date for settlement of the commitment.

Table 4.3-2: Summary of the attributes for export transaction

Attribute	Data Type	Description
AmountCancelled	Currency	Shows the Amount of money cancelled out of Permitted.
AmountDeclared	Currency	Shows the Amount of money repatriated out of Permitted.
Base of Shipment	Text	Shows the different type of base of shipments.
CountryName	Text	The country where the Items are exported to.
Currencycode	Text	The currency type used for the Payment.
DateOfIssued	Date	The date in which permit to export is given.
ExchangeRate	Number	Exchange rates of different Currencies.
ExporterName	Text	Name of the Exporter.
FobForeignCurrency	Currency	Shows the amount of money expected to be repatriated.
GradeDesc	Text	Description of Grade for some item.
HsCode	Text	Code given to each exported Items.
HsDescription	Text	Descriptions for Exported Items.
MethodOf Payment	Text	Method of Payment of the Exports.
OrganizationType	Text	It shows whether the business is private or non-private.
PermitNo	Text	Uniquely identifies the export transaction.
QualityDesc	Text	Description of quality for some items.
Quantity	Number	Amount of Items Exported.
UnitOfMeasurment	Text	Unit of Measurements for Each item.
UnitPrice	Currency	Price for Each Item.
ValidityDate	Date	Maturity date for settlement of the commitment.

4.4 Data Preparation/ Preprocessing

Data preprocessing is an important activity in data mining process. It can improve the accuracy and efficiency of the subsequent mining process. As stated by Ishwar [11], the entire preprocessing involves several activities, which can be classified as feature selection, data cleaning, data selection, data integration and transformation, etc. The main purpose of this phase is the production of the dataset used for modeling.

4.4.1 Feature Selection

All features (attributes) are not equally important for the mining purposes. Those features that are highly relevant to the mining task must be carefully selected. Accordingly based on the above attributes, a thorough discussion is made with the domain experts and selection of best attributes vital for the model building exercise has been done.

There are some attributes that are left out purposely by the researcher even though it is believed that they are important. For instance, the attributes that describes about the type of products have been excluded. The overwhelming reason for this omission is that sometimes in one permit a number of products can be listed and this makes it difficult to register product features for a specific product with this very permit number. Additionally, it has been observed that it is unmanageable to know the HS code for all the products that appear in the import/export transactions.

Likewise, other attributes related to the products such as UnitOfMeasurment, GradeDesc QualityDesc and Quantity are also left out as they do not provide the necessary information for the purpose of this research task.

As a financial institution data, the traders' names and their detail information are set aside for the sake of confidentiality. The researcher also finds that this information is not relevant for the theme of the research. However, due to concerns of confidentiality, the Bank was not willing to provide data on exchange rates. Although the researcher believes that information on exchange rates should be public so that anyone can access it even from the Internet, the alternative was to use the available mean values of the exchange rates for each type of currency involved. Having undergone this challenging process, the remaining attributes were extracted from different tables of the database and further analyzed using excels spreadsheets.

4.4.2 Data Cleaning

Real world data consists of incomplete, noisy, and inconsistent and most of the time with some errors, and irrelevant attributes which are not necessary for the goal of a data mining research at hand. Hence, data cleaning increase the quality of the data so as to reach to meaningful results.

There were many records that contain missing values, inconsistent data and wrong values. The following are some of the tasks accomplished for maintaining the quality of the data for the selected attributes,

- For some attributes whose values are critical and that has to be included, the researcher manually entered their values after collecting from other sources, such as the exchange rate amount.
- The entire record/s was deleted for some records where the values of the attributes are quite unordinary or difficult to predict the missing data. For instance, in some of the entries for currency attribute, the Ethiopian birr has been incorrectly stated in place of another currency. In addition, for Amount utilized attribute a zero value for attribute is

also meaningless. However, the cleaning of unnecessary or imperfect records is made with due consideration to the integrity of the overall data.

- Attribute Type of Business is dropped (removed). This attribute contains too many missing values (difficult to estimate or correct manually).
- Inconsistencies in data entry were observed such as for the same organization type, for some as PLC and for others as Private Limited Company. So it is adjusted in one format.
- Some wordings were corrected with their correct forms. For example, in MethodOfPayment attribute, such as LC-at-Sigh with LC-at-Sight, and ADVNCE-PAYMENT with ADVANCE-PAYMENT.
- Some missing values of few attributes were filled with constant values accordingly.

4.4.3 Data Selection

There are two data sets for this research. These are the import and export data sets. Because, different business rules apply for import and export transactions, these data sets are treated separately under this research. The default business rule for an import transaction to be classified as non delinquent is settling the import commitment in 120 days whereas for the export transaction the commitment has to be settled within 90 days starting from the issue date. As stated in earlier chapters, the source of the data set is the FEMoS database of the NBE.

Initially, the size of the collected data on import transactions for the year 2007 is about 30,000 records whereas for the export, it reaches around 26,000 records. However after reducing unnecessary records for both transactions, the size of the data for import reaches 20,613 and for

export reaches 20,183 records. This size of data is deemed sufficient to train machine-learning algorithms like decision tree and neural network.

4.5 Data Transformation

The task, according to CRISP-DM, includes constructive data preparation operations such as the production of derived attributes or transforming values for existing attributes. It is often necessary to construct new predicator variables derived from the raw data. Certain variables that have little effect alone may need to be combined with others, using various arithmetic or algebraic operations [26].

As mentioned in the aforementioned discussions, the data has been collected from various sources. Integration of these various sources into one coherent table or dataset helps for further processing. Accordingly, the researcher integrated the various sources of the data into one dataset for each transaction.

Under this phase of activity, new attributes have been constructed. The derived attributes are Validity_period, AmtInBirr, AmtInBirr_Range, Country_Range and Delinquency. The process is described below.

- Considering IssueDate and ValidityDate attribute separately does not significantly help for the research theme. Therefore, from IssueDate (when the permit is effective) and ValidityDate (when the permit is expected to settle), a new attribute called 'Validity_period' (period of time) is derived. This derived attribute shows the period over which the permit commitment has to be settled. The derivation of this attribute greatly contributes to the generation of best rule by showing which transaction types become

delinquent with respect to the period given for the transaction. The default business rule for a permit to settle is 120 days for import transaction and 90 days for export.

- For import transaction, the value of this attribute is further transformed into two categories of values such as "<=120 days" and ">120 days". This is based on the 120 days default business rule for import transaction.
- Likewise, for export transaction, the value of this attribute is further transformed into two categories of values, that is: "<=90 days" and ">90 days". This is a reflection of the 90 days default business rule for export transaction.

- The second derived attribute is AmtinBirr. This attribute shows the equivalent amount in birr of the foreign currency the importer used whereas for the exporter, it shows the equivalent amount in birr of the same the exporter has to repatriate.
  - For the import transaction this attribute is derived from the exchange rate and AmountPermitted
  - For the export transaction this attribute is derived from the exchange rate and FobForeign currency.

    The advantage of this derived attribute is that it enables the conversion of transactions recorded in different currencies into a uniform unit of measurement thereby facilitating data comparability.
- Further transformation was done for AmtinBirr attribute. The values of the attribute were discritized to ten ranges. This is Because of the difficulty of generation of rules using the original values.

- Likewise Country_Region attribute was derived by transforming the original value of CountryName attribute to a higher conceptualized class.
- For import transactions the target attribute (Delinquency_status) derived from AmountPermitted and AmountDeclared; whereas for the export transactions this attribute derived from FobForeignCurrency, AmountDeclared and AmountCancelled.
    - For import transactions, if the value of the AmountDeclared is less than the value of the AmountPermitted then it is delinquent otherwise non delinquent.
    - For export transactions, if the sum of the value of the AmountDeclared and AmountCancelled is less than the value of the FobForeignCurrency then it is delinquent otherwise non delinquent.

Having completed the tasks of data transformation, the selected attributes of the final data set for the research work are presented in Table 4.5-1 and Table 4.5-2.

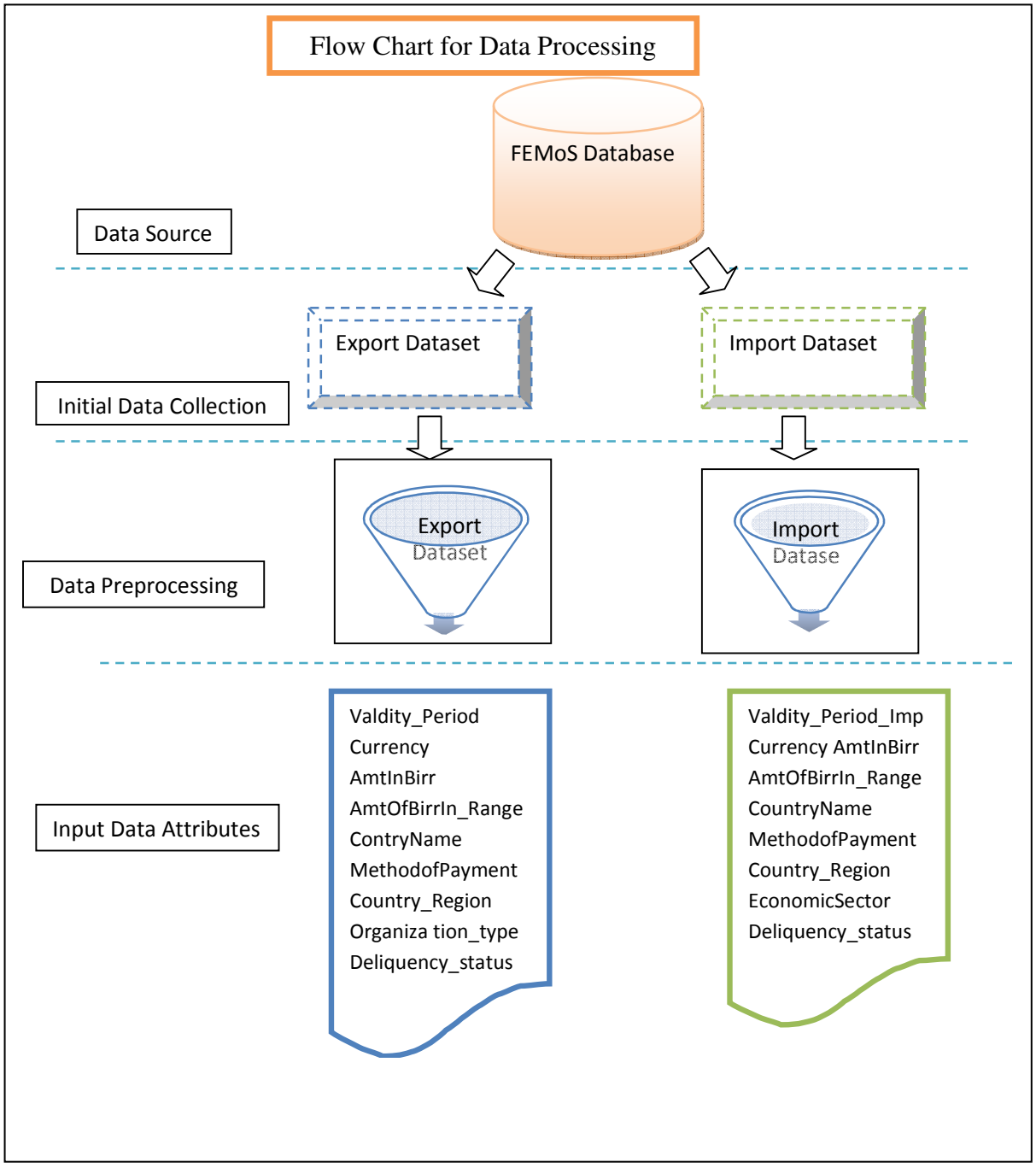
Table 4.5-1: Selected fields For Export Transaction after preprocessing the data

Attribute	Description
Valdity_Period	It shows the number of days given for a transaction to settle its commitment.
Currency	It shows the type of the foreign currency involved in the transaction.
AmtInBirr	The equivalent amount in Birr of the foreign currency the exporter repatriates.
AmtOfBirrIn_Range	This is a discritized value for the amount of the money that the transaction accomplished.
MethodofPayment	Method of payment for the export.
ContryName	The exact country where the product is exported to.
Country_Region	Shows that the continent where the product is exported to.
Organization_type	Describes the type of the organization, whether it is PLC, Ent, Gov, etc.
Base_of_Shipment	Shows the different type of base of shipments.
Delinquency_status	Reveals the delinquency status of the transaction.

Table 4.5-2: Selected fields For Import Transaction after preprocessing the data

Attribute	Description
Valdity_Period_Imp	It shows the number of days given for a transaction to settle its commitment.
Currency	It shows the type of the foreign currency involved in the transaction
AmtInBirr	The equivalent amount in Birr of the foreign currency the importer is expected to settle
AmtOfBirrIn_Range	This is a discritized value for the amount of the money that the transaction accomplished
MethodofPayment	Method of payment for the import
CountryName	The country from where the product is imported from
Country_Region	Shows the continent where the product is imported from
EconomicSector	Describes the type of the organization, whether it is private, public budgetary, public non budgetary
Delinquency_Status	Reveals the delinquency status of the transaction

Figure 4.5-1: Flow chart for data processing



CHAPTER FIVE

5 EXPERIMENTATION

5.1 Modeling

Modeling techniques in data mining have one important characteristic - they automatically generate new propositions (models), about relations among important variables in the data, which is an added value to traditional confirmatory statistical analyses. In this phase, selection of the modeling techniques, construction of the model, and the assessment of the model built have been undertaken.

5.1.1 Selecting Modeling Techniques.

In data mining, there are various techniques available for model construction. Based on the goal of the data mining task, the nature of data and the results needed, one technique may become more appropriate than the other.

Selecting the actual modeling technique is usually the initial step for model building. Here, decision tree and neural network techniques have been selected for the modeling exercise. The Decision tree used is built in J48, which uses information gain method for constructing the tree. The Neural Network, which is the second alternative technique, is built using MultilayerPerceptrone, which uses back propagation algorithm.

Decision trees are powerful and popular tools for classification (predicting what group a case belongs to), and for regression (predicting a specific value).The attractiveness of the tree based methods is due to the fact that, in contrast to neural networks, decisions trees represent rules. Rules can readily be expressed so that humans can understand them. In other words, the visual

presentation makes the decision tree very easy to understand and assimilate. As a result, the decision tree has become a very popular data mining technique [22].

As Ishwar [11] elucidated, the decision tree method encompasses a number of specific algorithms. These are classification and regression trees (CART), chi-squared automatic interaction detection (CHAID), and C4.5. CART requires less data preparation than CHAID, but produces only two- way splits. CHAID can produce tree with multiple sub nodes for each splits. The ID3 algorithm is a decision tree algorithm based on greedy attribute selection. It calculates information gain on the unselected attributes and splits the node using the attribute that has the highest information gain. This algorithm requires both the predictors and the class attribute to be nominal. It does not support training set that has missing values.

The J48 algorithm is derived from C4.5 algorithm, which is an enhancement of ID3. The reason for selecting J48 decision tree algorithm was it supports both numeric and nominal predictors, and nominal class attribute. It can also handle missing values. Additionally, J48 divides the source data by the attribute that most cleanly separates the dataset, one can use the resulting tree to see what attributes are the most important.

The second alternative data mining techniques selected was neural network. Neural networks or Artificial Neural Networks (ANN) are densely interconnected networks of simple computational elements. The elements of networks are called neurons.

Neural networks have broad applicability to real world business problems. They have already been successfully applied in many industries. Since neural networks are best at identifying patterns or trends in data, they are well suited for classification or prediction needs including; risk management, customer research, sales forecasting, data validation and target marketing [23].

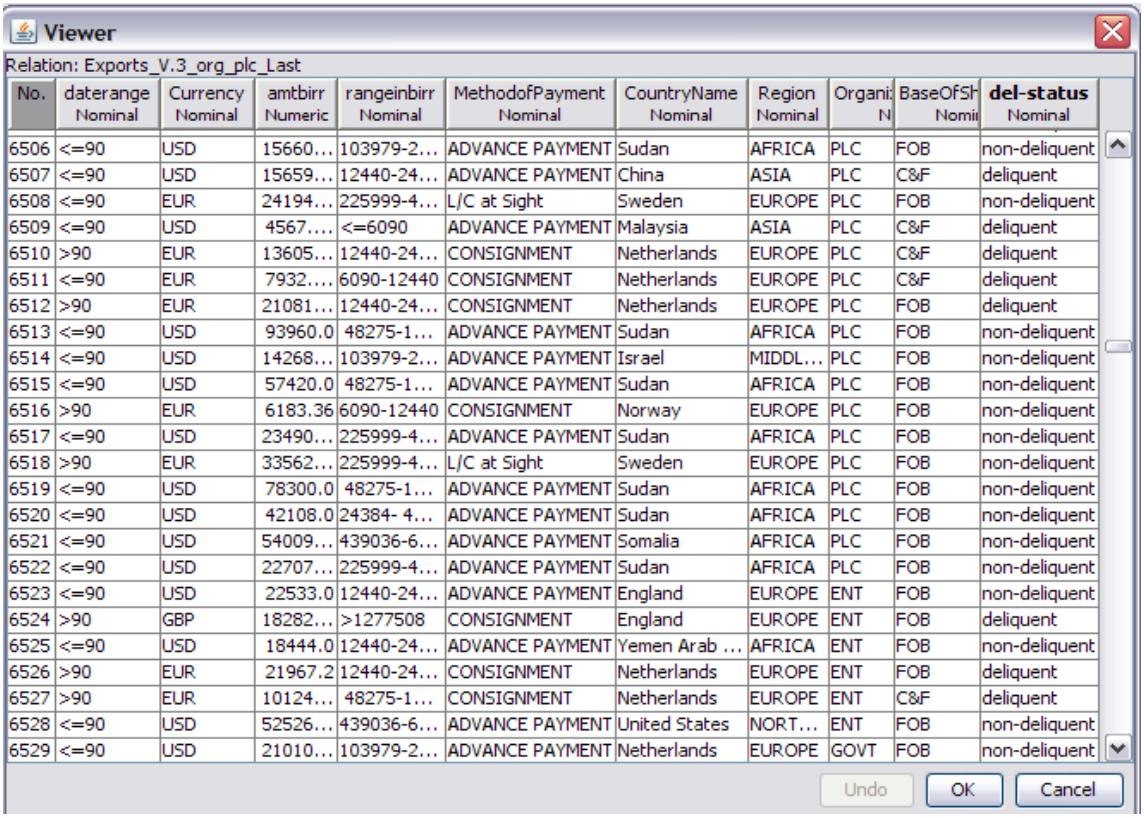
Although Neural Networks are not preferable for their black box behaviors and lack of reproducibility of achievements, they are good choices for most classification and prediction tasks when the results of the model are more important than understanding how the model works. These were the main reasons for the selection of the neural networks as the second alternative modeling techniques for this study so that it also helps to compare the accuracy of the result obtained from the two algorithms namely decision tree and neural network.

The most widely used neural classifier today is MultilayerPerceptron (MLP) network which has also been extensively analyzed and for which many learning algorithms have been developed. The MLP belongs to the class of supervised neural networks. MLP networks are general-purpose, flexible, nonlinear models consisting of a number of units organized into multiple layers. The complexity of the MLP network can be changed by varying the number of layers and the number of units in each layer.

5.1.2 Model Building

Once the modeling techniques were selected and the type of the desired algorithm identified, the next task was setting the necessary parameters for the algorithm. For all the techniques, the target class is delinquent_status with values of delinquent and non-delinquent, and the values of the rest predictors were numerical and nominal. Figure 5.1-1 and figure 5.1-2 provide sample datasets for export and import transaction respectively.

Figure 5.1-1: Sample Data Set for Export Transaction



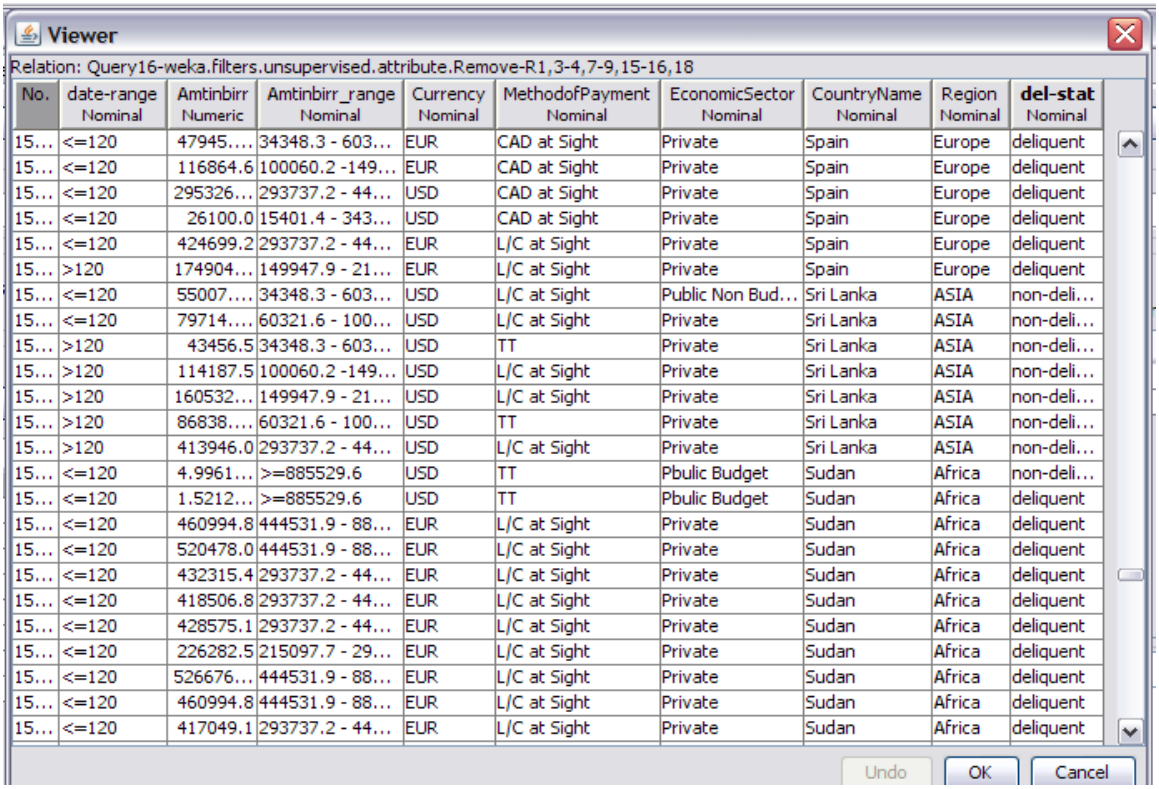
Viewer

Relation: Exports_V.3_org_plc_Last

No.	daterange Nominal	Currency Nominal	amtbirr Numeric	rangeinbirr Nominal	MethodofPayment Nominal	CountryName Nominal	Region Nominal	Organi: N	BaseOfSh Nomi	del-status Nominal
6506	<=90	USD	15660...	103979-2...	ADVANCE PAYMENT	Sudan	AFRICA	PLC	FOB	non-deliquent
6507	<=90	USD	15659...	12440-24...	ADVANCE PAYMENT	China	ASIA	PLC	C&F	deliquent
6508	<=90	EUR	24194...	225999-4...	L/C at Sight	Sweden	EUROPE	PLC	FOB	non-deliquent
6509	<=90	USD	4567....	<=6090	ADVANCE PAYMENT	Malaysia	ASIA	PLC	C&F	deliquent
6510	>90	EUR	13605...	12440-24...	CONSIGNMENT	Netherlands	EUROPE	PLC	C&F	deliquent
6511	<=90	EUR	7932....	6090-12440	CONSIGNMENT	Netherlands	EUROPE	PLC	C&F	deliquent
6512	>90	EUR	21081...	12440-24...	CONSIGNMENT	Netherlands	EUROPE	PLC	FOB	deliquent
6513	<=90	USD	93960.0	48275-1...	ADVANCE PAYMENT	Sudan	AFRICA	PLC	FOB	non-deliquent
6514	<=90	USD	14268...	103979-2...	ADVANCE PAYMENT	Israel	MIDDL...	PLC	FOB	non-deliquent
6515	<=90	USD	57420.0	48275-1...	ADVANCE PAYMENT	Sudan	AFRICA	PLC	FOB	non-deliquent
6516	>90	EUR	6183.36	6090-12440	CONSIGNMENT	Norway	EUROPE	PLC	FOB	non-deliquent
6517	<=90	USD	23490...	225999-4...	ADVANCE PAYMENT	Sudan	AFRICA	PLC	FOB	non-deliquent
6518	>90	EUR	33562...	225999-4...	L/C at Sight	Sweden	EUROPE	PLC	FOB	non-deliquent
6519	<=90	USD	78300.0	48275-1...	ADVANCE PAYMENT	Sudan	AFRICA	PLC	FOB	non-deliquent
6520	<=90	USD	42108.0	24384- 4...	ADVANCE PAYMENT	Sudan	AFRICA	PLC	FOB	non-deliquent
6521	<=90	USD	54009...	439036-6...	ADVANCE PAYMENT	Somalia	AFRICA	PLC	FOB	non-deliquent
6522	<=90	USD	22707...	225999-4...	ADVANCE PAYMENT	Sudan	AFRICA	PLC	FOB	non-deliquent
6523	<=90	USD	22533.0	12440-24...	ADVANCE PAYMENT	England	EUROPE	ENT	FOB	non-deliquent
6524	>90	GBP	18282...	>1277508	CONSIGNMENT	England	EUROPE	ENT	FOB	deliquent
6525	<=90	USD	18444.0	12440-24...	ADVANCE PAYMENT	Yemen Arab ...	AFRICA	ENT	FOB	non-deliquent
6526	>90	EUR	21967.2	12440-24...	CONSIGNMENT	Netherlands	EUROPE	ENT	FOB	deliquent
6527	>90	EUR	10124...	48275-1...	CONSIGNMENT	Netherlands	EUROPE	ENT	C&F	deliquent
6528	<=90	USD	52526...	439036-6...	ADVANCE PAYMENT	United States	NORT...	ENT	FOB	non-deliquent
6529	<=90	USD	21010...	103979-2...	ADVANCE PAYMENT	Netherlands	EUROPE	GOVT	FOB	non-deliquent

Undo OK Cancel

Figure 5.1-2: Sample Data Set for Import Transaction



Viewer

Relation: Query16-weka.filters.unsupervised.attribute.Remove-R1,3-4,7-9,15-16,18

No.	date-range Nominal	Amtinbirr Numeric	Amtinbirr_range Nominal	Currency Nominal	MethodofPayment Nominal	EconomicSector Nominal	CountryName Nominal	Region Nominal	del-stat Nominal
15...	<=120	47945....	34348.3 - 603...	EUR	CAD at Sight	Private	Spain	Europe	deliquent
15...	<=120	116864.6	100060.2 -149...	EUR	CAD at Sight	Private	Spain	Europe	deliquent
15...	<=120	295326...	293737.2 - 44...	USD	CAD at Sight	Private	Spain	Europe	deliquent
15...	<=120	26100.0	15401.4 - 343...	USD	CAD at Sight	Private	Spain	Europe	deliquent
15...	<=120	424699.2	293737.2 - 44...	EUR	L/C at Sight	Private	Spain	Europe	deliquent
15...	>120	174904...	149947.9 - 21...	EUR	L/C at Sight	Private	Spain	Europe	deliquent
15...	<=120	55007....	34348.3 - 603...	USD	L/C at Sight	Public Non Bud...	Sri Lanka	ASIA	non-deli...
15...	<=120	79714....	60321.6 - 100...	USD	L/C at Sight	Private	Sri Lanka	ASIA	non-deli...
15...	>120	43456.5	34348.3 - 603...	USD	TT	Private	Sri Lanka	ASIA	non-deli...
15...	>120	114187.5	100060.2 -149...	USD	L/C at Sight	Private	Sri Lanka	ASIA	non-deli...
15...	>120	160532...	149947.9 - 21...	USD	L/C at Sight	Private	Sri Lanka	ASIA	non-deli...
15...	>120	86838....	60321.6 - 100...	USD	TT	Private	Sri Lanka	ASIA	non-deli...
15...	>120	413946.0	293737.2 - 44...	USD	L/C at Sight	Private	Sri Lanka	ASIA	non-deli...
15...	<=120	4.9961...	>=885529.6	USD	TT	Pbulic Budget	Sudan	Africa	non-deli...
15...	<=120	1.5212...	>=885529.6	USD	TT	Pbulic Budget	Sudan	Africa	deliquent
15...	<=120	460994.8	444531.9 - 88...	EUR	L/C at Sight	Private	Sudan	Africa	deliquent
15...	<=120	520478.0	444531.9 - 88...	EUR	L/C at Sight	Private	Sudan	Africa	deliquent
15...	<=120	432315.4	293737.2 - 44...	EUR	L/C at Sight	Private	Sudan	Africa	deliquent
15...	<=120	418506.8	293737.2 - 44...	EUR	L/C at Sight	Private	Sudan	Africa	deliquent
15...	<=120	428575.1	293737.2 - 44...	EUR	L/C at Sight	Private	Sudan	Africa	deliquent
15...	<=120	226282.5	215097.7 - 29...	EUR	L/C at Sight	Private	Sudan	Africa	deliquent
15...	<=120	526676...	444531.9 - 88...	EUR	L/C at Sight	Private	Sudan	Africa	deliquent
15...	<=120	460994.8	444531.9 - 88...	EUR	L/C at Sight	Private	Sudan	Africa	deliquent
15...	<=120	417049.1	293737.2 - 44...	EUR	L/C at Sight	Private	Sudan	Africa	deliquent

Undo OK Cancel

5.1.3 Decision Tree Modeling

While building the model the following tasks were carried out:

- Both the export and import data sets were saved as comma-separated text format and imported to the WEKA software separately.
- Testing of all the partition methods in order to select the best method that results in the best accuracy measure.
- And finally, introduction of the algorithm (j48) of the selected model techniques into the experiment.

5.1.3.1 Modeling Results of the Experiments

5.1.3.1.1 Results for Export Transaction

The first experiment was conducted by considering the attributes that were not summarized such as the amount in birr and the country name. As described below, a 10-fold cross-validation partitioning method and the j48 classifiers was employed on 20,183 instances. The purpose of this experiment was to compare the results obtained from applying the summarized dataset with those from un-summarized dataset.

=== Run information ===

Scheme: weka.classifiers.trees.J48

Instances: 20183

Attributes: 8 Validity_Period, Currency, AmtInBirr, MethodofPayment, CountryName, Organization_type, BaseOfShipment, Delinquency_status

Test mode: 10-fold cross-validation

=== Summary ===

Correctly Classified Instances	19007	94.1733 %
Incorrectly Classified Instances	1176	5.8267 %

=== Confusion Matrix ===

a	b	<-- classified as:
6991	621	a = delinquent
555	12016	b = non-delinquent

The confusion matrix depicts that out of the total datasets (20,183) provided to the experiment, 19,007 (94.17%) datasets were classified correctly and 1,176 (5.83%) were classified incorrectly. A further look at the output results indicates that:

- Out of the 12,571 non-delinquent transactions, 12,016 (95.59%) datasets were identified as true non-delinquent transactions where as 555 (4.42%) of the dataset were incorrectly classified as delinquent.
- Likewise out of the 7,612 delinquent transactions, 6,991 (91.84%) datasets were identified as true delinquent transactions where as 621 (8.16%) of the dataset were incorrectly classified as non-delinquent.

Then, based on the first experiment results and further consultations with the relevant expert at the NBE, the value of the amount in Birr (AmtInBirr) and CountryName attributes were set to higher class concept. This was in an effort to arrive at a better alternative to the complexity of the tree result found from the first experiment, which makes it difficult to understand the rule derived therein. The size of the tree and the number of the leaves produced in the first experiment were

173 and 136 respectively. Thus, the second experiment was conducted by considering the summarized value for the above two mentioned attributes and setting similar values for the rest of the attributes as in the case of the first experiment. The result of the second experiment is given below.

=== Run information ===

Scheme: weka.classifiers.trees.J48

Instances: 20183

Attributes: 8 Validity_Period, Currency, AmtOfBirrIn_Range, MethodofPayment, Country_Region, Organization_type, BaseOfShipment, Deliquency_status

Test mode: 10-fold cross-validation

=== Summary ===

Correctly Classified Instances	18920	93.7423 %
Incorrectly Classified Instances	1263	6.2577 %

=== Detailed Accuracy by Class ===

Precision	Recall	Class
0.907	0.93	delinquent
0.957	0.942	non-delinquent

=== Confusion Matrix ===

a	b	<-- classified as:
7076	536	a = delinquent
727	11844	b = non-delinquent

The confusion matrix shows that out of the total datasets (20,183) used in the second experiment, 18,920 (93.74%) datasets were classified correctly and 1,263 (6.26%) were classified incorrectly. Further, out of the 12,571 non-delinquent transactions, 11,844 (94.23%) datasets were identified as true non-delinquent transactions whereas 727 (5.77%) of the dataset were incorrectly classified as delinquent. Similarly, out of the 7,612 delinquent transactions, 7,076 (92.96%) datasets were identified as true delinquent transactions while 536 (7.04%) of the dataset were incorrectly classified as non-delinquent.

Notwithstanding the slight decrease (by 0.43%) in the accuracy of the model, with the change introduced to the two attributes of the datasets, the size of the tree and the number of the leaves significantly reduced to 35 and 27 in the second experiment. This makes the process of generating rules from the model considerably easier.

In order to select the best dataset partitioning, various experiments have been made by applying different partition mechanisms. The results are summarized in Table 5.1-1 below.

Table 5.1-1: The result of dataset partitioning experiment on export transaction records

Partition Method	Delinquent	Non-Delinquent	Total	No. Leaves /Tree size
70%	93.36%	94.43%	94.02%	27 / 35
80%	92.77%	94.28%	93.68%	27 /35
90%	92.96%,	94.11%,	93.66%	27 / 35
10-fold cross-validation	92.96%	94.22%	93.74%	27 / 35

From Table 5.1-1, the best result is obtained at the 70 % partitioning split. The table further indicates the significance of undertaking various partitioning splits in order to reach at the best outcome.

The detail result for the 70% partitioning split is demonstrated below.

=== Run information ===

Scheme: weka.classifiers.trees.J48

Instances: 20183

Attributes: 8 Validity_Period, Currency, AmtOfBirrIn_Range, MethodofPayment, Country_Region, Organization_type, BaseOfShipment, Deliquency_status

Test mode: split 70.0% train, remainder test

=== Summary ===

Correctly Classified Instances	5693	94.0215 %
Incorrectly Classified Instances	362	5.9785 %

=== Detailed Accuracy By Class ===

Precision	Recall	Class
0.911	0.934	delinquent
0.959	0.944	non-delinquent

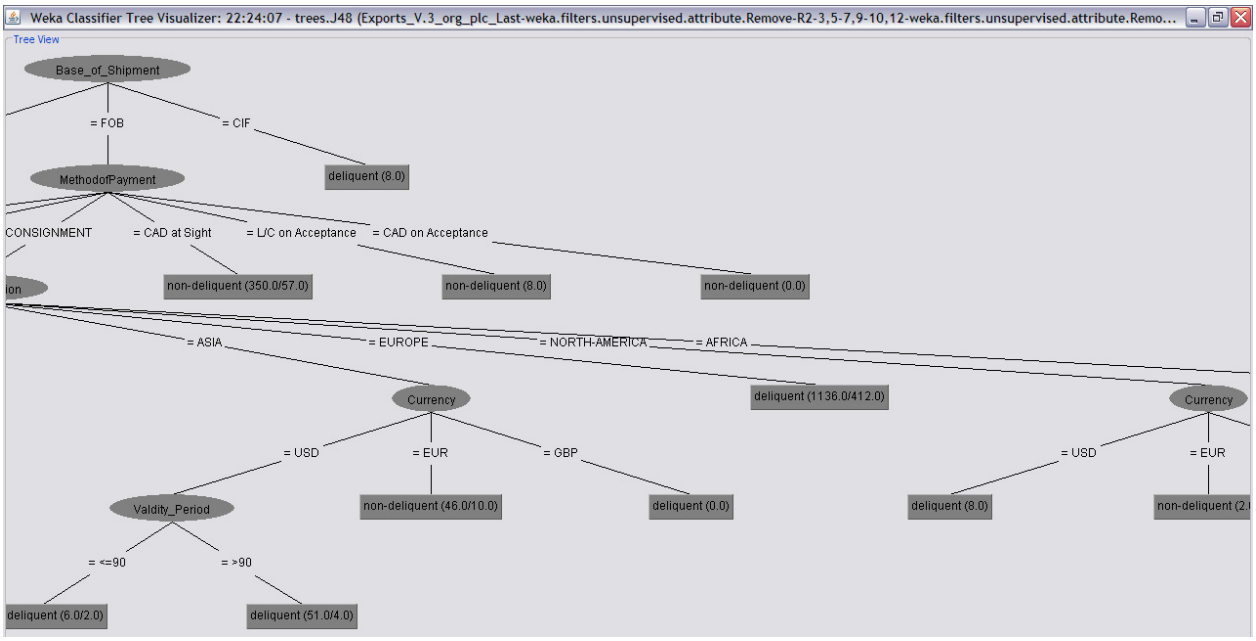
=== Confusion Matrix ===

a	b	<-- classified as:
2151	153	a = delinquent
209	3542	b = non-delinquent

A number of experiments were attempted by changing parameters like minNumObj (Minimum number of instances in a leaf) in order to improve the accuracy and the complexity of the tree of the model. However, the result shows that the accuracy of the model has decreased as the size of the tree reduced thereby omitting important attributes.

Hence, the decision tree model built using the summarized attributes and with 70% split percentage is considered for further comparison to the model built using the neural network approach.

Figure 5.1-3: Decision tree generated for export transaction.



By traversing the generated tree from root to leaf, it is possible to derive important rules that facilitate business decision-making. Some sample rules derived from the experiment are described as follows.

1. IF (BaseOfShipment = FOB) AND (MethodofPayment = ADVANCE PAYMENT)
   THEN non-delinquent

2. IF (BaseOfShipment = FOB) AND (MethodofPayment = CONSIGNMENT) AND (Country_Region = MIDDLE-EAST) AND ((AmtOfBirrIn_Range = 12440-24384) OR (AmtOfBirrIn_Range = 6090-12440) OR (AmtOfBirrIn_Range = 103979-225999))
   THEN delinquent

3. IF (BaseOfShipment = FOB) AND (MethodofPayment = CONSIGNMENT) AND (Country_Region = ASIA) AND (Currency = USD) AND (Validity_Period >=90)
   THEN delinquent

4. IF (BaseOfShipment = FOB) AND (MethodofPayment = CONSIGNMENT) AND (Country_Region = EUROPE)
   THEN delinquent

5. IF (BaseOfShipment = FOB) AND (MethodofPayment = CONSIGNMENT) AND (Country_Region = AFRICA
   THEN non-delinquent

6. IF (BaseOfShipment = FOB) AND (MethodofPayment = CAD at Sight)
   THEN non-delinquent

5.1.3.1.2 Results for the Import Transaction

The first two experiments (using summarized and unsummarized datasets) were carried out and the results showed a similar condition as in the case of the export transaction dataset. Accordingly, only the result of the experiment on the summarized dataset is discussed underneath.

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Instances: 20613

Attributes: 7 Valdity_Period_Imp, AmtOfBirrIn_Range, Currency, MethodofPayment, EconomicSector, Country_Region, Deliquency_status

Test mode: 10-fold cross-validation

=== Summary ===

Correctly Classified Instances	20207	98.0304 %
Incorrectly Classified Instances	406	1.9696 %

=== Detailed Accuracy by Class ===

Precision	Recall	Class
0.975	0.997	non-delinquent
0.993	0.942	delinquent

=== Confusion Matrix ===

a	b	<-- classified as:
14379	44	a = non-delinquent
362	5828	b = delinquent

From the confusion matrix, it is observed that out of the total datasets (20,613) used in the experiment, 20,207 (98.03%) datasets were classified correctly and 406 (1.97%) were classified incorrectly. Further output results are such that, out of the 14,423 non-delinquent transactions, 14,379 (99.69%) datasets were identified as true non-delinquent transactions, but 44 (0.31%) of the dataset were incorrectly classified as delinquent. Also, out of the 7,612 delinquent transactions, 5,828 (94.15%) datasets were identified as true delinquent transactions, while 362 (5.85%) of the dataset were incorrectly classified as non-delinquent.

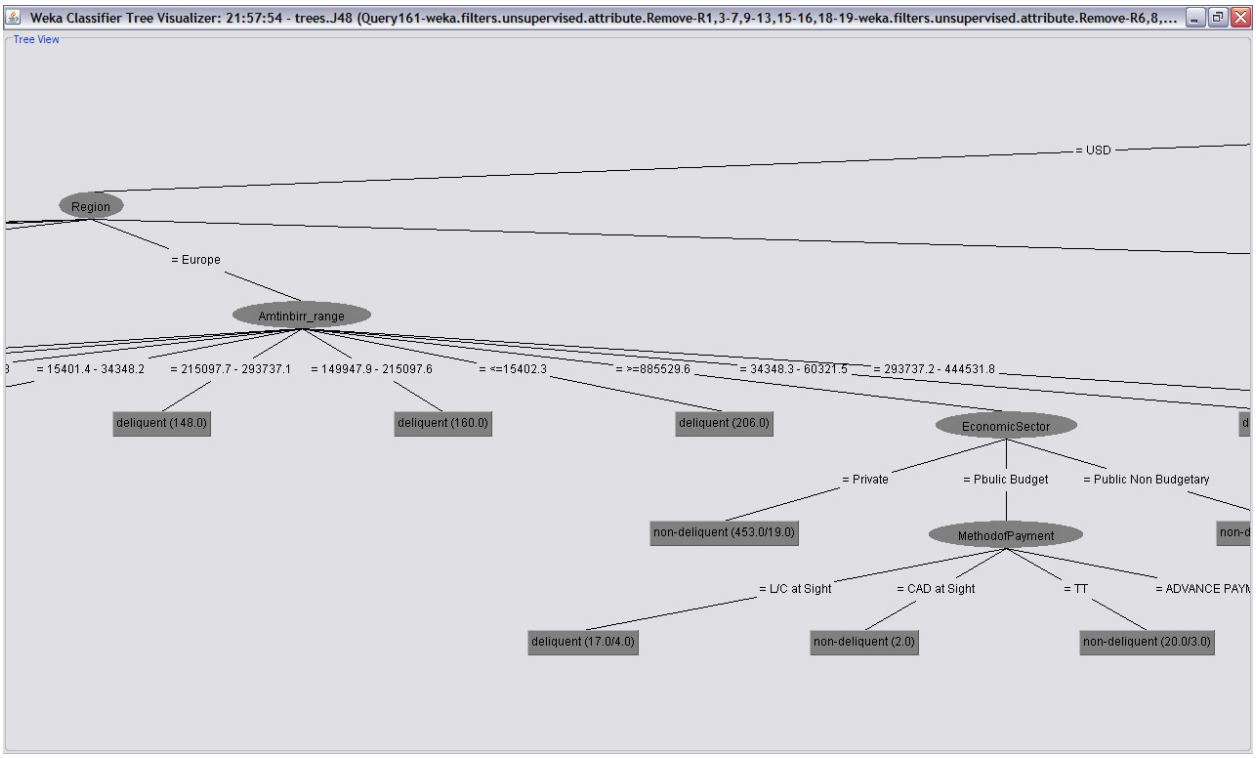
Similarly, several different attempts were made to improve the accuracy of the experiment by changing various parameters, such as the dataset partitioning method and the minNumObj. The results of the dataset partitioning experiments on import transaction records are presented in Table 5.1-2

Table 5.1-2: The result of the dataset partitioning experiment on import transaction records

Partition Method	Non-Delinquent	Delinquent	Total	No. Leaves /Tree size
70%	99.63%	93.68%	97.85%	46 / 58
80%	99.68%	94%	97.96%	46 / 58
90%	99.72%	94.05%	98.01%	46 / 58
10-fold cross-validation	99.7%	94.15%	98.03%	46 / 58

As Table 5.1-2 portrays, the best measure of delinquent transactions is captured under the 10-fold cross-validation method, which also signifies the highest total percentage value. Applying this dataset partitioning method, the researcher further tried to reduce the complexity of the tree by changing the minNumobj parameter. Nevertheless, the result was such that the complexity of the tree was reduced at the cost of losing important attributes, such as Validty_period_imp. Therefore, the default parameter for minNumobj (value 2) was considered to include the important attributes.

Figure 5.1-4: Decision tree generated for import transaction.



By the same token, navigating through the generated tree from root to leaf aids to construct relevant decision making rules. In what follows, examples of possible rules derived from the experiment are enumerated.

1. IF (Currency = USD) AND (Country_Region = ASIA) AND (EconomicSector = Private)
   THEN non-delinquent (8009.0/105.0)

2. IF (Currency = USD) AND (Country_Region = ASIA) AND (EconomicSector = Public Budget) AND (MethodofPayment = L/C at Sight OR TT)
   THEN delinquent (47/16)

3. IF (Currency = USD) AND (Country_Region = Middle-East)
   THEN non-delinquent (3988.0/75.0)

4. IF (Currency = USD) AND (Country_Region = Europe) AND (AmtOfBirrIn_Range = 60321.6 - 100060.1 OR 100060.2 -149947.8 OR 15401.4 - 34348.2 OR 215097.7 - 293737.1 OR 149947.9 - 215097.6)
   THEN delinquent (810)

5.1.4 Neural Network Model Building

The second alternative data mining techniques employed was neural network. This technique was selected for comparison of the accuracy result attained from the decision tree. Prior to the construction of the model, the dataset had to be converted to the appropriate format that fits into the neural network model.

Neural network techniques can only possess datasets when the values of attributes are between 0 and 1. Only the target class can have nominal value. Therefore, to make the dataset suitable for the neural network, the values of all the attributes except the target class were changed into numbers between the acceptable ranges, this process usually known as normalization [28]. The attributes value can be changed in to its equivalent normalized form using the following formula.

$$ai = (vi - \min vi) / (\max vi - \min vi) \quad \text{----------(5.1)}$$

Where vi is the actual value of attribute i, and the maximum and minimum are taken over all instance of the attribute i.

5.1.4.1 Experiment on Export Transaction Dataset

As the aim of this experiment was to look into the accuracy of results, the same dataset used for the experiment in the decision tree model was applied. After the conversion of the export dataset into normalized values, the dataset was saved as comma-separated text format. Then by employing the data import wizard, the dataset was imported to the WEKA software and a 10-fold

cross validation was set to partition the dataset. After partitioning, multilayerPerceptron algorithm was applied to train the model. In line with theoretical and experimental recognition, the modeling was carried out using the default parameter.

=== Run information ===

Scheme: Weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -E 20 -H a

Instances: 20183

Attributes: 8 Validity_Period, Currency, AmtOfBirrIn_Range, MethodofPayment,
Country_Region, Organization_type, BaseOfShipment, Deliquency_status

Test mode: 10-fold cross-validation

=== Summary ===

Correctly Classified Instances: 18853 93.4103 %

Incorrectly Classified Instances: 1330 6.5897 %

=== Confusion Matrix ===

a	b	<-- classified as:
7066	546	a = delinquent
784	11787	b = non-delinquent

The accuracy of this experiment was 93.41% which indicate that out of the total records (20,183) supplied 18, 853 records were classified correctly while the remaining 1,330 records (6.59%) were classified incorrectly. Moreover, the results of the experiment have shown that about 92.83% of the records in the class of delinquent were classified correctly while 93.76% of the records in the class of non-delinquent were classified correctly.

Also, an attempt was made to improve the accuracy of the experiment by changing various parameters of the algorithm. To this end, different result options were reached at by altering the hidden layer numbers starting from 1 to 9. Generally speaking, as the number of hidden layers increases the accuracy of the model also increases. The result of the experiment is shown in Table 5.1-3.

Table 5.1-3: Accuracy result for different hiddenlayer values for export transaction

Number of hiddenlayer	1	2	3	4 / a	5	6	7	8	9
Accuracy	92.96 %	93.18 %	93.43 5%	93.41 %	93.52 %	93.539 %	93.57 %	93.54 %	93.589 %

Further adjustments on various parameters were made in order to test for better accuracy results. For example, adjusting the split percentage of the dataset to training and test set as 70 and 30% respectively resulted in a better performance of 93.69% accuracy. Additional attempts in regulating some parameters such as the learning rate and the momentum of the model, also led to a further improvement in accuracy of 93.71%. The detail of the experiment leading to this superior result is explained below.

=== Run information ===

Scheme: Weka.classifiers.functions.MultilayerPerceptron -L 0.5 -M 0.2 -N 500 -E 20 -H 9

Instances: 20183

Attributes: 8 Validity_Period, Currency, AmtOfBirrIn_Range, MethodofPayment, Country_Region, Organization_type, BaseOfShipment, Deliquency_status

Test mode: split 70.0% train, remainder test

=== Summary ===

Correctly Classified Instances	5674	93.7077 %
Incorrectly Classified Instances	381	6.2923 %

=== Detailed Accuracy by Class ===

Precision	Recall	Class
0.898	0.941	delinquent
0.963	0.934	non-delinquent

=== Confusion Matrix ===

a	b	<-- classified as :
2169	135	a = delinquent
246	3505	b = non-delinquent

The result of the experiment depicts that out of the total test records (6,055) provided to the model 5,674 (93.71%) records were correctly classified whereas the rest records 381(6.29%) were incorrectly classified. The comparison of the results obtained from the decision tree model and the neural network model are discussed in the evaluation section of the paper.

5.1.4.2 Experiment on the Import Transaction Dataset

In this experiment, the same dataset provided for the decision tree model was fed into the neural network model. Similarly, the dataset was normalized to the format that is acceptable by the model. Moreover, the default parameter setting was used for the mulilayerPerceptron and the 10-fold cross validation was applied to the experiment.

=== Run information ===

Scheme: weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a

Instances: 20613

Attributes: 7 Valdity_Period_Imp, AmtOfBirrIn_Range, Currency, MethodofPayment, EconomicSector, Country_Region, Deliquency_status

Test mode: 10-fold cross-validation

=== Summary ===

Correctly Classified Instances	19986	96.9582 %
Incorrectly Classified Instances	627	3.0418 %

=== Confusion Matrix ===

a	b	<-- classified as :
14330	93	a = non-delinquent
534	5656	b = delinquent

The result of the experiment depicts that out of the total records (20, 163) provided to the model 19,986 (96.96%) records were correctly classified whereas the rest records 627(3.04%) were incorrectly classified.

As in the case of the export dataset, it was attempted to improve the accuracy of the experiment by changing parameter values of the algorithm as shown in Table 5.1-4. The result is such that increasing the value of the parameters is likely to render increased accuracy.

Table 5.1-4: Accuracy result for different hiddenlayer values for import transaction

Number of hiddenlayer	1	2	3	4	5	6	7	8	9
Accuracy	91.02 %	96.95 %	96.82 %	96.82 %	97.12 %	97.1 %	97.1 %	97.23 %	97.26 %

As it is observed from the result table, the best percentage of accuracy is measured at the hiddenlayer with parameter value 9. In neutral networks, a better measure of accuracy can be normally attained by trying to make adjustments on the various parameters involved. However, unlike the export dataset experiment, a further attempt to fine tune the learning rate and the momentum parameters did not result in improved accuracy. The details of the experiment with superior results follow.

=== Run information ===

Scheme: weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H 9

Instances: 20613

Attributes: 7 Valdity_Period_Imp, AmtOfBirrIn_Range, Currency, MethodofPayment, EconomicSector, Country_Region, Deliquency_status

Test mode: 10-fold cross-validation

=== Summary ===

Correctly Classified Instances 20049 97.2639 %

Incorrectly Classified Instances 564 2.7361 %

=== Detailed Accuracy by Class ===

Precision	Recall	Class
0.965	0.997	non-delinquent
0.991	0.917	delinquent

=== Confusion Matrix ===

a	b	<-- classified as
14374	49	a = non-delinquent
515	5675	b = delinquent

The result of the experiment depicts that out of the total records (20, 163) provided to the model 20,049 (97.26%) records were correctly classified whereas the rest records 564(2.74%) were incorrectly classified.

5.2 Evaluation

Now that the relevant models for the research have been built and experiments run, this section provides an assessment of performance between the decision tree and neutral network models. This assessment is meant to highlight a more practical approach of data mining with respect to greater accuracy of performance. There are various criteria for undertaking this kind of assessment work. For example, according to [9], classification models can be evaluated and compared in terms of accuracy, precision, recall value, confusion matrix, and interpretability.

In the experiment for export transaction, the best predictive classification models built, using the decision tree and neural network, were considered for comparison. While building the models, the same dataset was used for maintaining consistency. The decision tree had an accuracy result

of 94.02% and the neural network model with accuracy result of 93.71%. As one can observe from the output of the experiment, the decision tree model accuracy result exceeds slightly that of the neural network.

Table 5.2-1: Comparison table for the model built for export transaction

Decision Tree Model							
		Predicted					
		Delinquent	Non-Delinquent	Total	Result	Precision	Recall
Actual	Delinquent	2151	153	2304	93.36%	0.911	0.934
	Non-Delinquent	209	3542	3751	94.43%	0.959	0.944
	Total	2360	3695	6055	94.02%		
Neural Network Model							
		Delinquent	Non-Delinquent	Total	Result	Precision	Recall
Actual	Delinquent	2169	135	2304	94.14%	0.898	0.941
	Non-Delinquent	246	3505	3751	93.44%	0.963	0.934
	Total	2415	3640	6055	93.71%		

The result from the experiment depicts that the precision and recall value of the built models had almost similar figures.

Both models had incorrect predictions. The decision tree model incorrectly predict 153 records as non-delinquent and 209 records as delinquent whereas the neural network model wrongly predict 135 records as non-delinquent and 246 records as delinquent. When conducting an experiment with a large amount of datasets, such difference between models becomes negligible. Therefore, according to the evaluation of this specific criterion both the decision tree and the neural model show comparable performance.

In terms of interpretability, the decision tree is a more appropriate tool as it lends itself to further interpretation of results. The neutral network is however less so as it does not have a structural description that explicitly explains how the classification is done. Overall, in terms of accuracy, precision, recall and interpretability, the decision tree model has a better advantage over the neutral network

In the case of import transaction experiment, similar to the above the best predictive classification models were selected for comparison. The same dataset was used for both model techniques in order to maintain consistency required for the comparison of the two techniques. The decision tree had an accuracy result of 98.03% and the neural network model with accuracy result of 96.96%. As one can observe from the output of the experiment, the decision tree model accuracy result exceeds that of the neural network.

Table 5.2-2: Comparison table for the model built for import transaction

Decision Tree Model							
		Predicted					
		Non-Delinquent	Delinquent	Total	Result	Precision	Recall
Actual	Non-Delinquent	14379	44	14423	99.69%	0.975	0.997
	Delinquent	362	5828	6190	94.15%	0.993	0.942
	Total	14741	5872	20613	98.03%		
Neural Network Model							
		Non-Delinquent	Delinquent	Total	Result	Precision	Recall
Actual	Non-Delinquent	14374	49	14423	99.66%	0.965	0.997
	Delinquent	515	5675	6190	91.68%	0.991	0.917
	Total	14889	5724	20613	97.26%		

The result from the experiment depicts that the precision and recall value of the model built by the decision tree had performed better than that of the model built by neural network.

Both models had incorrect predictions. The decision tree model incorrectly predicts 44 records as delinquent and 362 records as non-delinquent whereas the neural network model also wrongly predicts 49 records as delinquent and 515 records as non-delinquent. Therefore, according to the evaluation of accuracy of the model the decision tree model performed better than neural net model.

Similarly, as the neural network doesn't have a structural description that explicitly described how the classification is done, this makes it less interpretable as compared to the decision tree model.

Thus, considering all the above criteria, the predictive model built by the decision tree performs better than the model built by the neural network.

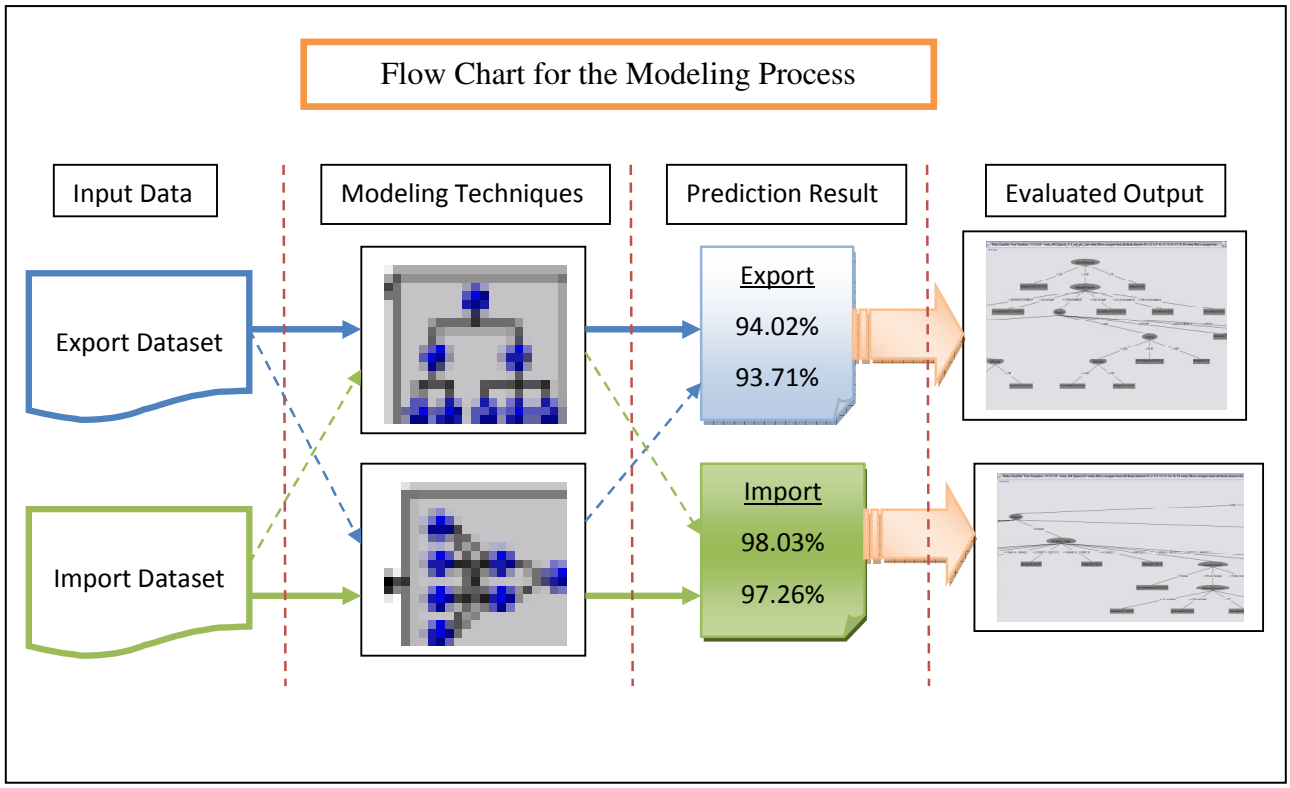
Finally, to evaluate the practical validity of the results obtained from the predictive models and the rules generated from the decision tree, the researcher has demonstrated the experiment to the relevant experts at the NBE. The work has been acknowledged as supportive of efforts at the NBE in establishing systems aimed at predicting the delinquency status of the importers and exporters, and thereby effectively manages the Country's foreign currency reserve.

5.3 Model Deployment

As described earlier, a model deployment needs an integrated effort of human and material resources such as domain experts, technology experts, necessary equipments and enough amount of money. Thus, the implementation of the predictive classification model of the study can be applied live on the Bank decision making process after a thorough evaluation and integrating of

the necessary resources mentioned above. In this respect the Bank Information Technology department can be used this study as a bridge for further work.

Figure 5.3-1: Flow Chart for the Modeling Process



CHAPTER SIX

6 CONCLUSION AND RECOMMENDATION

6.1 Conclusion

The application of data mining technology has increasingly become very popular and proved to be relevant for many sectors such as insurance, telecommunications, banking, and marketing. Particularly in the Banking Industry, data mining technology has been applied for predicting risks.

This study has tried to assess the application of data mining technology in predicting the delinquency status of importers and exporters at National Bank of Ethiopia by developing predictive model. Such a predictive model enables the Bank to predict the status of the importers and exporters well in advance and identify them as either potentially delinquent or non delinquent based on their transaction characteristics.

The CRISP-DM process model was adopted for conducting this study. The process model includes seven major parts namely: business understanding, data understanding, data preparation, data transformation, model building, evaluation and deployment. However, since a data mining task is an iterative process, these steps were not followed strictly.

A data set with 20,183 and 20,613 records for export and import transactions respectively were used to develop the classification models. Decision tree and neural network were used to develop the models. For decision tree, j48 and for neural network, a MultilayerPerceptron with back propagation algorithm were used to develop the models.

A number of experiments with various parameters were attempted in order to come up with best predictive classification model. The performance of the models built was compared and evaluated using the standard evaluation criteria such as accuracy, precision, recall and interpretability. In line with this, the predictive classification model identified as best for both export and import transactions were the models developed by the decision tree with 94.02% and 98.03% accuracy results respectively. Although the decision tree performs the best, the neural network model also depicts a promising result with 93.71 % and 97.26% accuracy for export and import transactions correspondingly.

The models built by the decision tree also show better results in precision and recall value. Moreover, in terms of interpretability, the decision tree can be explained further using the derived rule by showing the hidden relationship between the discriminate attributes. MethodofPayment, BaseOfShipment, Country_Region, AmtOfBirrIn_Range, Currency, Validity_period and EconomicSector were some of the best discriminate attributes to be considered. In the construction of the decision tree model for export transactions, the attribute Organization_type was not considered. This indicates that the attribute was not important in discriminating the records to the target class. The reason for this could be most of the transactions are PLC type.

In general, the decision tree appears to be more relevant tool for addressing the problem at stake as compared to the neural network.

In summary, the results of this research have shown the potential applicability of data mining technology at the NBE to classify importers and exporters into predefined classes (delinquent and non-delinquent) based on their transaction characteristics. Specifically, the NBE can apply a decision tree based predictive classification model with higher level of accuracy to support its

efforts in the prediction of importers and exporters delinquency with respect to the utilization of foreign exchange. Also the various important rules derived during the course of building this decision tree model can greatly facilitate the Bank's decision-making process.

6.2 Recommendation

This research work revealed the potential applicability of data mining technology in predicting the delinquency status of the importers and exporters at the National Bank of Ethiopia with respect to the utilization of the foreign currency. As this research work is an academic exercise, it should be considered as a preliminary effort to give insight on the applicability of the data mining technology for the specified area of the research problem. Thus, based on the findings of the study, the researcher would like to highlight the following recommendations for future work

- As stated earlier, the dataset used for conducting this research work was only a one-year data. Hence, it may be appropriate to consider data for more than one year as this gives broader opportunity to look into profound phenomena on the dataset. In addition, consideration of greater number of attributes such as commodity details, which contribute, their share for exhaustive examination of datasets.
- The Information Technology department of the NBE may need to investigate the applicability of the data mining technology in more details and other related comprehensive approach for strategic business needs of the Bank such as building data warehouse, customer relation management, etc.
- In the process of the study, the domain expert explained for the researcher that the FEMoS database is going to attach with the information system of the Revenue Authority. This will open an opportunity to broadly examine the delinquency status of importers and

exporters with the Bank by allowing the incorporation of more attributes into the system. It is therefore advisable that the NBE explore the possibility of putting in place a networked information system with other relevant institutions that are critical to the management of foreign currency, such as public and private commercial banks.

- In order to increase the accuracy of the predictive classification models for decision tree and neural network, further experiments should be made by incorporating and adjusting various parameters of the techniques not considered in this study.
- Drawing on the findings of the study, the researcher encourages business oriented organizations to work on the application of data mining technology to appreciate and employ classification techniques for different problems, and as a result gain a competitive advantage.

REFERENCES

1. Abenet T. (2005), Predictive Modeling using Data Mining Technology in Support of Cement Quality Assessment. Master Thesis, Addis Ababa University, Addis Ababa.
2. Asian Network for Scientific Information. (2006), Applying Data Mining Techniques: In Intrusion Detection System on Web and Analysis of Web Usage, Information Technology Journal 5 (1): 57-63, ISSN 1812-5638, [http://www.ansijournals.com/itj/2006/57-63.pdf]
3. Berry, M.J.A & Linoff, G. (2000), Mastering Data Mining: Art and Science of Customer Relationship Management. New York: Jhon Willey & Sons, Inc.
4. Berry, M.J.A & Linoff, G. (1997), Data Mining Techniques: For Marketing Sales and Customer Support. New York: Jhon Willey & Sons, Inc.
5. Daniel, T. (2005), Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley & Sons publishing.
6. Dorian, P. (1999), Data Preparation for Data Mining. Morgan Kaufmann San Francisco CA.
7. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996), From data mining to knowledge discovery in databases. AI Magazine, 17(3), pp. 37-54.
8. Frawley, W. J. et al. (1991), Knowledge Discovery in Databases. Menlo Park, CA: AAAI Press.
9. Han J. and Kamber M. (2001), Data Mining: Concepts and Techniques, Academic Press, USA.
10. Hand, D, Mannila, H, and Smyth, P. (2001), Principles of Data Mining, MIT Press.
11. Ishwar K. Sethi (n.d), Data Mining: An Introduction, December 2008, Oakland University, [http:// ww.cse.secs.oakland.edu]

12. Jeffrey W. Seifert. (2004), Data Mining: An Overview, The Library of Congress, October 2008, [http://epic.org/privacy/fusion/crs-dataminingrpt.pdf].
13. Kantardzic, M. (2003), Data Mining: Concepts, Models, Methods, and Algorithms. Speed Scientific School, University of Louisville, John Wiley & Sons publishing.
14. Klosgen W. and Zytkow J. M. (2002), "KDD: The Purpose, Necessity and Chalanges", Klosgen W. and Zytkow J. M. (Eds.), Handbook of Data Mining and Knowledge Discovery, Oxford University Press
15. Leul W. (2003), The Application of Data Mining in Crime Prevention: The Case of Oromia Police Commission. Master Thesis, Addis Ababa University, Addis Ababa.
16. Madan Lal Bhasin. (2006), Data Mining: A Competitive Tool in the Banking and Retail Industries, October 2008, [http://www.icai.org/resource_file/9935588-594.pdf]
17. Meretwork S. (2004), Possible Application of Data Mining Techniques in Supporting Credit Risk Assessment: The Case of NIB International Bank S.C. Master Thesis, Addis Ababa University, Addis Ababa.
18. Oded Maimon & Lior Rokach. (n.d), Decomposition methodology for knowledge discovery and data mining, Theory and Applications, October 2008, [http://www.worldscibooks.com]
19. Osmar R. Zaine. (n.d), Introduction to Data Mining, October 2008, [http://www.cs.ualberta.ca]
20. Paolo Giudici. (2003), Applied Data Mining: Statistical Methods for Business and Industry. England: John Wiley & Sons Ltd,
21. Peter, C, Julian, C, Randy, K, Thomas, K, Thomas, R, Colin, S, and Rudiger, W. (2000), CRISP–DM Step-by-Step Data Mining Guide.
22. Quinlan, J.R. (2003), C5.0 Online Tutorial, September 2008, [http://www.rulequest.com]

23. Richard, K and Eibe, F. (2006), WEKA Explorer User Guide for Version 3-5-3, University of Waikato, New Zealand.
24. Tamene T. (2006), Pattern Extraction from Telephone Line Fault Dataset Using Data Mining Techniques: The Case of Fixed Line at the Ethiopian Telecommunications Corporation, Master Thesis, College of Telecommunications and Information Technology, Addis Ababa.
25. Thearling K. (2003), An Overview of Data Mining Techniques, October 2008, [http://www.thearling.com/dmwhite/index.htm]
26. Two Crows Corporation (1999), Introduction to Data Mining and Knowledge Discovery, Third Edition, September 2008, [http://www.twocrows.com]
27. Two Crows Corporation. (2005), Introduction to Data Mining and Knowledge Discovery. Third Edition ISBN: 1-892095-02-5: U.S.A
28. Witten, I.H. and Frank, E. (2000), Data mining: Practical machine learning tools with Java Implementation. Morgan-Kaufman Publisher.
29. Zied Elouedil, et al, Classification with Belief Decision Trees, December 2008, [http://irida.ulb.ac.be]

APPENDICES

6.2.1 Annex A: Decision tree generated for export transaction

J48 pruned tree

Base_of_Shipment = C&F: delinquent (6203.0/142.0)
Base_of_Shipment = FOB
| MethodofPayment = ADVANCE PAYMENT: non-delinquent (6589.0/77.0)
| MethodofPayment = L/C at Sight: non-delinquent (5116.0/299.0)
| MethodofPayment = CONSIGNMENT
| | Country_Region = MIDDLE-EAST
| | | AmtOfBirrIn_Range = 48275-103979: non-delinquent (33.0/15.0)
| | | AmtOfBirrIn_Range = 439036-677999: non-delinquent (8.0/1.0)
| | | AmtOfBirrIn_Range >1277508: delinquent (4.0/2.0)
| | | AmtOfBirrIn_Range = 103979-225999: delinquent (29.0/11.0)
| | | AmtOfBirrIn_Range = 677999 - 1277508: non-delinquent (6.0)
| | | AmtOfBirrIn_Range = 24384- 48275: non-delinquent (30.0/13.0)
| | | AmtOfBirrIn_Range = 225999-439036: non-delinquent (11.0/2.0)
| | | AmtOfBirrIn_Range = 12440-24384: delinquent (130.0/59.0)
| | | AmtOfBirrIn_Range = 6090-12440: delinquent (172.0/60.0)
| | | AmtOfBirrIn_Range <=6090
| | | | Valdity_Period <=90: non-delinquent (14.0/4.0)
| | | | Valdity_Period >90: delinquent (56.0/22.0)
| | Country_Region = ASIA
| | | Currency = USD
| | | | Valdity_Period <=90: non-delinquent (6.0/2.0)
| | | | Valdity_Period >90: delinquent (51.0/4.0)
| | | Currency = EUR: non-delinquent (46.0/10.0)
| | | Currency = GBP: delinquent (0.0)
| | Country_Region = EUROPE: delinquent (1136.0/412.0)
| | Country_Region = NORTH-AMERICA
| | | Currency = USD: delinquent (8.0)
| | | Currency = EUR: non-delinquent (2.0)
| | | Currency = GBP: delinquent (0.0)
| | Country_Region = AFRICA: non-delinquent (167.0/47.0)
| MethodofPayment = CAD at Sight: non-delinquent (350.0/57.0)
| MethodofPayment = L/C on Acceptance: non-delinquent (8.0)
| MethodofPayment = CAD on Acceptance: non-delinquent (0.0)
Base_of_Shipment = CIF: delinquent (8.0)

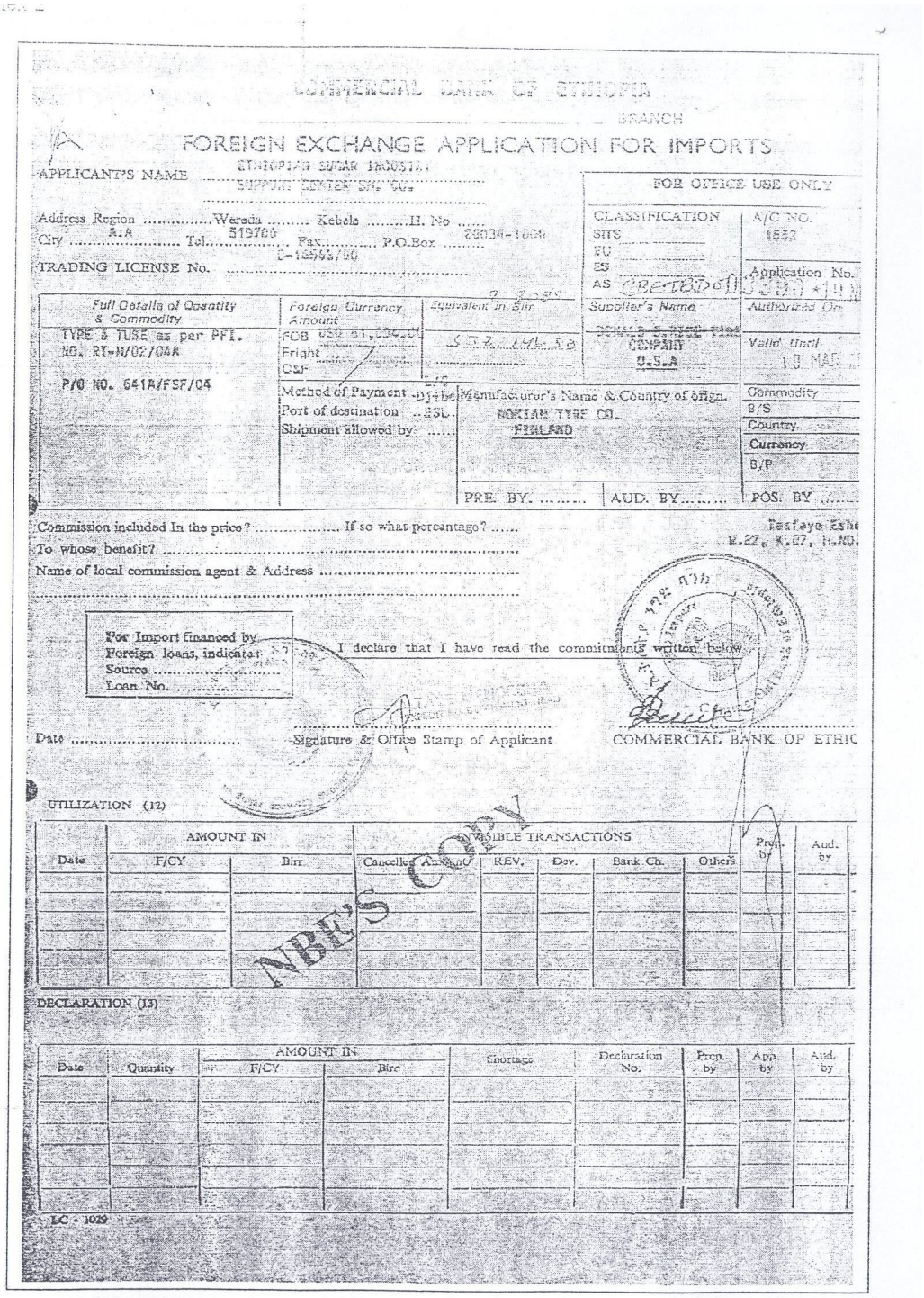
6.2.2 Annex B: Decision tree generated for import transaction

J48 pruned tree

Currency = USD
| Country_Region = North-America: non-delinquent (308.0/13.0)
| Country_Region = ASIA
| | EconomicSector = Private: non-delinquent (8009.0/105.0)
| | EconomicSector = Pbulic Budget
| | | MethodofPayment = L/C at Sight: delinquent (39.0/15.0)
| | | MethodofPayment = CAD at Sight: non-delinquent (22.0/1.0)
| | | MethodofPayment = TT: delinquent (8.0/1.0)
| | | MethodofPayment = ADVANCE PAYMENT: non-delinquent (0.0)
| | EconomicSector = Public Non Budgetary: non-delinquent (73.0/11.0)
| Country_Region = Australia
| | EconomicSector = Private: non-delinquent (32.0/1.0)
| | EconomicSector = Public Budget
| | | Valdity_Period_Imp >120: non-delinquent (2.0)
| | | Valdity_Period_Imp <=120: delinquent (2.0)
| | EconomicSector = Public Non Budgetary: non-delinquent (0.0)
| Country_Region = Middle-East: non-delinquent (3988.0/75.0)
| Country_Region = Europe
| | AmtOfBirrIn_Range = 60321.6 - 100060.1: delinquent (155.0)
| | AmtOfBirrIn_Range = 444531.9 - 885529.5: non-delinquent (243.0/100.0)
| | AmtOfBirrIn_Range = 100060.2 -149947.8: delinquent (159.0)
| | AmtOfBirrIn_Range = 15401.4 - 34348.2: delinquent (208.0)
| | AmtOfBirrIn_Range = 215097.7 - 293737.1: delinquent (148.0)
| | AmtOfBirrIn_Range = 149947.9 - 215097.6: delinquent (160.0)
| | AmtOfBirrIn_Range = <=15402.3: delinquent (206.0)
| | AmtOfBirrIn_Range >=885529.6
| | | EconomicSector = Private: non-delinquent (453.0/19.0)
| | | EconomicSector = Pbulic Budget
| | | | MethodofPayment = L/C at Sight: delinquent (17.0/4.0)
| | | | MethodofPayment = CAD at Sight: non-delinquent (2.0)
| | | | MethodofPayment = TT: non-delinquent (20.0/3.0)
| | | | MethodofPayment = ADVANCE PAYMENT: non-delinquent (0.0)
| | | EconomicSector = Public Non Budgetary: non-delinquent (38.0/10.0)
| | AmtOfBirrIn_Range = 34348.3 - 60321.5: delinquent (197.0)
| | AmtOfBirrIn_Range = 293737.2 - 444531.8: delinquent (193.0)
| Country_Region = Africa
| | EconomicSector = Private: non-delinquent (1516.0/11.0)
| | EconomicSector = Pbulic Budget
| | | AmtOfBirrIn_Range = 60321.6 - 100060.1: delinquent (0.0)
| | | AmtOfBirrIn_Range = 444531.9 - 885529.5: delinquent (1.0)
| | | AmtOfBirrIn_Range = 100060.2 -149947.8: non-delinquent (3.0)

| | | AmtOfBirrIn_Range = 15401.4 - 34348.2: non-delinquent (1.0)
| | | AmtOfBirrIn_Range = 215097.7 - 293737.1: delinquent (0.0)
| | | AmtOfBirrIn_Range = 149947.9 - 215097.6: delinquent (1.0)
| | | AmtOfBirrIn_Range <=15402.3: delinquent (1.0)
| | | AmtOfBirrIn_Range >=885529.6: delinquent (9.0/2.0)
| | | AmtOfBirrIn_Range = 34348.3 - 60321.5: non-delinquent (1.0)
| | | AmtOfBirrIn_Range = 293737.2 - 444531.8: delinquent (0.0)
| | EconomicSector = Public Non Budgetary: non-delinquent (23.0/2.0)
Currency = GBP
| Country_Region = North-America: delinquent (5.0)
| Country_Region = ASIA: non-delinquent (7.0)
| Country_Region = Australia: delinquent (1.0)
| Country_Region = Middle-East: delinquent (1.0)
| Country_Region = Europe: delinquent (387.0)
| Country_Region = Africa: delinquent (1.0)
Currency = EUR: delinquent (3973.0/11.0)

6.2.3 Annex C: Foreign Exchange Application for imports



COMMERCIAL BANK OF ETHIOPIA

BRANCH

FOREIGN EXCHANGE APPLICATION FOR IMPORTS

APPLICANT'S NAME ETHIOPIAN SUGAR INDUSTRY SUPPORT CENTER SHE Co.,

Address Region A.A Wereda 519709 Kebele H. No

City Tel. Fax P.O.Box 23034-1000

TRADING LICENSE No. 0-16595/90

FOR OFFICE USE ONLY

CLASSIFICATION SITS FU ES AS

A/C NO. 1552

Application No.

Full Details of Quantity & Commodity	Foreign Currency Amount	Equivalent in Birr	Supplier's Name	Authorized On
TYRE & TUBE as per PFI. NO. RI-M/02/04A P/O NO. 641A/FSF/04	FOB USD 81,050.00 Fright C&F		COMPANY U.S.A	Valid Until 10 MAR

Method of Payment D/P

Port of destination ESL

Shipment allowed by

Manufacturer's Name & Country of origin. NOKIAN TYRE CO. FINLAND

Commodity B/S Country Currency B/P

PRE. BY. AUD. BY. POS. BY

Commission included In the price? If so what percentage?

To whose benefit?

Name of local commission agent & Address

Tesfaye Eshe W.22, K.07, H.NO.

For Import financed by Foreign loans, indicate: Source Loan No.

I declare that I have read the commitments written below

Date Signature & Office Stamp of Applicant COMMERCIAL BANK OF ETHIC

UTILIZATION (12)

Date	AMOUNT IN F/CY	AMOUNT IN Birr	INVISIBLE TRANSACTIONS Cancelled Amount	REV.	Dev.	Bank Ch.	Others	Prep. by	Aud. by

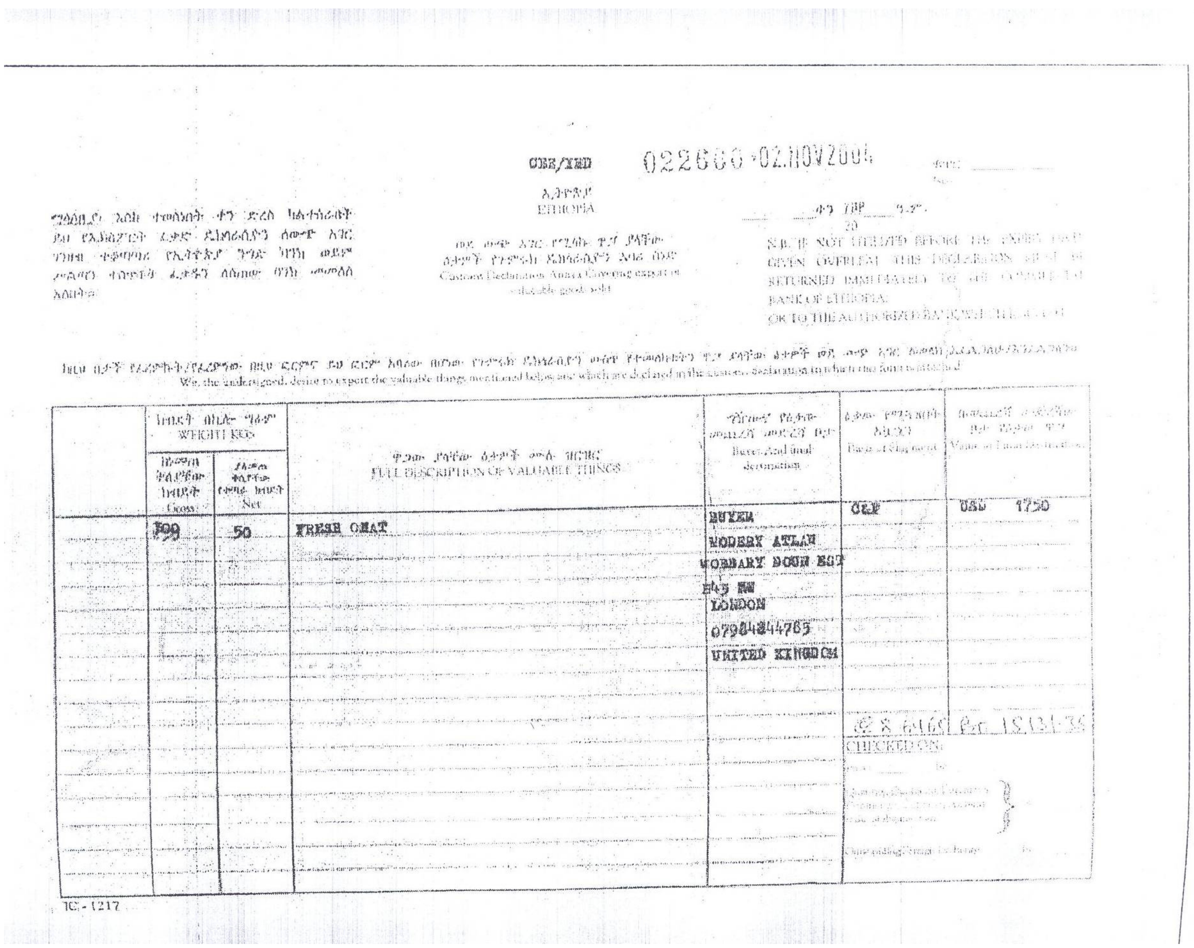
NBE'S COPY

DECLARATION (13)

Date	Quantity	AMOUNT IN F/CY	AMOUNT IN Birr	Shortage	Declaration No.	Prep. by	App. by	Aud. by

LC - 1029

6.2.4 Annex D: Custom Declaration



CBE/XED 022660 02 NOV 2004

ETHIOPIA

Customs Declaration Annex Covering export of valuable goods sold

N.B. IF NOT [illegible] BEFORE THE EXPIRY [illegible] GIVEN OVERLEAF THIS DECLARATION MUST BE RETURNED IMMEDIATELY TO THE [illegible] BANK OF ETHIOPIA OR TO THE AUTHORIZED BANK [illegible]

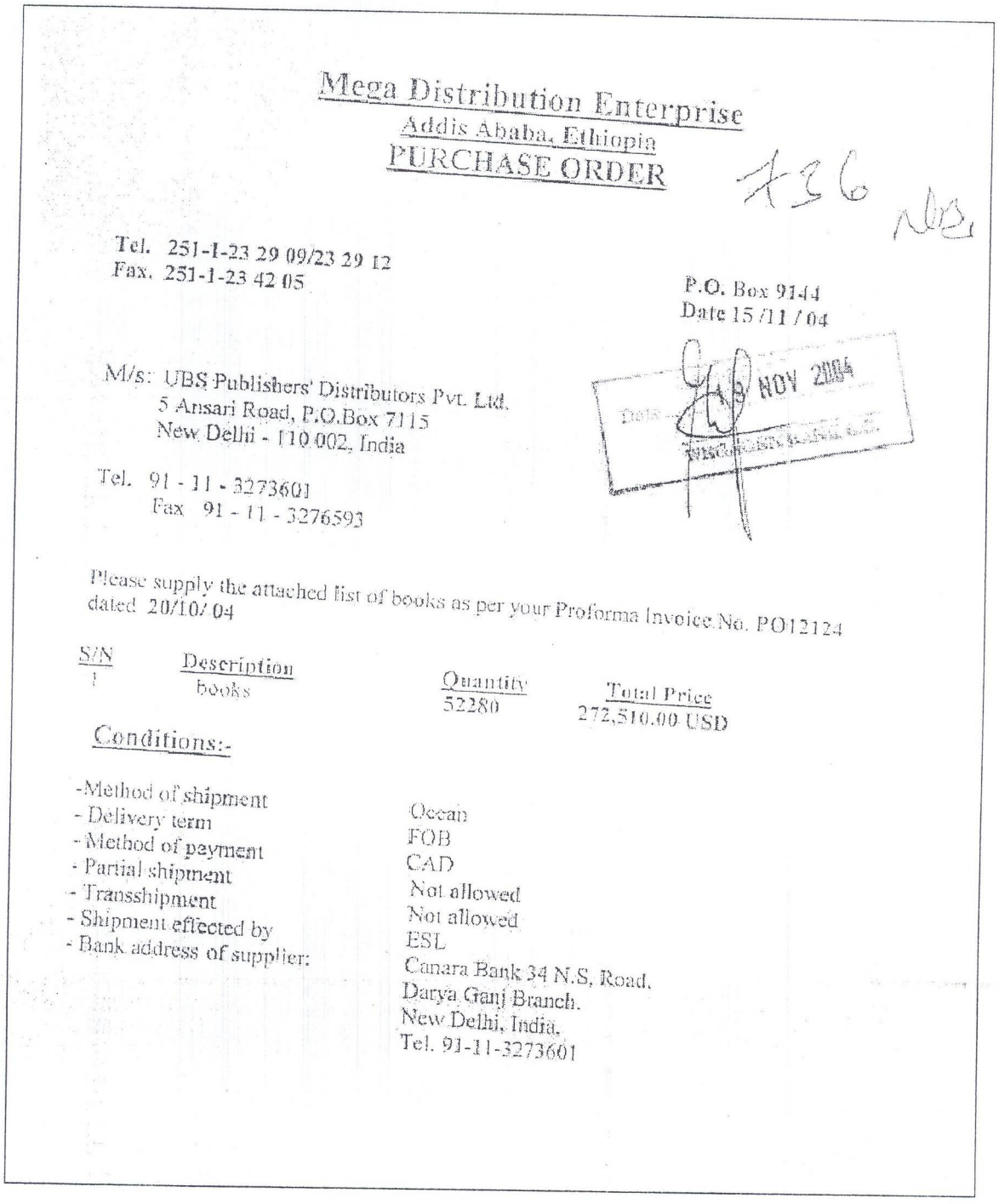
We, the undersigned, desire to export the valuable things mentioned below and which are declared in the customs declaration to which this form is attached

	WEIGHT KGS Gross	WEIGHT KGS Net	FULL DESCRIPTION OF VALUABLE THINGS	Buyer And final destination	Basis of Shipment	Value at Final Destination
	100	50	FRESH CHAT	BUYER MODERY ATLAN WORBARY DOUN EST E45 NW LONDON 07984844765 UNITED KINGDOM	C&F	USD 1750

CHECKED ON:

IC-1217

6.2.5 Annex E: Purchase order



Mega Distribution Enterprise
Addis Ababa, Ethiopia
PURCHASE ORDER

436 Nos.

Tel. 251-1-23 29 09/23 29 12
Fax. 251-1-23 42 05

P.O. Box 9144
Date 15 /11 / 04

M/s: UBS Publishers' Distributors Pvt. Ltd.
5 Ansari Road, P.O.Box 7115
New Delhi - 110 002, India

Tel. 91 - 11 - 3273601
Fax 91 - 11 - 3276593

19 NOV 2004

Please supply the attached list of books as per your Proforma Invoice No. PO12124 dated 20/10/ 04

S/N	Description	Quantity	Total Price
1	books	52280	272,510.00 USD

Conditions:-

-Method of shipment	Ocean
- Delivery term	FOB
- Method of payment	CAD
- Partial shipment	Not allowed
- Transshipment	Not allowed
- Shipment effected by	ESL
- Bank address of supplier:	Canara Bank 34 N.S. Road, Darya Ganj Branch. New Delhi, India. Tel. 91-11-3273601

6.2.6 Annex F: Clearance form

የኢትዮጵያ ብሔራዊ ባንክ
NATIONAL BANK OF ETHIOPIA
ADDIS ABABA

EGRAPHIC ADDRESS
ATIONBANK
TELEX 21020
CODES USED
ERSON 3rd & 4th ED.
TLEY'S 2nd PHRASE
B. C. 6th EDITION

PLEASE ADDRESS ANY REPLY TO
P. O. Box 5550
ADDIS ABABA

Addis Ababa

Date: 30/4/04
Clearance No. NFEMA/0077818

Dear Sirs,

Please be advised that M/S Shell Eth. Ltd. has settled foreign exchange commitments with us as of 30/4/04. Therefore, you can process import applications in accordance with the Directive to Transfer NBE's Foreign Exchange Functions to Commercial Banks No. FXD/07/1993 for the understated proforma invoice/s.

	Proforma Invoice No.		Proforma Invoice No.
1.	9042773	6.	
2		7.	
3.		8.	
4.		9.	
5.		10.	

Please note that Commercial Banks are hereby instructed not to process import applications without a bona-fide clearance certificate issued by the National Bank of Ethiopia and also advised to strictly check whether the proforma invoice has fulfilled the requirements stated in Directive FXD/07/1998 Article 5.1/b.

NATIONAL BANK OF ETHIOPIA

Authorized Signature

DECLARIATION

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

_____

Yishak Yilma

January 2009

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

_____

Dr. Manoj V.N.V

January 2009