ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCES

SCHOOL OF INFORMATION SCIENCE

DETECTING FRAUDULENT BANK CHEQUE CUSTOMERS USING DATA MINING TECHNOLOGY; THE CASE OF COMMERCIAL BANK OF ETHIOPIA

By

DAGNACHEW ASSEFA

June, 2017

ADDIS ABABA, ETHIOPIA

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCE

SCHOOL OF INFORMATION SCIENCE

DETECTING FRAUDULENT BANK CHEQUE CUSTOMERS USING DATA MINING TECHNOLOGY; THE CASE OF COMMERCIAL BANK OF ETHIOPIA

A Thesis Submitted to 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: Dagnachew Assefa

Advisor: Million Meshesa (PhD)

June, 2017

Addis Ababa, Ethiopia

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCE

SCHOOL OF INFORMATION SCIENCE

DETECTING FROUDALENT BANK CHEQUE USING DATA MINING TECHNOLOGY, THE CASE OF COMMERCIAL BANK OF ETHIOPIA

By: Dagnachew Assefa

Name and signature of Members of the Examining Board

Million Meshesa (PhD)	_____	_____
Advisor	Signature	Date
_____	_____	_____
Examiner	Signature	Date
_____	_____	_____
Examiner	Signature	Date

Declaration

This thesis has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree in any university.

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by citations giving explicit references. A list of references is appended.

Signature: _____

Dagnachew Assefa

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

Advisor's Signature: _____

Million Meshesa (PhD)

DEDICATION

I would like to dedicate this thesis to my father and my mother, Ato Assefa Mekonnen and W/o Alemnesh Siede.

ACKNOWLEDGMENT

Primary, my deepest gratitude is to the almighty God for giving me the ability to face challenges and complete this research work.

Next to this, I would like to express my sincerest gratitude to my advisor Dr. Million Meshesa for his critical guidance and valuable comments.

I would like to thank my colleagues from CBE application management and business teams specifically Wondesen Yeshitla who facilitating the data collection process and Philemon Kassa who provide me all the necessary documents and for the business guidance. I would also like to express my sincere gratitude to all Philemon Kassas' friends for their cooperativeness and for their availability whenever I need a support from domain experts.

My special thanks goes to my friend and classmate Beris Geremew, for his friendly and cooperative approach.

I have a special thank for my friends Ephrem Girmay, Girma Bulty, Mulat Belay & Mesfin Temesgen who encourage me every time. I also would like to thank my friends and others who contributed to this study.

Finally, I would like to give a special thank for my sister Sendu Assefa and my wife Daniyana Mehari for supporting me and pray for my achievement. They are the source of my inspiration and the reason of success in my life.

Abstract

The ultimate goal of any country wide organization like that of Commercial Bank of Ethiopia is to make maximum effort on delivering effective and efficient services to its' customer. To attain this objective, organization of such type should decide on the best approaches to adopt. However, this requires detailed and accurate information on the existing organizational systems. Unfortunately, such information might belong at different organization and the data may also need expert effort for using of it. So, for this type of concern data mining make use of techniques and tools to bring out the needed knowledge.

Most of the Banking sector have customers of their own. And, those customers are a user of the services given by the banks. Banks have procedures for every service they give in their branches. Although most customers obey the rules and served on the proper way, it is obvious that some other may move on the illegal way intentionally. This kind of customers will bring failure for the bank and a loss for other loyal customer of the bank. One of the service given by bank is cheque, which is exposed for this threat. This research, therefore, tried to figure out a way to handle this problem using DM technique to build a model which predict cheque fraudulent at the time of new customer registration for cheque users.

This research also strictly follow hybrid data mining process model to build predictive model that performs better having an aim of classifying fraudulent bank cheque by studying their patterns. The dataset for this study was taken from Commercial Bank of Ethiopia (CBE), a dataset of 6856 rows of customer data and 1154 rows of malicious customer data National Bank of Ethiopia (NBE) is considered for the mining task.

The heart of this research is bases on experiment conducted using different classification algorithms with diverse parameters. Basically, three classifications techniques are used such as J48, PART and SMO. For each algorithm 10 fold cross validation and percentage split was tried. A model with the best performance was registered using PART algorithm having accuracy of 86.4% with the default parameter.

The model drives interesting knowledge such as, although, customers registered in south Addis district who kept low balance in their account has found to be loyal for cheque transaction. Similarly a customer who is a male in west Addis district whose marital status is not mentioned (had missing value) at the time of registration is found to be cheque fraudulent.

The research can go for a better accuracy if a standard data has been given, and the research can move one step if the rules generated mapped to user interface for easy access of the model at hand.

Table of contents

Contents

Declaration i
DEDICATION ii
ACKNOWLEDGMENT iii
Abstract iv
Table of contents v
List of Tables vii
List of Figures viii
List of Acronyms and abbreviations ix
INTRODUCTION 1
1.1 Background of the Study 1
Cheque in banks 2
What is a fraud? 3
1.2 Statement of the Problem 4
1.3 Objective of the study 6
1.4 Scope and limitation of the Study 7
1.5 Significance of the study 7
1.6 Research Methodology 8
1.7 Organization of the thesis 11
REVIEW OF LITERATURE 12
2.1 Overview of Data Mining 12
Knowledge Discovery in Database (KDD) and Data mining 13
2.2 Knowledge Discovery Process Models (KDPM) 13
2.3 Data mining tasks 18
2.4 Data mining tools 23
2.5 Applications of Data Mining in Banking Sector 24
2.6 Impact of fraud 25
2.7 Studies on the area 28
2.7.1 Expert Studies on cheque fraud 29
2.7.2 Fraud by Bank Insiders 32
2.7.3 Applying data mining for cheque fraud analysis 35
2.7.4 Locally Related works 38

DATA PREPARATION 39

3.1 Understanding of the problem domain 39
3.2 Understanding of data 40
3.3 Data preparation 41
3.3.1 Source data selection 41
3.3.2 Data cleaning 42
3.3.3 Data constriction 44
3.3.4 Data Integration 44
3.3.5 Data Transformation 44
3.3.7 Attribute selection and ranking 45
3.4 Description of the selected attributes 47
3.5 Data Formatting 52
EXPERIMENTATION 53

4.1 Building the Model 54
4.1.1 Selecting Modeling Technique 54
4.2 Conducting Experiment 56
4.2.1 Model building using tree (J48 decision tree) 56
4.2.2 Model building using Rule Induction Algorithms (PART) 59
4.2.3 Model Building using function (SMO Algorithms) 62
4.3 Comparison of J48, PART and SMO models 65
4.4 Evaluation of Discovered Knowledge 66
4.5 finding of the study 68
CONCLUSION AND RECOMMENDATION 70

5.1 CONCLUSION 70
5.2 RECOMMENDATION 71
REFERENCES 73

Appendix: rules generated for PART algorithm 78

List of Tables

Table 1.1 Representation of confusion matrix.....19
Table 2.1 Top five financial crimes made in Australian.....37
Table 3.1 Initial data collected.....50
Table 3.4 The selected attributes and their data types.....55
Table 3. 5 Sector attribute description.....55
Table 3.6 Gender attribute description.....56
Table 3.7 Marital status attribute description.....56
Table 3.8 Customer status attribute description.....57
Table 3.9 Target attribute description.....57
Table 3.10 Balance attribute description.....58
Table 3.11 District attribute description.....59
Table 3.12 Located in Addis Ababa attribute description.....60
Table 3.13 Fraudulent attribute description.....60
Table 4.1 Different algorithm's parameter of WEKA.....63
Table 4.2 Results of J48 algorithm with four different parameters.....65
Table 4.3 Confusion matrix for predicting customer status (J48).....67
Table 4.4 Result of PART algorithm with four different parameters.....68
Table 4.5 Confusion matrix for PART predicting customer status.....70
Table 4.6 Result of SMO algorithm with two different parameter.....71
Table 4.7 Confusion matrix for SMO predicting customer status.....73
Table 4.8 Comparison of employed algorithms.....74

List of Figures

Figure 1.1 Cheque clearing process……..……..14
Figure 2.1 KDD processes model……..….23
Figure 2.2 The CRISP-DM process model……25
Figure 2.3 A hybrid mode of DM process model…….…….26
Figure 2.4 A simple decision tree……..28
Figure 2.5 Basic pseudo code for J48…….……...29
Figure 2.7 Proposed system for mobile cheque clearing system…….……..43
Figure 2.8 Electronic cheque processing…….……..44
Figure 3.1 The ranked attributes……58
Figure 4.1 A simple architecture for building a model ……..…61
Figure 4.2 Confusion matrix of J48……...62
Figure 4.3 Confusion matrix of PART……..69
Figure 4.4 Confusion matrix of SMO…….……...72

List of Acronyms and abbreviations

ACPS	Automatic cheque processing system
MICR	Magnetic Ink Character Recognition
ATM	Automated teller machine
KYC	Know your customer
SVM	Support vector machine
ECC	Electronic cheque clearing
CHAID	Chi-squared Automatic Interaction Detection
CART	Classification and Regression Trees
CBE	Commercial Bank of Ethiopia
NBE	National Bank of Ethiopia
CATS	Customers' Account and Transaction Services
CHAID	Chi-squared Automatic Interaction Detection
CART	Classification and Regression Trees
SEMMA	Sample, Explore, Modify, Model, Assess
KDPM	Knowledge Discovery Process Models
CRISP-DM	CRoss-Industry Standard Process for Data Mining
NRB	Nepal Rastra Bank
DM	Data mining
PCD	Photo Cheque Deposit
NSF	Non-sufficient fund
KD	Knowledge discovery
DSS	Decision support system
KDD	Knowledge discovery from database
SMO	Sequential minimal optimization
CRM	Customer relation manager
SMOTE	Synthetic Minority Oversampling Technique
ACFE	Association of Certified Fraud Examiners
ARFF	Attribute-Relation File Format

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

Nowadays Information is being a key resource for each and every activity of human beings. Starting from few decades ago, there has been massive production of data by different institutions. However, humans were not able to get important knowledge and hidden pattern out of it. Due to this, scientists were looking for solution to make maximum use of knowledge from a given massive data. Currently, there is abundance of information but, we are suffering from information overload, and too often, our systems deliver deafening noise without meaning. In future we would need 'world information order like, new world economic or political orders [1].

Currently, thanks to data mining, a merely collection of data has meaning by itself, because one can extract useful knowledge and unexpected patterns from it, which seems to be impossible by using humans minds processing capability. Data Mining (DM) is the process of extracting hidden knowledge from large volumes of data. The knowledge must be new, not obvious, and one must be able to use it. Data mining has been defined as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data" [1].

Advances in computer hardware and data mining software have made data mining systems accessible and affordable to many businesses [3]. Data mining techniques can be applied to a wide variety of data repositories including databases, data warehouses, spatial data, multimedia data, Internet or web-based data and complex objects [4]. It can be applied in many areas, including education, banking and insurance, medicine, security and communication [5].

DM has been applied in almost all sectors and one of the prominent applicable areas is banking industry. In Ethiopia there are a number of governmental and private banks. Currently, Commercial bank of Ethiopia (CBE) takes the leading position on the sector. CBE has been playing a prominent role in the development of the economy for the country since it was established 1942. The bank has more than 14 million customers [2]. To make the business leading continual, the bank has to design creative and technology based strategies which have to be reviewed frequently. Managing business the old way is no longer acceptable. The current situation demands a superior and flexible business process that enables a business organization to fulfill their customer expectations and provide them quality services so as to retain them and attract new ones [2].

Data mining technology can be applied by using the huge amount of bank data for extracting the new knowledge. In the present day environment, this massive electronic data is being maintained by banks around the globe. The huge size of these data bases makes it impossible for the organizations to analyze these data bases and to retrieve useful information as per the need of the decision makers [13, 14].

Cheque in banks

A Cheque (A bill of exchange) is one of the basic services banks provide to their customers [15]. Cheque is an instrument in writing containing an unconditional order, addressed to a banker, sign by the person who has deposited money with the banker, requiring him to pay on demand a certain sum of money only to or to the order of certain person or to the bearer of instrument [15]. A cheque account (current account) is a place where a customer put or deposit their money when they need to spend it soon i.e. a customer can pay a bills using cheque that is what it makes different from other type of account such as saving (this is experience from other developed countries). To make use of cheque, one has to first need to have a current account [15].

A current account is a form of deposit account that allows, among others, payment by way of cheques, and different banks may have different criteria for opening a current account [1]. When a customer open a current account, his/her bank requires customers' specimen signature. Then, the bank compares user signature on the cheque issued by customers against this specimen signature. If the signature is different, the bank refuses to pay the cheque and return the cheque on the ground that "signature differs". For a joint current account, all the account-holders' signatures must be provided together with specific instructions as to who can operate the joint account [1].

As CBE's customers' account and transaction services (CATS) procedure mentioned, a typical cheque bears the following elements across its face [1].

- Name of the bank
- Name of the branch
- Date of cheque
- Name of account holder and account number
- Name of payee/beneficiary
- Amount in words and figures
- Cheque number
- Signature of the account holder or Signatories

A cheque also contains four parties, the drawer (who writes /signs the cheque), the payee (Who receives the cash/money), the drawe bank and the collecting bank. But, cheque type of transaction involves at least three parties in case of, the drewee bank and the collecting bank are the same.

When a customer deposit a cheque into his account, customer's bank which is the collecting bank sends it to specific clearing centers to be sorted and dispatched to the drawee bank for payment. This process is called cheque clearing. The time taken for the cash to be credited into payee's account may need a few number of days (most of the time two or three days) required to clear the cheque. This process sometimes end up with problem so that, the payee may not get paid with expected time due to the following reasons which probably related to fraud [16].

- Non-sufficient fund (NSF) in customer's account to clear the cheque
- Missing signature(s)
- Signature(s) differ from specimen signature(s) in the bank's record
- Post-dated
- Stale cheque (cheque which is dated more than 6 months ago)
- Amount in words and figures differ
- Stop payment instruction has been issued
- Account closed or "legally blocked" account

What is a fraud?

A fraud is "a deliberate act of omission or commission by any person, carried out in the course of a banking transaction or in the books of accounts maintained manually or under computer system in banks, resulting into wrongful gain to any person for a temporary period or otherwise, with or without any monetary loss to the bank" [35].

Regarding this problem, different banks across the globe has different experience of tackling the problem. Most banks has strict rules to protect themselves and their customers from these situations.

Banks may use 'Overdraft Protection'; if a customer write a cheque more than the amount he/she has in his account, he/she bounces a cheque [15]. Overdraft Protection is a service that helps people avoids this mistake. With this service, if a customer writes a cheque for more money than he/she has in his/her cheque account, the bank takes money out of their savings account (or another account) to pay for the cheque. There will be also a fee whenever a customer commit overdraft as a penalty [15]. In this case a customer is obligated to deposit a cash in his account for this purpose.

Other banks may overcome this problem in different ways, since the clearing process in the banks take time, (which creates a time gap for payer to withdraw his cash before reaching to the other customer or payee) a number of studies conducted to shorten the time taking for clearing the cheque. Smart phone and internet browser are good tools for faster way of clearing the cheque in most of the developed countries. i.e. "An Alternative Method for Facilitating Cheque Clearance Using Smart Phones Application" is a study which brings a faster way of clearing cheque by means of Photo Cheque Deposit (PCD), uses a scanned version of front and back page of the paper cheque [17].

Even though, it is payer (drawee) responsibility to maintain sufficient funds in her/his account to cover the payment. Sometimes customers intentionally make insufficient fund on their current account not to conduct the payment for the other customer (payee) due to their personal cases. Considering other banks experience, if there are insufficient cleared funds in the customer's drawee account or the cheque exceeds overdraft limit, the bank dishonores customer's cheque, and charges a fee [17].

The mentioned way outs, penalty after the problem and making the cheque clearing process fast are tried, and literatures show significant improvement has been seen, this research also be conducted to add up other way of tackling the problem.

This research tried to study the fraudulent customer behavior and use it for predicting them by using a massive history data of customers. This was done by means of building a model using a machine learning classification algorithms.

1.2 Statement of the Problem

One of the important means of efficient fund transfer is through cheque clearance [16]. A lot of customers in banks might need to use cheque since, it is easy to understand and use, uniform in appearance, requires no special equipment to accept and process and it is also widely accepted [16]. Although, cheque is one of a good optional way for making transaction, it is exposed for fraud. Since the payer can be confidential only after making sure of his current account has sufficient cash at the time of cheque clearance process.

Let us consider a scenario on the existing practice, where a customer (payer) of Bank of Abyssinia issues a cheque to a customer (payee) of another bank, Commercial Bank of Ethiopia. The payee on receipt of cheque form the payer has to go to the bank holding his/her account to deposit the physical cheque so that the amount specified in the cheque is credited in his/her account. It takes about 2-3 days' of time for a cheque to get cleared and cash to be transferred. During this process Commercial Bank of Ethiopia communicates with Bank of Abyssinia to which the cheque belongs and it is only after the acceptance of the cheque by the Bank of Abyssinia that the cash specified on the cheque is credited in the account of the Payee [16].

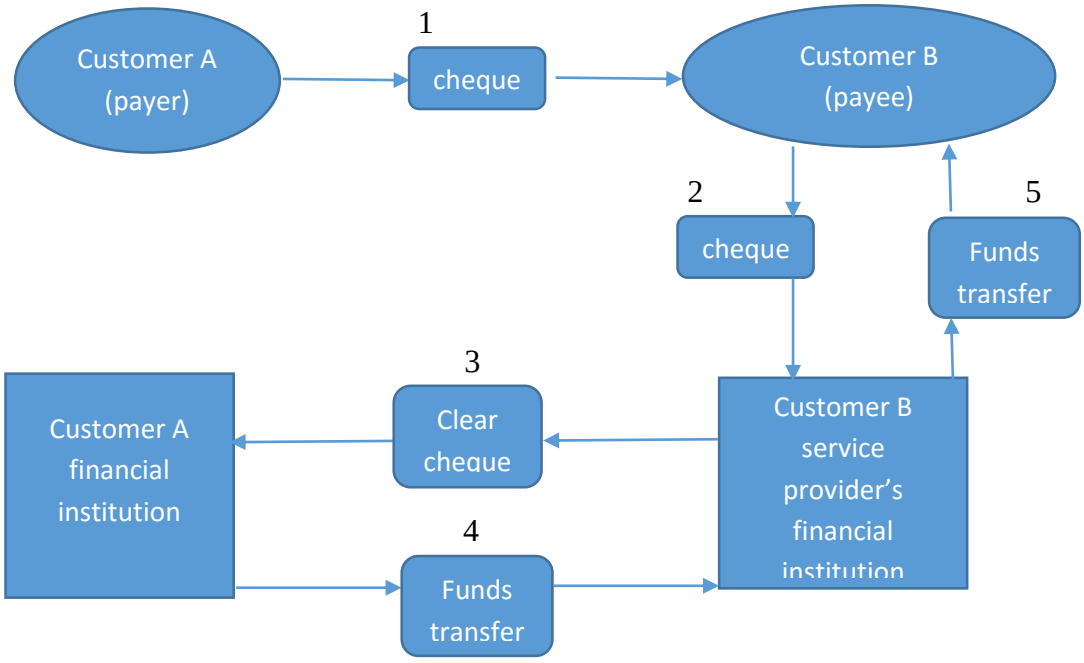

Figure 1.1 cheque clearing process

Everyday banks receive a large number of cheques; these cheques are checked and submitted to their corresponding clearing banks for clearance. After verification the clearing banks send signed instructions back to the bank stating whether the cheque is cleared or rejected [17].

The problem appears when the cheque is rejected. Although, there are a number of reasons for a cheque to be rejected, (listed on the introduction part) the main one is due to 'non-sufficient funds' (NSF) that means there is not enough cash in the account on which the cheque is drawn. As the domain experts told, this problem shows a constant increase in the banking sectors. Accourding to their information, although there is no exact recorded figure in terms of number, but the malicious customer has significantly increased 2016/2017 compared to the previous and recent consecutive years.

In the case of Commercial Bank of Ethiopia(CBE), if a customer write a cheque without sufficient cash in his/her account, the bank would give first time warning for the customer and push the case to be resolved legally (if it is necessary). If the same customer repeat this three times, the bank immediately reports the personal detail of that customer to the National Bank of Ethiopian (NBE). So, NBE distributes the detail to the other banks throughout Ethiopia with order to close the customer account and also not to let him open other account in other banks for unlimited time as

a penalty for his fraud. In Ethiopia it is known that all the public and private banks are under the control of National Bank of Ethiopia (NBE).

A better way of handling such issues comes from researches and different experiments done by domain experts. A deep study on the issue can always bring a better solution for the problem at hand. The need for conducting research on the sector help financial institutions to take competitive advantage on business, one over the other.

Considering the problem existence, there is a need for segmenting the loyal and fraudulent customers from the bank side. So, this study tried to classify fraudulent cheque customers by using CBE's dataset of customer information by making use of a model, which is built by implementing standard data mining technology with different experimentation to choose a better predictor. Therefore, it has been attempted to obtain answer for the following research questions throughout this paper:

- What are the main determinate attributes that enable to know whether the customer is cheque fraudulent or not?
- What kind of rules be formulated to predicting fraudulent cheque customers?
- To what accuracy level the applied predictive model identified bank cheque fraudulent customer?

The current information systems regarding cheque fraud in CBE is not designed as decision support systems (DSS) that would help management make effective decisions to manage resources, compete efficiently, and enhance customer satisfaction and service.

So, this research is conducted by considering the increasing in number of cheque fraud attempt in banks. And due to this gap, the bank need somehow a reliable way of tackling the problem. On the other hand, the researcher is tried to provide one optional solution through this research using personal information from customers' cheque details.

1.3 Objective of the study

1.3.1 General objective

The general objective of this study is to investigate the application of data mining technology for identifying and predicting fraudulent cheque users, using CBE's customer data record.

1.3.2 Specific objectives

In order to achieve the above stated general objective, the research conducted the following specific objectives.

- ❖ To review literatures related to this study, and identify them.
- ❖ To collect relevant data, that can be an input for this research.
- ❖ To pre-process the data (clean data, by fill in missing values, smooth noisy data, and resolve inconsistencies)
- ❖ To conduct different experimentation with different scenarios.
- ❖ To compere and select appropriate classification data mining algorithm that suites for the problem domain.
- ❖ To evaluate the performance of the model.

1.4 Scope and limitation of the Study

The scope of this research is limited to design a system using classification models for the purpose of detecting fraudulent customers by applying data mining technology. Although cheque fraud is done all over the banks in Ethiopia, the researcher has stake only with CBE's customer data due to incapability of accessing input data for this research from other banks. This mean at least the payer account should exist in CBE's database to be include in the study.

Although, customers, account & malicious account tables have a lot of personal details, such as customer name, account number & customer number the researcher have not access to those personal information, instead more general attributes like that of age, sex, sector, marital status, district and so on was used.

On the other hand, use of the discovered knowledge or in the case of this study, making usable the discovered rules using suitable interface is very important, but due to different reasons such as a time took to collect the appropriate data, the time took to clean up the data had cause time constraint for making interface for the model at hand.

1.5 Significance of the study

It is known that every research is conducted to add-up on the existing knowledge and/or solve a problem, this research also try to address the gap on cheque transaction by predicting the dishonest customer while they come to one of the branch to open current account for being cheque user.

If a system like this is implemented for detecting the cheque fraudulent customer, then the bank can take a proper follow-up on the suspected ones from the beginning or from the time of registration, and able to protect the damage they might bring to the bank.

After this research is conducted and implemented, not only the bank can benefit from this but also the customer of the bank too. This is because, as the dishonest customers decrease the transaction becomes healthier, so the payee can get the payment.

The researcher thinks this may add up something for the bank and the loyal customer as well. This may save time, energy and money losses. Customer satisfaction can also increases, which is going to meet one of the major objective of banks. The model which was built for this research has ability to identify a suspicious fraudulent cheqhe customer, so that a proper follow up for the suspicious once will benefit a lot for the bank.

1.6 Research Methodology

The word methodology refers to a documented approach which is used to perform activities in a manner which is coherent, consistent, accountable and repeatable. Methodology is a process that mainly consists of intellectual activities. Usually only the end goal of the methodological process is manifested as the product or result of the physical work [11].

1.6.1 Research design

As mentioned above, methodology is structural steps and procedures that, one can follow to achieve the objectives and research questions of a certain study. In order to achieve the specified objectives of the current research, the researcher had tried to conduct an experimental research. This is because, a researcher planned to build a model that predict dishonor customers, for this reason, experimentation is very important for having a better predictive model.

This research incorporated the typical stages that characterize a data mining process. This study has been done according to hybrid processing model, and the reason to choose a hybrid model for this research is to assess both the academic and industrial features, in addition to use the advantage of its powerfulness even on complicated data. The development of academic and industrial models has led to the development of hybrid models or a combine aspects of both [18]. This model has advantage over the others for this study since it provide research-oriented description of the steps, it follows a data mining step rather than the modeling step and the knowledge discovered at last for a particular domain may be applied in other domains. The model comprises the bellow six steps [18].

1.6.2 Understanding of the problem domain

This initial step requires working closely with domain experts to define the problem and determine the project goals, identifying key people, and learning about current solutions to the problem. It also involves learning domain-specific terminology. A description of the problem, including its restrictions, is prepared. Finally, project goals are translated into DM goals, and the initial selection of DM tools to be used later in the process is performed.

In this research also various ways were used for understanding the business, interview, observation, document analysis can be mentioned primarily. CBE staffs specifically working on cheque clearing process played a vital role through discussion and deeper interview on the workflow and business understanding.

Literature were very important source of knowledge for conducting this research. Literature review is conducted by refereeing to manuals, procedures, books, journal articles, conference paper and other sources about data mining and its application and also cheque service in banks, how fraud is committed in cheque and related transactions.

1.6.3 Understanding of the data

The research primarily make use of a data from two different database CUSTOMER table from CBE which used to determine the customer behavior and MALICIOUS CUSTOMER (also belong in CBE which is distributed by NBE) from the NBE which is used as a class for this study.

A careful analysis of CUSTOMER and MALICIOUS CUSTOMER data and its structure is done together with domain experts by evaluating the relationships of the data with the problem at hand and the DM tasks to be performed.

1.6.4 Preparation of the data

Since this part is a key for the better result of a given study, a researcher did give a lot of his time for following a standard steps for data preparation. The researcher tried to collect features of customer from CUSTOMER table of CBE database, on the other hand malicious customers list is used as a class for classifying the given cheque customers.

1.6.5 Data mining

Here is the step where a data miner uses various DM methods to derive knowledge from preprocessed data. A classification technique particularly decision tree, rule induction and support vector machine techniques were selected and used for prediction of fraudulent customers. The employed algorithms are chosen due to their easiness for understanding and interpretation. On the other hand, the researcher used only classification technique because the research dataset has clear and simplified labeled class.

Basically MS-ACESS-2013 were used for data integration purpose (NBE's data with CBE's data) MS-EXCEL-2013 had also a significant contribution while conducting the data preparation. On the other hand, WEKA version 3.7.5 data mining tools: has a major role in help of success of this research, algorithms such as decision tree, PART and SMO had been used for experimentation and building the model.

1.6.6 Evaluation of the discovered knowledge

Analysis of the decision tree models are made in terms of detailed accuracy of the classifier on the training dataset. The confusion matrix is a useful tool for analyzing how well the classifier can recognize tuples of different classes. Confusion matrix shows four important numerical quantities namely: True-Positive, False-Positive, False-Negative and True-Negative as shown below.

		Predicted customer status		Total
		Normal	fraudulent	
Actual customer status	Normal	TP	FN	TP+FN
	Fraudulent	FP	TN	FP+TN
Total		TP + FP	FN+TN	TF+FP+FN+TP

Table 1.1 representation of confusion matrix

Contextually, what each variable stands for is stated as follows, TP: The number of actually loyal customer that are classified as loyal. FN: The number of actually loyal customer that are classified as fraudulent. FP: The number of actually fraudulent customer that are classified as loyal customer. TN: The number of actually fraudulent customer that are classified as fraudulent. TP+FN: The total number of actual loyal customer. FP+TN: The total number of actually fraudulent customers. TP + FP: The total number of predicted loyal customer. FN+TN: The total number of predicted fraudulent customers. TN + FP + FN + TP: The total number of all customers.

As shown in table 1.1, the model has basically two classes, fraudulent and loyal. The fraudulent customers are from NBE which has a list of customers collected from different branches of CBE having multiple (three) cheque fraud record by their name. On the other hand, loyal customers are collected from CBE who do not have cheque dishonest record on their file.

1.6.7 Use of the discovered knowledge

This final step consists of planning where and how to use the discovered knowledge. The application area in the current domain may be extended to other domains. A plan to monitor the implementation of the discovered knowledge is created and the entire project documented. Finally, the discovered knowledge is deployed.

1.7 Organization of the thesis

These research has five chapters and organized as shown below.

The first chapter deals with, the basic overview of this paper including background, statement of the problem, objective, scope of the study, methodology, scope of the study, significance of the study and thesis organization.

In the second chapter, in detail literature review (review of related work) are done in areas of bank industry, cheque service & cheque fraud and data mining applications at the end.

The third Chapter works on data preprocessing tasks. Through this chapter the major data preprocessing tasks were applied to the data at hand.

The forth chapter deals with experimentations and result interpretations. In this chapter building of model with training dataset and validating the result with testing datasets, and interpretation of the result of the experimentation were the major concern. A comparison of the algorithms in order to choose the better is done also in this chapter.

Finally, in the last chapter conclusion and recommendations are presented.

Chapter two

REVIEW OF LITERATURE

In this chapter, a researcher tried to assess the main concepts of payment systems related to cheque, data mining concepts and finally related works from different sources of literatures that can help as an input for supporting this research.

2.1 Overview of Data Mining

Over the past years, data mining became a matter of considerable importance due to the large amounts of data available in the applications belonging to various domains. Data mining, a dynamic and fast-expanding field, that applies advanced data analysis techniques, from statistics, machine learning, database systems or artificial intelligence, in order to discover relevant patterns, trends and relations contained within the data, information impossible to observe using other techniques [30]. The overall goal of the data mining process is to extract information from a dataset and transform it into an understandable structure for further use [23]. The banks who have realized the importance of data mining are in the process of reaping huge profits and considerable competitive advantage [39].

Data mining tools can answer business questions that traditionally were time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations [24]. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line [34]. For instance When Chase Manhattan Bank in New York began to lose customers to competitors, it began using data mining to analyses customer accounts and make changes in its account requirements, thereby allowing the bank to retain its profitable customers [46].

Data mining is also being used by Fleet Bank, Boston, to identify the best candidates for mutual fund offerings. The bank mines customer demographics and account data along different product lines to determine which customers may be likely to invest in a mutual fund, and this information is used to target those customers. Bank of America's West Coast customer service call center has its representatives ready with customer profiles gathered from data mining to pitch new products and services that are the most relevant to each individual caller. Mortgage bankers are also concerned with retaining customers. The program uses leading edge Internet technologies, predictive models, and customer-direct marketing to enable lenders to identify new customers and retain those that they already have [46].

Data mining solutions implement advanced data analysis techniques used by companies like that of huge financial institutes for discovering unexpected patterns extracted from vast amounts of data, patterns that offer relevant knowledge for predicting future outcomes. The availability and affluence of data belonging to various domains make data analysis a matter of significant importance and necessity today.

Data mining consists of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns (or models) over the data [31]. Data mining helps finding knowledge from raw, unprocessed data. Using data mining techniques allows extracting knowledge from the data mart, data warehouse and, in particular cases, even from operational databases [32].

Knowledge Discovery in Database (KDD) and Data mining

It is known that, the term knowledge discovery in databases (KDD) is synonymously used with data mining. KDD refers to the overall process of discovering useful knowledge from data and that data mining refers to a particular step in the process. Furthermore, data mining is considered as the application of specific algorithms for extracting patterns from data.

The basic tasks of data mining and KD are to extract particular information from existing databases and convert it into understandable or sensible conclusions, also states that knowledge discovery (KD) is the process of transforming data into previously unknown or unsuspected relationships that can be employed as predictors of future action [67].

2.2 Knowledge Discovery Process Models (KDPM)

There are different process models originating either from academic or industrial environments.
All the processes have in common the fact that they follow a sequence of steps which more or less resemble between models.

According to Santos & Azevedo (2008) the followings are commonly used Data mining process model or methodology ; Knowledge Discovery in Database (KDD), CRoss-Industry Standard Process for Data Mining (CRISP-DM), Sample, Explore, Modify, Model, Assess (SEMMA) and Hybrid model.

The main reason for introducing process models is to formalize knowledge discovery projects within a common framework, a goal that result in cost and time savings, and improve understanding, success rates, and acceptance of such projects.

2.2.1 Knowledge discovery in database (KDD) process model

Generally there are five stages for knowledge discovery in database. Data mining is one of the tasks in the process of knowledge discovery from the database (KDD) [40].

Figure 2.1 shows the process of knowledge discovery. The steps involved in Knowledge discovery are [41]:

1. Data Selection: The data relevant to the analysis is decided and retrieved from the various data Locations.

2. Data Preprocessing: In this stage the process of data cleaning and data integration is done.
- *Data Cleaning*: It is also known as data cleansing; in this phase noise data and irrelevant data are removed from the collected data.

- *Data Integration*: In this stage, multiple data sources, often heterogeneous, are combined in a common source.

3. Data Transformation: In this phase the selected data is transformed into forms appropriate for the mining procedure.

4. Data Mining: It is the crucial step in which clever techniques are applied to extract potentially useful patterns. The decision is made about the data mining technique to be used.

5. Interpretation and Evaluation: In this step, interesting patterns representing knowledge are identified based on given measures.

The discovered knowledge is visually presented to the user. This essential step uses visualization techniques to help users understand.

Figure 2.1 KDD processes model

2.2.2 The CRISP-DM process

CRISP-DM (CRoss-Industry Standard Process for Data Mining) is a data mining project compromises a multi-step, iterative process. It consists on a cycle that comprises six stages.

1. Business understanding: this initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a DM problem definition and a preliminary plan designed to achieve the objectives.

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.

3. Data preparation: the data preparation phase covers all activities to construct the final dataset from the initial raw data.

4. Modeling: in this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values.

5. Evaluation: at this stage the model (or models) obtained are more thoroughly evaluated and the steps executed to construct the model are reviewed to be certain it properly achieves the business objectives.

6. Deployment: creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained needs to be organized presented in a way that the customer can use it.

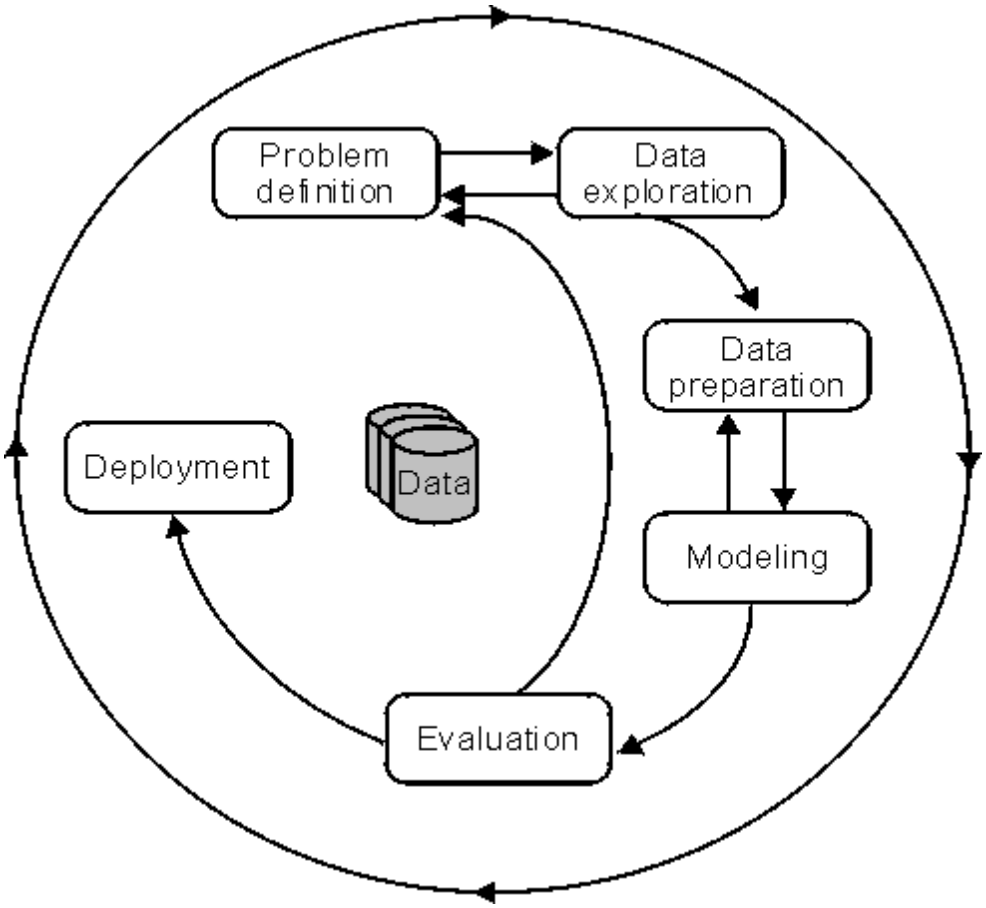

Figure 2.2 the CRISP-DM process model

2.2.3 Hybrid Model

This model was developed, by adopting the CRISP-DM model to the needs of academic research community. The model consists of six steps [19]. The description of this process model was presented in chapter one, design research part

1. Understanding of the problem domain
   This initial step requires working closely with domain experts to define the problem and determine the project goals, identifying key people, and learning about current solutions to the problem. It also involves learning domain-specific terminology. A description of the problem, including its restrictions, is prepared. Finally, project goals are translated into DM goals, and the initial selection of DM tools to be used later in the process is performed.

2. Understanding of the data
   This step includes collecting sample data and deciding which data, including format and size, is being needed. Background knowledge can be used to guide these efforts. Data are checked for completeness, redundancy, missing values, plausibility of attribute values, etc.

Finally, the step includes verification of the usefulness of the data with respect to the DM goals.

3. Preparation of the data
   This step concerns deciding which data has to be used as input for DM methods in the subsequent step. It involves sampling, running correlation and significance tests, and data cleaning, which includes checking the completeness of data records, removing or correcting for noise and missing values, etc. The cleaned data may be further processed by feature selection and extraction algorithms (dimensionality) and by derivation of new attributes (discretization).

4. Data mining
   Here is the step where a data miner uses various DM methods to derive knowledge from preprocessed data.

5. Evaluation of the discovered knowledge
   Evaluation includes understanding the results, checking whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. Only approved models are retained, and the entire process is revisited to identify which alternative actions could have been taken to improve the results.

6. Using the discovered knowledge
   This final step consists of planning where and how to use the discovered knowledge. The application area in the current domain may be extended to other domains.

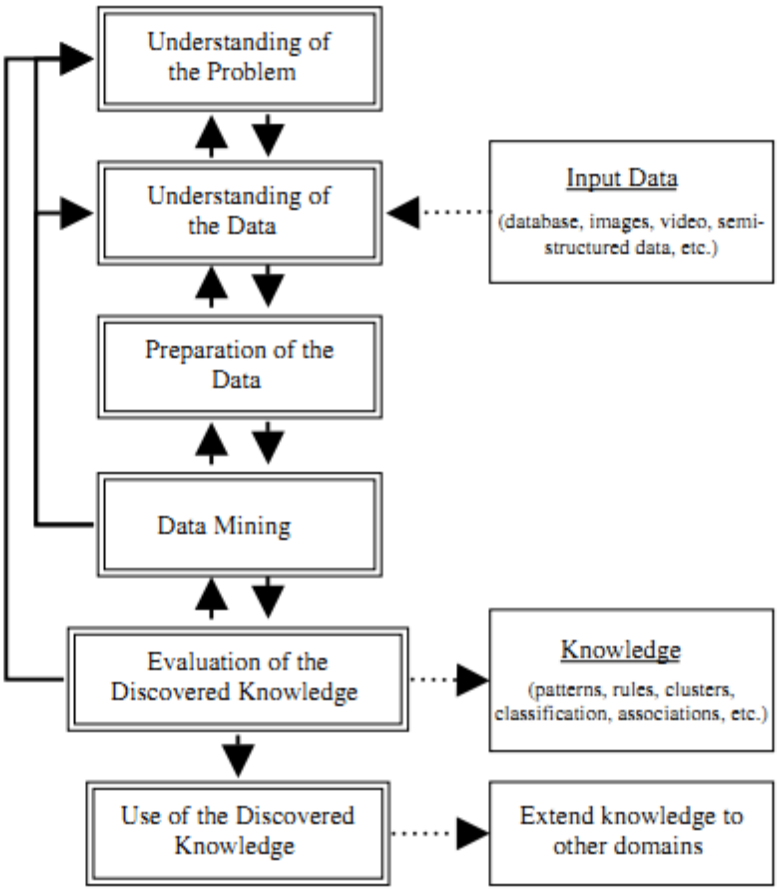

Figure 2.3 a hybrid of DM process model

2.3 Data mining tasks

The data mining methods, used for extracting hidden patterns from the data, are basically classified into the following two categories: description methods and prediction methods. Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks.

2.3.1 Predictive modeling

Predictive modeling permits the value of one variable to be predicted from the known values of other variables. Classification, regression, prediction etc. are some examples of predictive modeling. Many of the data mining applications are aimed to predict the future state of the data [48]. Prediction is the process of analyzing the current and past states of the attribute and prediction of its future state. Classification is a technique of mapping the target data to the predefine groups or classes. It is a supervised learning because the classes are predefined before the examination of the target data. The data mining stage of KDD process consists of two main types of data mining methods: verification oriented (the system verifies user's hypothesis) and discovery oriented (the system finds new rules and patterns autonomously) [61].

Classification

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree, rule induction, support vector machine, logistic regression, case-based reasoning and neural network-based classification algorithms [62]. The data classification process involves learning and classification.

In classification, test data are used to estimate the accuracy of the classification rules [42]. If the accuracy is acceptable, the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis.

A variety of techniques have been used to obtain good classifiers. Some of the more widely used and well known techniques that are used in data mining are used for this study is discussed below.

Decision tree

Data Mining uses machine-learning methods using decision trees to classify objects based on the dependent variable. Various decision tree algorithms such as CHAID (Chi-squared Automatic Interaction Detection), C4.5/5.0, CART (Classification and Regression Trees), J48 and any with less familiar acronyms, produce trees that differ from one another in the number of splits allowed at each level of the tree, how those splits are chosen when the tree is built, and how the tree growth is limited to prevent over-fitting [49].

Today's data mining software tools allow the user to choose among several splitting criteria and pruning rules, and to control parameters such as minimum node size and maximum tree depth allowing one to approximate any of these algorithms [48].

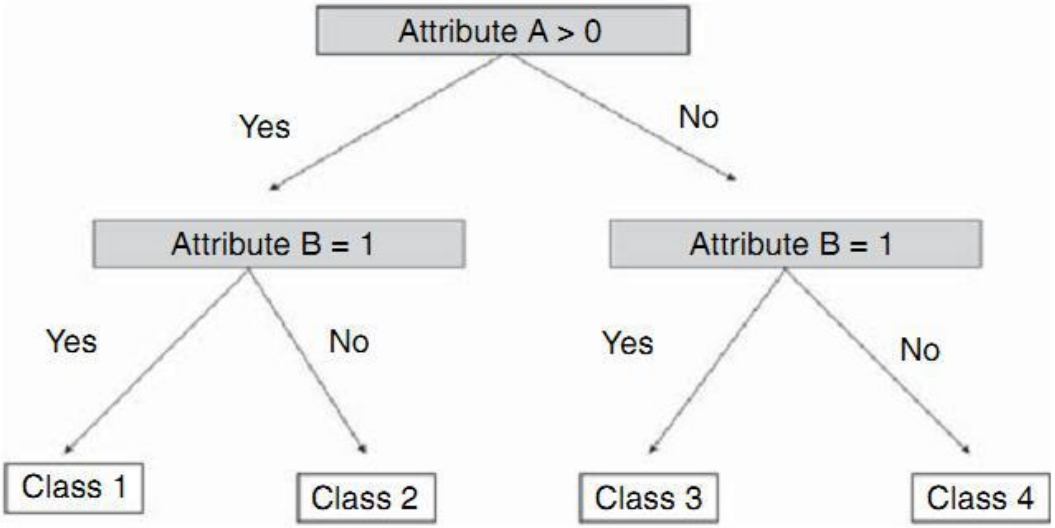

Figure 2.4 a simple decision tree

In decision tree induction, the nodes of the tree are labeled with attribute names, the edges of the tree are labeled with possible values for the attributes and the leaves of the tree generate decision tree from a given set of attribute-value tuples. An important quality of decision trees is that they handle non-numeric data very well.

Decision tree Algorithms

J48 decision tree algorithms adopt an approach in which decision tree models are constructed in a top-down recursive divide-and-conquer manner. Most algorithms for decision tree induction also follow such a top-down approach, which starts with a training set of tuples and their associated class labels [55].

Algorithm: Generate decision tree.
Input: Sets of training dataset (D), Attribute list ,Attribute selection method;
Output: A decision tree.
Procedures:(1) Create a node N; (2) If tuples in D are all of the same class, C then
(3) Return N as a leaf node labeled with the class C;
(4) If attribute list is empty then
(5) Return N as a leaf node labeled with the majority class in D; // majority voting
(6) Apply Attribute selection method (D, attribute list) to find the "best" splitting criterion;
(7) Label node N with splitting criterion;
(8) If splitting attribute is discrete-valued and multi way splits allowed then // not restricted to
binary trees
(9) Attribute list ← attribute list- splitting_ attribute; // remove splitting attribute
(10) For each outcome j of splitting criterion // partition the tuples and grow sub trees for each
partition
(11) Let Dj be the set of data tuples in D satisfying outcome j; // a partition
(12) If Dj is empty then (13) attach a leaf labeled with the majority class in D to node N;
(14) Else attach the node returned by Generate decision tree (Dj, attribute list) to node N; End for
(15) Return N;

Figure 2.5 Basic pseudo code for J48 decision tree algorithms

Rule Induction

Rule induction is one of the techniques most used to extract knowledge from data, since the representation of knowledge as if/then rules is very intuitive and easily understandable by problem-domain experts [63]. It is an area of machine learning in which formal rules are extracted from a set of observations.

if (attribute − 1; value − 1) and (attribute − 2; value − 2) and ……and (attribute − n; value − n)then (decision; value):

In supervised learning, all cases are pre classified by an expert [64]. In other words, the decision value is assigned by an expert to each case.

Rule induction on a database is a process undertaking using intelligent software where all possible patterns are systematically pulled out of the data. In this process the accuracy is added to them that tell the user how strong the pattern is and how likely it is to occur again [65].

Rule induction is also known as **Separate-And-Conquer method**. This method apply an iterative process consisting of first generating a rule that covers a subset of the training examples and then removing all examples covered by the rule from the training set. This process is repeated iteratively until there are no examples left to cover. The final rule set is the collection of the rules discovered at every iteration of the process [66]. Rule induction supports a number of algorithm two of them are discussed below

PART

PART is a separate-and-conquer rule. The algorithm producing sets of rules called 'decision lists' which are ordered set of rules. A new data is compared to each rule in the list in turn, and the item is assigned the category of the first matching rule a default is applied if no rule successfully matches [59].

Support Vector Machine

A support vector machine (SVM) is a concept in statistics and computer science for a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis. Some of the advantage of this algorithm are it has better empirical performance than other classification technique, the SVM algorithm is its accuracy due to its excellent learning performance, and the SVM algorithm scales well to relatively large datasets he SVM algorithm is flexible due in part to the robustness of the algorithm itself.

Sequential minimal optimization

SMO is widely used for training support vector machines. SMO is an iterative algorithm for solving the optimization problem by breaking the problem into a series of smallest possible sub-problems, which are then solved analytically.

2.3.2 Descriptive modeling

The descriptive model identifies the patterns or relationships in data and explores the properties of the data examined.

Clustering

Clustering can be said as identification of similar classes of objects. Clustering, also known as cluster analysis or unsupervised classification, is a general term to describe methodologies that are designed to find natural groupings or clusters based on measured or perceived similarities among the items in the clusters using a multidimensional dataset. There is no need to identify the groupings desired or the features that should be used to classify the dataset.

This is the technique of combining the transactions with similar behavior into one group, or the customers with same set of queries or transactions into one group. Classification approach can also be used as effective mean of distinguishing groups.

So clustering can be used as preprocessing approach for attribute subset selection and classification [41]. For Example: The customer of a given geographic location and of a particular job profile demand a particular set of services, like in banking sector the customers from the service class always demand for the policy which ensures more security as they are not intending to take risks, like-wise the same set of service class people in rural areas have a the preferences for some particular brands which may differ from their counterparts in urban areas [41].

This information helps the organization in cross-selling their products, Instead of mass pitching a certain "hot" product, the bank's customer service representatives can be equipped with customer profiles enriched by data mining that help them to identify which products and services are most relevant to callers.

This technique will help the management in finding the solution of 80/20 principle of marketing, which says: Twenty per cent of your customers will provide you with 80 per cent of your profits, then problem is to identify those 20 % and the techniques of clustering will help in achieving the same [39].

Association rule

Association and correlation is usually to find frequently used data items in the large dataset. It is the technique of finding patterns where one event is connected to another event. This type of findings help businesses to make certain decisions regarding pricing, selling and to design the strategies for marketing, such as catalogue design, cross marketing and customer shopping behavior analysis [42]. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little value [42].

2.4 Data mining tools

There are a lot of data mining tools, we can basically classify them as free in the market for academic purpose and the others are developed for business purpose.

2.4.1 Free data mining tools

One of the well-known free data mining software is called WEKA (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. WEKA is free software available under the GNU General Public License. The WEKA workbench contains a collection of visualization tools and

algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality [24].

The key features responsible for Weka's success are [68]:

- ✓ It provides many different algorithms for data mining and machine learning
- ✓ It is open source and freely available
- ✓ It is platform-independent
- ✓ It is easily useable by people who are not data mining specialists
- ✓ It provides flexible facilities for scripting experiments
- ✓ It has kept up-to-date, with new algorithms being added as they appear in the research literature.

2.5 Applications of Data Mining in Banking Sector

The banking industry across the world has undergone tremendous changes in the way of business conducted. With the recent implementation, greater acceptance and usage of electronic banking, the capturing of transactional data has become easier and simultaneously, the volume of such data has grown considerably. It is beyond human capability to analyses this huge amount of raw data and to effectively transform the data into useful knowledge for the organization. Data Mining can help by contributing in solving business problems by finding patterns, associations and correlations which are hidden in the business information stored in the data bases.

By using data mining to analyze patterns and trends, bank executives can predict, with increased accuracy, how customers reacts to adjustments in interest rates, which customers likely to accept new product offers, which customers is higher risk for defaulting on a loan, and how to make customer relationships more profitable [45].Some business areas that successfully apply data mining: [25]

There are various areas in which data mining can be used in financial sectors like customer segmentation and profitability, credit analysis, predicting payment default, marketing, fraudulent transactions, ranking investments, optimizing stock portfolios, cash management and forecasting operations, high risk loan applicants, most profitable Credit Card Customers and Cross Selling. [20].The main examples of applications of the data mining techniques in the banking industry are the following:

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.

***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. For example the examples of both 'good' and 'bad' loan applicant's histories can be used to develop a profile for a good and bad new loan applicant.

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. A score is thus produced that allows a lender to evaluate the customer and decide whether the person is a good candidate for a loan, or if there is a high risk of default. Customers who have been with the bank for longer periods of time, remained in good standing, and have higher salaries/wages, are more likely to receive a loan than a new customer who has no history with the bank, or who earns low salaries/wages. By knowing what the chances of default are for a customer, the bank is in a better position to reduce the risks [46].

Fraud Detection: One system that has been successful in detecting fraud is Falcon's called "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" [46]. As the above and many other literatures show, data mining is popular for the purpose of detecting fraud.

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 can able to retain existing customers by offering incentives that are individually tailored to each customer's needs [45].

2.6 Impact of fraud

The government of one country is committed to ensuring that the payment system is innovative, efficient and effective, and that they meet the needs of end users. It is known that Payment system sits at the heart of countries economy. Payment System May be defined as: a set of instruments, banking procedures and, typically, interbank funds transfer systems that ensure the circulation of money [27]. They are the mechanisms that allow money to flow continually among and between financial institutions, households and businesses. While dealing with payment system specially on

banking sector, securing the bank and its' customer from fraud is quite cumbersome task. But, the vulnerability of banks to fraud has been heightened by technological advancements in recent times. Fraud can be conceived of as a deception intended to achieve financial or personal gain. In practice, however, this description can take the form of a wide range of criminal behaviors including: counterfeiting currency, forgery, identity theft, credit card fraud, corporate misappropriation and so on [47].

To answer the question, "**how fraud is impacting the payment system?"** we need to see and refer researches as well reports done on area. The impact of frauds on entities like banks, which are engaged in financial activities, is more significant as their operations involve intermediation of funds. The economic cost of frauds can be huge in terms of likely disruption in the working of the markets, financial institutions, and the payment system. Besides, frauds can have a potentially debilitating effect on confidence in the banking system and may damage the integrity and stability of the economy. It can bring down banks, undermine the central bank's supervisory role and even create social unrest, discontent and political upheavals [35].

Statistics quoted in a recent report by the Association of Certified Fraud Examiners' (ACFE) 2012 titled "**Report to the Nation on Occupational Fraud and Abuse**" may have some Answer on the wideness of the case. The report has estimated that a typical organization loses 5% of its revenues to fraud each year and cumulative annual fraud loss globally during 2011 could have been of the order of more than $3.5 trillion. The ACFE report further mentions that as in the previous years, banking and financial services industry continues to be among the most commonly victimized sectors as far as fraud is concerned.

Other report published by "**Bureau of crime statistics and research"** consoled this concept as 'fraud is starting to accelerate'. The reason for this conclusion is, in the five years to December 2013(between 2008 and 2013) incidents of fraud grew on average 8.5 percent per year. In the 24 months to December 2013 (2013-2014), however, the increase was 13.2 per cent [47].

What the above reports and some other studies reveal is that the frequency, volume and the gravity of instances of fraud across various sectors, particularly in the financial sector, has gone up tremendously over the past few years. Broadly, the frauds reported by banks can be divided into three main sub-groups [35]. Technology related, KYC related (mainly in deposit accounts) and advances related.

A closer examination of the reported fraud cases has revealed that around 65% of the total fraud cases reported by banks were technology related frauds (covering frauds committed through at internet banking channel, ATMs and other alternate payment like that of cheque)

If the banks are not able to proactively manage the technology risks in their delivery systems, they may have to face litigations on customer protection and also incur the wrath of the regulators and customer interest groups [35]. Apart from enlisting active co-operation from their technology vendors, banks must look to build a close rapport with other banks, investigative agencies and regulators to ensure that there is prompt and coordinated exchange of information, whenever required [35]. With the spread of mobile banking, banks would also need to closely engage with the telecom service providers for reducing the technology related fraud risk. Banks could also consider seeking insurance coverage as a risk transfer tool and mitigate for the financial losses arising from technology induced fraudulent customer transactions [35].

If the existence of the problem conformed, we need a solution that minimize or eliminate the problem. According to literatures wrote on the area, we can understand that data mining is one of the best mechanisms for dealing with such a case. And, here is where, data mining play its part by using its powerful tools and techniques for the purpose of detecting fraud.

One 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 [45].

Data mining is becoming strategically important area for many business organizations including banking sector. It is a process of analyzing the data from various perspectives and summarizing it into valuable information. Data mining assists the banks to look for hidden pattern in a group and discover unknown relationship in the data [38].

Data Mining techniques can be of immense help to the banks and financial institutions in this arena for better targeting and acquiring new customers, fraud detection in real time, providing segment based products for better targeting the customers, analysis of the customers' purchase patterns over time for better retention and relationship, detection of emerging trends to take proactive approach in a highly competitive market adding a lot more value to existing products and services and launching of new product and service bundles [39].
Banking industry by itself provides individuals and business with a choice of payment methods that adequately serve their needs, and currently cheques are still one of the basic and prominent modes of payment throughout the world [27].

'**Bureau of crime statistics and research**' reported top five financial crimes made in Australian in December, 2014. According to this report cheque fraud belong to fives position as shown in the below table [47]. The aim of this study is to provide an understanding of the nature of fraud incidents recorded and to find out which fraud types are contributing to the increase in recorded fraud.

Number and average cost of fraud incidents summarized					
Fraud Description	2008- 2009	2012- 2013	Total	Average cost ($) per incident**1**	Estimated total cost ($) for year to Sep 2013**5**
Card fraud	219	237	456	940	16,207,377
Fuel drive-offs	186	210	396	62	928,340
Identity theft	31	32	63	8,317	19,811,952
Embezzlement / Theft by a person of trust	28	21	49	36,588	67,788,305
Cheque Fraud	21	23	44	12,413	20,651,398
Total for five most common offences	485	523	1008	3,290	125,387,372

Table 2.1 top five financial crimes made in Australian in December, 2014

The report finally concluded that Fraud is a growing problem and will likely continue to rise with new technologies and payment options. This study founded that the fraud types with the highest recorded prevalence are quite different to those which have the greatest overall cost implication [47].

According to the information on the table 2.1 we can understand how the problem is widening even in developed countries. It is also problem for non-developed countries too including Ethiopia but, consolidated report which take few years are not adapted yet. In my opinion a lot has to be done on the area in the future.

The concern of this study is detecting a fraudulent chuqhe payment transaction using the data mining tools and techniques. With this area somehow, a lot of researches has done in other foreign countries especially in U.S.A and Western Europe. But, the fact in Ethiopia is different only a procedure published by banks can be found but not a research, although this is the case, related or similar works like that of "loan risk analysis", "detecting fraudulent tax payment" and some other has been tried.

2.7 Studies on the area

Most of the researches conducted in this area fall under one of the two categories either they use benefit of data mining techniques and to get a hidden knowledge behind or uses expert on the field to analyze causes and effects of cheque fraud to find a solution. Referring the two pillories is going to get the research in a better position to understanding most of the situation about cheque fraud and data mining. For this purpose, some literatures are discussed below under both categories.

2.7.1 Expert Studies on cheque fraud

Cheque Fraud Working Group is an international organization suited in U.S.A, one of their publication called cheque fraud "**a guide to avoiding losses**" describe as follow.
The publication starts by indicating cheque is an important way of payment system but exposed for fraud. Cheque fraud is one of the largest challenges facing financial institutions. Technology has made it increasingly easy for criminals, either independently or in organized manner, to create increasingly realistic counterfeit and fictitious cheque's as well as false identification that can be used to defraud financial institutions [29].

A statistical information show the problem is wide even in developed countries. A study conducted U.S.A on cheque fraud indicated that about billions of dollars is being lost annually [29]. Although this study is for a purpose of general guide, financial institutions can, however, get details on bankers, tellers, operations personnel, and security officers to think about the problem and show how they can help protect their institutions from cheque fraud.

To protect the banking industry and its customers from cheque fraud, financial institutions must become familiar with common cheque fraud schemes. This booklet describes some of those schemes and presents tactics for use in combating cheque fraud. It cannot describe comprehensively all types of cheque fraud or cheque fraud schemes, because the variations are limitless.

As this study indicated, fraud schemes involving cheques take many forms. Cheque's may be [29]:
• Altered, either as to the payee or the amount.
• Counterfeited.
• Forged, either as to signature or endorsement.
Check fraud criminals may be financial institution insiders, independent operators, or organized gangs. The methods they use to further cheque fraud include getting customer information from financial institution insiders or Stealing financial institution statements and cheque's or even working with dishonest employees of merchants who accept payments by cheque.

This study tried to get detailed information with different scenarios, Descriptions of some common cheque fraud schemes, with information on what makes them successful, and what bankers can to do avoid them are discussed.

Altered Cheque

Altered cheques are a common fraud that occurs after a legitimate maker creates a valid cheque to pay a debt. A criminal then takes the good cheque and uses chemicals or other means to erase the amount or the name of the payee, so that new information can be entered. The new information can be added by typewriter, in handwriting, or with a laser printer or cheque imprinter, whichever

seems most appropriate to the cheque. Scenarios for fraud using altered cheque include the following [29].

Scenario 1
A daily laborer works for a day and the employer agreed with 69.99 birr. The employer pays by cheque, writing 69.99 birr to the far right on the line for the amount in figures, and the words "sixty-nine" to the far right of the amount in the text line. The criminal uses the blank spaces on both lines to alter the cheque by adding "9" before the numbers line, and the words "Nine Hundred" before the text line. The 69.99 birr cheque is now a fraudulent cheque for 969.99 birr, which the criminal cash.

Scenario 2
A small company that provides service to several small clients is paid by cheques payable to "Daniel CO." or "Daniel Company." Criminals steal a number of those payment cheques and use a chemical solution to erase the word Co. or Company. They then type in the word Assefa, and subsequently cash the cheques using false identification.

Scenario 3
A criminal steals a wallet, with a cheque in it, from the glove compartment of a car. The criminal uses the signatures on the identification in the wallet to forge the endorsement. Then, using the identification in the wallet, altered if necessary, the criminal cash the cheque at the payee's financial institution.

Although, there may be some more ways to altering cheque schemes can be successful however, the main reason is when customers are careless and financial institutions fail to cheque payee identification properly.

Counterfeit Checks

Counterfeit cheques are presented based on fraudulent identification or are false cheques drawn on valid accounts. Scenarios for fraud using Counterfeit Cheque include the following [29].

Scenario 1
A group of criminals open cheque accounts and cash counterfeit cheques using fraudulent drivers' licenses, kebele's identification card, and other identification. They use information from personal and corporate trash to produce the identification with computer technology.

Scenario 2
A financial institution insider identifies corporate accounts that maintain large balances, steals genuine corporate cheques, counterfeits them, and returns the valid cheques to the financial institution. The financial institution insider may associate with a group of criminals that distributed the counterfeit cheques throughout the area and cashes those using fictitious accounts.

Counterfeit cheque schemes can be successful when criminals are skillful in their use of technology to create false documents or have access to information and supplies from financial institution insider.

Identity Assumption

Identity assumption in cheque fraud occurs when criminals learn information about a financial institution customer, such as name, address, financial institution account number, kebele's identification card, home and work telephone numbers, or employer, and use the information to misrepresent themselves as the valid financial institution customer. These schemes may involve changing account information, creating fictitious transactions between unsuspecting parties, or preparing cheques drawn on the valid account that are presented using false identification [29].

Scenario 1

A financial institution customer pays a bill in the normal course of business. An employee of the payee copies the cheque and provides it to a partner in crime who contacts the financial institution and, using information from the cheque, pretends to be the account holder. The criminal tells the financial institution that he or she has moved and needs new cheques sent to the new address quickly. When the financial institution complies, the forged cheques are written against the customer's account.

Scenario 2

A gang member steals a statement for an account at financial institution A and another steals a box of new cheques for a different person's account at financial institution B. The gang pre pares the stolen cheques to be payable to the valid account at financial institution A. Using fraudulent identification, one of the criminals poses as the payee to cash the cheques at drive through windows at financial institution A. Because the criminals know that sufficient cash exists in the account to cover the cheque, they can ask safely for immediate cash.

Scenario 3

A criminal uses customer information, sometimes from a financial institution insider, to order cheques from a cheque printer or to create counterfeit cheques and false identification. The criminal then writes fraudulent cheques and presents them for deposit into the customer's account, requesting part of the deposit back in cash. The cash-out from the transaction represents the proceeds of the crime. This is also known as a split-deposit scheme.

Identity assumption schemes can be successful when a financial institution:

Accepts account changes over the telephone. Or, Careless in requiring and reviewing identification presented for cash-out transactions.

2.7.2 Fraud by Bank Insiders

Often cheque fraud schemes depend on information provided by bank insiders. In addition to schemes discussed elsewhere, which may involve access to information about one account or relationship, frauds based on insider knowledge are often broader because they are based on the knowledge of the bank's operations and access to many accounts [17].
Fraud by insiders can be successful when customer account information is not kept secure.

The paper finally put a solution for minimizing the current problem of cheque fraud from both customer of the bank and insiders, in addition it highly recommended that effective training and education are important in preventing cheque fraud losses.

Method for Facilitating Cheque Clearance Using Smart Phones Application

Another research conducted in the area, "**An Alternative Method for Facilitating Cheque Clearance Using Smart Phones Application**", has emphasis on minimizing the time taken by cheque clearing process [17]. As the authors said, cheques are chequeed and submitted to their corresponding clearing banks for clearance. After verification the clearing banks send signed instructions back to the bank stating whether the cheque is cleared or rejected. Then the money is transferred from the payee a/c to the receiver's a/c. This whole process takes up to 3 working days just for inter-bank national cheques [17].

Therefore there is a need to come up with an alternative method to shorten the lengthy process of depositing physical copy of the cheques in banks and its clearance making the entire cheque clearing process much simpler and speedier. This application allows users to deposit cheques without any need of depositing the physical cheques to banks.

As this paper show, the application uses Photo Cheque Deposit (PCD) which verifies a set of values (e.g. cheque number, payer's account number, date of cheque, cheque amount, payee's name and account number, payer's signature along with authentication that the cheque itself is not a counterfeit etc.). The authors also believe the system ensures that like paper cheques and photo cheques can't be counterfeited or altered.

On requisition for paper cheques by an account holder, a set of physical cheques components consisting of unique cheque number, account number of payer, CTS code and logo of the Bank are issued. These parameters help to identify the cheque. However any system can generate these details using false or incorrect values which automatic cheque processing system or ACPS (which uses Magnetic Ink Character Recognition (MICR)) helps to catch and reject the cheque outright. Cheques may also originate from fraud sources doing serious damages to individuals, companies and financial institution if mistakenly cleared.

This research brings a new and faster way of clearing cheque using just smart phone. The processing time of cheque deposit using smart phone is less when compared to physical cheques as the verification of e-cheque details and the signature are done electronically , there is no physical transfer of cheque from place to place and then to bank [17].

From the smart phone side, Photo Cheque Deposit (PCD) is likely to be a disruptive technology and driving force in consumer's adoption of mobile financial services [17].

In the study, a prototype is developed. The proposed system is initiated by proper login of the consumer; the system validates and authorize the consumer. The consumer then deposits its cheque for clearance indicating the account number to which the money is to be deposited and amount of the cheque. The system notify the status of cheque using a unique reference number provided after the successful process of cheque deposit.

The researchers consider a scenario where a payer issues a cheque to a payee who is a customer of another bank. There are four role model namely payer, payee, cheque submitting bank and cheque clearing bank. Here the payer issues a cheque to a payee. The payee submits that cheque using his smartphone in his servicing bank 'A', who submits that cheque for clearance to payer's servicing bank, B. If the cheque is cleared, funds are transferred from payer's account to payee's account.

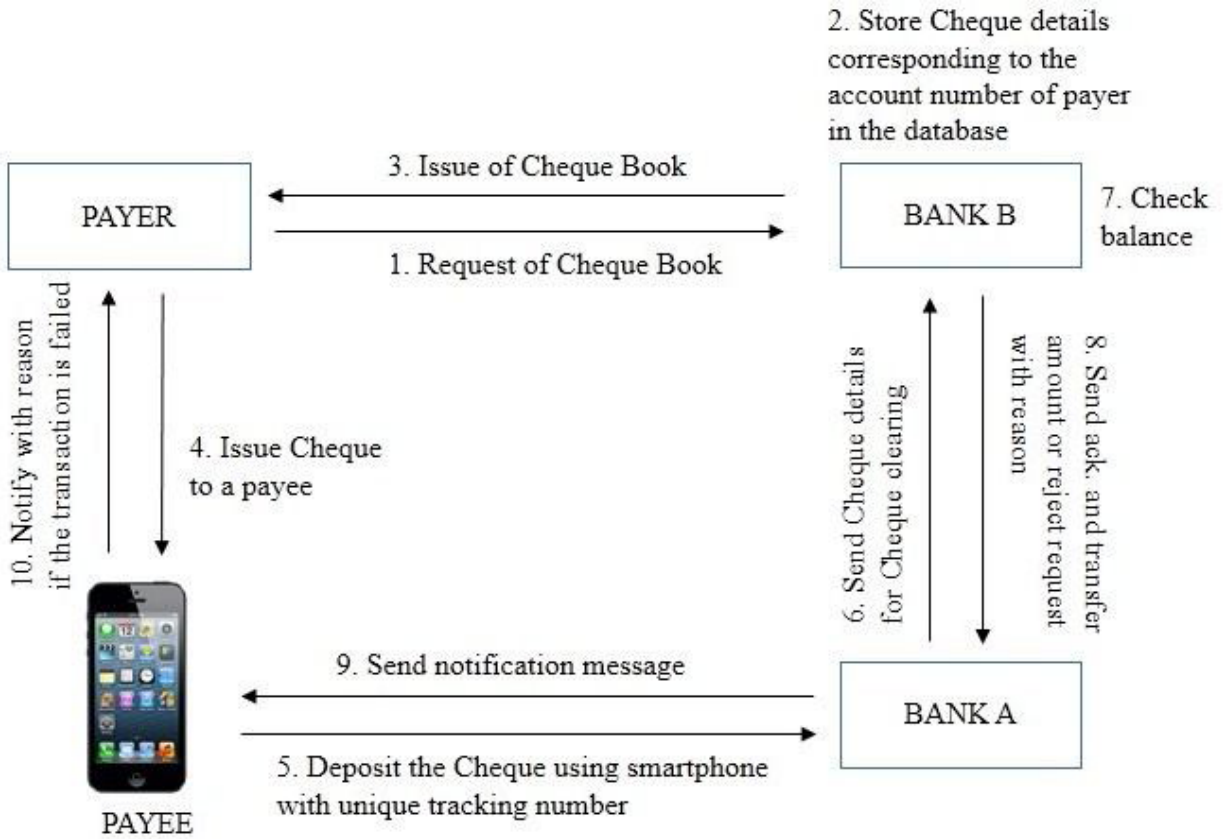

Figure 2.7 Mobile cheque clearing system [17]

Benefits of this research is, to minimize the time taken by manual work of clearing cheque, avoiding duplication of cheque and reduce fraud risk can be mentioned primary.

Duplication Cheque Detection: When cheques are submitted to a bank teller, a paper cheque is easily removed from circulation. However, a cheque scanned by a smart phone device might not be taken out of circulation after conversion to a cheque image. The same paper cheque can be accidentally or intentionally submitted for deposit multiple times. Therefore, a robust duplicate detection mechanism for cheques deposited through all customer channels are needed for fraud prevention and cost containment.

Fraud Risk: Certain aspects of fraud can be elevated in photo cheque deposit environment for both consumer and corporate customers. One such fraud risk is the presentment of a counterfeit or altered item, which can be more difficult to detect as a scanned image [9]. Many of the cheque security features, such as watermarking and micro-printing in the signature line, can be eluded when the cheque is clicked [9].

Electronic cheque clearing system

A Rule Book issued by the (Nepal Rastra Bank) NRB has published a document as a guide line for **electronic cheque clearing system**, which is discussed as below.

ECC uses Magnetic Ink Character Recognition MICR Line is the information printed on a designated space (bottom) of the cheque that is read automatically at the scanning stage.

ECC relies on transferring the electronic data and captured cheque images, rather than moving the physical cheques.The Operating Rules can help also the Members to perform their daily cheque clearing operation using electronic cheque clearing (ECC) system more safely and efficiently [36].

Electronic Cheque Clearing System do more than simply fasting the daily cheque transaction but it rather an interbank (which works across the bank) cheque clearing solution [36].

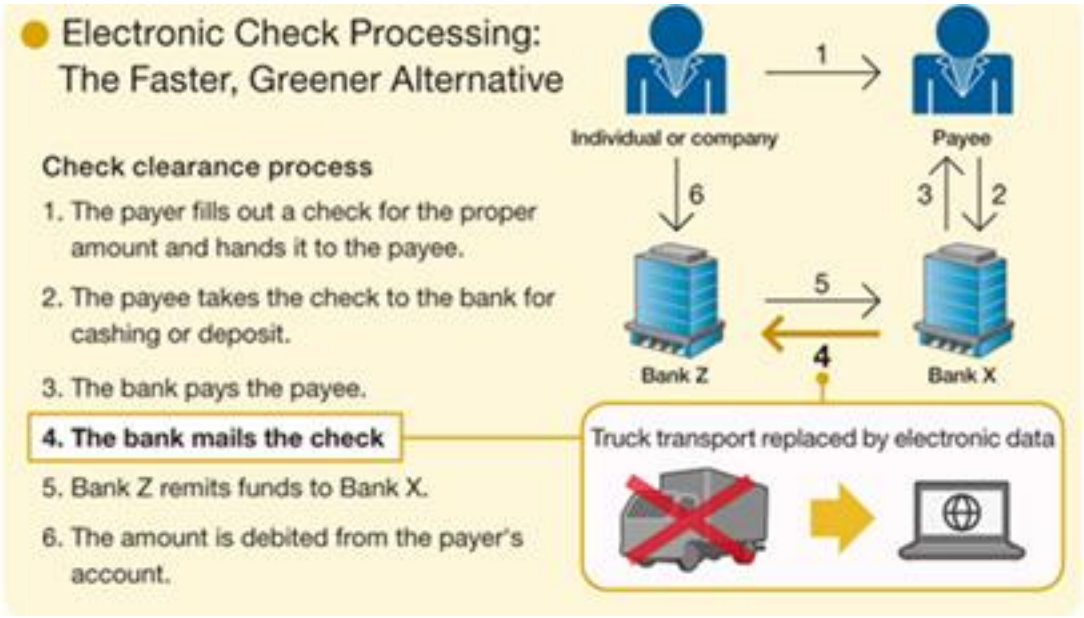

Figure 2.8 electronic cheque processing

Figure 2.8 shows a how the time took for chques clearing process can be significantly reduced just by replacing the physical exchange of the cheque between banks with electronic mails.

2.7.3 Applying data mining for cheque fraud analysis

While dealing with banks, the customers and the banks have the chances of falling on an easy prey to the frauds. So both the parties wish to be secure while dealing with each other. The data mining techniques can help them to detect and hence prevent frauds. The data mining techniques will help the organization to focus on the ways and means of analyzing the customer data in order to identify the patterns that can lead to frauds [42].

Study of Data Mining Techniques used for Financial Data Analysis

A research conducted titled on '**Study of Data Mining Techniques used for Financial Data Analysis**' which study about loan default risk analysis, Type of scoring and different data mining techniques like Bayes classification, Decision Tree, Boosting, Bagging, Random forest algorithm and other techniques.

As the study show, Loan default risk assessment is one of the crucial issues which financial institutions, particularly banks are faced and determining the effective variables is one of the critical parts in this type of studies. In this Credit scoring is a widely used technique that helps financial institutions evaluates the likelihood for a credit applicant to default on the financial obligation and decide whether to grant credit or not [23].

As the author said credit scoring models are known as statistical models which have been widely used to predict the default risk of individuals or companies. These are multivariate models which

use the main economic and financial indicators of a company or individuals' characteristic such as age, income and marital status as input, assign them a weight which reflect its relative importance in predicting default.

The result is an index of credit worthiness that is expressed as a numerical score, which measures the borrower's probability of default. The goal of a credit scoring model is to classify credit applicants into two classes: the "good credit" class that is liable to reimburse the financial obligation and the "bad credit" class that should be denied credit due to the high probability of defaulting on the financial obligation [23].

The classification is contingent on characteristics of the borrower (such as age, education level, occupation, marital status and income), the repayment performance on previous loans and the type of loan [23]. The study has tried to categorize scoring as below.

1) Application Scoring: This kind of scoring consists of estimation credit worthiness of a new applicant who applies for credit. It estimates financial risk with respect to social, demographic financial conditions of a new applicant to decide whether to grand credit to them or not.
2) Behavioral Scoring: It is similar to application scoring with difference that it involves existing customers, so the lender has some evidences about borrower's behavior to dynamic management of portfolio process.
3) Collection Scoring: Collection scoring classifies customers into different groups according to their levels of insolvency. In the other words, it separates the customers who need more decisive actions from those who do not require to be attended to immediately. These models are used in order to management of delinquent customers from the first signs of delinquency.
4) Fraud Detection: It categorizes the applicants according to the probability that an applicant be guilty.
5) Finally, this paper explains different types of data mining techniques which are suitable for financial institute with their respective advantage.

Image Data Mining of Cheque Transactions In Support of Customer Relationship Management at Banks

The other research conducted in U.S.A under the title of **"Image Data Mining of Cheque Transactions In Support of Customer Relationship Management at Banks"** for the purpose of retrieving customer information automatically at the time a user makes transaction at the bank using the cheque image [26]. It uses technology of image data mining for extracting data from cheque transactions to extract various fields which are explained below.

These fields and the information they convey about the bank customer could be summarized as follows:

- Payer field: Who is making payments to my customer?
- Payee field: Whom is my customer paying (business, or individual)?

- Date filed: When and how frequent are certain payments made by the customer?
- Memo field: What is the reason for the payment?
- Courtesy amount field: The payment amount in numerals.
- Legal amount field: The payment amount in words.
- Payee Signature: A security feature.
- The Code Line: The branch and account number of the payer at his/her home bank.

As of this paper witnessed 81 percent of U.S people are user of cheque account, and the author thinks this is good opportunity for real time identification of customer using their cheque image which also minimizes fraud as the same time, since, the bank can treated suspicious customers differently. We are witnessing that there is a rapid transition from a paper-based economy to a digital-based one. As this transition progresses, it becomes increasingly easier for businesses to capture customer-related information and use it to support a more sophisticated approach to direct marketing and customer relationship management [26].

The author further show the need for this research, since, Banks are faced on a daily basis with a huge number of cheques that must be processed and reconciled as soon as possible in conformity with state and/or federal regulations. This situation has motivated many banks to invest in automated cheque processing systems based on imaging in order to make their cheque processing operations more efficient [26].

A major application of image processing at banks is the automated reading of financial document amounts. This application relies mainly on the technologies of image capture, image processing & document analysis and pattern recognition. The primary motivating factors driving this technology is the demand for reduced labor costs and increased processing speeds for cheque processing operations.

A customer must first obtain a grayscale image of the paper document then convert the grayscale image to a binary (black and white) image through a binarization algorithm that suppresses the background noise (e.g. graphics) and keeps the important image pixels representing actual check transaction information [26].

Finally, the cheque image data mining system described in the previous section is envisioned to work against cheque images stored in an image archive or in a multimedia data warehouse (Integrating Cheque Image Data Mining with the Data Warehouse).
The paper also mentioned that the major challenge was the missing values of customer attribute at the time of using the collected data.

Application of clustering techniques and genetic algorithm in making decision trees for classification of banks customers

Khan Babayee [44] conducted a research titled: "**application of clustering techniques and genetic algorithm in making decision trees for classification of banks customers**". They used banks customers clustering for validation of banks customers and granting credit loans to them. He analyzed data obtained from one thousand bank transactions made by bank customers including cheque's status, time period, credit background, credit value, saving status, work experience, number of installments, personal status, age, gender, residence area, assets and properties, present credit status, job, number of children and so on by means of K-means and Two-steps clustering techniques and genetic algorithm and specified a routine for identification of valuable customers for granting bank loans [44].

2.7.4 Locally Related works

Fraud occurs in the following areas: Credit card fraud, Internet transaction fraud / E-Cash fraud, Insurance fraud and health care, Money laundering, Intrusion into computers or computer networks, Telecommunication fraud, Voice Over IP (VOIP) fraud, and Subscription fraud/Identity theft. Fraud detection is concerned with the detection of fraud cases from logged data of system and user behavior [69]

In Ethiopia, there are works done to assess the application of DM in different sectors like Banks, Airlines, Insurance, HealthCare, and Custom Authority. Since this study focused on banking sector, the few following paragraphs has tried to assess the works done in the sector.

A study conducted by 'Tariku' initiated with the aim of exploring the potential applicability of the data mining technology in developing model that can detect and predict fraud suspicious in insurance claims with a particular emphasis to African Insurance Company, employed clustering algorithm to identify the natural classes and then applied classification techniques[70].

A study conducted by 'Lule' "the role of data mining technology in electronic transaction expansion at Dashen bank S.C" has used KRISP-DM as a data mining technology and employed Artificial neural network for classification after using K-mean algorithm for getting the natural grouping of the classes [72].

Other study by 'Tilahun' has tried to assess the possible application of DM techniques to target potential VISA card users in direct marketing at Dashen Bank As the study said, the model built was very encouraging [71].

CHAPTER THREE

DATA PREPARATION

Through this chapter, the researcher tried to follow the first three steps of hybrid data mining model, which are problem domain understanding, data understanding and data mining. The other steps covered on the next chapter of this paper. Each details on dataset and data pre-processing are discussed.

3.1 Understanding of the problem domain

This step need iterative work with domain experts. As discussed in chapter two, cheque is one of the very basic service given by banks to their customers. This mean the domain experts are employee of the bank who specifically doing on cheque related works and faces different scenarios. They have procedure to follow derive by the bank [9]. So, communicating with those experienced staff, reading procedures, manuals and magazines which is written specifically on cheque is being used for understanding and defining the problem. Although the procedures and manuals do not mention the problem, magazines (NOV 12,2016) and discussion with domain experts helps a lot.

As the researcher tried to put cheque clearing process in chapter one, statement of the problem section, a cheque user can write a certain amount of cash for other person(payee) in different bank, but the payee will not able to immediately receive the specified amount due to clearing process which currently conducted by National Bank of Ethiopia. NBE is a trusted government financial institution having the role of cheque clearing process (generally uses debit and credit concept) being between different banks.

Under normal circumstance, customers who need to use cheque services, must first come at one of the CBE's branch to open current account, if that customer has fulfilled the requirements, a current account will be opened for him/her based on the their information. Then, he/she will be a user of cheque service. A scenario of fraud attempt by the customer may change the case, if a customer misuse the cheque for theft purpose, he/she will be in charge of what he does legally plus his/her account will be blocked (will be in the list of malicious account) for a certain period in any bank of Ethiopia.

It is known that a customer is a king for a given bank, and a bank has responsibility to make safe environment for their customers. One way of security is protecting customers from bad ones. In the case of cheque, a list is being kept in branches to identify the misbehaved customer, specifically

for those who misuse the cheque. For instance, if a given customer wrote a cheque without having enough balance in his account, his/her action will be recorded as his/her first record in his/her own branch. If the same customer commit similar trick again, this will be his/her final warning and second record of fraud attempt. Third time he/she found being guilty, his/her name will be sent to the NBE. So that he/she will no longer be a customer to any bank as a penalty for unspecified period of time.

As discussed early in this paper, the main aim of this research is to analyze and study the behavior of customers who make fraud using cheque and building a better model which capable of identifying them before they made any fraud on the bank and customer of the bank.

3.2 Understanding of data

As discussed in chapter one, the data for this research is collected from two different databases. Customer and account tables belong in CBE database but list of malicious account collected from NBE, but this list will be updated on CBE and all other banks' database as well. This is a kind of network, that enforce a bank/branch should not open any type of account for an individual or organization whose name is on the list.

Customer and account table has direct relationship and malicious cheque account has no direct relation with either of them. This means there is a need to differentiate the malicious cheque account which belongs inside of current account. The researcher believes that discussing about the tables and their respective attributes is very helpful for the reader.

Customer:
A Customer shall be defined as any legal entity individual or corporate with whom the bank agrees to conduct business. Detailed customer information is basic in business dealings. A customer can be current and saving account holders, correspondent banks, borrowers, guarantors and the like. A customer can be either prospect, individual or corporate [9].

Current account:
Current account is one of the prominent type of account which is opened for the purpose of cheque use. It has its own criteria for opening and managing it, comparing to other type of accounts. For instance an individual customer need to have at least five hundred birr and for corporate 1000 birr in addition to the basic information for opening cheque account.

Malicious cheque account:
These are current accounts which are closed due to fraudulent act by the holder of the account to penalize the customer. Which is kept and distributed by NBE to protect the customer from opening any other account in other bank or branch. This is made possible by combining and studying the

behavior of customer from two different database a customer table (which contain the detail of every customer) and the malicious customer table (who made fraud using cheque).

Having the above data, SQL database tool is used for combining this two different tables(current account with that of malicious account), on both of the instance, a name field is exist so that, using the SQL command, common full name is retrieved after cleaning the data. Then a special tool for bulk data called WEKA has been used for analyzing and building a model. WEKA provide different option for manipulating dataset. Classification has chosen by the researcher since the data contain pre-defined class of malicious customers.

3.3 Data preparation

A good result after applying data mining depends on an appropriate data preparation in the beginning. Important elements of the provided data have to be detected and filtered out. These kinds of things are settled in the Pre-processing phase [52]. Data preparation is a critical step on data analysis part. And it uses a number of sub-steps for getting a refined dataset. Data preparation may include data selection, data cleaning, data construction, data integration, data formatting, and attribute selection [52]. On the next few section the researcher show how the data was preprocessed following the above procedures.

3.3.1 Source data selection

A reliable data source is very important for the good result of any research, for this research the data has been taken from CBE tables reside on oracle database. The malicious customers are collected from NBE has a total number of 1,500 customer. From 1,500 customers about 1208 customers found in CBE's customer. Currently CBE has more than 15,000 current account. Notice that, the malicious customers belong in current account table but only in the history file, since their account is closed as soon as they found to be cheque fraudulent.

Data collected type	Primary Data Source	Number of data collected
Customer who has Current account	CBE	15,215
Malicious customers	NBE	1,501

Table 3.1 initial data collected

Although most of it is used for cheque purpose, the collected dataset also contain irrelevant and unnecessary data. For instance, attributes having more than 80% of their value null, are considered as irrelevant for this study (those attributes are non-mandatory at the time of registration, so that a user may let them null). Therefore all the data were not used as a dataset. A certain procedure has been followed while clearing the data, which will be presented later.

3.3.2 Data cleaning

Data cleaning is the other step which need very careful attention, for the purpose of cleaning the data the researcher use EXCEL and ACCESS after retrieving the data from ORACLE database. As we discussed previously all the necessary data for this research were collected from two different tables so, removing incomplete and missing attributes and values were applied on both current account table and malicious customer. There are also attributes which has no relation with the current problem at hand for example we can mention Account_number of a customer. And some other are too personal information which are not included in the dataset Customer_Name can be an example for this. On the other hand different attributes may have the same meaning. This features are discussed broadly on attribute selection part.

As discussed in the above section, the study need to investigate two tables in depth in order to clearly understand what data is used for this research. They are tried to be discuss in the below section.

3.3.2.1 Data cleaning for customer table

The final amount of data the researcher got is 6,856. This is because although there are about 15,215 customer who uses current account, only 7,415 found to a cheque user. The other 7,800 users are not yet cheque users, so they are not relevant for this study. On the other hand, there were about 461 missing and outlier values which cannot be replaced using missing value handling mechanism since three and more than three values in a given raw are missing. Finally 98 duplicated names were found since the researcher used name filed to join the normal(loyal) customer with that of malicious customer, this values need to be removed for the purpose of one to one relationship so that, the final dataset left with 6856 records.

On the other hand, Customer table has more than 15 fields which supposed to be filled by customer at the time of registration. But only some are mandatory fields (must be filled by customer) and the researcher also use those fields considering that using attributes with too many missing values may lead to wrong conclusion. Although a customer has more than 15 attributes the mandatory attributes are; Account.ID, Customer.ID, Account Title, Working Balance, Branch, Sector, Industry, Cheque Issued, Address, Gender, Marital Status and Owner.

- ✓ **Account ID**, **customer ID** and **Account title** (name of the customer) are not included in the dataset since this attributes are more personal and not helpful for the problem at hand.
- ✓ Gender and **owner** (customer title) express similar things (sex) in different way for instance if the customer is “male” on customer title it could be “Mr” customer x. so gender is chosen since it is more clear.
- ✓ Similarly sector and **industry** has almost the same meaning and the experts has mentioned they may use interchangeably when inputting a record.

- ✓ **Customer address** is the other attribute which has too many missing values (more than 50% of the whole data) on it and the values are not properly filled even for the ones which has a value on it and the formats are not the same.
- ✓ **Cheque issued** is attribute to make sure if the customer is using cheque service or not. In this paper only yes values are used since the study is on cheque fraud.
- ✓ Finally the chosen attributes only from this table are: SECTOR, GENDER, MARITAL_STATUS, CUSTOMER_STATUS, TARGET and BALANCE.

3.3.2.2 Data cleaning for malicious customer

This paragraph tried to show how the classifier data (which classify cheque fraudulent customer) is being pre-processed. Although we had about 1501 malicious customer of data at the beginning, about 293 of the customers are found to be other bank customers (not belong in CBE). And five of customer were duplicated (who has the same information). And finally out of 1,203 customers, 1154 can be found on customer table while combining this two different tables (current account & malicious account) using SQL query by customer name. This means about 47 records from the malicious customer can't be joined to the current account table. Accourding to the experts' explanation, this customers are not cheque users or current account holders, instead their record may be suspected by other type of fraud such as tread service.

On the other hand, malicious account table had 9 attributes malicious.ID, customer full name, blocked date, reporting date, bank, branch, reference id, tin no, and address in which all the information is kept primarily by NBE.

- ✓ From this table more than one thousand branches are grouped in to 15 districts for easier out-put. From Bank attribute only CBE customers are chosen, since the researcher has access to CBE customer data only.
- ✓ The other attribute which is taken from this table is a class called "fraudulent cheque customer". Notice this data also kept by CBE.
- ✓ Currently more than one thousand and five hundred malicious customer are kept in CBE and 1154 have been found on the customer table which will be used as a class for fraudulent.

As the researcher tried to put the figures, the total no of dataset has about 6,856 record. From which 5,702 are normal customer and the remaining 1,154 customers are found to be fraudulent by CBE. And notice that the combination of this two tables are made possible with the help of SQL server.

3.3.2.3 Missing values and outlier handling

The data collected is relatively in a good position since the selected attributes are mandatory (enforce the inputter (data encoder) to fill a value on) at the time of customer registration. For instance, CUSTOMER STATUS, BALANCE, DISTRICT, LOCATED IN ADDIS and IS

FRAUDULENT are free from any missing value and outliers. But other attributes like that of SECTOR, MARTIAL STUTS, TARGET and GENDER has some.

- ✓ "GENDER" has some missing values and it is handled manually using the customer name. If the owner name is male the missing value will be replaced with male the same as the female and corporate.
- ✓ "MARITAL STATUS" has less than 0.5% data of missing values, and the researcher and the domain experts find it difficult to guess the value. So that the missing value are handed as "Other"
- ✓ "TARGET" and "SECTOR" has not more than 0.5% of missing values and handled the same way with "Other"

3.3.3 Data constriction

Data construction is another important step in data preparation. In this stage new fields were derived from the existing ones. On this part of data preparation, the researcher has changed/drive the following two attributes.

- ✓ Currently CBE has more than 950 branches, and this attribute has to be re-grouped to more general to reduce the values, based on this concept 951 branches changed to their respective 15 districts. A list of districts are described on table 3.11
- ✓ Similarly, based on the discussion with domain experts one additional filed called "is located in Addis Ababa" is derived from the districts.

3.3.4 Data Integration

This part of data preprocessing is the one that we integrate our data from different sources if we have any. As we discussed in the previous section the data are collected from two different databases. Customer from CBE oracle database and malicious customer from NBE excel.

- ✓ A new class attribute called "IS FRAUDULENT" is derived integrating the two tables, namely customer and malicious customer, in which the commonly found customers from both tables are considered to be fraudulent (1154 records) and the other which do not belong in the malicious customer are non-fraudulent (5,702 records).

3.3.5 Data Transformation

Data transformation is the data are transformed or consolidated into forms appropriate for mining. Discretization is the process of transforming continuous space valued series x = [x1, x1, x….xn} into a discrete valued series Y = {y1, y2…..yn} [53]. Major discretization methods which are used can be majorly categorized as unsupervised and supervised discretization.

This paper implemented unsupervised discretization method called binning. Binning method is implemented due to the fact that it is the simplest methods to discretize a continuous-valued attributed by creating a specified number of bins [52].

In our case customer balance is continuous value which need to discretize. Before using binning we have one fact that is a minimum balance that a customer need to save in his account at the time of cheque is 500 birr for PLC and 1000 birr for corporation [9]. Based on the discussion with domain experts, any cash less than 500 birr can be taken as "lower balance". Then they mentioned that it will be preferable to use two other values other than lower balance. Which is medium and higher balance, and for this purpose binning is chosen.

Equal width Discretization is a simples discretization methods that divides the range of observed values for a feature into k equal sized bin [54]. The process involves finding values as the minimum (Min) and maximum (Max). The interval is computed by dividing the range of observed values for the variable into k number of equally sized bins using the formula, Int= (max-min/k).
Max balance we have is 5,831,678 birr
Min balance we have is 501 birr
Our k is 2 (Since we planned to divide it in two categories)
We can calculate the average balance and higher balance using binning concept.

- ✓ Bin one (lower_balance) 0-500 birr
- ✓ Bin two (normal_balance) 501 to 2,915,588 birr
- ✓ Bin three (higher_balance) 2,915,588 birr to 5,831,678 birr

3.3.7 Attribute selection and ranking

There are many reasons that make the number of attributes to have a significant decrease from the original collection. Some of the reasons are, it uses to speed up the learning process, it makes simple to understand the generated rules and a better accuracy of the predictive model for the problem at hand. Due to the above aims, the below attribute attributes are made to be important for understanding cheque fraudulent customer behavior.

As shown in the table below, the researcher being with domain experts, the below 9 attributes were selected, from which one more attribute is added based on the discussion with domain experts, according to their explanation most of the branch belong in Addis Ababa are fraud suspicious.
So having the selected eight independent attributes (SECTOR, GENDER, MARITAL_STATUS, CUSTOMER_STATUS, TARGET, BALANCE, DISTRICT, LOCATION_IN_ADDIS) and one dependent attribute (FRAUDULENT) will be discussed in the next couple of sections.

No	Name	Data type	Description
1	SECTOR	nominal	A sector, a customer works on
2	DISTRICT	nominal	Is a customer's branch located in Addis
3	GENDER	nominal	Gender of the customer
4	MARITAL_STATUS	nominal	Marital status of the customer
5	CUS_STATUS	nominal	The financial capital of that customer defined at the time of registration
6	TARGET	nominal	Type of customer for the bank
7	BALANCE	nominal	Latest Balance the customer has in his/her account
8	IS LOCATED_IN_ADDIS	nominal	Where a customer's branch resides? In Addis or outside Addis Ababa
9	IS_CHEQUE FRAUDULENT	Nominal	Is the customer fraudulent

Table 3.4 the selected attributes and their data types

Figure 3.1 shows the attribute rank using a WEKA tool. As it shown the most influential attributes are CUS_STATYS, GENDER & BALANCE.

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 9 NO_OF_REC):
        Correlation Ranking Filter
Ranked attributes:
 0.1251    4 CUS_STATUS
 0.0955    2 GENDER
 0.0667    6 BALANCE
 0.0607    3 MARITIL_STATUS
 0.0587    7 DISTRICT
 0.058     8 ADDIS
 0.0391    5 TARGET
 0.0355    1 SECTOR

Selected attributes: 4,2,6,3,7,8,5,1 : 8
```

Figure 3.1 the ranked attributes

3.4 Description of the selected attributes

Sector:

Sector is a customer's working field area, as shown in the table below most of the customers fall under individual which means the current account opened is used for their personal purposes.

No	Attribute value	How often (%)	Description
1	Individual	96	The owner is just an individual
2	Import & export	0.2	The owner is registered to be importer or exporter
3	Bldng & const	0.7	The owner is registered on sectors of building and construction
4	Domestic trade	0.4	The owner is registered as a domestic trader
5	Government	1.1	The owner of the account is hold by governmental organ
6	Other	1.2	This values are either missing or outliers

Table 3.5 Sector attribute description

Gender:

Gender is simply the sex of the account holder. In some cases as it can be seen in the below table corporate may be in place of male or female

No	Attribute value	How often (%)	Description
1	Male	83.5	When the gender male
2	Female	14.5	When the gender is female
3	Corporate	2	When the customer come as corporate

Table 3.6 Gender attribute description

Marital status:

Marital status is stand for describing the status of a customer whether he/she is single or married, else a different value is included in other option. We can observe from the table that most of the customer are married.

No	Attribute value	How often (%)	Description
1	Single	35.5	When the customer is registered as single
2	Married	49	When customer is registered as married
3	Other	15.5	When the customer marital status is not mentioned or if it is missing or if it is outlier

Table 3.7 Marital status attribute description

Customer status:

Customer status is the one which describes the customer financial capacity in the bank. and most of the customer fail under financial small. As the domain experts tell, this filed and other most attributes took their values at the time of customer registration.

No	Attribute value	How often (%)	Description
1	Financial small	88.5	When the owner is classified as financially small by the bank
2	Financial medium	5.2	When the owner is classified as financially medium by the bank
3	Financial large	6	When the owner is classified as financially large by the bank

Table 3.8 Customer status attribute description

Target:

This is the other customer filed where customers tell their target, according to the table shown below more than 63 present of the customer are business customer.

No	Attribute value	How often (%)	Description
1	Business customer	63.5	When the customer is registered as a general business organ
2	Retail customer	16	When the customer is retail customer
3	Commercial customer	2.5	When a customer is registered as commercial customer
4	Other customer	17.5	When the customer is other than the above or missed this as filed blank

Table 3.9 Target attribute description

Balance:

The balance is the one the researcher took a snapshot at the time of peaking the dataset from the database. The reader need to understand that this balance is not customer's opening balance but instead it is the actual balance.

No	Attribute value	How often	Description
1	Low	48.5	When a customer have less than 500 ETB balance in his/her account
2	Medium	30	When a customer have 500 to 2,915,588 ETB balance in his/her account
3	High	22.5	When a customer has more than 5,831,678 balance in his account

Table 3.10 Balance attribute description

District:

District is the location of branches grouped in to districts, as the researcher mentioned before more than 950 branches were grouped in to 16 districts.

Unlike the other attributes, district can't directly describe the behavior of a customer, rather the researcher and domain experts include it in order to gain knowledge about which branches/district are more exposed to cheque fraud. This may indirectly tell us the customer behavior since, probably, most customers use their nearer branches to their home or office.

No	Attribute value	How often appears (%)	Description
1	ADAMA DISTRICT	3	Branch located at Adama
2	BAHIR DAR DISTRICT	4	Branch located at Bahir dar
3	DESSIE DISTRICT	1.75	Branch located at Dessie
4	DIRE DEWA DISTRICT	1.8	Branch located at dire Dewa
5	EAST ADDIS DISTR	27.7	Branch located at east of Addis Ababa
6	GONDER DISTRICT	3.5	Branch located at Gonder
7	HAWASSA DISTRICT	1.1	Branch located at Hawassa
8	JIMMA DISTRICT	2.5	Branch located at Jimma
9	MEKELE DISTRICT	6	Branch located at Mekele
10	NEKEMTE DISTRICT	0.9	Branch located at Nekemet
11	NORTH ADDIS DISTRICT	18.3	Branch located at north of Addis Ababa
12	SHASHEMENE DISTRICT	1.4	Branch located at Shashemene
13	SOUTH ADDIS DISTRICT	13.5	Branch located at South of Addis Ababa
14	WEST ADDIS DISTRICT	12	Branch located at West of Addis Ababa
15	WOLITA SODO DISTRICT	2.2	Branch located at Wolita Sodo

Table 3.11 District attribute description

Located in Addis Ababa:

This filed tell us that the branch is located in Addis Ababa or not. as we can see more than 71 present of the branches located in Addis Ababa

No	Attribute value	How often (%)	Description
1	Yes	71.5	if the branch of a customer located in Addis Ababa
2	No	28.5	If the branch of a customer doesn't located in Addis Ababa

Table 3.12 located in Addis Ababa attribute description

Fraudulent:

This is the class attribute, or we use malicious customer table to classify the customers who are loyal with the bank with that of cheque fraudulent.

No	Represented as	How often (%)	Description
1	0	83.5	When a customer did not do a fraud 3 times
2	3	16.5	A customer is said to be fraudulent doing it 3 times

Table 3.13 Fraudulent attribute description

3.5 Data Formatting

This study implemented WEKA as a data mining tool. The available dataset should be prepared in a format and data type which is suitable for WEKA. WEKA data mining tool accept ARFF (Attribute-Relation File Format) file formates. This format contained attribute values whose values are separated by comma. The file extension for the file format ARFF is ".arff "[52].

Initially the data collected from the original database exported to an excel file format. Then the excel file format converted to common delimiter (csv) file format. Next the CSV file format opened with weka and then saved with arff file extension.

CHAPTER FOUR

EXPERIMENTATION

In this chapter, the remaining parts of hybrid process model have been described. The researcher also discuss the techniques that have been used in developing a model to predict fraudulent customer. This research incorporated the typical stages that characterize a data mining process. The researcher also discuss the experimentation process by relating the steps followed, the choice made , the task accomplished , the result obtained, evaluation of the model and results, and present it in a way that the organization can easily understand and use it.

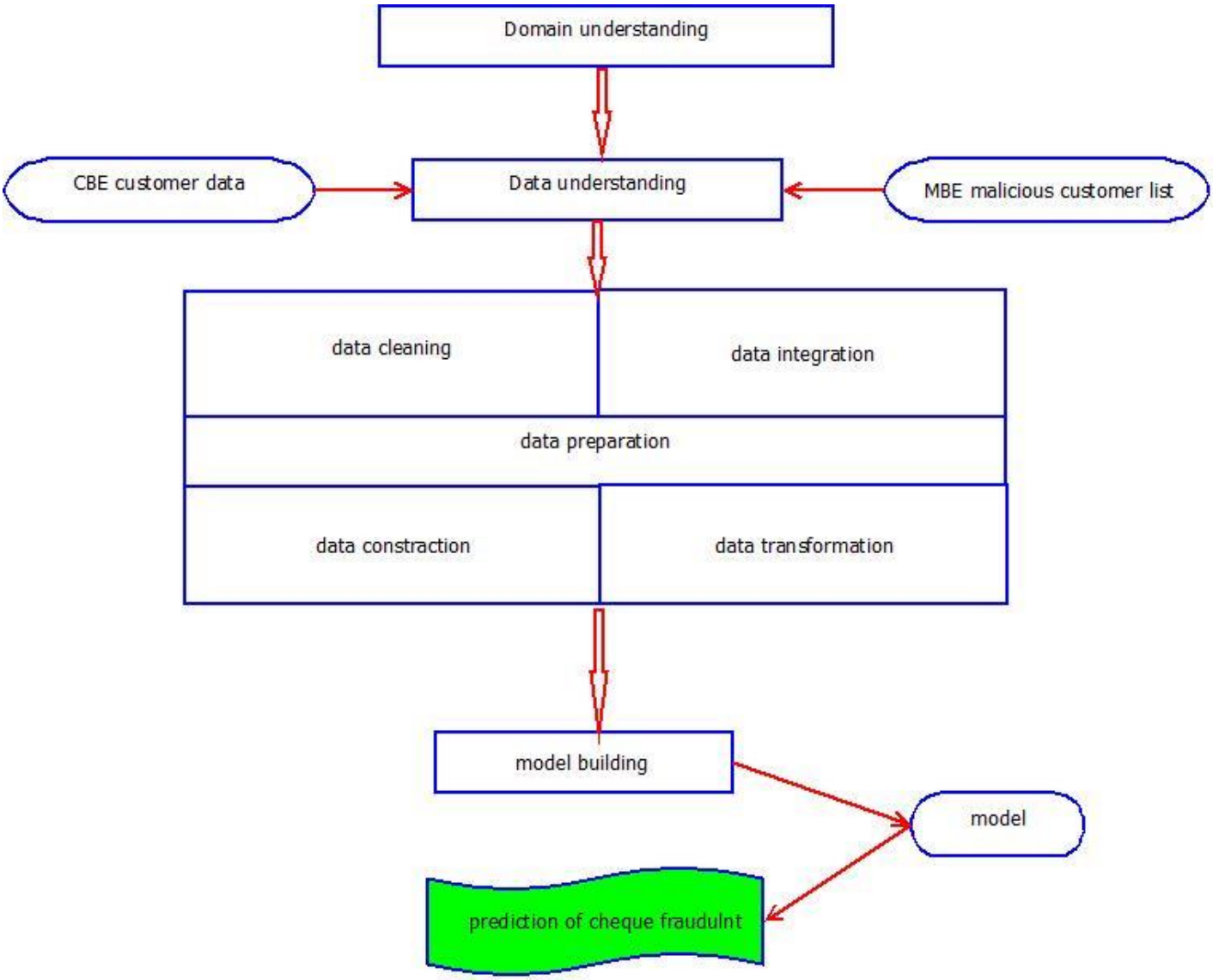

Figure 4.8 a simple architecture for model building

4.1 Building the Model

Modeling is one of the major tasks which is undertaken under the phase of data mining in hybrid methodology [55]. In this phase several data mining techniques are applied and their parameters are adjusted to optimal values. Several tasks need to be followed during this stage some of the basic tasks include selecting the modeling technique, conduct experimentation, building a model and finally evaluating the model.

4.1.1 Selecting Modeling Technique

In order to meet the objectives of these research three classification techniques has been selected for model building. The analysis was performed using WEKA tool. Among the different available classification algorithms three of them demonstrated using the dataset J48, PART and SMO are used for experimentation of this study. Although the researcher has a lot of other option, he selected the mentioned algorithms, since they are easy to understand and interpretation of the result is also be direct and forward.

J48 is one of the most common decision tree algorithms that are used today to implement classification technique using WEKA. It is WEKAs implementation of C4.5 top-down decision tree learners proposed by [58]. For a better result most of the time researchers change the default parameters of different algorithms. This parameters avail in WEKA is advantageous since trying with different value may obtain a better performance. And the algorithms having this features such as J48 and PART are preferable. For instance parameter in J48 algorithms decision tree and PART are listed in the following table (Table 4.1).

Option	Description
binarySplits	Whether counts at leaves are smoothed based on Laplace
confidenceFactor	The confidence factor used for pruning(smaller values incur more pruning)
debug	If set to true, classifier may output additional info to the console
minNumObj	The minimum number of instances per leaf
numFolds	Determines the amount of data used for reduced-error pruning, one fold is used for pruning and the rest for growing the tree
reducedErrorPruning	Whether reduced-error pruning is used instead of C.4.5 pruning
saveInstanceData	Whether to save the training data for visualization
Seed	The seed used for randomizing the data when reduced-error pruning is used
subtreeRaising	Whether to consider the sub-tree raising operation when pruning
unpruned	Whether pruning is performed
useLaplace	Whether counts at leaves are smoothed based on Laplace

Table 4.1 different algorithm's parameter of WEKA

On the other hand, since SMO popularity in 1998, it controls complexity and over fitting issues, so it works well on a wide range of practical problems. Due to this reason and better accuracy results, the researcher selected to implement SVM also in this study.

The test design specifies that the dataset should be separated into training and test set. The process of building predictive models requires a well-defined training and validation protocol in order to insure that most accurate and robust prediction [57].

In this research, 6,852 instances are used for training and testing. WEKA software has used to set up and measure the quality, validity and test of the selected model. For purpose of this study k-fold (10-folds) cross validation and percentage split test options (70%) is used. The specified percentage split is used since a better result is scored compared to the default option. 10-fold cross validation has been proved to be statistically good enough in evaluating the performance of the classifier [59]. According to 10-fold, the datasets are randomly partitioned equally into ten parts. Hence, 90% of the dataset is for training and 10 % for testing.

4.2 Conducting Experiment

On the next couple of sections, each experiment has been presented and the resultant models are properly interpreted and their performance are also explained.

4.2.1 Model building using tree (J48 decision tree)

J48 decision tree algorithm is a predictive machine learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data. It can be applied on discrete data, continuous or categorical data [60]. This algorithms is used in this study to classify cheque fraudulent customer.

On the other hand, as we have been discussing so far, we have about 6875 pre-processed records in our dataset. And about 1154 of them is fraudulent. The remaining figure is loyal customer who have been served based on the rule of the bank. So here, the J48 decision tree can serve as a model for classification as it generates simpler rules and remove irrelevant attributes at a stage prior to tree induction.

Since we need to do this experiment with different parameters, such as unpruned enabled, and disabled, 10-fold cross validation, percentage split (70%) is covered and compered each other to take the best of J48 model. The outputs of the task is compiled on the below table.

J48 model attributes	testing options with J48 algorithm			
	1st	2nd	3rd	4th
Testing approach	10 folds	10 folds	Percentage split	Percentage split
Unpruned	FALSE	TRUE	FALSE	TRUE
Accuracy	86.3%	86.2%	85.26%	85.9%
Time taken in seconds	0.16	0.03	0.04	0.15
No of leaves	41	176	41	176
Size of tree	53	226	53	226
Avg.TPR	0.863	0.862	0.856	0.864
Avg.FPR	0.671	0.637	0.738	0.660
Avg.PR	0.878	0.856	0.868	0.864
Avg.RR	0.863	0.862	0.856	0.830
Avg.ROC	0.711	0.774	0.578	0.761
CCI	5918	5911	1754	1768
ICI	938	945	303	289
Total no instance	6856	6856	2331	2331

Table 4.2 results of J48 algorithm with four different parameters

Keys: CCI: Correctly classified Instance, ICI (Incorrectly classified Instance), Accuracy: Registered performance of model, Avg: Average, TPR: True Positive Rate. FPR: False Positives Rate, ROC: Relative Optical character curve, PR: precision rate, RR: Recall rate

As we have seen on the above table, we are witnessed for having J48 with different option may also improve the performance of a given model. One can observe the 1st experimental scenario had scored better. Although we have tried to change some of the parameters we can still see the default is better in our dataset.

On the other hand a minimum number of leaves and tree is shown on 1st and 3rd scenario, where unpruned is set to be “false” (default value). If we try to see from the time perspective, we can observe that the 2nd scenario took the minimum time for building our model.

In order to choose the best model from the given ones, we need to consider it from different angles such as what resource do we have, what is the best we need and so on. So before we came up with a decision, we need to take in to consideration the above scenarios.

In this case, the time it take for building the model might not be significant but the size of the tree and number of leaves (as shown in table 4.2) have huge difference between unpruned and pruned testing approaches. Therefore, tracing the leaves might be a challenge. Based on this, the researcher intended to choose the 1st model as a best model from this experimental scenarios taking in consideration of all the factors.

The detailed confusion matrix as a result of unpruned set to be "false", J48 algorithm with all attribute has been discussed below

```
=== Summary ===

Correctly Classified Instances        5918               86.3186 %
Incorrectly Classified Instances       938               13.6814 %
Kappa statistic                          0.2821
Mean absolute error                      0.2235
Root mean squared error                  0.3356
Relative absolute error                 79.806  %
Root relative squared error             89.6854 %
Coverage of cases (0.95 level)          99.6791 %
Mean rel. region size (0.95 level)      93.7135 %
Total Number of Instances            6856

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.999    0.807    0.859      0.999    0.924      0.399    0.711     0.912     0
                 0.193    0.001    0.970      0.193    0.322      0.399    0.711     0.469     3
Weighted Avg.    0.863    0.671    0.878      0.863    0.823      0.399    0.711     0.838

=== Confusion Matrix ===

    a    b   <-- classified as
 5695    7 |    a = 0
  931  223 |    b = 3
```

Figure 4.2 confusion matrix of J48

		Predicted customer status		total
		LOYAL	FRAUDULENT	
Actual customer status	LOYAL	5695	7	5702
	FRAUDULENT	931	223	1154
Total		6626	230	6,856

Table 4.3 confusion matrix for predicting customer status (J48)

As we have seen in chapter one, what each variable stands for is stated as follows, TN: we predicted LOYAL and they are actually LOYAL. FN: we predicted LOYAL but it actually FRAUDULENT. TP: we predicted FRAUDULENT and it actually FRAUDULENT. FP: we predicted as FRAUDULENT but it actually LOYAL. FP + TN: The total number of actually LOYAL. TP + FN: The total number of actually FRAUDULENT. TN + FN: The total number of predicted as LOYAL. FP+ TP: The total number of predicted as FRAUDULENT. TN + FP + FN + TP: The total number of all customers.

Based on the above explanation, the interpretation of the real scenario will be discussed here.

- The no of customers who are classified as loyal and they are actually loyal is: 5695
- The no of customers who are classified as loyal and they are actually fraudulent is: 931
- The no of customers who are classified as fraudulent and they are actually fraudulent is: 223
- The no of customers who are classified as loyal but they are actually fraudulent is: 7
- The total number of actual loyal customer is: 5702
- The total number of actual fraudulent customer is: 1154
- The total number of predicted loyal customer is: 6626
- The total number of predicted fraudulent customer is: 230 and
- The total number of instance used is: 6856

We need also see other measurements, the performance of the experiments is measured in the following manner as depicted below:

- Sensitivity (TPR): indicated by the proportion of fraudulent customers that are correctly classified as fraudulent i.e. TPR= TP/ TP + FN which is: 0.19
- Specificity (TNR): The proportion of normal customer that are correctly classified as normal. TNR= TN/ TN+FN = 0.86
- False Negative Rate (FNR): The proportion of fraudulent customer that are erroneously classified as normal i.e. FNR= 1 – TPR or FNR = FN/ TP + FN, which is: 0.81

4.2.2 Model building using Rule Induction Algorithms (PART)

The next data mining classification technique used for this research was PART Rule induction algorithms. Although there are many rule induction algorithms but the researcher selected PART for the reason that PART has the ability to produce accurate and easily interpretable patterns/ rules

that helps to achieve the research objectives. PART is a separate-and-conquer rule learner like that of decision tree and proposed by [59].The algorithm applies an iterative process and produces set of (decision lists) which is ordered set of rules. It works by generating a rule that covers a subset of the training examples and then removing all examples covered by the rules from the training set. This process is repeated until there are no examples covered left to cover. The final rule set is the collection of the rules discovered at every iterations of the process [55]. The rules are based on IF-THEN rules. The concept had discussed more broadly in chapter two section 2.2

A similar procedure is being followed here also, as the former experiment. The same dataset has used without any amendment in our data, four different parameters are experimented (unpruned "enabled" and "disabled", 10-fold used, percentage split used), and we have to make sure the results are not biased by human intervention in our dataset.

The results of the four different scenarios using rule induction specifically "PART" are presented in the below table.

PART model attributes	Testing options with PART algorithm			
	1st	2nd	3rd	4th
Test approach used	10 folds	10 folds	Percentage split	Percentage split
Unpruned	FALSE	TRUE	FALSE	TRUE
Accuracy	86.4%	85.95%	85.8%	85.5%
Time taken in seconds	0.56	2.18	0.25	2.06
No of rules	35	307	35	307
Avg.TPR	0.864	0.860	0862	0.862
Avg.FPR	0.648	0.641	0.682	0.665
Avg.P	0.866	0.849	0.859	0.852
Avg.R	0.864	0.860	0.862	0.862
Avg.ROC	0.739	0.770	0.769	0.755
CCI	5923	5893	1765	1760
ICI	933	963	292	297
Total no instance	6856	6859	2331	2331

Table 4.4 result of PART algorithm with four different parameters

Keys: CCI: Correctly classified Instance, ICI (Incorrectly classified Instance), Accuracy: Registered performance of model, Avg: Average, TPR: True Positive Rate. FPR: False Positives Rate, ROC: Relative Optical character curve, PR: precision rate, RR: Recall rate

We can easily see the performance of the model has better ability on the 1st (10-fold cross validation with unpruned set to be false) experimental scenario in different criteria's. Performance wise it has 86.4% accuracy and time wish it took (0.56 second) a better position next to 3rd which took only 0.25 seconds to build the model. The other thing we need to see is the number of rules, again 1st & 3rd experimental scenario has a minimum number of rules.

Based on the above discussion the 1st experimental scenario is chosen to be the best from the other three PART experiments.

```
=== Summary ===

Correctly Classified Instances        5923               86.3915 %
Incorrectly Classified Instances       933               13.6085 %
Kappa statistic                          0.3074
Mean absolute error                      0.216
Root mean squared error                  0.3302
Relative absolute error                 77.121  %
Root relative squared error             88.2627 %
Coverage of cases (0.95 level)          99.7958 %
Mean rel. region size (0.95 level)      95.1356 %
Total Number of Instances             6856

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.994    0.778    0.863      0.994    0.924      0.400    0.739     0.919     0
                 0.222    0.006    0.880      0.222    0.354      0.400    0.739     0.519     3
Weighted Avg.    0.864    0.648    0.866      0.864    0.828      0.400    0.739     0.852

=== Confusion Matrix ===

    a    b   <-- classified as
 5667   35 |    a = 0
  898  256 |    b = 3
```

Figure 4.3 confusion matrix of PART

		Predicted customer status		total
		LOYAL	FRAUDULENT	
Actual customer status	LOYAL	5667	35	5702
	FRAUDULENT	898	256	1154
Total		6565	291	6,856

Table 4.5 confusion matrix for PART predicting customer status

Since we have seen all the appropriate information which are important for our data interpretation on the above section (J48 algorithm) the research only see the interpreted data below for PART algorithm.

- The no of customers who are classified as loyal and they are actually loyal is: 5667
- The no of customers who are classified as loyal and they are actually fraudulent is: 898
- The no of customers who are classified as fraudulent and they are actually fraudulent is: 256
- The no of customers who are classified as loyal but they are actually fraudulent is: 35
- The total number of actual loyal customer is: 5702
- The total number of actual fraudulent customer is: 1154
- The total number of predicted loyal customer is: 6565
- The total number of predicted fraudulent customer is: 291
- The total number of instance used is: 6856

We need also see other measurements, the performance of the experiments is measured in the following manner as depicted below:

- Sensitivity (TPR): indicated by the proportion of fraudulent customers that are correctly classified as fraudulent i.e. TPR= TP/ TP + FN which is: 0.193
- Specificity (TNR): The proportion of normal customer that are correctly classified as normal. TNR= TN/ TN+FN = 0.863
- False Negative Rate (FNR): The proportion of fraudulent customer that are erroneously classified as normal i.e. FNR= 1 – TPR or FNR = FN/ TP + FN, which is: 0.807

Some sample rules are taken from the PART generated algorithm for demonstration purpose and they are discussed in the Evaluation of the discovered knowledge part.

4.2.3 Model Building using function (SMO Algorithms)

The third classification task used in this research was Support vector machine. SVM algorithms that available in WEKA software is sequential minimum optimization algorithms. It works based on risk minimization principles [55].

Although, this algorithm has capacity to accept data with their missing and nosy values (capability of WEKA), But this research has proceed with the former dataset we have been using for building the J48 and PART model, since we had passed the step of preprocessing of the data on chapter 3

Here also the same methods is being employed for experimentation as used in the above two models except the pruned parameter. So, only two experimentation has conducted using this algorithm. These are the 10-fold cross validation and the percentage split with the same number of instance.
The result of SMO model is presented in the below table.

SMO model attributes	Testing options with SMO algorithm	
	1st	2nd
Test approach used	10 folds	Percentage split
Accuracy	83.4%	82.8%
Time taken in seconds	20.19	20.39
Avg.TPR	0.834	0.833
Avg.FPR	0.703	0.728
Avg.P	0.796	0.787
Avg.R	0.834	0.833
Avg.ROC	0.566	0.794
CCI	5718	1941
ICI	1138	390
Total no instance	6856	6859

Table 4.6 result of SMO algorithm

Keys: CCI: Correctly classified Instance, ICI (Incorrectly classified Instance), Accuracy: Registered performance of model, Avg: Average, TPR: True Positive Rate. FPR: False Positives Rate, ROC: Relative Optical character curve, PR: precision rate, RR: Recall rate

According to table 4.6 the 1st experimental scenario has a better performance in classification of the elements. On the time wise, as we can see there is no significant time difference between the two models.

So, 10 fold cross validation has better accuracy with 83.4%. But compering to the other previous models relatively low accuracy has been observed.

In general, SMO has scored the worst time (about 20 seconds) to build the model compared to the other two models, which were not took more than a second.

```
=== Summary ===

Correctly Classified Instances        5718               83.4014 %
Incorrectly Classified Instances      1138               16.5986 %
Kappa statistic                          0.1814
Mean absolute error                      0.166
Root mean squared error                  0.4074
Relative absolute error                 59.2706 %
Root relative squared error            108.8905 %
Coverage of cases (0.95 level)          83.4014 %
Mean rel. region size (0.95 level)      50      %
Total Number of Instances             6856

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.970    0.839    0.851      0.970    0.907      0.222    0.566     0.850     0
               0.161    0.030    0.522      0.161    0.246      0.222    0.566     0.225     3
Weighted Avg.  0.834    0.703    0.796      0.834    0.796      0.222    0.566     0.745

=== Confusion Matrix ===

    a    b   <-- classified as
 5532  170 |    a = 0
  968  186 |    b = 3
```

Figure 4.4 confusion matrix of SMO

		Predicted customer status		Total
		LOYAL	FRAUDULENT	
Actual customer status	LOYAL	5532	170	5702
	FRAUDULENT	968	182	1150
Total		6500	352	6856

Table 4.7 confusion matrix for SMO predicting customer status

As we have observed from the above confusion matrix, the following detail has been gained from the SMO model.

- The no of customers who are classified as loyal and they are actually loyal is: 5532
- The no of customers who are classified as loyal and they are actually fraudulent is: 968
- The no of customers who are classified as fraudulent and they are actually fraudulent is: 182
- The no of customers who are classified as loyal but they are actually fraudulent is: 170
- The total number of actual loyal customer is: 5702
- The total number of actual fraudulent customer is: 1154
- The total number of predicted loyal customer is: 6500
- The total number of predicted fraudulent customer is: 352
- The total number of instance used is: 6856

We need also see other measurements, the performance of the experiments is measured in the following manner as depicted below:

- Sensitivity (TPR): indicated by the proportion of fraudulent customers that are correctly classified as fraudulent i.e. TPR= TP/ TP + FN which is: 0.16
- Specificity (TNR): The proportion of normal customer that are correctly classified as normal. TNR= TN/ TN+FN = 0.851
- False Negative Rate (FNR): The proportion of fraudulent customer that are erroneously classified as normal i.e. FNR= 1 – TPR or FNR = FN/ TP + FN, which is: 0.84

4.3 Comparison of J48, PART and SMO models

In order to meet our objective, we need a model that can classify our dataset with a better performance from different viewpoints. So the below table has compare the output of all the three mode based on accuracy of the model, the time it take to build the model and correctly classified instances we have been conducting experiment on.

Model type	Accuracy of models	Time taken in seconds	Correctly classified
J48	86.3%	0.16	5918
PART	86.4%	0.56	5923
SMO	83.4%	20.19	5718

Table 4.8 comparison of used algorithms

From the above table we can identify the best model from the listed data mining algorithm by considering characteristics they have. For instance J48 has a better accuracy performance (86.3%), in terms of time taken for building a model and still J48 has scored less second (0.56) for building the model. But, the number of correctly classified instance is also higher in PART (5923) records are correctly classified, which make the accuracy level somehow higher (86.4%). Time taken by PART is little bit more compared to J48, but we are more sensitive for the accuracy level. On the other hand, SMO has scored lowest accuracy and even for building the model, it is seems to be time taking. Therefore, based on the criteria's seen above, PART algorithm has chosen to be the best model.

4.4 Evaluation of Discovered Knowledge

As we have discussed on chapter one, evaluation of the discovered knowledge is 5th stage on hybrid data mining process model. Evaluation needs understanding the results. The researcher also need to check whether the discovered knowledge is novel and interesting. The interpretation of the results has to be evaluated by domain experts and checking the impact of the discovered knowledge is important. Then only a researcher can get approved model [18]. And finally the entire process is revisited to identify which alternative actions could have been taken to improve the results [18].

In this research also, starting from having the right data, and then pre-processing the data and handing it to WEKA is the most important part of this study and the researcher also give a best attention for this steps in order to get the best knowledge from the dataset. The detailed step by step data pre-processing is shown in chapter three.

The researcher also tried to show some of the interesting results for the domain experts and has gain a positive feedback. After the discussion in which the author had with experts, one of them personally recommended using the model at the time of new customers come for registration may bring a new way of handling customers.

It is important to see the side effect when a new model is build. If the researcher only focuses on the accuracy and other positive side of the model, he/she may end up with bad model after implementation.

The impact of this model has tolerable effect on the customers since out of the total 6858 customers the model thought only 35 (less than 0.52%) them as fraudulent, while they are actually loyal ones. This can be also minimized (approaches to 0%) when a number of data included in the dataset increases in the future.

PART appears to be the best model, among the chosen three different algorithms. As we have been discussing these three classification algorithms (J48, PART and SMO) are selected due to their easiness for use and interpretation of their result as well. They are also compatible each other for comparison purpose. Some rules of PART retrieved are discussed below, the rules are selected based on the discussion with domain expert which are somehow interesting rules.

RULE – 1
CUS_STATUS = FINANCIAL_LARGE AND
SECTOR = INDIVIDUAL AND
BALANCE = HIGHER AND
GENDER = MALE: 0 (48.0)

The first rule tell us that if the given customer is an individual whose gender is a male and registered with financially high customer and kept relatively high amount in his account then he is 100 % loyal customer. The rule retrieved 48 customer with similar characteristics and it found all of them loyal.

RULE – 2
CUS_STATUS = FINANCIAL_SMALL AND
TARGET = OTHER_CUSTOMER AND
BALANCE = LOWER: 0 (111.0/26.0)

Similarly the second sample rule tell us if a customer registered with financially low and whose target is not defined (missing) and having low balance in his account shows more than four customer out of five other customer with similar character is loyal cheque customer.

RULE - 3
GENDER = MALE AND
TARGET = RETAIL_CUSTOMER AND
MARITIL_STATUS = OTHER: 0 (59.0/26.0)

The third rule also tell that if a customer is male and registered having the target of retail customer and whose marital status is not mentioned then one person with similar behavior can be fraudulent out of three.

RULE - 4
ADDIS = YES AND
MARITIL_STATUS = SINGLE AND
CUS_STATUS = FINANCIAL_SMALL AND
SECTOR = Individual: 0 (256.0/78.0)

A corporate customer which owned by government and the account is registered as financial large out of 38 government corporation all are found to be loyal cheque customer.

RULE - 5 CUS_STATUS = FINANCIAL_LARGE AND
SECTOR = Government: 0 (38.0)

The last sample rule show that, a customer who is a male and whose financial status is registered to be financial medium but his balance in his account is less than 500, then all 160 customers are found to be fraudulent.

Notice that,"0" is assigned for loyal cheque customer and "3" is assigned for cheque fraudulent customer and the numeric values which appeared in the bracket next to the class label indicates the number of correctly and incorrectly classified records respectively.

4.5 finding of the study

As discussed previously on chapter one, on statement of the problem part, there is a need of classification of customers on their cheque usage behavior from the bank side. This can be made true by studying the patterns of the huge data given. As a result customers with a certain behavior may be risk free and some others might not be, so that it is easier for the bank to treat them in different ways. For instance the model show that a customer which is a "corporate" and registered with "high" financial status owned by government are risk free. So, the bank may conduct a payment transaction from those specific customers without checking the balance for fastening the service.

On the other hand, a customer who is a male and whose account show "low balance" then one out of three customer with similar character is cheque fraudulent. So, for such type of customer, a bank need to check their account before conducting cheque payment.

As PART rule in general reveals, cheque fraud is relatively more in Addis Ababa (within four district) rather than out-side of Addis Ababa (eleven other district). A customer having "lower" balance is more exposed to commit cheque fraud than those, who has "high" balance.

The domain experts also explained the above rules with a reason as shown next.
A frequent attempt of cheque fraud is happening on west addis district (addis ababa), Tekle Haimanot, Atena Tera, Aba Koran, Addis Ketema, Anwar, Mehal Gebaye are to mention some. According to their explanation, this branches have very huge number of transaction per day. People has been using cheque as a common way of payment for a bigger trading using those branches. The domain experts were not excite until they observe, this is only be true when that customer did not tell his marital status and also keeps "average" balance in his account.

The other interpretation of experts from the rule is that since, most of the fraud is done due to "insufficient funds transfer", and this happening on the individuals who had kept a low balance in

their account. As they told, this is the most common way of a customer to be recorded in the malicious customer list.

The domain experts were expecting a very tight relationship between customers who had low balance in their account with malicious customer. On the other discussion we find out that, another scenario, a customer who is about to stop using his/her account may withdraw all the cash in it also. That account may not be closed for several years. Obviously this should increase the no of customer who has very less cash in their account. This may cause an ambiguity of getting the desired result out of our model.

Although they expected under balanced customers are relatively fraudulent, they found it surprising that most of the customer who did this was not mention their marital status at the time of registration.

There are also other cases which are not easily give a meaning by the experts, for instance a customer who is registered as financially middle class and who kept normal balance in his account found to be cheque fraudulent customer. This was somehow, difficult for them to give a meaning for a situation and it may be a hidden pattern.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1 CONCLUSION

The effective use of information and technology is crucial for banking industry to stay competitive in today's dynamic and evolving environment. Recently, commercial banks in Ethiopia are working on new area of business, to take the leading in business and to maximize the profit they get from their customer. In the prevailing dynamic and competitive business environment, excellence in customer services is crucially important to maintaining a sustainable business growth [9].

Different researches on the area and other countries experience hold the major role on revealing previously unknown direction in business. This may include expanding the services they were giving to the customers, securing the ways of transaction, increase their customer satisfaction and so many other. Those successes cannot merely be achieved unless this ideas are supported by researches on the domain.

On the other side, data mining is a machine learning field of study, which is capable of discovering a knowledge having a massive data on hand by using different algorithms. A data mining tool called WEKA gives a lot of technical options, so, one can chose his algorithm depending on the data he/she has at hand. Application of data mining techniques may help in answering several important and critical questions by extraction of useful knowledge that enables support for cost-savings and decision making [55].

In this research also, an attempt has been made to apply the data mining technology for predicting fraudulent cheque customers consuming CBE's customer data and NBE's fraudulent list by making use of three different classification techniques namely decision tree, rule induction and support vector machine. Data mining Algorithms J48, PART and SMO were used for building the model by means of six step data mining methodology called hybrid model which is very help full on fixing the problems going back and forward. Attributes were been selected being with domain experts and nine of them are SECTOR, GENDER, TARGET, MARITAL STATUS, CUSTOMER STATUS, LOCATED IN ADDIS ABABA, DISTRICT & CHEQEU FRAUD. Through this study, one of the research question at the beginning was to find out the main factors for defining the customer if he/she is cheque fraudulent.

The methodology has strictly been followed while undertaking the experimentation starting with business domain understanding on chapter four and end up with evaluation of the model on chapter five.

Experimentation has been done using different scenarios for the purpose of choosing a model with a better performance. To mention the changes, unpruned set to be “true” and also with the default value (”false”). On the other hand 10-fold cross validation and percentage split has been also tried. So for each algorithm 4 experiments has been conducted except SMO which is shown using only two experiments.

This study concluded by PART rules which are generated by the model and the rules was interpreted by the domain experts, they found most of rules have interesting patterns. The kind of rules generated answered the other research question of this study.

A result show that a PART rule induction algorithms registered batter performance with 86.4% accuracy running with 10-fold cross validation with default parameters than any experimentation done for this research purpose. The accuracy level of the model answered the last research question of this study.

Other models such as J48 has registered 86.3% accuracy which is only 0.10% less than that of PART algorithm with a default parameter. SMO was scored the least with 83.4%. As discussed on the previous chapter PART’s better accuracy is not the only one which make it closable but also time taken for building the model and the feasibility of the rules are liked by domain experts.

5.2 RECOMMENDATION

Although this study is conducted on the bases of academic research purpose, a researcher thinks it may help a lot on the banking industry specifically for CBE. Since, CBE needs to keep the customers as well us transaction be healthy and safe. This can’t be achieved easily, because it need a deep study on customer behavior, and identify characteristics of customer which led them to commit fraud. This research also tried to help on predicting this characteristics of customers using the history database at the end. The researcher believe if this model is deployed on the place (before new customer registration), customers who have similar behavior with previously fraudulent may be detected, so the bank may think a proper take care for itself and the other side customers.

This research used number of record as a class, (3) for those who committed the fraud three times and (0) who don’t. A customer must commit fraud 3 times to be in the list of malicious customer, then only that customer should be sent to NBE, then NBE also distributes the information of a customer to all the other branches. A researcher recommend that, if there is a way to collect customers with only one record and two records, from all over the branches. Then after, a better

performance may be scored on the predictive model since, we can able to see what kind of behavior lead to make the customer to commit a cheque fraud.

This research tried to group branches (about 950) to their respective districts (16). The researcher thinks there may be a hidden knowledge under districts which can be directly related to some specific branches. In the future a research may be needed to reveal the knowledge on branches of CBE by consolidating the dataset.

The other gap the researcher and the domain expertise found out finally is, as discussed on the conclusion part, a customer is more suspicious for cheque fraud having low balance in his/her account, this is a general truth in branches, and also a near result has find out using the model, but still the expertise were not satisfied. After a while, the cause of the problem was found through the discussion with the group. "The first act of any customer, who need to close or change his account will primarily withdraw his balance and let his/her account to be empty" this type of customers are too many and they are not necessarily cheque fraudulent. On the other hand, there are customers who make cheque transaction and make their account empty deliberately to commit fraud. The above two type of customers categorized under one group with "low balance". The researcher recommend, a research which can handle this issue on separate category will definitely bring a better result. As I discussed above, customers may have other reason to withdraw all their balance (not necessarily for fraud), but this may somehow biased our model accuracy level.

In this study as the researcher had explained only CBE customer data has been employed, but it is quite important to have a system which is country wide. For such kind of study the dataset has to be collected from all banks (including private banks) of Ethiopia. So that, the research could have power to positively influence the financial institutions all over the country.

The research conducted using 6852 rows of dataset. Currently, this number is growing very fast, (Since the number of customer and the dataset has a direct relation) as a result one can improve the performance of the model using bigger dataset.

In chapter three, experimentation was conducted using only three algorithms J48, PART and SMO, they are selected due to their easiness for understanding and interpretation in addition to their similar characteristics which is helpful for comparing each other. Although this techniques scored nearly similar results, other researcher may discover better result by doing further experimentation whether by changing different parameters or using different algorithms.

REFERENCES

1. Alvin Toffler, (1981). “The Third Wave”, Bantam Books, United States of America
2. Commercial Bank of Ethiopia manual (2012).” CBE CATS Technical Training manual”, PP 37-38
3. Singapore Institute of Management (2002), “Data mining and customer relationship”, vol.10, pp. 34-38.
4. Lori, A. (2006). ”Data Mining for Information Professionals’, vol.10, pp.120-128
5. Deshpande, S. and Thakare, V (2010). “Data Mining System and Applications”, International Journal of Distributed and Parallel System (IJDPS), vol.1, pp. 32-44.
6. Rajanish, D (2002). ”Data Mining in Banking and Finance: A Note for Bankers”, Indian Institute of Management, Ahmadabad.
7. Rene, T (2010). “Applying Data Mining to Banking”, Business Management Articles, vol. 5, pp. 3-7.
8. Srivastava, J. (1991). “Data Mining for Customer Relationship Management”, vol. 12, pp. 4-8
9. CBE manual (2012) “Customer Account and Transaction Service Process”, vol.0.1, pp. 6-9
10. Rene T. Domingo (2003), “Applying Data Mining to Banking” head of information technology department international Black Sea University
11. http://irn.uit .tufts.edu/research _planner/documents/6/methodology_tips.pdf
12. Hillol Kargupta, Anupam Joshi, Krishnamoorthy Siva Kumar, Yelena Yesha, (2005) "Data Mining: Next Generation Challenges and Future Directions", Publishers: Prentice-Hall of India, Private Limited.
13. S.R. Mittal, Report of Committee on Internet Banking (2001), Constituted by Reserve bank of India, Chairman of the Committee
14. Rajanish Dass, "Data Mining in Banking and Finance: A Note for Bankers", Indian Institute of Management Ahmadabad.
15. "Basic Banking Services and Checking Accounts ", High-Intermediate/Advanced
16. Scotiabank, "Day-to-day banking companion booklet" july 2016.

17. Mohit Bhansali1, Praveen Kumar, 'An Alternative Method for Facilitating Cheque Clearance Using Smart Phones Application ', Volume 2, Issue 1, January 2013
18. http://www.springer.com/978-0-387-33333-5
19. Robson, C. (1993) Real world research: A resource for social scientists and practitioner-researchers. Blackwell: Oxford; Cambridge.
20. Vikas Jayasree and Rethnamoney Vijayalakshmi Siva Balan, "A Review on Data Mining In Banking Sector", (AJApS), Vol. 10, Issue 10, 2013, ISSN 1554-3641, pp. 1160-1165.
21. A. B. Devale and Dr. R. V. Kulkarni, Applications of data mining techniques in life insurance, International JDM&KMP, Vol.2, Issue 4, July 2012, ISSN 2230-9608, pp. 31-40.
22. Rexer Analytics, Annual Rexer Analytics Data Miner Survey, Commercial Tools – Ranked by Primary Tool, 2013, Available
23. Abhijit A. Sawant and P. M. Chawan,"Study of Data Mining Techniques used for Financial Data Analysis",(IJESIT),Volume 2, Issue 3, May 2013
24. Dr. Sudhir B. Jagtap,Dr. Kodge B. G.,"Census Data Mining and Data Analysis using WEKA",(ICETSTM – 2013)
25. Ruxandra PETRE,Data Mining Solutions for the Business Environment,Database Systems Journal vol. IV, no. 4/2013
26. Khaled Hassanein,Image Data Mining of Check Transactions In Support of Customer Relationship Management at Banks,Michael G. Degroote School of Business,Canada
27. Rosie Neilson, Speed up cheque payments, UK, 2014
28. RBI, IMPORTANT GUIDELINES ON FRAUD –CLASSIFICATION AND REPORTING, MASTER CIRCULAR
29. Check Fraud, (1999) "a guide to avoid loses",
30. Ruxandra PETRE,"Data Mining Solutions for the Business Environment",Database Systems Journal vol. IV, no. 4/2013
31. Usama Fayyad, Gregory Piatetsky- Shapiro and Padhraic Smyth, From Data Mining to Knowledge Discovery in Databases, AI Magazine, Vol. 17, Issue 3, 1996, ISSN 0738-4602, pp. 37-54.

32. Ion Lungu and Adela Bâra – “Improving Decision Support Systems with Data Mining Techniques”, “Advances in Data Mining Knowledge Discovery and Applications”, Croatia, 2012.
33. Oded Maimon and Lior Rokach, Data Mining and Knowledge Discovery Handbook. Second Edition, Springer Publishing Company, USA, 2010.
34. George M. Marakas, Modern Data Warehousing, Mining, and Visualization, Pearson Education, New Delhi, 2005.
35. K C Chakrabarty, Fraud in the banking sector – "causes, concerns and cures", New Delhi, July 2013
36. Electronic Cheque Clearing Operating Rules,Kathmandu, Nepal, June 2011
37. "Check 21 - FAQs." http://www.ffiec.gov/exam/cheque21/faq.htm.
38. Dr. K. Chitra, B. Subashini, Data Mining Techniques and its Applications in Banking Sector IJET&AE,August 2013
39. Vivek Bhambri, Application of Data Mining in Banking Sector, indea, IJCST Vol. 2, Iss ue 2, June 2011
40. S.P. Deshpande, Dr. V.M. Thakare, "Data Mining SystemAnd Applications: A Review".
41. Hillol Kargupta, Anupam Joshi, Krishnamoorthy Siva Kumar, Yelena Yesha, "Data Mining: Next Generation Challenges and Future Directions",India, 2005.
42. S. S. Kaptan, N S Chobey, "Indian Banking in Electronic Era", Sarup and Sons, Edition 2002.
43. Majid Sharahi, Mansoureh Aligholi, Classify the Data of Bank Customers Using Data Mining and Clustering Techniques,February 2015
44. Khan Babayee, Muhamamd, use of clustering technique and genetic algorithm in making decision trees for optimal classification of banks customers, Iran, 2009
45. M. P. Thapliyal,Data Mining: A Tool for Banking Industry,IJERM&T,April 2015
46. B. Rajdeepa, D. Nandhitha, Fraud Detection in Banking Sector using Data mining, IJSR, 2013
47. Wayne Macdonald, Jacqueline Fitzgerald, Understanding fraud, Bureau of crime statistics and research, Australian, December 2014
48. Asia nesredin, mining patientes’ data for effective Tuberculosis diagnosis, AAU, Ethiopia, JUNE, 2012

49. Berry M. and Linoff G., ‘Data Mining Techniques for Marketing, Sales and Customer Relationship Management’, 2004.
50. Bhargavi P. and Jyothi S., ‘Applying Naïve Bayes Data Mining Technique for Classification of Agricultural Land Soils’, (IJCSNS), 2009
51. Rashmi, 2010, ‘Clustering in Data Mining’.
52. Jiawei H, Jian P, (2015) "Data Mining Concepts and Technology" University of Llinois third Edition.
53. Jasdeep Singh Malik, Prachi Goyal, Mr. Akhilesh K Sharman 2012 A Comprehensive Approach towards Data Preprocessing Techniques and association Rules. International Journal of Emerging Technology and advance engineering Technology
54. P. Chaudhari, D. P. Rana, R. G. Mehta, N. J. Mistry, M. M. Raghuwanshi 2014 Discretization of Temporal Data A survey. Internation Journal ofComputer science and information security
55. Girma Aweke, (2012) “Predicting HIV Infection risk factor using voluntary counseling and testing data: A CASE OF AFRICAN AIDS INITIATIVE INTERNATIONAL (AAII)”, Addis Ababa University, Ethiopia
56. Kantardzic ,Mehmed(2003). Data mining: concepts, models, methods, and algorithms. The University of California; Wiley-Inter science
57. Two crows Corporation (1999), Introduction to data mining and knowledge discovery (2nd Ed).by Edelstein, H., A. Potomac, MD: Two Crows Corp.
58. Quinlane , J.R. (1986). Induction of Decision Trees. Kluwer Academic publisher. Boston (1), 81-106
59. Frank,E, and Witten ,Ian .(1998).Generating Accurate Rule Sets Without Global Optimization. In: Fifteenth International Conference on Machine Learning, 144-151
60. Bharti, K Jain, S and Shukla, S (2010) Fuzzy K-mean Clustering Via J48 for Intrusiion Detection System.International Journal of Computer Science and Information,1(4).
61. Fayyad, Usma, Piatetsky-shpiro, G. and smyth, padharic.(1996). “From Data Mining to knowledge Discovery in Databases” http://citeseer.nj.nec.com/fayyad96from.html
62. Moshkovich,Helen M., Mechitov,Alexander I., and Olson,David L. (2002) .Rule induction in data mining: effect of ordinal scales. Expert systems with applications. Elsevier Science Ltd. Technical University of Crete.

63. Wong, M. L., Leung, K. S.(2000). Data Mining using Grammar Based Genetic Programming and application. Springer
64. Grzymala-busse, jerzy w. (n.d.) Rule Induction. University of Kansas Retrieved on February, 2012 from people.eecs.ku.edu/~jerzy/c95-perugia.pdf
65. Berson, Alex, Smith, Stephen, and Thearling, Kurt (n.d.) An Overview of Data Mining Techniques (Feb, 2012)http://www.thearling.com/text/ dmtechniques/dmtechniques.htm
66. My Data Mining Weblog (MDMW). (2008).Rule learner (or Rule Induction) Retrived Febrary 2012 from: http://mydatamine.com/?tag=induction-algorithms
67. Trybula, Walter J. (1997) Data Mining and Knowledge Discovery. Ed. Williams, Martha E. Annual Review Information Science and Technology, Vol. 32, Information Today, Inc. New Jersey, USA: Medford
68. Ingrid Russell, Zdravko Markov, (2005), University of Hartford, "An Introduction to the WEKA Data Mining System"
69. Trnka, A, 2010, 'Six Sigma Methodology with Fraud Detection', Advances in Data Networks Communications, Computers, Trnava, Slovakia, pp 162-165
70. Tariku, A 2011, 'Mining Insurance Data for Fraud Detection: The Case of Africa Insurance Share Company', Master of Science thesis, school of Information Science, Addis Ababa University, Addis Ababa, Ethiopia
71. Tilahun, M 2009 'Possible Application of Data Mining Techniques to Target Potential VISA Card users in Direct Marketing: The Case of Dashen Bank S.C.' , Master of Science thesis, school of Information Science, Addis Ababa University, Addis Ababa, Ethiopia
72. Lule, B, 2011, "the role of data mining technology in electronic transaction expansion at Dashen bank S.C", Master of Science thesis, school of Information Science, Addis Ababa University, Addis Ababa, Ethiopia

Appendix: rules generated for PART algorithm

PART rules

CUS_STATUS = FINANCIAL_LARGE AND
ADDIS = NO: 0 (85.0/1.0)

CUS_STATUS = FINANCIAL_LARGE AND
SECTOR = Individual AND
BALANCE = HIGHER AND
GENDER = MALE: 0 (48.0)

CUS_STATUS = FINANCIAL_SMALL AND
DISTRICT = NORTH_ADDIS AND
TARGET = BUSSINESS_CUSTOMER AND
SECTOR = Individual: 0 (1082.0/105.0)

CUS_STATUS = FINANCIAL_LARGE AND
SECTOR = Individual: 0 (215.0/38.0)

CUS_STATUS = FINANCIAL_LARGE AND
SECTOR = Government: 0 (38.0)

CUS_STATUS = FINANCIAL_SMALL AND
DISTRICT = EAST_ADDIS AND
MARITIL_STATUS = SINGLE: 0 (753.0/47.0)

CUS_STATUS = FINANCIAL_LARGE AND
DISTRICT = EAST_ADDIS: 0 (12.0)

CUS_STATUS = FINANCIAL_SMALL AND
ADDIS = NO: 0 (1699.0/171.0)

CUS_STATUS = FINANCIAL_SMALL AND
DISTRICT = SOUTH_ADDIS: 0 (700.0/87.0)

CUS_STATUS = FINANCIAL_SMALL AND
DISTRICT = EAST_ADDIS AND
SECTOR = Individual AND
BALANCE = HIGHER: 0 (262.0/26.0)

GENDER = FEMALE AND
MARITIL_STATUS = MARRIED: 0 (174.0/14.0)

CUS_STATUS = FINANCIAL_LARGE AND
DISTRICT = SOUTH_ADDIS: 0 (8.0)

CUS_STATUS = FINANCIAL_SMALL AND
DISTRICT = EAST_ADDIS AND
SECTOR = Individual AND
TARGET = BUSSINESS_CUSTOMER: 0 (267.0/61.0)

CUS_STATUS = FINANCIAL_SMALL AND
SECTOR = Other: 0 (27.0/3.0)

CUS_STATUS = FINANCIAL_SMALL AND
TARGET = OTHER_CUSTOMER AND
BALANCE = LOWER: 0 (111.0/26.0)

BALANCE = HIGHER AND
CUS_STATUS = FINANCIAL_MED AND
GENDER = MALE: 0 (62.0)

CUS_STATUS = FINANCIAL_MED AND
BALANCE = LOWER AND
GENDER = MALE: 0 (160.0)

ADDIS = YES AND
MARITIL_STATUS = SINGLE AND
CUS_STATUS = FINANCIAL_SMALL AND
SECTOR = Individual: 0 (256.0/78.0)

ADDIS = YES AND
MARITIL_STATUS = MARRIED AND
DISTRICT = WEST_ADDIS AND
BALANCE = LOWER: 0 (152.0/59.0)

ADDIS = NO: 0 (49.0/6.0)

DISTRICT = SOUTH_ADDIS: 0 (22.0/3.0)

MARITIL_STATUS = MARRIED AND
DISTRICT = EAST_ADDIS: 0 (149.0/37.0)

SECTOR = Bldng_const: 0 (26.0/6.0)

MARITIL_STATUS = MARRIED AND
BALANCE = HIGHER AND
DISTRICT = WEST_ADDIS: 0 (87.0/25.0)

DISTRICT = NORTH_ADDIS AND
BALANCE = HIGHER: 3 (24.0/6.0)

DISTRICT = NORTH_ADDIS: 0 (36.0/4.0)

BALANCE = HIGHER AND
GENDER = MALE: 0 (28.0/4.0)

DISTRICT = WEST_ADDIS AND
MARITIL_STATUS = OTHER AND
GENDER = MALE: 3 (55.0)

SECTOR = Individual AND
TARGET = BUSSINESS_CUSTOMER: 0 (110.0/48.0)

CUS_STATUS = FINANCIAL_MED AND
BALANCE = MEDIUM: 3 (5.0)

CUS_STATUS = FINANCIAL_LARGE: 3 (4.0)

DISTRICT = EAST_ADDIS AND
TARGET = OTHER_CUSTOMER: 0 (37.0/8.0)

GENDER = MALE AND
BALANCE = LOWER: 3 (10.0/3.0)

GENDER = FEMALE AND
CUS_STATUS = FINANCIAL_SMALL: 3 (4.0)

GENDER = MALE AND
TARGET = RETAIL_CUSTOMER AND
MARITIL_STATUS = OTHER: 0 (59.0/26.0)

DISTRICT = WEST_ADDIS: 3 (29.0/12.0)