



**UNIVERSITY OF NAIROBI**

**CREDIT EVALUATION MODEL USING NAÏVE BAYES CLASSIFIER**

*A Case of a Kenyan Commercial Bank*

**BY**

**CHOGE JOSPHAT KIPCHUMBA**

**P58/70982/08**

**SUPERVISOR**

**Dr. Robert Oboko**

This report is submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Science

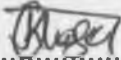
**SCHOOL OF COMPUTING AND INFORMATICS**

**UNIVERSITY OF NAIROBI**

**OCTOBER 2012**

## DECLARATION

I hereby declare that this report is my own work and effort and that it has not been submitted anywhere for any award. Where other sources of information have been used, they have been acknowledged.



.....  
Choge Josphat Kipchumba

P58/70982/2008



.....  
Date

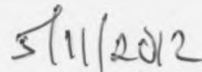
**Supervisor:**



.....  
Dr. Robert Oboko

School of Computing and Informatics,

University of Nairobi



.....  
Date

## DEDICATION

I dedicate this project to my family for their support and encouragement. Above all I dedicate it to God for His blessings all through.

## ACKNOWLEDGEMENTS

Firstly, all praises and glory be to God Almighty for His continued Love and Mercy.

I deeply thank my supervisor, Dr. Robert Oboko, whose advice, contributions and assistance has been invaluable in undertaking this study.

My special gratitude goes to my beloved parents and family for their love, continued support, encouragement and understanding and always there for me when I need them.

This undertaking would not have been possible without the support of my friends. I express my gratitude to all my friends who offered their valuable help.

Finally, to whom I failed to mention, who contributed to this project. Thank you very much.

## ABSTRACT

With the increasing demand for credit facilities for the purpose of development, more and more financial institutions are being established to cater for the need. Acquiring these facilities from the institutions sometimes prove slow and inefficient due to the model adopted for credit evaluation. Reliance on traditional methods for instance, a checklist of bank rules, conventional statistical methods and personal judgment in evaluating credit worthiness makes the process slow and such judgments could be biased. Effective models are required to help mitigate these day-to-day challenges. This study examines the relevance of Naïve Bayes Classifier as an enabling tool in credit decision that can automatically evaluate credit applications based on customer's biographic, demographic and behavioural characteristics. Data used is obtained from one of the commercial banks in Kenya. Feature selection is performed on the data in order to eliminate redundant and less relevant variables. A model using Naïve Bayes Classifier algorithm is developed and its classification performance evaluated. Results show that Naïve Bayes Classifier can be used as a credit decision tool that can speed up and improve efficiency of the process. It also shows that using significant variables improves the model's classification performance. The classification accuracy obtained indicates that the classifier has ability to correctly classify credit applications thereby identifying "bad" credit applications at an early stage hence reducing loss of revenue. Implementation of such model in Kenyan Commercial banks can be helpful for the decision making process.

## TABLE OF CONTENTS

DECLARATION .....	ii
DEDICATION .....	iii
ACKNOWLEDGEMENTS .....	iv
ABSTRACT .....	v
LIST OF TABLES.....	viii
LIST OF FIGURES.....	ix
LIST OF ABBREVIATIONS.....	x
CHAPTER 1 .....	1
INTRODUCTION .....	1
1.1 Problem Statement.....	1
1.2 Objectives .....	2
1.3 Significance of the Project .....	2
1.4 Scope of the Project .....	3
1.5 Definition of Terms.....	3
LITERATURE REVIEW .....	4
2.1 Introduction.....	4
2.2 Credit Evaluation .....	4
2.3 Credit situation in Kenyan Commercial Banks.....	5
2.4 Related Works.....	7
2.5 Naïve Bayes Classifier .....	9
2.6 Advantages of Naïve Bayes Classifier .....	11
CHAPTER 3 .....	12
SYSTEM DEVELOPMENT METHODOLOGY .....	12
3.1 Agile Methodology overview .....	12
3.2 Agile Unified Process .....	13
3.3 Phases of Agile UP Methodology.....	14
3.4 Tools and Materials.....	18
CHAPTER 4 .....	19

<b>RESEARCH METHODOLOGY</b> .....	19
4.1 Introduction.....	19
4.2 Naïve Bayes Model.....	19
4.3 Data Analysis.....	21
4.4 Data Preprocessing and Level of Measurement.....	22
4.5 Feature Selection.....	25
4.6 Model Validation.....	26
4.7 Methodology Assumptions.....	27
<b>CHAPTER 5</b> .....	28
<b>RESULTS AND ANALYSIS</b> .....	28
5.1 Introduction.....	28
5.2 Feature Analysis and Selection.....	28
5.3 NBC Model Results.....	31
5.3.1 NBC validation mode results without feature selection.....	31
5.3.2 NBC validation mode results with feature selection.....	33
5.4 NBC test mode results.....	34
5.5 Comparison of results with previous research.....	35
<b>CHAPTER 6</b> .....	36
<b>DISCUSSIONS, CONCLUSION AND RECOMMENDATIONS</b> .....	36
6.1 Introduction.....	36
6.2 Summary of research findings.....	36
6.3 Conclusion.....	36
6.4 Recommendations.....	37
6.5 Limitations of the Study.....	37
6.6 Suggestions for future work.....	37
<b>REFERENCES</b> .....	38
<b>APPENDICES</b> .....	42

## LIST OF TABLES

Table 1: Model test results .....	18
Table 2: Customer Characteristics.....	21
Table 3: Data Partition Set .....	21
Table 4: Dataset Description and Level of measurement .....	22
Table 5: Chi-Statistic Test Results .....	28
Table 6: Classification Performance on the Validation Dataset.....	34
Table 7: NBC test mode data results .....	35
Table 8: NBC test mode classification performance metrics .....	35



## LIST OF FIGURES

Figure 1: Structure of Naïve Bayes Classifier .....	11
Figure 2: Agile Unified Process (AUP) Lifecycle.....	14
Figure 3: Credit Evaluation Use Case diagram .....	15
Figure 4: Architecture of Credit Evaluation Model using Naïve Bayes Classifier.....	16
Figure 5: User Interface (UI) Flow Diagram.....	17
Figure 6: Schematic of Credit Evaluation Model using Naïve Bayes Classifier .....	19
Figure 7: Confusion Matrix .....	26
Figure 8: Confusion Matrix of validation mode without feature selection .....	31
Figure 9: Performance Metrics of validation mode without feature selection.....	32
Figure 10: Confusion Matrix of validation mode with feature selection .....	33
Figure 11: Performance Metrics of validation mode with feature selection .....	33
Figure 12: Confusion Matrix of NBC test mode .....	34

## **LIST OF ABBREVIATIONS**

- AC** – Accuracy
- ANN** – Artificial Neural Network
- AUP** – Agile Unified Process
- CBK** – Central Bank of Kenya
- FN** – False Negative
- FP** – False Positive
- DT** – Decision Tree
- LDA** – Linear Discriminant Analysis
- LR** – Logistic Regression
- LRA** – Logistic Regression Analysis
- LS-SFM** – Least Squares Support Feature Machine
- MAP** – Maximum a posteriori
- NB** – Naïve Bayes
- NBC** – Naïve Bayes Classifier
- RUP** – Rational Unified Process
- SVM** – Support Vector Machine
- TN** – True Negative
- TP** – True Positive
- UP** – Unified Process
- VBDTM** – Vertical Bagging Decision Tree Model

# CHAPTER 1

## INTRODUCTION

In the currently fast evolving developments especially in the field of Information and Communication Technology, end-users call for better and efficient ways of handling certain processes, for example, banking processes. Most of the human life has been revolving around queuing in order to get some of the processes done especially in the banking industry, where one has to line up to be served. This also applies to obtaining credit facilities in which case an application has to go through several steps before making a sufficient decision on whether to award it or not. Credit facilities constitute a cornerstone of banking industry. The granting of loans by banks is one of the key areas concerning decision problems that need subtle care (Handzi et.al, 2003). Therefore, proper management of these facilities as well as processes involved in acquiring them is vital in ensuring stability and profitability of a bank. Moreover, rapid and continued changes in business environment, banking and credit regulations, marketing strategies of the banks, emerging competition in the lending strategies in banks and the borrowing patterns adopted by customers demand for efficient methods for handling credit processes.

Traditional models are unable to exhibit uncertainty in the banking loan environment (Shachmurove, 2002). A set of credit scoring models that correctly classify loan applications have been developed to support traditional judgmental methods (Malhotra and Malhotra, 2003). Furthermore, a challenge for today's commercial banks is their ability to understand large amounts of information and reveal useful knowledge to improve decision-making. Mitchell and Pavur (2002), notes that modern bank managers are flooded in data and the sustainability of their banks depends on their capabilities to sift through large volumes of data, to extract useful knowledge and enforce this knowledge in their decisions.

To counter these challenges, application of self-training systems that can learn through experience in making decisions, for instance, a system that learns through experience with respect to history of an applicant can be designed to make such decisions. To help eliminate or reduce the time taken to complete these processes and improve efficiency, the end-user is given control to own the process by providing an interface with the capability to provide feedback whether or not a customer is eligible for the credit. The tools and techniques that have been developed in this field have contributed immensely towards decreasing costs and increasing productivity. Savings have been created through decreased task time, fewer user errors, greatly reduced user disruption, reduced biases and reduced burden on support staff among others.

### 1.1 Problem Statement

Credit facilities are of great importance especially in the developing countries. Access to these facilities is vital in ensuring continued growth and development in the various dimensions of life such as social and economic aspects. With respect to the emergence and rapid growth of financial institutions, credit lending has increased significantly both in banking as well as in micro-financing institutions. In spite of the increase in consumer loans and competition in the banking market, most of the commercial banks in Kenya have been reluctant to use machine learning methods

in their decision making. Credit officers rely on traditional methods as a guide in evaluating worthiness for instance, a checklist of bank rules, conventional statistical methods and personal judgment are used to evaluate loan applications. An application submitted for processing goes through the entire process of evaluation which sometimes proves slow due to the model adopted for evaluation. As a result, the applicant waits for some time for a decision to be made with regards to the application. This calls for a quicker or nearly instant evaluation and feedback process.

Further, credit officers need to sift large volumes of data concerning a customer as part of the decision making process. This might in some way be biased because after some time of experience, the credit officers develop the ability to judge the worthiness of a loan decision. Given the absence of objectivity, such judgment is biased, ambiguous, and nonlinear and humans have limited capabilities to discover useful relationships or patterns from a large volume of data.

Thus, effective tools are required to help mitigate these day-to-day challenges. Utilizing the capability of Naïve Bayes Classifiers can help handle the complexity of these processes. In view of this, presented here is a model based on the use of Naïve Bayes Classifier to help speedup and improve efficiency of credit evaluation procedures.

## **1.2 Objectives**

- i. To examine the role of Naïve Bayes Classifier as an enabling tool in Credit evaluation
- ii. To Eliminate bias in credit evaluation process using Naïve Bayes Classifier
- iii. To design a conceptual model that improves the quality of credit evaluation process.

## **1.3 Significance of the Project**

With the increasing demand for credit facilities for the purpose of development, more and more financial institutions are being established to cater for the need. Acquiring these facilities from the institutions sometimes proves inefficient. Thus, this study develops a model that demonstrates the capabilities of Naïve Bayes Classifier as a tool used in credit evaluation to help improve efficiency of the process. The main challenge during credit evaluation process is due to the fact that most of the activities revolve around manual processes performed by credit officers. Given large volumes of data for example, applications submitted, a credit officer needs to sift through history of a given applicant in order to make conclusion, which sometimes might be biased. At the end of the process, much time will have already been used. This study helps identify an efficient tool to mitigate these issues.

Given competitiveness in the financial and banking industry, businesses will be required to adopt automatic credit evaluation methodology to remain competitive in their ability to meet customer needs and expectations. Credit evaluation offers great benefits to the credit department such as accuracy through exemption from human error, consistency in terms of using same set of rules in evaluation and significant reduction in time required to undertake the process.

#### 1.4 Scope of the Project

The study examines the application of Naïve Bayes Classifier and its relevance to the Credit Evaluation domain in a Kenyan commercial bank. It examines significant factors for credit evaluation and applies them in the classification of the credit in the decision process. Further, the study will lead to the development of the model using Naïve Bayes Classifier.

#### 1.5 Definition of Terms

**Credit Evaluation (Approval)** - is the process a business or an individual must go through to become eligible for a loan.

**Naïve Bayes Classifier (NBC)** - A probabilistic classifier based on applying Bayes' Theorem with strong (naive) independence assumptions.

**Conditional Independence** – A simplifying assumption that attribute values are independent given a target value.

**Maximum a posterior (MAP)** – This is the maximally probable hypothesis from amongst a set of generated hypotheses

**Confusion Matrix** – It is an  $n$ -dimensional square matrix, where  $n$  is the number of distinct target values

**Training set** - A set of examples used for learning. It is used to obtain the pattern in data

**Validation set** - A set of examples used to tune the parameters of a classifier

**Testing Set** - A set of examples used only to assess the performance (generalization) of a fully-specified classifier

**Sensitivity** - Measures the proportion of actual positives which are correctly identified as such (i.e. accuracy on the class Positive)

**Specificity** - Measures the proportion of negatives which are correctly identified (i.e. accuracy of the class Negative)

**Accuracy (AC)** - is the proportion of the total number of predictions that were correct.

## CHAPTER 2

### LITERATURE REVIEW

#### 2.1 Introduction

Credit evaluation is an important process in credit lending institutions. Given the increasing demand for credit facilities as well as rapid growth of financial institutions especially in Kenya, has led to the need for effective tools and techniques of handling credit evaluation procedures. These tools should also be capable of speeding up these processes to ensure quick and quality delivery of service to the customers in the financial sector. The institutions employ credit officers to make credit decisions or recommendations. The lending officer usually makes approval for applications that are worthy of giving a loan. These officers are given some guidelines to direct them in evaluating the worthiness of credit applications depending on the borrowers' characteristics. Further, credit officers sifting large volumes of data concerning a customer as part of the decision making process, which might in some way be biased can significantly slow down the process. This practice is inefficient, inconsistent and non-uniform.

In the recent past, researchers have incorporated Naïve Bayes Classifier techniques into easy-to-use toolsets in turn enabling the development of decision support systems in a diverse set of application domains some of which include medical diagnosis, safety assessment, equipment fault diagnosis, credit analysis, software quality and procurement. With the technological advancement, it follows that end-users, rather than just researchers are now able to develop and deploy their own Naïve Bayes Classifier solutions. As a result, these methods are becoming acceptable in mainstream business practice with recent studies providing impressive return on investment from these techniques. For an effective management of processes in day-to-day activities, end-users require systems that provide a nearly instant response to their queries. Credit evaluation by financing institutions is one of the key areas with decision problems that need efficient methods of handling the processes. Credit evaluation method such as in the proposed model has not been commonly used in the Kenyan financial institutions. Therefore, this study provides a credit evaluation model that utilizes capability of Naïve Bayes Classifier as an enabling tool for decision making.

#### 2.2 Credit Evaluation

One of the key decisions to make in financial institutions is to decide whether or not to grant a loan to a customer and this form credit evaluation as the final stage of consumer credit process. This decision basically boils down to a binary classification problem which aims at distinguishing good payers from bad payers. The decision must be evaluated on a case-by-case basis whereby the institution needs to gather and evaluate financial information, decide whether to grant credit and if so how much, and communicate the decision to the customer in a timely manner. Credit evaluation and approval is the process a business or an individual must go through to become eligible for a loan or to pay for goods and services over an extended period. It also refers to the process businesses or lenders undertake when evaluating a request for credit (Hillstrom and Hillstrom, 2002). This process includes collecting, analysing and classifying different credit elements and variables to assess the credit decisions. The quality of bank loans is the key determinant of competition, survival and profitability. Granting credit approval depends on the willingness of the creditor to lend money in the current economy and that same lender's assessment of the ability and

willingness of the borrower to return the money or pay for the goods obtained – plus interest – in a timely fashion. Typically, small businesses must seek credit approval to obtain funds from lenders, investors, and vendors, and also grant credit approval to their customers (Hillstrom and Hillstrom, 2002).

The objective of credit evaluation is to:

- a) Provide the best tailored loan for the customer and a quality loan for the lender
- b) Ensure compliance with regulations and bank policy
- c) Keep the goodwill of the consumer; and
- d) Ensure that the level of risk is acceptable

In order for banking institutions to realize these objectives and ensure their competitive edge in the dynamic society, measures need to be taken into consideration especially in credit evaluation procedures. Banks have credit policies that guide them in the process of awarding credit. Credit control policy is the general guideline governing the process of giving credit to bank customers. The policy sets the rules on who should access credit, when and why one should obtain the credit including repayment arrangements and necessary collaterals. The method of assessment and evaluation of risk of each prospective applicant are part of a credit control policy. Credit evaluations are not based on a single factor but upon how an applicant matches up to a set of lending criteria laid down by the lender. For instance, certain information is gathered about the credit applicant and such include name and address, bank or trade references, employment and income information and financial statements among others. The goal is to form an assessment of the character, reputation, financial situation and collateral circumstances of the applicant.

A firm's credit policy may be lenient or stringent. In the case of a lenient policy, the firm lends liberally even to those whose credit worthiness is questionable. This leads to high amount of borrowing and high profits, assuming full collections of the debts owed. With the stringent credit policy, credit is restricted to carefully determined customers through credit appraisal system. This minimizes costs and losses from bad debts but might reduce revenue earning from loans, profitability and cash flow (Bonin and Huang, 2001).

### **2.3 Credit situation in Kenyan Commercial Banks**

While commercial banks in Kenya have faced difficulties over the years for a multitude of reasons, the major cause of serious financial problems continues to be directly related to credit standards for borrowers, poor portfolio risk management or lack of attention to changes in the economic circumstances and competitive climate (Central Bank of Kenya Annual Supervision Report, 2000; Central Bank of Kenya Risk management guidelines, 2005). Results from the risk management survey done in 2010 by Central Bank of Kenya (CBK) in Kenyan banking institutions show that market risk (in this context, comprising of equity risk, interest rate risk, currency risk and commodity risk) was the risk facing most institutions, having been identified as a principal risk by all forty three respondents (100%), followed by credit and operational risks which were identified by 95% and 93% of respondents respectively. In the 2004 survey, credit risk was the most widely identified risk (97% of respondents). The current survey's results show

that credit risk is still essentially as significant as it was in the earlier survey, which may be attributed to the large proportion of banks' asset portfolio made up of loans and advances to customers (CBK Risk management survey, 2010). This means that credit management is still a challenge facing Kenyan banking institutions.

The credit decision should be based on a thorough evaluation of the risk conditions of the lending and the characteristics of the borrower. With the introduction and implementation of Basel II accord in Kenyan banking institutions, proper policies regarding credit management need to be developed. The Basel Committee on Banking Supervision, with its revised capital adequacy framework (Basel Committee on Banking Supervision, 2005) – commonly known as Basel II – proposes a more flexible capital adequacy framework to encourage banks to make ongoing improvements in their risk assessment capabilities.

Basel II consists of three pillars:-

1. **Pillar I** - Minimum Capital Requirements – Banks are required to set aside adequate capital to cover the credit, market and operational risks they face.
2. **Pillar II** - Supervisory Review – Under this pillar, banks are required to set aside adequate capital to cover all the business risks they face including those outside credit, market and operational risks. Supervisors on the other hand are required to assess the internal capital adequacy process that banking institutions have put in place.
3. **Pillar III** – Market Discipline – Under this pillar, banks are required to disclose to the public, their risk management frameworks i.e. how they identify, measure and mitigate the risks they face.

The Basel Committee believes that all banks should be subject to a capital adequacy framework comprising minimum capital requirements, supervisory review, and market discipline. The objective is reached by giving banks a range of increasingly sophisticated options for calculating capital charges. Banks will be expected to employ the capital adequacy method most appropriate to the complexity of their transactions and risk profiles. For credit risk, the range of options begins with the standardized approach and extends to the internal rating-based (IRB) approach. The standardized approach is similar to the current Accord: banks will be expected to allocate capital to their assets based on the risk weights assigned to various exposures. It improved on the original Accord by weighting those exposures based on each borrower's external credit risk rating. Clearly, the IRB approach is a major innovation of the New Accord: bank internal assessments of key risk drivers are primary inputs to the capital requirements. For the first time, banks will be permitted to rely on their own assessments of a borrower's credit risk. The close relationship between the inputs to the regulatory capital calculations and banks' internal risk assessments will facilitate a more risk sensitive approach to minimum capital. Changes in a client's credit quality will be directly reflected in the amount of capital held by banks (Angelini et al, 2007).

It is worth noting that the Central Bank of Kenya (CBK) has made considerable progress in fulfilling the prerequisites of Basel II and it is now appropriate to begin engaging the banking sector on the new capital framework. Further, a strong banking sector is essential to fulfill the national development aspirations encapsulated



in Kenya's current development blueprint, Vision 2030. Under the Vision, the banking sector is expected to play a catalytic role in mobilizing the substantial resources required to push Kenya to "new frontiers" of development. The Vision also seeks to transform Kenya into a "financial services hub" for the Eastern Africa Region. This will in turn require the formulation of a "world class" enabling legal and regulatory framework. In the current supervisory context, "world class" is set by the Basel Committee and it is anticipated that Basel II will in the near term become the global supervisory standard in the same way that Basel I was accepted and adopted by more than 100 countries in the world (CBK Basel II Implementation Survey, 2008).

#### **2.4 Related Works**

In the last decade, various quantitative methods have been applied by banking and financing institutions to the process of evaluating consumer loans as ways of improving accuracy. This increased need for efficient methods of credit evaluation is as a result of economical crisis causing the institutions to have more attention to credit risk (Wang et al, 2011). Credit evaluation is a binary classifier used to classify the applicants into two types; applicants with good and bad credit (Chen et al, 2010). Applicants with good credit have great possibility to repay financial obligation and those with bad credit have high possibility of defaulting (Wang et al, 2011). To reduce risks as a result of this, various credit evaluation models have been developed to support credit decisions (Hsieh and Hung, 2010).

Linear Discriminant Analysis (LDA) is the earliest model used for the credit scoring. LDA finds a linear combination of features which characterizes or separates two or more classes of objects or events. Utilization of LDA to credit evaluation has often been criticized due to the assumptions of linear relationship between input and output variables, which seldom holds, and it is sensitive to deviations from the multivariate normality assumption (West, 2000). Logistic Regression (LR) is another common model used in credit scoring. It is a technique for predicting a discrete outcome from a set of variables that may be continuous, discrete, and dichotomous or a mix of any of these. In contrast to LDA, the relationship between the predictor and response variables is not a linear function. The advantages of this method are that the logistic regression does not assume linearity of relationship between the independent variables and the dependent, does not require normally distributed variables. The main challenge in using this model is that the independent variables need to be linearly related to the logit of the dependent variable.

Support Vector Machine (SVM) is a classifier technique involving three elements namely; a score formula which is a linear combination of features selected for the classification problem, an objective function which considers both training and test samples to optimize the classification of new data and an optimizing algorithm for determining the optimal parameters of training sample objective function. The advantages of this method are that SVM requires no data structure assumptions such as normal distribution and continuity, they can perform a nonlinear mapping from an original input space into a high dimensional feature space and also capable of handling both continuous and

categorical predictions. However, its weaknesses are that, it is difficult to interpret unless the features interpretable and standard formulations do not contain specification of business constraints (Ravi, 2008).

Artificial Neural Networks have been proved to outperform conventional statistical model in terms of classification accuracy and the limitation of reduction. However, structure selection is always the difficult part of these methods since unsuitable structure will lead to local minima or over-fitting. Baesens *et al* (2003) pointed out that the main reason behind the lack of applying neural networks methods in credit risk evaluation industry is the lack of explanatory capabilities of these techniques and therefore the enhancement of the transparency of neural networks is one of the key factors of their successful deployment. This explanatory capability plays a pivotal role in credit-risk evaluation as the evaluator may be required to give justification as to why a certain credit application is approved or rejected. Steiner and Neto (2006) use Artificial Neural Network tools to make credit risk evaluation. They emphasize the importance of universal approximation property and its high prediction accuracy rate. The main challenge identified is that it is not easy to understand how the model reached the solution. Computational burden in Artificial Neural Network is another challenge as it requires large amount of data for training which also lead to longer training time.

Abramowicz (2003) develops an idea of decision support tool in credit scoring domain using Bayesian network. The study was aimed at applicability of Bayesian belief networks within the procedures of working capital credit scoring conducted in commercial banks. The challenge of this approach is that all the branches in the network must be calculated in order to calculate the probability of any one branch. This means that generation and maintenance of conditional probability tables in the model can be computationally costly. Further, this cost can also be attributed to the fact that each node in a Bayesian network can have multiple parent and child nodes, and thus multiple ancestor and descendant nodes, thus evaluating Bayesian networks is more complex than performing a single calculation with Bayes' theorem. Lai *et al* (2006) assessed credit risk with Least Squares Support Feature Machine (LS-SFM). The results obtained showed that the LS-SFM is an effective classification tool for credit assessment. In addition, LS-SFM can deliver the users how important each feature is.

Recent research works have also indicated that ensemble classification models or hybrid models often deliver better results for credit evaluation. A hybrid approach by Chen *et al* (2010), where they combined SVM with other tools including conventional statistical Linear Discriminate Analysis (LDA), Decision tree, Rough sets and F-score approaches, showed that LDA with SVM outperforms SVM alone. Zhang and Zhou (2010), propose a novel vertical bagging decision trees model (VBDTM) built on several base learners that consist of various types of machine learning single models, rule extraction models, two stages models, hybrid models and aggregation models and experimental results indicate that the novel ensemble method outperforms other methods.

Further, an ensemble model developed by Wang *et al* (2011) comprised of three methods namely Logistic Regression Analysis (LRA), Decision Tree (DT), Artificial Neural Network (ANN) and Support Vector Machine

(SVM). From these hybrid approaches, it is evident that a level of accuracy can be achieved by using ensemble models. However, hybrid or ensemble models often times have a drawback being that they involve a fair amount of computational complexity which subsequently increases processing times (time costs). Moreover the evidence is inconclusive that they do obtain better results than single models.

Models such as Artificial Neural Networks (Hsieh, 2005; Tsai et al, 2008), decision trees (Huang et al, 2007) and Zhang et al (2010), genetic programming On et al (2005), Support Vector Machines (Wang et al, 2011; Chen et al, 2010; Huang et al, 2007; TunLi et al, 2006) and logistic regression analysis (Wang, 2011; Huang et al, 2007; Thomas 2000) have been developed and used to better the accuracy of evaluation but the choice of which model is the best still remains an issue in financial research. Previous studies that have been done concerning performances of these models indicate that some have proven higher levels of accuracy (Zhang et al, 2010; Leea et al, 2006; Tsai et al 2008; Hsieh et al, 2010) while other classifiers may not have. Huang and Chen (2009) note that this variation is as a result of the nature of the dataset as some classifiers behave differently with different dataset. Further, Wang et al (2011) records that the choice of the best method is related on the details of the problem, the data structure, the used characteristics, the extent to which it is possible to segregate the classes by using those characteristics and the objective of the classification.

In this study, a model using Naïve Bayes Classifier is proposed to help overcome these overheads. Naïve Bayes (NB) models are popular in machine learning applications, due to their simplicity in allowing each attribute to contribute towards the final decision equally and independently from the other attributes. This simplicity equates to computational efficiency, which makes NB techniques attractive and suitable for many domains including credit evaluation. The conditional independence assumption, even when violated, does not degrade the model's predictive accuracy significantly and this makes NB-based systems offer quick training, fast data analysis and decision making, as well as straightforward interpretation of test results. The other key feature is that NB models are less data intensive meaning that they don't require a lot of data for training.

## 2.5 Naïve Bayes Classifier

A Naïve Bayes Classifier (NBC) is a simple probabilistic classifier based on Bayes' rule with strong (naive) independence assumptions. Naïve Bayes Classifier is used mainly for performing classification tasks. Considering  $D$  to be the data we've seen so far and  $h$  being a possible hypothesis, then Bayes' theorem definition is given by:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Where:

$P(h)$ : Prior probability of hypothesis  $h$  - Prior

$P(D)$ : Prior probability of training data  $D$  - Evidence

$P(D|h)$ : Conditional Probability of  $D$  given  $h$  - Likelihood

$P(h|D)$ : Conditional Probability of  $h$  given  $D$ - Posterior probability

In the general case, we have  $K$  mutually exclusive and exhaustive classes;  $h_i, i = 1 \dots K$ ;  $P(D|h_i)$  is the probability of seeing  $D$  as the input when it is known to belong to class  $h_i$ . The posterior probability of class  $h_i$  can be calculated as-

$$P(h_i|D) = \frac{P(D|h_i)P(h_i)}{P(D)} = \frac{P(D|h_i)P(h_i)}{\sum_{i=1}^K P(D|h_i)P(h_i)}$$

Source: (Alpaydin, 2004)

The above formula can be summarized as:

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}.$$

Since the denominator of the fraction does not depend on the class variable, the numerator is considered and thus the latter is equivalent to the joint probability model. This is represented as:

$$P(D|h_i)P(h_i)$$

The Naive Bayes Classifier is based on the simplifying assumption that the attribute values are conditionally independent given target value, Mitchell (1997). This assumption is called class conditional independence. It is made to simplify the computation involved and this is why it is considered "naïve". In other words, the assumption is that given the target value of the instance, the probability of observing the conjunction  $a_1, a_2, \dots, a_n$  is just the product of the probabilities for the individual attributes:

$$P(a_1, a_2, \dots, a_n|v_j) = \prod_i P(a_i|v_j)$$

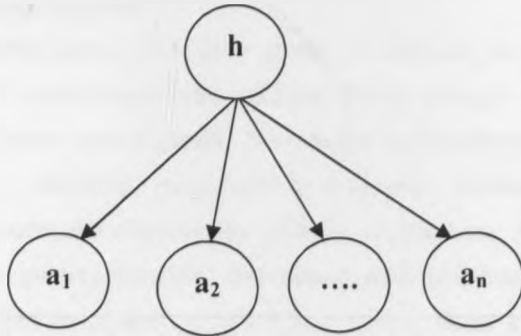
Source: (Mitchell, 1997)

Naïve Bayes Classifier incorporates a decision rule. One common rule is to pick the maximally probable hypotheses from amongst the set of generated hypotheses. This is known as Maximum a posteriori (MAP) decision rule. It is noticed that in a Naive Bayes classifier the number of distinct  $P(a_i|v_j)$  terms that must be estimated from the training data is just the number of distinct attribute values times the number of distinct target values. This is a much smaller number than if we were to estimate the  $P(a_1, a_2, \dots, a_n|v_j)$  terms as needed for Bayesian theory.

Incorporating the assumption, the Naïve Bayes Classifier is given by:

$$v_{NB} = \operatorname{argmax}_{v_j \in V} P(v_j) \prod_i P(a_i | v_j)$$

Figure 1 shows the structure of Naïve Bayes Classifier.



**Figure 1:** Structure of Naïve Bayes Classifier

NBC is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods (Eyheramendy *et al*, 2003 and Rish, 2001). In some domains its performance has been shown to be comparable to that of neural network and decision tree learning. The assumption in using NBC makes it an effective classification tool that is easy to interpret. It is best employed when faced with the problem of ‘curse of dimensionality’, when the number of predictors is very high.

### 2.6 Advantages of Naïve Bayes Classifier

Suitability and extensive use of NBC as an enabling tool for financial business decisions such as credit evaluation have been attributed to certain contributing factors some of which include the following:

- Easy to implement - It requires a small amount of training data to estimate the parameters necessary for classification
- The decoupling of the class conditional feature distributions means that each distribution can be independently estimated as a one dimensional distribution. Naïve assumption of class conditional independence helps reduce computational cost.
- Highly practical Bayesian learning method and is particularly suited when dimensionality of the input is so high
- Its operation is simple and intuitive, relying only on basic laws of probability
- It accommodates limited information as encountered in the problem domain, thus allows a broader set of model parameters to be used, since the model does not require observations for all independent variables.
- Being explicitly probabilistic, it reports results in a form that can easily be interpreted.
- It is robust to outliers
- It can account for information received at varying points in time.

## CHAPTER 3

### SYSTEM DEVELOPMENT METHODOLOGY

System development methodology adopted in the design of the model is Agile methodology.

#### 3.1 Agile Methodology overview

Agile software development refers to a group of software development methodologies based on iterative development, where requirements and solutions evolve through collaboration between self-organizing cross-functional teams. Conboy and Fitzgerald (2004) define Agile software development as the continual readiness of an entity to rapidly or inherently, proactively or reactively, embrace change, through high quality, simplistic, economical components and relationships with its environment. Agile methodologies provide a collection of foundational values, guiding principles and development practices that address the challenges of prescriptive, weighty methodologies to produce software in a lighter, faster and more people-centric manner. Thus, this methodology aims to develop and implement software quickly in close cooperation with the customer in an adaptive way so that it is possible to react to changes set by the changing business environment and at the same time maintain effectiveness and efficiency.

Agile software development methodology is guided by an Agile Manifesto. The Manifesto articulates core values and principles that guide agile methodologies. The Manifesto for Agile software development is highlighted as follows:

*We are uncovering better ways of developing software by doing it and helping others to do it. Through this work we have come to value:*

- **Individuals and interactions** over processes and tools
- **Working software** over comprehensive documentation
- **Customer collaboration** over contract negotiation
- **Responding to change** over following a plan

*That is, while there is value on the items on the right, we value the items on the left more (Agile Alliance, 2002; Cockburn, 2001).*

This implies that Agile methodology is focused on promoting an environment of adaptation, teamwork, self-organization, rapid delivery and client focus. In Agile methodology, there are defined 12 principles that further explicate what it is to be Agile. They include:

1. Our highest priority is to satisfy the customer through early and continuous delivery of valuable software.
2. Welcome changing requirements, even late in development. Agile processes harness change for the customer's competitive advantage.
3. Deliver working software frequently, from a couple of weeks to a couple of months, with a preference to the shorter timescale.
4. Business people and developers must work together daily throughout the project.

5. Build projects around motivated individuals. Give them the environment and support they need, and trust them to get the job done.
6. The most efficient and effective method of conveying information to and within a development team is face-to-face conversation.
7. Working software is the primary measure of progress.
8. Agile processes promote sustainable development. The sponsors, developers, and users should be able to maintain a constant pace indefinitely.
9. Continuous attention to technical excellence and good design enhances agility.
10. Simplicity--the art of maximizing the amount of work not done--is essential.
11. The best architectures, requirements, and designs emerge from self-organizing teams.
12. At regular intervals, the team reflects on how to become more effective, then tunes and adjusts its behavior accordingly. (Agile Alliance, 2002; Cockburn, 2001)

The selection of this methodology for the model development in this study is attributed to the following factors:

- Its basis is on an iterative and incremental approach. This means that the requirements and the solutions evolve together.
- It is a conceptual framework that promotes foreseen interactions throughout the development cycle.
- Adaptability to change – Allows changes and/or new ideas to be incorporated easily
- Allows for easy bug identification and reporting to ensure early elimination. This means that at a given stage of development, testing can be done.

### 3.2 Agile Unified Process

Agile method used in the design of the model is Agile Unified Process (AUP). Agile Unified Process (UP) is a streamlined approach based on IBM's Rational Unified Process (RUP). RUP provides a disciplined approach to assigning tasks and responsibilities within a development organization. Its goal is to ensure the production of high-quality software that meets the needs of its end-users, within a predictable schedule and budget. Lifecycle of Agile UP is serial in the large, iterative in the small, delivering incremental releases over time. Agile UP describes a simple, easy to understand approach to developing business application software using agile techniques and concepts (Scott, 2005).

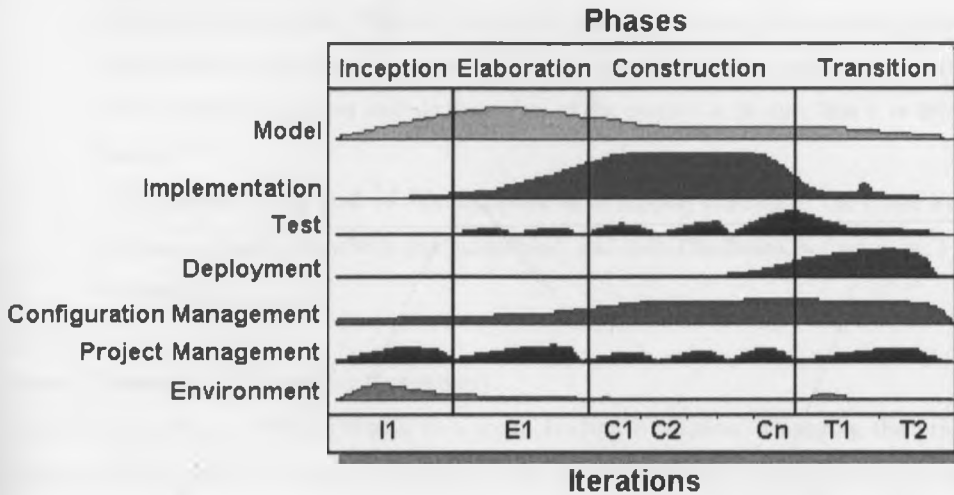
Agile UP is based on the following principles:

1. **Your staff knows what they're doing.** People aren't going to read detailed process documentation, but they will want some high-level guidance and/or training from time to time.
2. **Simplicity.** Everything is described concisely using a handful of pages, not thousands of them.
3. **Agility.** The Agile UP conforms to the values and principles of the Agile Alliance.
4. **Focus on high-value activities.** The focus is on the activities which actually count and not every possible thing that could happen to you on a project.

5. **Tool independence.** You can use any toolset that you want with the Agile UP.
6. **You'll want to tailor this product to meet your own needs.**

### 3.3 Phases of Agile UP Methodology

Agile UP methodology has four main phases as shown in Figure 2. They include Inception, Elaboration, Construction and Transition.



Source: (Scott, 2005)

Figure 2: Agile Unified Process (AUP) Lifecycle

The serial nature of Agile UP is captured in its four phases:

1. **Inception** - Domain understanding. This is the initial step in which a problem to be solved is defined. The goal is to identify the initial scope of the project and a potential architecture for the system. It also defines high level system requirements that shall guide the development process.
2. **Elaboration** - Model the solution to the problem. The goal is to prove the architecture of the system from the perspective of specified requirements.
3. **Construction** - Implement the solution to the problem. The goal is to build working software on a regular, incremental basis which meets the highest-priority needs of the project stakeholders.
4. **Transition** - Verify the solution. The goal is to validate and deploy the system into a production environment.

Figure 2 also shows disciplines that define activities performed in the defined phases. The disciplines are:

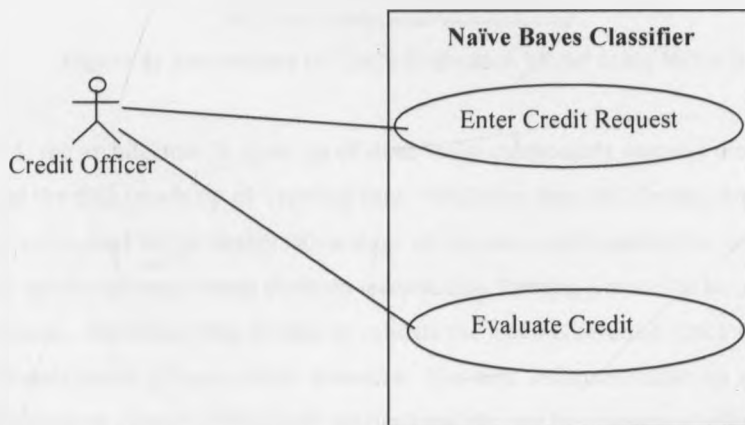
- **Model** - The goal of this discipline is to understand the business of the organization, the problem domain being addressed by the project, and to identify a viable solution to address the problem domain.
- **Implementation** - The goal of this discipline is to transform your model(s) into executable code and to perform a basic level of testing, in particular unit testing.



- **Test** - The goal of this discipline is to perform an objective evaluation to ensure quality. This includes finding defects, validating that the system works as designed, and verifying that the requirements are met.
- **Deployment** - The goal of this discipline is to plan for the delivery of the system and to execute the plan to make the system available to end users.
- **Configuration Management** - The goal of this discipline is to manage access to your project artifacts. This includes not only tracking artifact versions over time but also controlling and managing changes to them.
- **Project Management** - The goal of this discipline is to direct the activities that takes place on the project. This includes managing risks, directing people (assigning tasks, tracking progress, etc.), and coordinating with people and systems outside the scope of the project to be sure that it is delivered on time and within budget.
- **Environment** - The goal of this discipline is to support the rest of the effort by ensuring that the proper process, guidance (standards and guidelines), and tools (hardware, software, etc.) are available for the team as needed.

### Phase 1: Domain understanding (Inception)

Credit evaluation is a critical process in a credit lending institution. Managing the processes involved in credit evaluation is vital and thus requires a proper model designed that is able to improve process efficiency. The aim of this study is to design a conceptual model that uses Naïve Bayes Classifier to evaluate credit applications. The model provides capability for the credit officers to select features for credit approval and apply them in classifying credit applications. Figure 3 shows a generalized use case diagram for the model.



**Figure 3:** Credit Evaluation Use Case diagram

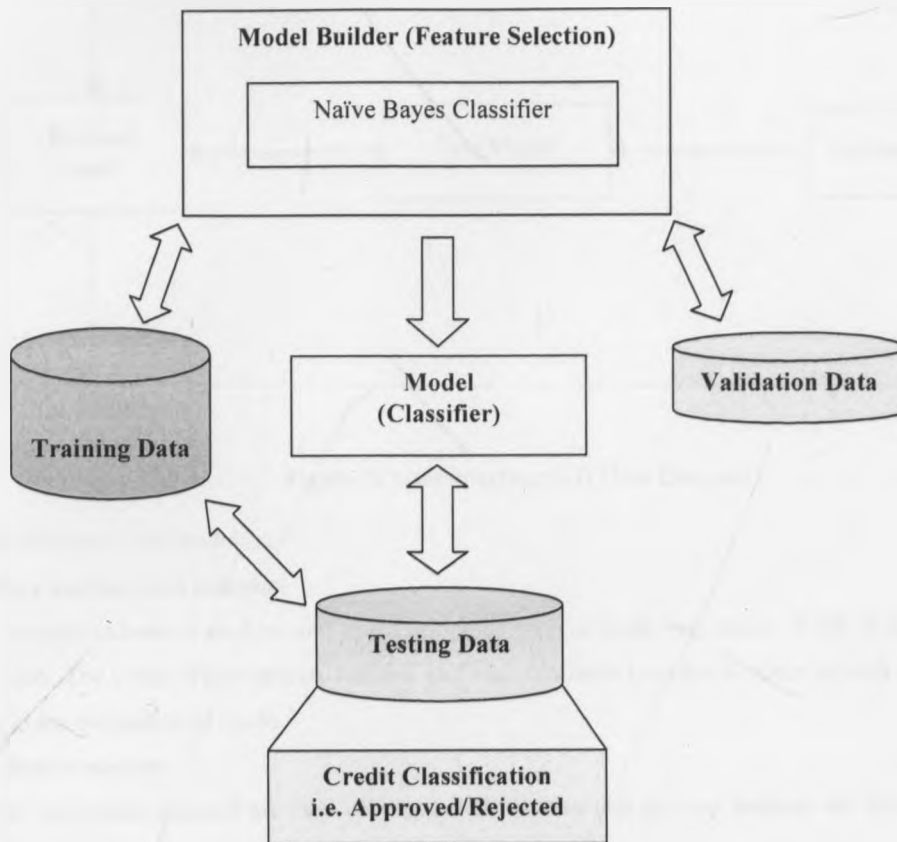
From Figure 3, the main actor is the credit officer and the summary of the flow of events are as follows:

**Goal:** Evaluate credit request

1. Credit officer enters details of a credit application
2. The system evaluates the credit request
3. System generates feedback of the result of evaluation

## Phase 2: Model the solution to the problem (Elaboration)

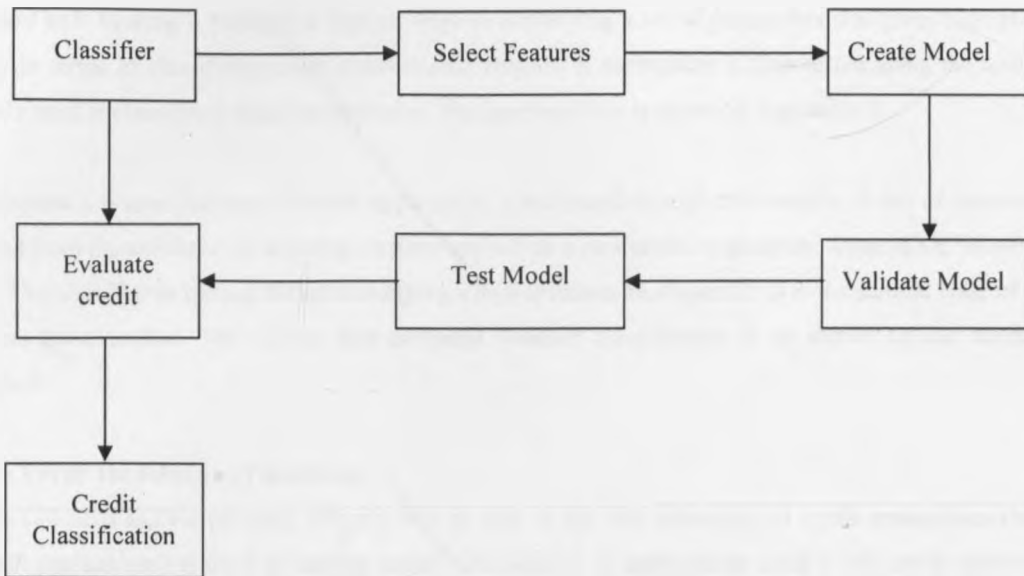
The overall architecture of the system is as shown in Figure 4.



**Figure 4:** Architecture of Credit Evaluation Model using Naïve Bayes Classifier

From Figure 4, the architecture is made up of three main components namely; the Model builder, validated model (classifier) and the data (made up of Training data, Validation data and Testing data). The Model builder is used to build a model to be used in the classification stage where new credit application instances are being processed. The model builder selects optimum credit decision features thus forming a model to be used in the actual classification of credit applications. Validation data is used to validate the features selected. Once the model has been validated, it's used in the classification of new credit instances. The new instances make up the Testing dataset. The model performs classification of each of the credit applications into one two classes as either approved or rejected.

Figure 5 is a User Interface (UI) flow diagram showing the overall architecture of the system. It explores the flow of the system between the major user interface elements.



**Figure 5: User Interface (UI) Flow Diagram**

The main modules in the model are:

**1. Feature analysis and selection**

This module is used to analyze and select attributes used in credit evaluation. A list of features or attributes is provided. The credit officer selects features and analyzes them in order to come up with a significant set to be used in the evaluation of credit.

**2. Validation module**

Model parameters selected are then validated. This ensures that optimal features are determined and validated for use in the classification of credit applications.

**3. Testing module**

The model provides a capability to test selected optimal features. Optimal parameters are applied in the evaluation of a new credit application.

**4. New Classification**

A new credit application is entered and evaluated through this function. The system applies the selected and tested features to classify a new credit instance.

**Phase 3: Model Construction**

This step entails the actual design of the model. System specifications from phase 2 are used to develop the model. Below are user interfaces of the major modules of the model.

**Feature analysis** - The interface provides the capability to select features for credit evaluation. The module generates an analysis of the attributes with respect to their significance in credit decision. Appendix A shows the screen display of this feature.

**Validation and Testing** – Validation feature helps in identifying a set of parameters that gives high predictability accuracy in terms of classifying credit applications. This set of parameters is then tested using the testing module before it's used in classifying credit applications. The user interface is shown in Appendix B.

**Classification** – A new instance of credit application is evaluated through this module. A set of features validated and tested from the validation and testing step are applied on a new credit application. Appendix C shows the screen display. The classifier is trained before classifying a new instance. In Appendix D is the sample code of the routine that trains the classifier. The routine that performs instance classification is as shown by the sample code in Appendix E.

**Phase 4: Verify the solution (Transition)**

Model is validated and tested using different sets of data. A test data consisting of credit applications (both “good” and “bad” applications) is used to test the model. The number of applications used is 707 credit applications with 204 “bad” applications (Rejected applications) and 503 “good” applications (Approved applications). The aim of the test is to obtain the proportion of the credit applications correctly classified to their respective classification. For instance, those “bad” applications correctly classified by the model as such and “good” applications classified as good. Table 1 shows the results of the test performed.

**Table 1: Model test results**

Good Applications	Correctly Classified	503	452
	Misclassified		51
Bad Applications	Correctly Classified	204	139
	Misclassified		65
		<b>707</b>	<b>707</b>

The results show that out of 503 approved applications, the model is able to classify 452 as belonging to the class approved and misclassifying 51 applications. On the other hand, out of 204 rejected applications, the model is able to classify 139 of them as rejections and wrongly classifying 65 applications as rejections. This means that the model is able to classify credit applications and thus can be used in credit decision making process.

**3.4 Tools and Materials**

Tools and materials used for the development of the model include:

- Microsoft Visual Basic (2008)
- Microsoft SQL Server 2008 – Express Edition

## CHAPTER 4

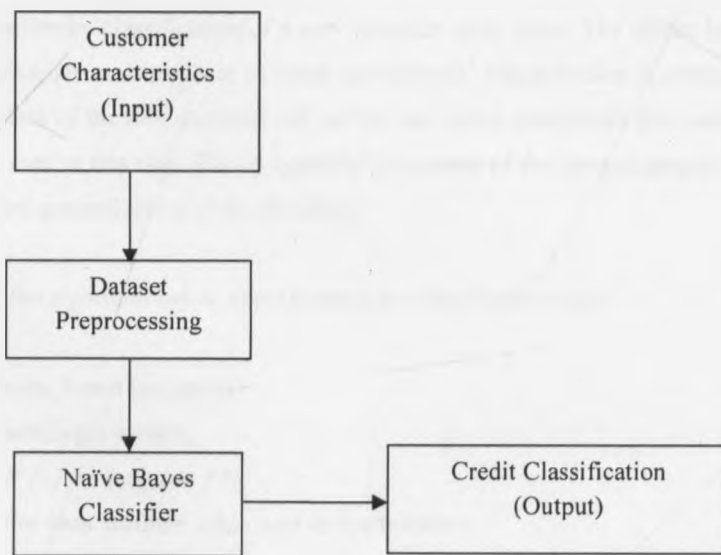
### RESEARCH METHODOLOGY

#### 4.1 Introduction

Credit evaluation is one of the key functions in the financing institutions. Proper tools and methods are needed to ensure that the process is undertaken in an efficient manner. Since this is part of the initial steps in acquiring credit facilities, it will greatly help reduce related credit risks; for example, determining the credit strength of borrowers, approximating the probability of default and reducing the risk of non-payment to an acceptable level. Using credit evaluation models such as Naïve Bayes Classifier, as in this study, have the potential in reducing the inconsistency of credit decisions as well as adding effectiveness to the assessment of lending risks. Such models provide the ability to evaluate a set of variables to determine creditworthiness of similar credit applications.

#### 4.2 Naïve Bayes Model

In this study, a model that uses Naïve Bayes Classifier has been developed. To achieve this, the model examines the guiding factors for credit evaluation. The model is a desktop application to be used by the credit officers. Figure 6 shows the schematic structure of the model.



**Figure 6:** Schematic of Credit Evaluation Model using Naïve Bayes Classifier

As shown in Figure 6, the customer data containing behavioural (such as past credit history), biographic and demographic (for example, gender and marital status) information are pre-processed and fed as inputs to the Naïve Bayes classifier. The classifier undergoes through three main phases; Training phase, Validation and Testing phases.

### 4.2.1 Training Phase

During the training phase the input data along with the known classification are cycled through the classifier and in this process the classifier learns the mapping relationship between the input data attributes and the class associated with it. This means that the classifier establishes relationships between the consumer input attributes and their credit classification. The model computes the probability of each class based on the probability distribution in the training data by taking into account the probability of each attribute by multiplying their probabilities.

The data used in this mode is known as the Training dataset. This is typically 70 percent of the sample dataset.

### 4.2.2 Validation Phase

The next phase after training is the Validation phase. To determine the optimal model to be used for testing, the classifier is validated using a different dataset that has not been encountered during the previous stage (Training stage). This ensures that the performance of the model in actual classification is guaranteed. Validation phase requires a validation dataset. The validation dataset is typically 10 percent of the total sample dataset. This is a set of data that is used to optimize the classifier to fit the specific classification problem, in this case the credit evaluation.

### 4.2.3 Testing Phase

This is the stage whereby classification of a new instance takes place. The model has been trained and validated, thus it is used to evaluate new instances of credit applications. Classification involves computation of the posteriori probability. The class of the new instance will be the one which maximizes this probability. A different dataset, a testing dataset, is used at this step. This is typically 20 percent of the sample dataset. This dataset is used to assess the performance and generalization of the classifier.

The model adopts the algorithm below in performing the classification task.

Naïve\_Bayes\_Learn (*examples*)

For each target value  $v_j$

$P^{\wedge}(v_j) \leftarrow \text{estimate } P(v_j)$

For each attribute value  $a_i$  of each attribute  $a$

$P^{\wedge}(a_i|v_j) \leftarrow \text{estimate } P(a_i|v_j)$

Classify\_New\_Instance( $x$ )

$$v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_i P(a_i|v_j)$$

Credit evaluation is a classification problem and therefore, the target output is a classification of an application into two classes namely; Approved (value 1) and Rejected (value 0).

### 4.3 Data Analysis

Data used in this study has been obtained from one of the banking institutions in Kenya. The dataset includes customer's financial history, biographic as well as demographic details as shown in Table 2.

**Table 2: Customer Characteristics**

	Description
1	Age of applicant
2	Period at Present employment
3	Period at present residence
4	Number of dependants
5	Residential Status
6	Account Type
7	Gender
8	Marital Status
9	Net Income
10	Number of Credit Cards
11	Income paid to account
12	Credit reference bureau opinion
13	Security offered
14	Credit Amount
15	Relationship with Bank (Time with the Bank in years)
16	Time with the Account

The dataset received consists of 3533 credit applications and all these records are used as the target population for this study. The dataset is partitioned into three categories namely the Training dataset, Validation dataset and Testing dataset. The training dataset has 2473 records, 353 records for validation and 707 for testing as shown in Table 3.

**Table 3: Data Partition Set**

Data Partition Set	Records	Percentage (%)
Training Set	2473	70
Validation Set	353	10
Testing Set	707	20
<b>Total</b>	<b>3533</b>	<b>100</b>

#### 4.4 Data Preprocessing and Level of Measurement

The dataset shall be analyzed to eliminate any anomalies in the data. Preprocessing of the data is also vital and this shall be carried out before the data is fed onto the classifier. This ensures that the significant parameters used during credit evaluation shall be used to come up with optimal variables as inputs. For instance, the customer data including behavioral (for example, past credit history), biographic and demographic information (for example age, gender and marital status) are pre-processed and fed to the classifier. The two major kinds of data are categorical and quantitative. Categorical variables take only a finite number of possible values, and there are usually several or more cases falling into each category. Categorical variables may have symbolic values (e.g. "male" and "female") that must be encoded into numbers before being input to the classifier. On the other hand, quantitative variables are numerical measurements of some attribute. The measurements must be made in such a way that at least some arithmetic relations among the measurements reflect analogous relations among the attributes of the objects that are measured.

Naïve Bayes Classifier relies on counting techniques to calculate probabilities and as a result, the main level of measurement used in this study is categorical. Therefore, some of the variables shall be scaled into categories. The resulting model in using this type of data is simpler, more robust and generalizes better with changes in the data. Numerical variables are discretized to categorical before being fed into the classifier. Further, the categorical variable can be classified as Nominal or Ordinal. The dataset is encoded into their respective values before being input to the classifier for example, males will be categorized as 1 and females as 0. Table 4 shows the description and the level of measurement of the dataset.

**Table 4:** Dataset Description and Level of measurement

	Variable Code	Description	Value	Level of Measurement
1	AccountType	Account Type	Savings Account = 1 Fixed Deposit = 2 Current Account = 3 No Information = 4	Nominal
2	Age	Age of applicant	0 – 10 = 1 11 – 20 = 2 21 – 30 = 3 31 – 40 = 4 41 – 50 = 5 51 – 60 = 6 61 – 70 = 7 71 – 80 = 8 81 – 90 = 9	Ordinal



3	CRB Opinion	Credit Reference Bureau Opinion	Adverse = 1 Favourable = 2	Nominal
4	Credit Amount	Credit Amount (x Ksh.1000)	0 – 10 = 1 11 – 20 = 2 21 – 30 = 3 31 – 40 = 4 41 – 50 = 5 51 – 60 = 6 61 – 70 = 7 71 – 80 = 8 81 – 90 = 9 91 – 100 = 10 >100 = 11	Ordinal
5	Gender	Gender	Male = 1 Female = 0	Nominal
6	Income Paid To Account	Income Paid to Account	Yes = 1 No = 2 No Information = 3 No Response = 4	Nominal
7	Length Acct Relship	Time with the Account	Years <1 = 1 Years 1-3 = 2 Years 3-5 = 3 Years 5-7 = 4 Years >7 = 5 No Response = 6 No Information = 7	Nominal
8	Marital Status	Marital Status	Single = 1 Married = 2 Separated = 3 Divorced = 4 Widowed = 5 No Response = 6 No Information = 7	Nominal
9	Net Income	Net Income (x Ksh.1000)	No Information = 0 0 – 20 = 1 21 – 40 = 2	Ordinal

			41 – 60 = 3 61 – 80 = 4 81 – 100 = 5 101 – 150 = 6 151 – 200 = 7 201 – 250 = 8 251 – 500 = 9 501 - 1000 = 10 1001 – 1250 = 11 1251 – 1500 = 12 1501 – 2000 = 13 > 2000 = 14	
10	NoOfCreditCards	No of Credit Cards	None = 1 2 = 2 2-5 = 3 5 = 4	Nominal
11	NumberOfDependants	Number of Dependants	0 Kids = 1 1-3 Kids = 2 >3 Kids = 3 No Response = 4 No Information = 5	Nominal
12	PeriodPresentEmployment	Period at Present Employment	Unemployed = 1 Self-Employed = 2 Pensioner = 3 Employed <1 = 4 Employed 1-2 = 5 Employed 2-5 = 6 Employed 5-10 = 7 Employed >10 = 8 No Response = 9 No Information = 10	Nominal
13	PeriodPresentResidence	Period at present Residence	<1 Yrs = 0 1-3 Yrs = 1 3-8 Yrs = 2 8-11 Yrs = 3 >11 Yrs = 4	Nominal

			No Information = 5 No Response = 6	
14	RelationshipWithBank	Relationship with Bank	Years < 1 = 1 Years 1-3 = 2 Years 3-5 = 3 Years 5-7 = 4 Years > 7 = 5 N/A = 6 No Response = 7 No Information = 8	Nominal
15	ResidentialStatus	Residential Status	Owner With Mortgage = 1 Owner Without Mortgage = 2 Living With Parents = 3 Renting = 4 Employer Owned = 5 No Response = 6 No Information = 7	Nominal
16	SecurityOffered	Security Offered	None = 1 Legal Charge = 2 Bank Guarantee = 3 FDR = 4	Nominal

#### 4.5 Feature Selection

Feature selection methods are used to identify input variables that are not useful and do not contribute significantly to the performance of the classifier. The removal of insignificant inputs helps improve the generalization performance of a classifier. To select the features, the relationship between the customer characteristics and their significance to credit evaluation is analyzed using Chi-Square Test of Independence. A Chi-Square ( $X^2$ ) statistic is used to investigate whether distributions of categorical variables differ from one another. Therefore, the test statistic is chosen because the variables (both dependent and independent variables) used in this study are categorical. For each variable in the dataset, its relationship with credit evaluation is determined. To establish the relationship between the variables and credit risk, the hypotheses below are used.

Null Hypothesis:

$H_0$ : No relationship between variable X and Y (i.e. zero association)

Alternative Hypothesis:

$H_1$ : There is a relationship between the two variables (non-zero association)

A Contingency table is used to determine the Chi square statistic. Chi Square equation used is:

$$\text{Chi Square } (x^2) = \text{the sum of all the } (f_o - f_e)^2 / f_e$$

Where  $f_o$  denotes the frequency of the observed data and  $f_e$  is the frequency of the expected values.

The test statistic is converted to a conditional probability called a p-value. A p-value is the probability of observing data as or more extreme as the actual outcome when the null hypothesis is true. A small p-value provides evidence against the null hypothesis because it indicates that the observed data are unlikely when the null hypothesis is true. P-values are determined for each of the selected variable. To test the hypothesis, a level of significance (Alpha ( $\alpha$ )), is used. Alpha ( $\alpha$ ) is a probability threshold for a decision. As a result, if  $p \leq \alpha$ , we will reject the null hypothesis. Otherwise it will be retained for want of evidence. The conventional level of significance (Alpha  $\alpha$ ) of 5% or 0.05 has been used in this study. This means that when the computed Chi-Square statistic exceeds the critical value in the table for a 0.05 probability level, then we can reject the null hypothesis of equal distributions. Therefore, if the P-value is less than or equal to the significance level ( $p \leq 0.05$ ), then the null hypothesis is rejected and alternative hypothesis is accepted (That is, there is a relationship between the variables). If the p-value is greater than the significance level ( $p > 0.05$ ), then accept the null hypothesis (There is no relationship between the variables).

#### 4.6 Model Validation

To validate the results of the model, a confusion matrix will be used. A confusion matrix is an n-dimensional square matrix, where n is the number of distinct target values (in this case n will be 2 since the target values are two – Approved and a Rejected). Figure 7 shows a confusion matrix. The row indexes of a confusion matrix correspond to actual values observed and used for model testing; the column indexes correspond to predicted values produced by applying the model to the test data. For any pair of actual/predicted indexes, the value indicates the number of records classified in that pairing. A confusion matrix provides a quick understanding of model accuracy and the types of errors the model makes when scoring records. It is the result of a test task for classification models (Badgerati, 2010).

		Prediction	
		0	1
Actual	0	TN	FP
	1	FN	TP

Source: (Badgerati, 2010)

Figure 7: Confusion Matrix

In the Confusion Matrix, for each cell in the matrix we have fields as True Positives (TP), False Positives (FP), False Negatives (FN) and True Negatives (TN). These are defined as:

- False Positive (FP): Falsely predicting a label (or saying that a Rejection is an Approval).
- False Negative (FN): Missing an incoming label (or saying an Approval is a Rejection).
- True Positive (TP): Correctly predicting a label (or saying an Approval is an Approval).
- True Negative (TN): Correctly predicting the other label (or saying a Rejection is a Rejection).

Since the model is a binary classification, two equations called **Sensitivity** and **Specificity** shall be derived to evaluate the classifier. **Sensitivity** measures the proportion of actual positives which are correctly identified as such (i.e. accuracy on the class Positive). **Specificity** measures the proportion of negatives which are correctly identified (i.e. accuracy of the class Negative). **Accuracy** (AC) is the proportion of the total number of predictions that were correct. Equations for computing these measures of performance from the confusion matrix are as shown below.

$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{Specificity} = \text{TN} / (\text{TN} + \text{FP})$$

$$\text{Accuracy} = (\text{TN} + \text{TP}) / (\text{TN} + \text{TP} + \text{FN} + \text{FP})$$

A model can be very specific without being sensitive, or it can be very sensitive without being specific. Both factors are equally important.

#### 4.7 Methodology Assumptions

The main assumption in the methodology used is that the dataset is made up of features that are independent of each other given the class.

**CHAPTER 5  
RESULTS AND ANALYSIS**

**5.1 Introduction**

The main aim of this chapter is to outline the analysis of the research findings. Variables are described and their contributions to credit evaluation decision process are also detailed.

**5.2 Feature Analysis and Selection**

The dataset obtained is made up of 16 attributes defined as categorical in this study. Feature or variable selection is an important step as it determines the performance of the model. Before the input data is fed into the classifier, variables are selected in order to come up with optimal features for the model. Chi-Square test of independence is applied to establish the relationship between customer's behavioral, biographic and demographic characteristics and their contributions towards credit decision process. Table 5 shows the results of the test of independence using Chi-Square statistic for each of the variable.

**Table 5: Chi-Statistic Test Results**

	<b>Feature</b>	<b>Alpha(<math>\alpha</math>)</b>	<b>Degree of Freedom (d. f)</b>	<b>Chi Statistic (<math>X^2</math>)</b>	<b>P-Value</b>
1	AccountType	0.05	3	34.083	1.90284396872954E-07
2	Age	0.05	8	28.731	0.0004
3	CRBOpinion	0.05	1	0.027	0.8695
4	CreditAmount	0.05	10	273.821	5.23409674400891E-53
5	Gender	0.05	1	4.081	0.0434
6	IncomePaidToAccount	0.05	3	9.615	0.0221
7	LengthAcctRelship	0.05	6	77.19	1.35810108706184E-14
8	MaritalStatus	0.05	6	4.955	0.5496
9	NetIncome	0.05	14	301.018	7.25034579270144E-56
10	NoOfCreditCards	0.05	3	9.847	0.0199
11	NumberOfDependants	0.05	4	1.437	0.8377
12	PeriodPresentEmployment	0.05	9	26.898	0.0015
13	PeriodPresentResidence	0.05	6	13.221	0.0397
14	RelationshipWithBank	0.05	7	8.653	0.2785
15	ResidentialStatus	0.05	6	34.083	6.48362796163272E-06
16	SecurityOffered	0.05	3	2.045	0.5631

### 5.2.1 Hypothesis Testing

The conventional level of significance (Alpha  $\alpha$ ) of 5% or 0.05 has been used in this study. To determine the relationship between the dependent and the independent variables, the null and alternative hypotheses are tested using Chi-Square test statistic. The hypotheses are defined as:

Null Hypothesis:

$H_0$ : No relationship between variable X and Y (i.e. zero association)

Alternative Hypothesis:

$H_1$ : There is a relationship between the two variables (non-zero association)

The data in Table 5 reflects the following information about the relationship between customer characteristics and credit decision.

**Account Type:** A p-value of 1.9046949002265E-07 and level of significance of 0.05 forces the rejection of the null hypothesis resulting in acceptance of the alternative hypothesis. The p-value is less than the significance level ( $p < 0.05$ ) and therefore there is evidence against null hypothesis. This implies that Account Type is a contributing factor to the credit decision. Thus, Account Type is used as a feature that will be fed to the Naive Bayes classification model.

**Age:** A significant p-value of 0.0004 (i.e.  $p < 0.05$ ) which is less than the significance level (0.05) indicates that there is a relationship between Age and credit evaluation. The small p-value provides strong evidence to reject the null hypothesis.

**Credit Reference Bureau Opinion:** Since the p-value (0.8695) is not significant (it is greater than the significance level), the null hypothesis is not rejected. Thus, it seems unlikely that Credit Reference Bureau opinion will be a differentiating factor that can assist the classifier to categorize credit application. Interdependencies with other variables may be useful for categorizing.

**Credit Amount:** Credit amount has a small p-value ( $p < 0.05$ ); therefore, it forces the rejection of null hypothesis. As a result, there is evidence that credit amount is a determining factor in credit evaluation.

**Gender:** The p-value (0.0434) is significant and the null hypothesis that there exists no relationship between gender and credit decision is rejected. Gender is thus a contributing factor to credit evaluation.

**Income Paid to Account:** Whether or not income is paid to the account is a feature that can assist the Naive Bayes Classifier in making credit decision. The p-value of 0.0221 (that is  $p < 0.05$ ) means that the null hypothesis that there is no relationship between payment of income to the account with credit decision is rejected.

**Time with the Account (LengthAcctRelship):** The p-value is less than the level of significance ( $p < 0.05$ ). The null hypothesis that there is no relationship between time with the account and credit decision is therefore rejected. There is a relationship between the length of time a customer has an account with and credit decision. Time with the account is therefore used as an input to the Naïve Bayes model.

**Marital Status:** Marital Status has a p-value of 0.5496 ( $p > 0.05$ ). Since the value is greater than the significance level, we fail to reject the null hypothesis (There is no relationship) and conclude that Marital Status is not a differentiating factor that can be used by the classifier in credit decision.

**Net Income:** The significant p-value ( $7.25034579270144E-56$ ) forces the rejection of null hypothesis that there is no difference on credit decision amongst the different income levels.

**Number of Credit Cards:** The significant p-value of 0.0199, forces the rejection of the null hypothesis that there is no difference between the number of credit cards and credit decision. The number of credit cards is a feature that can be used by the Naïve Bayes model in credit decision.

**Number of Dependants:** With a p-value of 0.8377, which is greater than the level of significance (0.05), indicates that we accept the null hypothesis that there is no relationship between the number of dependants and credit decision. It's unlikely that the number of dependants can assist the classifier in categorizing credit applications.

**Period at Present Employment:** The p-value (0.0015) is significant and the null hypothesis that there exists no relationship between the period at present employment and credit decision is rejected. Period at employment is thus a contributing factor to credit evaluation.

**Period at Present Residence:** The p-value is significant and the null hypothesis is rejected. This means that there is a relationship between credit decision and the period at present residence.

**Relationship with Bank:** The p-value (0.2785) is not significant; therefore, we fail to reject the null hypothesis. There is no relationship between customer's relationship with the bank and credit decision.

**Residential Status:** The p-value is significant and the null hypothesis that there is no relationship between the residential status and credit decision is rejected. Residential status is thus a contributing factor to credit decision.

**Security Offered:** Significant p-value is greater than the level of significance and thus, null hypothesis is not rejected. It is therefore unlikely that security offered will be a differentiating factor that can be used by the model.



The results of establishing the relationship between the customer's characteristics and their influence in credit decision have shown that, while there is no relationship between credit reference bureau opinion, marital status, number of dependants, security offered and the customer's relationship with the bank and credit decision, there is a relationship between account type, age, gender, income paid to the account, time with the account, net income, credit amount, number of credit cards, period at present employment, period at present residence and residential status with credit decision. These features could be used to train the Naïve Bayes model to classify credit applications.

The next section describes the results obtained when using the characteristic data (demographic, biographic and behavioural) to train and test the Naïve Bayes model.

### 5.3 NBC Model Results

Training, validation and testing is performed on the model by using the customer characteristics data. The dataset acquired contained 3533 records. This is split into three categories; training set, validation set and testing set with the ratio 70 percent, 10 percent and 20 percent respectively. Before the model is subjected to new labels in the testing set, the model's performance on the validation set is first carried out. An initial performance is measured with all the 16 features without feature selection. Then, the features identified as significant in the previous section are used and its performance also measured.

#### 5.3.1 NBC validation mode results without feature selection

The Naïve Bayes model is trained on the training set. The training set consists of examples with known labels for each class. Training is done to help the classifier learn the patterns or correlations between the features and the class in the training data. To determine the optimal model to be used for testing and classification of new instances, the identified characteristic features are used on the validation dataset. The validation dataset contains a total of 353 records. Initially, all the 16 features are used to design the model before the final optimal model is created and the performance results obtained are as presented on the confusion matrix as shown in Figure 8.

		Predicted	
		Rejected	Approved
Actual	Rejected	45	57
	Approved	50	201

**Figure 8:** Confusion Matrix of validation mode without feature selection

To measure the performance of the Naïve Bayes Classifier model without feature selection, performance metrics are calculated from the confusion matrix as shown in Figure 9.

Sensitivity:	0.80
Specificity:	0.44
Accuracy:	0.70

**Figure 9:** Performance Metrics of validation mode without feature selection

If the instance is positive and it is classified as positive, it is counted as a true positive else if it is classified as negative, it is counted as a false negative. On the other hand, if the instance is negative and it is classified as negative, it is counted as a true negative; if it is classified as positive, it is counted as a false positive. Therefore, from Figure 8, it follows that:

- There are 102 credit applications classified into class Rejected in which 45 of these are correctly classified into the class Rejected while 57 are wrongly classified into class Approved.
- There are 251 credit applications in the class Approved. Out of these, 50 are incorrectly classified as class Rejected while 201 are correctly classified as Approved.

From these, in order to determine the accuracy of the classifier, Sensitivity and Specificity measures of performance are defined. Whereby, Sensitivity measures the accuracy of the class positive (in this case, the class Approved).

Sensitivity is calculated as:

$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{Sensitivity} = 201 / (201 + 50) = 0.80$$

The result above indicates a Sensitivity of 0.80 (80%). This means that the Naïve Bayes classifier is 80% accurate in classifying credit applications as Approved (class positive).

Specificity measures the proportion of negatives which are correctly identified as such. This is computed as follows:

$$\text{Specificity} = \text{TN} / (\text{TN} + \text{FP})$$

$$\text{Specificity} = 45 / (45 + 57) = 0.44$$

Therefore, on the validation dataset, the Naïve Bayes model has an accuracy of 0.44(44%) in classifying negative instances as negative. That is it correctly classifies Rejected credit applications into the class Rejected.

Accuracy defines the proportion of the total number of predictions that were correct. This is given using the equation:

$$\begin{aligned}
 \text{Accuracy} &= (TN + TP) / (TN+TP+FN+FP) \\
 &= (45 + 201) / (45 + 201 + 50 + 57) \\
 &= 0.70
 \end{aligned}$$

This indicates that the classifier has an accuracy of 70% in correctly classifying the credit applications on the validation set.

### 5.3.2 NBC validation mode results with feature selection

Feature selection is performed and selected features are applied on the validation set. There are a total of 11 features selected as significant out of the 16 features. Results of the model are then generated in a confusion matrix as shown in Figure 10.

		Predicted	
		Rejected	Approved
Actual	Rejected	43	59
	Approved	43	208

**Figure 10:** Confusion Matrix of validation mode with feature selection

Figure 11 shows the performance measures extracted from the confusion matrix. Sensitivity ratio of 83% implies that by using feature selection, the model is 83% accurate in classifying positive (in this case, class Approved) instances while 42% for the class negative (Rejected applications). Overall accuracy of the model with feature selection is 71%.

Sensitivity:	0.83
Specificity:	0.42
Accuracy:	0.71

**Figure 11:** Performance Metrics of validation mode with feature selection

Classification performance of the classifier with and without feature selection is summarized in Table 6.

**Table 6: Classification Performance on the Validation Dataset**

Performance Measure	Accuracy without Feature Selection (%)	Accuracy with Feature Selection (%)
Sensitivity	80	83
Specificity	44	42
Accuracy	70	71

From Table 6, it is worth noting that the classification accuracy of the classifier increased after performing feature selection. Feature selection helps in identifying redundant attributes that might affect the classification performance of a classifier. The classification accuracy increased from 70% without feature reduction to 71% if feature selection is performed. This indicates that feature selection is important in improving classification accuracy of the Naive Bayes Classifier. The model with reduced features is thus selected to be used for testing.

#### 5.4 NBC test mode results

Testing the model is necessary in order to examine its ability to generalize in a real world data or unseen instances. A separate dataset (testing dataset) from the ones previously applied is used for testing. The ability of the classifier is being evaluated in this phase as the model has not been previously exposed to the test dataset during training and validation. The test dataset contains 707 credit applications (204 “bad” credit applications and 503 “good” applications). The results from the test mode are presented in the confusion matrix shown in Figure 12.

		Predicted	
		Rejected	Approved
Actual	Rejected	65	139
	Approved	51	452
		116	591

**Figure 12: Confusion Matrix of NBC test mode**

Table 7 and Table 8 show the summary of the performance measures extracted from the confusion matrix in Figure 9. Test results for the Naive Bayes Classifier are optimal as earlier predicted by the validation results.

**Table 7: NBC test mode data results**

Performance Measure	Counts
True Positive	452
False Positive	139
True Negative	65
False Negative	51

**Table 8: NBC test mode classification performance metrics**

Performance Measure	Rate (%)
Sensitivity	90
Specificity	32
Accuracy	73

#### **5.4.1 Analysis of test NBC results**

NBC with 11 attributes has a sensitivity rate of 90% and specificity of 32%. The sensitivity rate implies that the model is able to correctly classify 90% “good” applications. Out of 503 “good” applications, NBC is able to classify 452 correctly as such and misclassify 51. On the other hand, it can correctly classify 32% of “bad” credit applications. This follows that, out of 204 credit applications of the class rejected (“bad” applications), it correctly classify 65 and misclassify 139.

#### **5.5 Comparison of results with previous research**

The results obtained in this study are comparable to that achieved by Novakovic and Rankov (2011). In their work, two datasets were used; an Australian credit dataset and a German credit dataset. The Australian dataset had 1000 credit records and 20 attributes and the Naïve Bayes achieved an accuracy of 77.7%. On using the German dataset with 690 records with 14 attributes, the Naïve Bayes performance was 75.4%.

Baesens et al (2002) evaluate the performance of Naïve Bayes Classifier on three different datasets. Firstly, the classification performance on German credit with 1000 observations with 15 attributes is 71.52%. The same classifier is also tested on two datasets obtained from Benelux financial institutions, one with 3123 observations and 23 input attributes and the other with 7190 observations and 28 attributes. Results on the first and second Benelux datasets achieved accuracy of 69.54% and 69.09% respectively. The Naïve Bayes Classifier used in this study performs with greater accuracy of 73%.

## CHAPTER 6

### DISCUSSIONS, CONCLUSION AND RECOMMENDATIONS

#### 6.1 Introduction

This chapter presents the findings of the study, communicate the recommendations and conclusion and suggest areas of further research. The first section provides a summary of the research findings including the achievements accomplished by conducting this study. The second section of this chapter outlines the recommendations and conclusion. The aim is to prove that the suggested recommendations and conclusion are logically derived from the analysis of the findings. Limitations of the study are also identified. The last section is a list of suggestions for further research.

#### 6.2 Summary of research findings

It is important to note that the objectives of this undertaking have been realized. One of the objectives was to examine whether Naïve Bayes Classifier can be applied accurately to consumer credit evaluation. The results of the study have shown that Naïve Bayes Classifier can be used to evaluate credit applications. The classification accuracy obtained indicates that the Naïve Bayes Classifier has the ability to correctly classify credit applications as either “good” or “bad”. Identifying “bad” credit applications at an early stage ensures that there is reduced loss of revenue to the credit lending institution.

Hypotheses testing have been utilized to show that features in the data, for instance, net income, age, number of credit cards, the time with account, can affect the performance of the classification system. As a result, feature selection has proven to be vital in improving the performance accuracy of the classifier.

A model has been designed and used to evaluate the credit data. This ensures efficiency in credit evaluation processes that is free of bias and ensuring that the results are obtained in a short period of time.

#### 6.3 Conclusion

Efficient management of credit is significant for the banking industry in general and Kenya in particular. This study has developed a decision tool that is able to adapt to the dynamic changes in the financial sector and therefore, assisting Kenyan commercial banks in evaluating consumer credit hence reducing loss of revenue. Reliance on credit officers by Kenyan commercial banks in making credit decisions might affect their evaluation capabilities. By utilizing such technique as Naïve Bayes Classifier in credit evaluation as developed in this study, it will reduce any bias or emotional intention that can distort the decision process thereby leading to reduced cost of credit processing and improved quality of customer service. Furthermore, the high accuracy of this model proves that Naïve Bayes Classifier can be useful for classifying credit applications.

#### **6.4 Recommendations**

Results obtained in this study have shown that Naïve Bayes Classifier is an effective tool in classifying credit applications at a high accuracy level. This means that the implementation of such model in Kenyan commercial banks can be helpful for the decision making process. However, to gain more trust in this model, banks can use more cases of credit applications (both “good” and “bad” applications) from their databases.

#### **6.5 Limitations of the Study**

Obtaining comprehensive set of actual data from banking institution is difficult as such information is considered confidential and thus should be hidden from competitor in the industry.

#### **6.6 Suggestions for future work**

This study has presented a credit decision tool using Naïve Bayes Classifier to assist credit officers make better decisions in evaluating customer credit applications. With emergence of E-banking and M-banking, services have been brought to the customer’s fingertips. Therefore, as a guideline for further study, the model can be extended to be available to the customer by providing an interface where the customer is able to perform self-evaluation since this study restricted the scope to the credit officers.

## REFERENCES

- Abramowicz, W., Nowak, M., Szykiel, J. (2003). *Bayesian Networks as a Decision Support Tool in Credit Scoring Domain*. Idea Group Publishing.
- Agile Alliance. (2002). *Agile Manifesto*. <http://www.agilealliance.org>. (Accessed on 23 July, 2012).
- Alpaydin, E. (2004). *An Introduction to Machine Learning*. The MIT press, Cambridge, Massachusetts, London, England.
- Angelini, E., Giacomo, D. T., Andrea R., (2007). *A neural network approach for credit risk evaluation: The Quarterly Review of Economics and Finance*, doi:10.1016/j.qref.2007.04.001
- Badgerati, (2010), <http://computersciencesource.wordpress.com/2010/01/07/year-2-machine-learning-confusion-matrix> (Accessed on: 10 March, 2012)
- Baesens, B., Egmont-Petersen, M., Castelo, R.; Vanthienen, J. (2002). *Learning Bayesian network classifiers for credit scoring using Markov Chain Monte Carlo search*
- Baesens, B., Setiono, R., Mues C., and Vanthienen J., *Using Neural Network Rule Extraction and Decision Tables for Credit Risk Evaluation*, Computer Journal of Management Science, vol. 49, no. 3, pp. 312-329, 2003.
- Basel Committee on Banking Supervision. 2005. *International convergence of capital measurement and capital standards. A revised framework*. Basel: Bank for International Settlements.
- Bonin J., Huang Y. P. (2001), "Dealing With the Bad Loans of the Chinese Banks", Journal of Asian Economies, Summer.
- CBK, (2000), *Bank Supervision Annual Reports*, Central Bank of Kenya.
- CBK, (2005), *Risk management guidelines 2005*, Central Bank of Kenya.
- CBK, (2008), *Basel II Implementation Survey Results 2008*, Central Bank of Kenya.
- CBK, (2010), *Risk management survey for the Banking sector 2010*, Central Bank of Kenya



Chen, F., Li, F., (2010). *Combination of feature selection approaches with SVM in credit scoring*. Expert Systems with Applications, vol. 37 pp. 4902-4909.

Cockburn, A. (2001). *Agile software Development*. Addison-Wesley, Boston, MA.

Conboy, K., Fitzgerald, B., (2004). *Toward a Conceptual Framework of Agile Methods: A Study of Agility in Different Disciplines*. Proceedings of the 2004 ACM Workshop on Interdisciplinary Software Engineering Research. ACM 37-44, Newport Beach, CA. November 2004. <http://portal.acm.org/citation.cfm?id=1029997.1030005>.

Eyheramendy, S., Lewis, D. and Madigan, D (2003). On the naive bayes model for text categorization. *Proceedings Artificial Intelligence Statistics*.

Handzic M., Tj and Rawibawa F. and Yeo J. (2003). *How Neural Networks Can Help Loan Officers to Make Better Informed Application Decisions*, Informing Science Insite.

Hillstrom, K., Hillstrom, L. C., (2002). *Encyclopedia of small business*. 2nd ed. Detroit, MI: Gale Group

Hsieh, N. C., (2005). *Hybrid mining approach in the design of credit scoring models: Expert system with applications* vol.28, pp. 655-665.

Hsieh N. C., Hung L. P. (2010). *A data driven ensemble classifier for credit scoring analysis* *Expert Systems with Applications*, 37 (1), pp. 534-545.

Huang, C. L., Chen, M. C., Wang, C. J., (2007). *Credit scoring with datamining approach based on support vector machine*, Expert System with application vol.37, pp. 847-856.

Lai, K. K. Yu, L. Wang, S. & Zhou, L. 2006a. *Credit Risk Analysis Using a Reliability-Based Neural Network Ensemble Model*. In: Kollias, S. ed. *ICANN 2006, Part II, LNCS 4132*. Springer-Verlag: Berlin. 682 – 690.

Leea, T., Chiub, Ch.Ch,Y.-Ch. Chouc,Ch, J. Lud. (2006). *Mining the customer credit using classification and regression tree and multivariate adaptive regression splines*, Expert system with applications, vol.50, pp. 1113-1130.

Malhotra, R. and D.K. Malhotra, 2003. *Evaluating consumer loans using neural networks*. Omega, 31: 83-96. DOI: 10.1016/S0305-0483(03)00016-1

Mitchell, D. and Pavur, R., (2002). *Using modular neural networks for business decisions*. Manage. Dec. 40/1 58-63. DOI: 10.1108/00251740210413361.

Mitchell, T. (1997). *Machine learning*. McGraw Hill, USA.

Novakovic J., Rankov S. (2011). *Classification Performance Using Principal Component Analysis and Different Value of the Ratio R*: International Journal of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844 Vol. VI (2011), No. 2 (June), pp. 317-327.

On, C. S., Jeng, J., Huang, G., HshiangTzeng (2005). *Building credit scoring model using genetic programming: Expert System with application*, vol.29, pp. 41-47.

Ravi, V. (2008), *Advances in banking technology and management: impacts of ICT and CRM: information science reference*.

Rish, I. (2001). An empirical study of the naive bayes classifier. *Proceedings of IJCAI-01*, 2001.

Scott W. A., (2005). The Agile Unified Process. <http://www.ambyssoft.com/unifiedprocess/agileUP.html>.

Shachmurove, Y., (2002). Applying artificial neural networks to business, economics and finance. <http://ideas.repec.org/p/cla/penntw/5ecbb5c20d3d547f357aa130654099f3.html>

Steiner, M., Neto, P., Soma, N., Shimizu T., Nievola J., (2006) Using Neural Network Rule Extraction for Credit-Risk Evaluation: IJCSNS International Journal of Computer Science and Network Security, VOL.6 No.5A, May 2006

Thomas, L. C., (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers, international journal of forecasting vol.16, pp.149-172.

Tsai, Ch. F., Wu, J. W., (2008). Using neural network ensembles for bankruptcy prediction and credit scoring, Expert Systems with Applications, vol.34, pp. 2639-2649.

TunLi, S., Shiue, W., Huang, M. H., (2006)"The evaluation of consumer loans using support vector machines," Expert System with application, vol.30, pp.772-782.

Wang, G., Hao J., Ma J., Jiang.H., (2011). *A comparative assessment of ensemble learning for credit scoring*, Expert system with applications vol.38, pp.223-230.

West, D. 2000. Neural Network Credit Scoring Models. *Computer & Operations Research*, 27, 1131–1152.

Zhang, D., X. Zhou et al (2010). *Vertical bagging decision trees model for credit scoring*, Expert system with applications, vol. 37, pp. 7838-7843.

## APPENDICES

### Appendix A – Feature selection screen

**Feature Analysis**

Select	Attribute
<input checked="" type="checkbox"/>	Account Type
<input checked="" type="checkbox"/>	Age
<input checked="" type="checkbox"/>	CRBOpinion
<input checked="" type="checkbox"/>	Credit Amount
<input checked="" type="checkbox"/>	Gender
<input checked="" type="checkbox"/>	Income Paid To Account
<input checked="" type="checkbox"/>	Length Acct Relship
<input checked="" type="checkbox"/>	Marital Status
<input checked="" type="checkbox"/>	Net Income
<input checked="" type="checkbox"/>	NoOfCreditCards
<input checked="" type="checkbox"/>	NumberOfDependants
<input checked="" type="checkbox"/>	Period Present Employment
<input checked="" type="checkbox"/>	Period Present Residence
<input checked="" type="checkbox"/>	Relationship With Bank
<input checked="" type="checkbox"/>	Residential Status
<input checked="" type="checkbox"/>	Security Offered

Select All Features

**Analysis Summary**

Attribute	Alpha (α)	df	ChiStatistic	PValue
Account Type	0.05	3	34.083	1.90284396873954E-07
Age	0.05	8	28.731	0.000353469089724352
CRBOpinion	0.05	1	0.027	0.86948178281955
Credit Amount	0.05	10	273.821	5.23409674400891E-53
Gender	0.05	1	4.081	0.0433679907920083
Income Paid To Account	0.05	3	9.615	0.0221389052469437
Length Acct Relship	0.05	6	77.19	1.35810108706184E-14
Marital Status	0.05	6	4.955	0.549597575284037
Net Income	0.05	14	301.018	7.25034579270144E-56
NoOfCreditCards	0.05	3	9.847	0.0199124814896483
NumberOfDependants	0.05	4	1.437	0.837739419338136
Period Present Employment	0.05	9	26.898	0.00145414655343115
Period Present Residence	0.05	6	13.221	0.0396576462229915
Relationship With Bank	0.05	7	8.653	0.278538277148676
Residential Status	0.05	6	34.083	6.48362796163272E-06
Security Offered	0.05	3	2.045	0.563119516396454

Significant Features

### Appendix B – Model validation and testing screen display

**Feature**

- Age
- Credit Amount
- Gender
- Income Paid To Account
- Length Acct Relship
- Net Income
- NoOfCreditCards
- Period Present Employment
- Period Present Residence
- Residential Status
- Security Offered

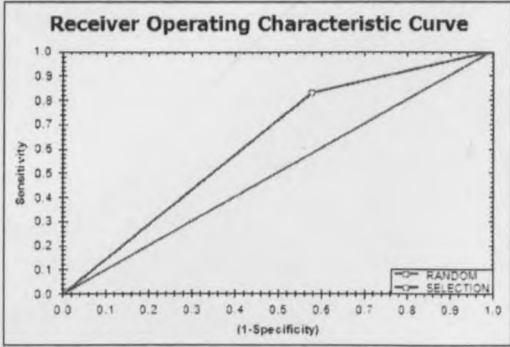
**Confusion Matrix**

		Predicted	
		Rejected	Approved
Actual	Rejected	43	59
	Approved	43	208

**Results from Confusion Matrix**

- True Positive (TP): 208
- False Positive (FP): 59
- True Negative (TN): 43
- False Negative (FN): 43
- Sensitivity: 0.83
- Specificity: 0.42
- Accuracy: 0.71

**Receiver Operating Characteristic Curve**



### Appendix C – Screen display of New Classification module

Select the Model to use for Classification

Model:  Description: With Feature Selection

Attribute	Value Label	
Age	21 - 30	▼
CreditAmount	41 - 50	▼
Gender	Female	▼
IncomePaidToAccount	Yes	▼
LengthAcctRelshp	Years 5-7	▼
NetIncome	61 - 80	▼
NoOfCreditCards	None	▼
PeriodPresentEmployment	Employed 1-2	▼
PeriodPresentResidence	1-3 Yrs	▼
ResidentialStatus	Employer Owned	▼
SecurityOffered	Bank Guarantee	▼

Result Analysis

Approve 0 0000005831499885366

Reject 0 0000000814767008109

Result Approved

Buttons: Classify, Load, Save, Print, Close

### Appendix D – Sample code of the routine that Trains the Naïve Bayes Classifier

```

Sub GetClassTotal()
Try
numReject = 0
numAccept = 0
strSQL = "select count(*) as num from CreditApplication where Result=0 union select count(*) as num from
CreditApplication where Result=1"
dsCredit = objDB.FillDataset(strSQL)

If Not dsCredit Is Nothing Then
With dsCredit
If .Tables.Count > 0 Then
'Get the Total per Target output - Approved and Rejected
If .Tables(0).Rows.Count > 0 Then
If Not .Tables(0).Rows(0) Is Nothing Then
numRejected = .Tables(0).Rows(0).Item("num")
Else
numRejected = 0
End If

If Not .Tables(0).Rows(1) Is Nothing Then
numApproved = .Tables(0).Rows(1).Item("num")

```

```

        Else
            numApproved = 0
        End If

        numExamples = (numApproved + numRejected)
        numReject = FormatNumber((numRejected / numExamples), 4)
        numAccept = FormatNumber((numApproved / numExamples), 4)

        End If
    End If
End With
End If
Catch ex As Exception
    Throw ex
End Try
End Sub

```

### Appendix E – Sample code that performs classification of a new instance

```

'Sub-routine to perform Classification of a New Instance(x)
Sub ClassifyNew(ByVal dtTable As DataTable, ByVal strColumns As String, ByVal IsUpdate As Boolean, ByVal
UpdateDB As String)
Try
    Dim i As Integer = 0, j As Integer = 0
    Dim numTotal As Long = 0, strName As String = ""
    Dim Reject As Integer = 0, Approval As Integer = 0
    Dim intTempReject As Decimal = 0, intTempApprove As Decimal = 0
        numReject = 0
    numAccept = 0
    decAppResult = 0
    decRejResult = 0

    numReject = FormatNumber((numRejected / numExamples), 4)
    numAccept = FormatNumber((numApproved / numExamples), 4)

    'For each instance to classify
        For Each dtRow As DataRow In dtTable.Rows
            dsTemp = New DataSet
            intTempReject = numReject
            intTempApprove = numAccept

            'get probability for each attribute
            For c As Integer = 0 To dtTable.Columns.Count - 2
                strSQL = "Select " & dtTable.Columns(c).ColumnName & ",COUNT(*) as num,Result from
                CreditApplication where " & dtTable.Columns(c).ColumnName & "='" & dtRow.Item(c) & "' GROUP BY
                " & dtTable.Columns(c).ColumnName & ",Result order by " & dtTable.Columns(c).ColumnName & "
                asc,Result asc"
                dsTemp = objDB.GetTempData(strSQL)

            Using dsTemp
            With dsTemp
            If Not dsTemp Is Nothing Then

            'Apply Bayes Theorem to get the MAP Hypothesis

```

```

If .Tables.Count > 0 Then
  If .Tables(0).Rows.Count > 0 Then
    If .Tables(0).Rows.Count > 1 Then
      Reject = .Tables(0).Rows(0).Item("num")
      Approval = .Tables(0).Rows(1).Item("num")
    ElseIf .Tables(0).Rows.Count = 1 Then
      If .Tables(0).Rows(0).Item("Result") = 0 Then
        Reject = .Tables(0).Rows(0).Item("num")
        Approval = 0
      Else
        Approval = .Tables(0).Rows(0).Item("num")
        Reject = 0
      End If
    Else
      Approval = 0
      Reject = 0
    End If
  End If

  If Approval <> 0 And Reject <> 0 Then
    intTempReject *= (Reject / numRejected)
    intTempApprove *= (Approval / numApproved)

    ElseIf Approval = 0 And Reject = 0 Then
      'Missing values - Omit the attribute from calculation

  Else
    'Apply Laplacian Correction - deal with zero probability values
    Dim dr() As DataRow = Nothing, intNum As Integer = 0

    If Not dsAttribValues Is Nothing Then
      dr = dsAttribValues.Tables(0).Select("Attribute='" & dtTable.Columns(c).ColumnName
      & "'", "ValueLabel")
    End If

    If Approval > 0 Then
      intTempApprove *= (Approval / numApproved)
    Else
      If Not dr Is Nothing Then
        If dr.Length > 0 Then
          intNum = dr.Count
          'Add 1
          Approval = 1 + Approval
          'Add the total counts to the denominator
          intTempApprove *= (Approval / (intNum + numApproved))
        End If
      End If
    End If

  End If

  If Reject > 0 Then
    intTempReject *= (Reject / numRejected)
  Else
    If Not dr Is Nothing Then
      If dr.Length > 0 Then
        intNum = dr.Count
      End If
    End If
  End If

```

```

                'Add 1
                Reject = 1 + Reject
                'Add the total counts to the denominator
                intTempReject *= (Reject / (intNum + numRejected))
            End If
        End If
    End If
    End If
    Else 'Missing values - Omit the attribute from calculation
        strValue = ""
    End If
End If
End If
End With
End Using
Next

'Get the Class of the new instance as either Approved or Rejected
If intTempReject > intTempApprove Then
    Result = "Rejected"
    intResult = 0
Else
    If intTempReject = intTempApprove Then
        strValue = ""
    End If
    Result = "Approved"
    intResult = 1
End If

decAppResult = intTempApprove
decRejResult = intTempReject

Next

Catch ex As Exception
    Throw ex
End Try
End Sub

```





