# KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY, KUMASI, GHANA



## Enhancing Direct Marketing and Loan Application Assessment Using Data Mining: A Case Study of Yaa Asantewaa Rural Bank Limited

**By**

**Anokye Acheampong Amponsah**

**(BSc. Information Technology Education)**

**A Thesis submitted to the Department of Computer Science, College of Science in partial fulfillment of the requirements for the degree of**

**MASTER OF PHILOSOPHY**

**APRIL, 2016**

# CERTIFICATION

I, Anokye Acheampong Amponsah hereby declare that this submission is my own work towards the award of MPhil Information Technology and that, to the best of my knowledge, it contains no material previously published by another person, nor material which has been accepted for the award of any other degree of university, except where due acknowledgements have been made in the text.

Anokye Acheampong Amponsah …….. ……..

PG2574814 (20368649) Signature Date

**Certified by:**

Kwaku Agyepong Pabbi …….. ……..

(Supervisor) Signature Date

**Certified by:**

James Ben Hayfron - Acquah (Dr.) …….. ……..

(Head of Department) Signature Date

# DEDICATION

I dedicate this work to my family (Mr. Benjamin Franklin Amponsah, Mrs. Theresa Owusu, Nana Kusi Amponsah and Owusu Amponsah) and also to my future wife and children of Anokye Acheampong Amponsah.

# ACKNOWLEDGEMENT

I wish to express my profound gratitude, appreciation and thanks to the Almighty God who has granted me good health and wisdom to produce this work. My sincere thanks go to my family for their support and motivation all this time. Also I say thank you to my supervisor Mr. Kwaku Agyepong Pabbi of Kwame Nkrumah University of Science and Technology, Kumasi. Without his guidance, patience, constructive criticism and motivation this work would not have been a reality.

I owe much gratitude to all lecturers who handled me in my course of program. I say God bless you all.

I also want to say thank you to the General Manager, Branch Manager, loan officer, and Information Technology head for their support, cooperation and kindness.

Finally, my thanks go to all my mates who have taken greater patience and care to be with me throughout our program especially Obaweya Seyi, Fuseini Inusah, Kwabena Antwi Boasiako, Darkwa Michael, Gideon Evans Norvor and Michael Osei Boakye.

# ABSTRACT

Due to the wide availability of the web and the internet, ever increasing processing power and the continuous decline in the cost of storage devices, data are being generated massively today than it were decades ago. With fields like the banking industry, data are being generated massively on regular basis. So Managers and Administrators are finding ways to turn these data into very beneficial information to their advantage. Marketing risk and Credit risk are considered the most serious problems of every bank. Fortunately all marketing promotions are highly dependent on the data about customers stored in electronic format. Past events contain patterns that are hidden in the data that records them. Data mining, machine learning and artificial intelligence promises the use of models to go through and analyze huge data that are very difficult for human experts. Data mining also possesses incredible power to detect hidden relationships, correlations and associations in data. This study was conducted to demonstrate with practical methods, experiments and datasets that data mining can be used to assist in direct marketing and analyze credit risk assessment. The two algorithms used were decision tree and random forest. The University of Waikato data mining open source software (Weka) was used to simulate all the experiments. Confusion matrix was used to calculate the accuracy, sensitivity and specificity which were used to evaluate the performance of the models. Receiver Operating Characteristic curve was also used to pictorially display the performance of the models. The results of the experiment showed with high precision that the models can be used in detecting prospects for marketing campaigns and

also detecting risky loan applicants. The decision tree classifier produced an accuracy of 92.5% and the random forest classifier produced an accuracy of 76.4 %

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATIONS

| | | |
|---|---|---|
| CRISP-DM | – | Cross Industry Standard Process For Data Mining |
| DM | – | Data Mining |
| FN | – | False Negative |
| FP | – | False Positive |
| TP | – | True Positive |
| TN | – | True Negative |
| ID3 | – | Iterative Dichotomiser 3 |
| KNN | – | K-Nearest Neighbors |
| ROC | – | Receiver Operating Characteristics |
| NN | - | Neural Networks |
| ANN | - | Artificial Neural Networks |
| MLPNN | - | Multilayer Perceptron Neural Network |
| LR | - | Logistic regression |
| SVM | - | Support Vector Machine |
| NB | - | Naïve Bayesian |
| KDD | - | Knowledge Discovery in Data |
| NA | - | Not Applicable |
| DT | - | Decision Tree |
| RF | - | Random Forest |
| FPR | - | False Positive Rate |
| TPR | - | True Positive Rate |

# CHAPTER ONE

# INTRODUCTION

This chapter presents introduction of the main contents of this thesis. It presents the background of the study, problem statement and justification, the purpose and objectives of study, the research questions which the study attempts to answer, significance, scope, limitation and organization of the study.

## 1.1 BACKGROUND TO THE STUDY

All over the world companies and organizations generate huge amount of data on daily basis which are accumulated automatically or manually by clearing a customer with barcode reader, filling a form on a webpage, clicking on a button to save the entered data, clicking on a link to other websites or accessing corporate databases using an android app on a smart phone. Ghana is considered one of the developing countries in West Africa, with numerous organizations and institutions that generate a lot of data on daily basis as part of their operations. These institutions/organizations include banking, financial, educational institutions, supermarkets, shopping malls, insurance companies, publicity, Non-governmental organizations and security agencies. These institutions are operating in a different realm with the development of ICT (Du, 2006).

Du (2006) writes that company managers are now in possession of gigantic data. These companies and institutions could be more fruitful and increase their productivity when they have in their possession tools that can turn these huge data into useful information. With these tools companies can make their customers the number one priority and desist from product oriented to customer oriented notion so as to survive in today's competitive market places where customer demands are

constantly changing. With this change over, organizations could channel resources in the direction which would be more productive. Customers are always looking forward to new and innovative ways of doing business where satisfying ideas, inventions, products and services must be constantly introduced to them. Therefore, corporate operations must regularly change to satisfy such expectations.

To be able to satisfy almost all customers there is the need for a competent system that would be used as a business tool to pinpoint, select, acquire and grow the loyal and most profitable customers. Such systems when available would help corporate organizations to analyze huge amounts of data and would be more helpful even when applied on transactional data. To expand customer acquisition and retention, organizations can also hit into the territory of unstructured data such as personal daily life, behaviours on social media, customer suggestions or feedbacks and support requests where conclusions are virtually impossible for the human experts.

The marketing departments of organizations play significant role in the introduction of products and services to the outside world. The marketing department is an equally powerful section of any organization and makes significant use of organizational resources. So the fact is if there can be a drastic decrease in the expenditure in the marketing department then the annual profits of the entire company can also increase. Also banks and other financial institutions are faced with the problem of frauds and bad loans where loans granted by the banks are not paid back as agreed.

Yaa Asantewaa Rural Bank ltd is located at Ejisu in the Ashanti region of Ghana. The Bank has been in operation since 2012 and has been serving its customers with a number of loan products. Aside the loan products there are different types of accounts that can be opened and used by their customers. The bank does direct marketing of its products to customers with respect to the available loan products and accounts. There is a marketing team (staffs) which develops a weekly “root

plan" that guides them as to who to contact, what product to sell and whom to sell to, the route to take and so on. This process is aimed at selling the Bank's products to the community people whiles expanding the customer base. In other words the ultimate aim of this direct marketing process is to acquire more customers and also increase sales of products. Although the Bank cannot state the exact amount of money spent on this campaign as there is no specific documentation covering the activities, it is considered to be expensive since prospective customers who reject the offered products are contacted several times more until there is no reason to contact them anymore. The Loan Manager gave two reasons for stopping to contact the customers and these are either the customers eventually buy a product or the marketing team is convinced about the impossibility to attract that customer. There are other rural and private banks in the district which also dispatch their mobile bankers to sell products. With this, it is quite obvious that marketing decisions on how to pinpoint customers who have high tendency of purchasing a loan so as to limit the number of contacts per person is very crucial for the bank in the district. In order to stay on top of the competition, this study is undertaken to use data mining to enhance the process of marketing and also to analyze loan assessment to prevent bad loans.

Figure 1 is a graph depicting the number of approved loans between October, 2012 and September, 2015.

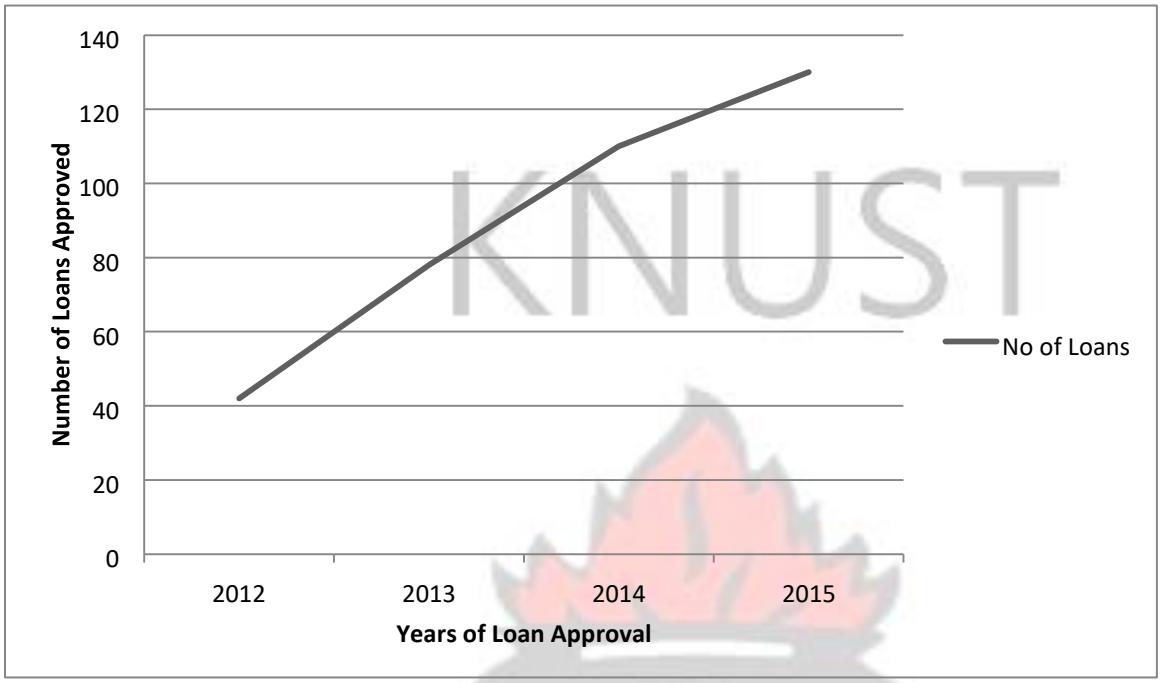



**Figure 1. 1: The Number of Loans Approved (Oct 2012 - Sept 2015) Source: Yaa Asantewaa Rural Bank**

## 1.2 WHY DATA MINING

According to Linoff and Berry (2004), data is being and will always be produced regularly by organizations and DM has the power to make sense of these huge data. A significant benefit of Data mining is the discovering of hidden pattern and the identification of relationships among variables and also assisting the companies in seeing each of their customers individually (Kiron et al, 2012).

DM has the capacity to operate on data that are extracted from several sources which include but not limited to Web, databases and data warehouses (Han and Kamber, 2012). The concept of data warehousing has made it possible for the integration of data from different sources and data warehouses are programmed and configured to support data mining (Linoff and Berry, 2004).

The tools and techniques available in DM could be used by the business to make helpful decisions. By using these data mining techniques, organizations and institutions can mine information that does not exist about their customers and products and by this can simply define the ideals of customers and forecast their future behaviours and requirements (Bhambri, 2011).

Data mining uses several techniques to achieve its purpose. These techniques are grouped into Supervised and Unsupervised learning (Zhaohan, 2009). Each of these tools consists of a number of algorithms and has its unique benefits whiles others cut across. In literature, prediction and classification techniques of data mining are found to be very helpful in any organization. Prediction has been proven to accurately predict the success or otherwise of a product by estimating the number and identifying the characteristics of customers who may probably be interested in that product. In a typical retail shop, transactions of every customer are stored in a database. Such database can be mined and prediction can be made on a similar product in terms of cross selling and up selling (Chopra et al 2011). Other authors also state the importance of the prediction technique in accurately identifying customers with the tendency of not paying back a loan with the agreed interest rates and time frame. In other words, it is very good at preventing risky loans. Classification works best in categorizing datasets into their respective groups. With this profitable and non-profitable customers can be identified which suggest to bank executives which group of customers to focus their attention on (Bhambri, 2011, and Chopra et al 2011).

It is from this corridor that the project stems to implement data mining techniques as efficient and cost effective way of marketing loan products and analyzing credit risk of loan applicants in the Yaa Asantewaa Rural bank Ltd. in Ghana.

## 1.3 PROBLEM STATEMENT AND JUSTIFICATION

The fact is, more businesses fail and are eventually shut down because of unsatisfactory customer services and more. Examples include Solyndra in America (Gattari, 2012 and Dick, 2013). Business authorities spend time and huge sums of money to import products that are of low significance to their current customers. Many more financial institutions (credit unions, savings and loans, microfinance and banks) collapse because of high risk in granting loans to non-creditworthy customers and focusing attention on non-profitable customers. Typical examples are Work Up Microfinance Company Limited, Noble Dream financial service, Lord Winners Microfinance Westbanc Capital Group and many more in Ghana (Boateng et al, 2016).

Financial institutions spend more time and money on designing and implementing services and advertising them on the media but they do not get the anticipated returns.

Yaa Asantewaa Rural Bank has a team of marketers (staffs) which introduces the products and services of the Bank to the community people of Ejisu. They develop a weekly "root plan" that enhances their movements. This method is not efficient and cost-effective as about 50 percent or less of the prospects approached only respond by purchasing a product. This means 50 percent or more reject the offer. Out of the 50 percent or less prospects that accept to buy a product, about 10 to 20 percent are found to be risky if granted loans. In this study two major problems of the bank are solved using data mining models (algorithms). First the models are applied on existing prospective customer data on loans marketing to classify and predict customer behavior in order to accurately identify potential customers to direct marketing efforts to and second to make sure loans are approved for non-risky customers. It was anticipation that the study would ensure that adequate returns are made on every cedi spent on marketing and granted loans are repaid as scheduled.

## 1.4 PURPOSE AND OBJECTIVES OF STUDY

### 1.4.1 General Objectives

The main objective of this work is to investigate into how to reduce marketing cost of rural banks in Ghana by using data mining techniques to identify potential customers and also using data mining to assess the credit worthiness of loan applicants.

### 1.4.2 Specific Objectives

The specific objectives are

- To determine how data mining model can be used to anticipate the number of probable customers to be interested in a particular loan product so as to:
  - Limit the scope of marketing and enhance administrative focus.
  - Ensure customer-tailored loan products.
- To design, implement and test a data mining model for determining the credit worthiness of customers so as to grant loans to non-risky and profitable customers.

## 1.5 RESEARCH QUESTIONS

- How can the scope of marketing be limited and administrative focus enhanced when prospective customers to be interested in a loan product are anticipated using data mining model?
- Can credit worthiness of customers be determined by using mining model so as to grant loans to worthy and profitable customers?

## 1.6 SIGNIFICANCE OF THE STUDY

In general terms if the bank is able to drastically reduce the cost of marketing to the general public and increase their income, automate loan approval procedures and approve loans for nonrisky customers, the local people can receive loans which are tailored to their individual needs at the appropriate time, thereby allowing them to embark on their dreams and intentions. Since most of the local people are business people, it is expected that there will be booming and flourishing of businesses in the local community and living standards will be improved. The following are Specific significance:

- It was expected that, by the end of the study the bank institution would be able to predict the future behaviours of their existing and prospective customers and forehand devise solutions to posing threats. Specifically the information identified by the DM model could help in significantly reducing marketing cost in that the characteristics of prospects to be contacted in the Bank's "root plan" would be known and resources would be channelled towards marketing loan products to prospects that match such characteristics.
- Also, it was expected that profitable and non-profitable customers would be identified after implementing data mining techniques on existing customer data. With this, the institution would be able to personalize the marketing experience. Customized services would be implemented based on such findings.
- Furthermore, it was expected that the institution would approve loan application of credible customers thereby reducing the number of bad loans or portfolios at risk where customers would not be able to pay back the granted loans or at the agreed interest rates within the stipulated time.

- There is a community of researchers who are interested in using data mining to enhance marketing processes and credit risk analysis. The findings of this study are expected to contribute to that of the community.
- Lastly students and researchers who would want to research into similar topics of using Data mining in marketing, Loan approval systems, credit risk assessment may find this work useful and as a reference guide to their work.

## 1.7 SCOPE OF THE STUDY

The study was confined to only one selected financial institution – Yaa Asantewaa Rural Bank ltd in Ejisu in the Ashanti region of Ghana. The data set used in the model was collected, corrected and modified to suit the environment of the bank. Therefore the generalization, conclusion and recommendations would not be made to include any other financial institution in the region/country. In view of this, a more complete generalized study would have to be conducted and this should include more financial institutions in Ghana.

## 1.8 LIMITATION OF THE STUDY

Time and financial constraints were the major limitation factors as far as this study is considered. Also, sensitivity, confidentiality and privacy issues accredited to customer data also prevented some information from being included in the study.

Bank of Ghana policies and directives restricted the exposure of certain documents. Also the Bank's security policies prohibited access to certain aspects of the organizational assets (Building, IT rooms, Banks Database and Certain information).

## 1.9 ORGANIZATION OF THE STUDY

The study under discussion is structured into six chapters. The Chapter one is the introduction which has the background of the study, Why Data Mining, Statement of the problem and justification, Purpose of the study, Research questions, Significance of the study, Scope of the study, Limitations of the study, and Organization of the study as its sub-headings.

In Chapter two, related literature was reviewed. All the reviewed literature was in the light of data mining in the banking sector. The sections are Data mining and data mining techniques (Classification, Estimation, Prediction, Affinity grouping or association rules, Clustering, Description and visualization), Processes involved in data mining (KDD), Importance and application of data mining, How data mining can assist in marketing in the bank (Customer relationship management, Marketing, Risk management, Fraud detection, Present industry status) and Ethics in data mining marketing. Chapter three embodies the methodology of the study. The following headings are discussed. Research design (descriptive research design, openended interviewing, case study), the case (rural bank), description of the data set, model evaluation methods (cross validation, confusion matrix, ROC curve), Weka (waikato environment for knowledge analysis) software, the cross-industry standard process for data mining (CRISP-DM) (business understanding, data understanding, data preparation, system modeling, system evaluation, deployment) and privacy and confidentiality of customer data. Chapter four presents system modeling and performance evaluation procedures. It discusses how the two models were developed and evaluated. Two main headings are decision tree modeling and random forest construction. Chapter five deals with results and analyses, two main headings under which there are subsections are J48 decision tree classifier and random forest classifier.

Chapter six is about the summary, conclusion, recommendations and future works.

# CHAPTER TWO

# LITERATURE REVIEW

## 2.0 INTRODUCTION

This chapter discusses the relevant literature of other authors who researched into areas where Data mining techniques are very helpful in the Banking sector. This chapter is grouped under the following sub-headings:

- DATA MINING AND DATA MINING TECHNIQUES
    - Classification ○ Estimation ○ Prediction
    - Affinity Grouping Or Association Rules
    - Clustering ○ Description and visualization
- IMPORTANCE AND APPLICATION OF DATA MINING
- HOW DATA MINING CAN ASSIST IN MARKETING IN THE BANK SECTOR
    - Customer Relationship Management ○ Marketing ○ Risk Management ○ Fraud Detection ○ Present industry status ○ Customer Churn/Attrition ○ Pattern Management
- EMPIRICAL EVIDENCE
    - Enhancing Direct Marketing using Data Mining ○ Enhancing Loan Application Assessment using Data mining

- ETHICAL AND PRIVACY IN DATA MINING AND MARKETING
- SUMMARY OF THE CHAPTER

## 2.1 DATA MINING AND DATA MINING METHODS

With the advancement of technology and ever increasing doubling power of information technology, huge amounts of data are generated on daily basis which has called for the need to turn these data into useful information. Data mining (DM) has as its main objective as to fish out unidentified patterns in huge data. DM offers a platform for extracting non-existing pattern or relationship in huge data.

Raju et al (2014) cites several authors in defining Data Mining as follows;

Cabena et al defines Data mining "*as the process of extracting previously unknown, valid, and actionable information from large databases and then using the information to make crucial business decisions*".

Fabris also describes Data mining "*as the automated analysis of large amounts of data to find patterns and trends that may have otherwise gone undiscovered*".

Because the computer hardware manufacturing and software development companies are always coming out with new and advanced technological tools, computational power, programming logic and huge storage devices, it has made the generation, processing and storage of data very easy. And with this data warehousing and data mining are acquiring more recognition. Another concept that is very useful in the world of business is ***what if analysis*** that makes it possible for decision makers to change values in order to make proper predictions and estimations. Data Mining empowers companies to touch their clients and customers with the correct item, at the right charge and at the exact time. (Chopra et al, 2011).

Bill Inmon gave a definition of DM which was cited by Chopra et al (2011), "*it is a subjectoriented, integrated, time variant and non-volatile collection of data in support of management's decision making process*".

Data Mining can be defined as the method of evaluating big data from different sections and generating useful information. The basic purpose of DM is helping companies to make good use of their data by discovering buried links among a group of customers and also identify fresh and innovative connection between data (Chopra et al, 2011).

According to Doug Alexander, in his article "Data Mining", DM involves the process of using the computer to mine and analyze huge sets of data and then transforming them into very useful data. DM presents tools that are used to forecast performances and upcoming events. This would allow businesses to prepare in advance. DM tools can move through large records of data and solve business challenges that are impractical or impossible for the human brain or other computer related techniques. They also search databases for unseen correlation or relationship, discovering prognostic information that experts would not consider because it is not interesting to them (http://www.laits.utexas.edu/~anorman/BUS.FOR/course.mat /Alex/).

To Han and Kamber (2012) "*Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically."*

To make better and concrete business judgments which are very necessary for the survival of the business, Singh et al (2015) discovers that, Data mining techniques provide a better business decision making scheme by discovering patterns in customer behavior. These patterns can be very helpful in designing and implementing strategic marketing policies, sales approaches and

purposefully distinguishing between profitable and non-profitable customers and also instituting inexpensive customer support structures.

There are several types of patterns or relationships that can be revealed after conducting data mining and they deeply rest on the data mining tasks used. Zaïane (1999) summarizes them into two main types of tasks. "*Descriptive data mining*" tasks which tend to give general descriptive information about the current data set, and "*predictive data mining*" tasks which also tend to make forecasts focusing on existing data set.

### 2.1.1 Groupings of data mining methods

There are numerous ways data mining can take to achieve the expected results. They can be grouped by objects, techniques and tasks. By categorizing data mining methods using objects, they can be classified as relational database, spatial database, object oriented database, temporal database, multimedia database, heterogeneous database, text data sources and web source (Du, 2006). By grouping data mining methods using techniques, they can be grouped into machine learning, statistical methods and Neural networks. This literature combines the grouping by task and techniques.

### 2.1.2 Grouping data mining methods by tasks and techniques.

Han and Kamber (2012) as well as Zaïane (1999) again identified the practical usefulness of data mining to consist of "*characterization and discrimination, extracting frequent patterns, association, correlation, classification and prediction, cluster analysis, outlier analysis and evolution analysis*". The functions of data mining can be grouped into six (6) main classes depending on their activities according to Anand et al (2014) and Radhakrishnan et al (2013). These are;

- Classification
- Estimation
- Prediction
- Affinity grouping or association rules
- Clustering
- Description and visualization



#### 2.1.2.1 Classification

Classification involves the process of grouping datasets into their respective categories. This makes use of class labels to mark items in the data using machine learning (Zaïane, 1999). Classification is used to estimate or predict the likelihood of a variable belonging to a group or class. For example in a typical financial institution, the manager may want to know the number of customers who would be able to pay back their loan with the agreed interest rate and duration and those who may not. The classification method could place the customers into Not-Risky, Less-Risky and Very-Risky classes. Fraudulent cases and credit risk detection are best area to apply classification (Zaïane, 1999).

In an online document made by Voznika and Viana, Classification involves revealing unknown variables using already known variables as input. The algorithm or classification model works on a set of data called training data which contains some characteristics and their associated results. The associated results which are normally called goal or prediction attribute together with the other characteristics of the training data is used to train the algorithm in detecting relationships. After training the algorithm or in other words after the algorithm learns from the data, it is further fed with different set of data with the same attributes as the training data (predictive or test data). Then based on the detected relationship, values are predicted for the test data. A test data is a set of data that does not depend on the training data, but has the same probability distribution as the training

data". The only difference between the training data and the prediction (test) data is the prediction attribute which is unknown for the prediction (test) data. The algorithm processes the input (prediction/test data) and brings out a prediction. In order to determine the effectiveness and efficiency of the algorithm and desist from *overfitting*, DM model performance characteristics like sensitivity, specificity, accuracy, F-measure and more need to be calculated.

Finding a model is the main idea behind Classification. It is the model that labels and separates the classes. The generation of the model involves the study of training data by the algorithm – thus variables for which labels are known, in order to predict the variables for which labels are unknown (Han and Kamber, 2012). The variables to be grouped are usually extracted as records from database tables or files and the process of classification involves the addition of a new attribute called the class labeller or class code (Radhakrishnan et al, 2013).

Singh et al (2015) and Han and Kamber (2012) indicate that, classification methods/models uses scientific methods like classification rules (if-then rules ), statistics, linear programming, neural networks, and decision trees. Neelamegam and Ramaraj (2013) also add K-Nearest Neighbor, Support Vector Machines and Naive Bayesian Classification as the most common algorithms used for classification in data mining.

| AGE GROUP | AGE DESCRIPTION | ACCOUNT | INSURANCE |
|---|---|---|---|
| Below 20 | Young | Savings | No |
| Below 20 | Young | Savings | No |
| 21 – 40 | Medium | Salary | Yes |
| Above 40 | Aged | Salary | Yes |
| Above 40 | Aged | Salary | Yes |
| 21 – 40 | Medium | Savings | No |

**Table 2. 1: Example of Training Data for a Bank**

| AGE GROUP | AGE DESCRIPTION | ACCOUNT | INSURANCE |
|---|---|---|---|
| 21 – 40 | Medium | Savings | ? |
| Below 20 | Young | Savings | ? |
| 21 – 40 | Medium | Salary | ? |
| Above 40 | Aged | Salary | ? |
| Above 40 | Aged | Salary | ? |
| Above 40 | Aged | Salary | ? |

**Table 2. 2: Example of Test Data for a Bank**

#### 2.1.2.1.1 Decision Trees

*"A decision tree is a flowchart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions"(*Han and Kamber 2012). A DT is among the most commonly implemented methods in data mining. It accepts numerous inputs and based on these inputs makes prediction. In a decision tree each internal node contains an input variable. Each internal node can have two or more children. The branch from an internal node to its children represents a condition which must be satisfied before the child is selected. After moving through a decision tree from the root node, the outcome is decision nodes and leaf nodes. A decision node which can further have two or more branches becomes internal node. Leaf node without any further branches represents a classification or decision. Decision trees can process categorical and numerical data (http://www.saedsayad.com/decision_tree.htm ). For a decision tree to be considered comprehensive, it has to be less complicated. Furthermore, Breiman *et al*. (1984) indicates that the complexity of the tree has a critical influence on the accuracy. The complexity of the tree is handled by the stopping criteria used and pruning method to be used. Under normal circumstance the complexity of the tree is determined by these subsequent factors: the total of nodes, leaves and

attributes used and the tree depth. Each single route (path) from the parent (root) node to the leaf node can be converted into a simple classification rules (IF-THEN statements). From the table 2.1 and 2.2 the corresponding decision tree may be:

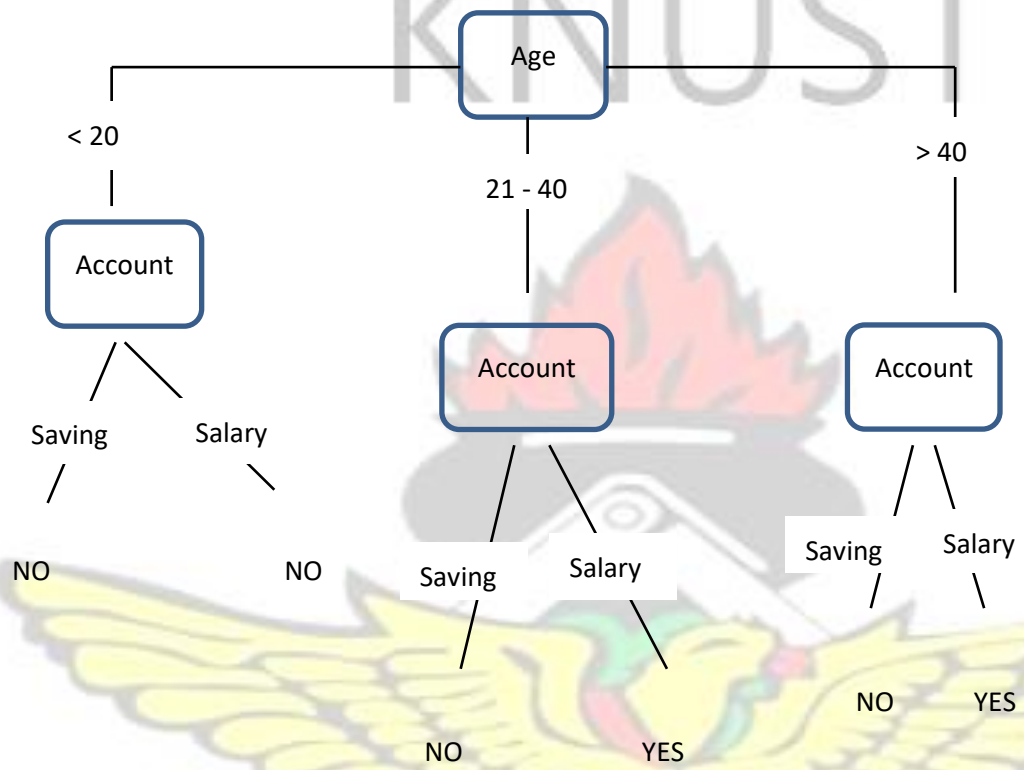



**Figure 2. 1: Decision Tree for Bank based on Table 2.1 and 2.2**

Figure 2.1 is a decision tree generated based on the values in table 2.1 and 2.2 only. Some paths in the tree may be ignored for this example.

#### 2.1.2.1.2 Neural Networks

In the view of Sakshi and Khare (2015), neural networks which are usually called artificial neural networks, is a mathematical model that has its root from biological neural networks. It possesses the capability of processing data in a non-linear form using appropriate statistical tools.

Neural networks contain a number of linked parts usually called neurons, units, or nodes. The units in neural networks collaborate with one another to come out with an output function. Thus the

units of neural networks are interconnected together in such a way that they execute input information by adopting a connectionist style of calculation (Sakshi and Khare, 2015).

Neural networks possess the capability of identifying relationship between a set of variables (data) that are extremely difficult to discover using the human brain or some other methods using computer. For the reason that neural networks use connectionist style of calculation, they are able to work even when some units are not working properly.

Neural networks mostly make precise predictions. However on the black side, one of their shortcomings is that they offer "black-box" style of research which means that they do not give a deeper understanding of the phenomena understudy (Voznika and Viana). The capability of neural networks to change its internal arrangements and configuration when exposed to external and internal data makes them adaptive systems (Sakshi and Khare, 2015).

In Nikam's work, he indicates that NN is sourced from appreciating and imitating the human brain to replicating human abilities like speech. He also writes that NN can be applied in many fields like banking, legal, medical, news and other computer programs. Inputs to a node can be output from other nodes or input from data sets (Nikam, 2015).

**2.1.2.1.3 Classification Rules**

Classification rules are also referred to as prediction rules. They are identified to be one of the common algorithms used in classification to define knowledge. A set of **IF-THEN** statements are the easiest form for hypotheses. The **IF** part is followed by condition(s) and the **THEN** part is followed by a set of variables that are predicted or classified based on the satisfaction of the condition (Voznika and Viana). Using table 2.1 and 2.2 classification rules may be the following;

**IF (Age Group <= 20 AND Account = Savings) OR (Age Group <=40 AND Account = Savings) THEN Insurance = No.**

**IF (Age Group <=40 AND Account = Salary) THEN Insurance = Yes.**

In this example the outputs (prediction) of the prediction/classification rule (Yes or No) is highly dependent on the satisfaction of the conditions. There is the tendency of having extensive statements than the above. With the use **OR** each variable that satisfies the condition in all attributes are captured. On the other hand using **AND** only captures variables that satisfy the condition in both attributes.

#### 2.1.2.1.4 K-Nearest Neighbor

Sakshi and Khare (2015) writes that, K-Nearest Neighbor (K-NN) is also a useful and simplest method used in machine learning usually in classification and regression. Data input for both classification and regression is the same but the results differ. Thus the data to be processed are the same but the result of that input after processing determines whether it was classification or regression. The difference is, in classification done by K-NN means the outcome is class membership but in regression the outcome is property value in object (Sakshi and Khare, 2015).

It is very beneficial to give a value that the neighbors in K-NN can contribute in classification and regression, when this happens the closer neighbors can donate more points on the average than the farther neighbors. According to Nikam (2015) T. Bailey and A. K. Jain improved K-NN by letting it focus on weights. A typical weighting scheme can be assigning every neighbor a value of *1/d* where *d* stands for the distance between k and the neighbor. Objects with known classes are termed as neighbors for classification in *k*-NN or object property value for regression in *k*-NN.

Forecasting can be made for unidentified data sets using K-NN. Euclidean distance and overlap metric are used to calculate space metric for continuous variables and discrete variables respectively. A problem with k-NN algorithm is its sensitivity to local data structure. Nikam (2015) also indicates that, the training data are used to find out which group a given data belongs. These

data sets are stored in memory during computation hence the name "memory based technique". Because K-NN adopts this strategy, processing power and memory storage are some of the challenges. To limit the impact of the challenges Nikam (2015) writes that, the iterative patterns and information focus which do not contribute any additional information to the data set and influence the outcome are eliminated. Different schemes could be adopted to solve the memory problem in K-NN. These include the use of orthogonal search tree, k-d tree, principal axis search tree, nearest feature line (NFL) and ball tree.

**K** → number of nearest neighbors

**For each** object X in the test set **do**

Calculate the distance D(X, Y) between X and every object Y in the training set

neighborhood ! the k neighbors in the training set closest to X

X.class → SelectClass (neighborhood)

**End for**

#### 2.1.2.1.5 Support Vector Machines

According to Cortex and Vapnik (1995), the first algorithm of Support Vector Machine (SVM) was developed by Vladimir N Vapnik and Alexey Ya. Chervonenkis (1963). A new way was proposed to make non-linear classifiers in 1992 by Bernhard E. Boser, Isabelle M. Guyon and Vladimir N Vapnik. The creation of the non-linear classifiers was done by applying the kernel trick to maximum-margin hyperplanes. The kernel presents standard incarnation (soft margin) that was proposed by Corinna Cortes and Vapnik (1993).

SVM handles pattern classification, thus it is not used for object (dataset) classification. Currently, there are two forms of patterns - Linear and non-linear. Patterns that are considered to be linear are

identifiable easily. They have several hyperplanes that can classify the data. Nonlinear patterns are not easily separable and for this reason they need additional processing before for easy distinguishability.

The basic idea behind SVM is the maximization of the margin so that it can accurately group the given patterns. It has been identified that the larger the margin size the more accurate the classification of the patterns is done (Archana and Elangovan, 2014).

According to Nikam (2015), the efficiency of the classification that is centered on SVM does not depend on the size of the categorized objects. Despite the fact that SVM is considered the most reliable classification technique, it has a number of pitfalls. It is computationally costly since SVM employs convex quadratic programming in its data analysis. This kind of approach requires huge matrix manipulations and mathematical calculations.

This is an example of a hyperplane;

Hyper plane, $aX + bY = C$

**2.1.2.1.6 Naive Bayesian Classification**

In machine learning, this technique uses the approach of set of simple probabilistic classifier with the application of Bayes' theorem on strong (naive) independence assumptions. The core probability model could be termed as "independent feature model" (Archana and Elangovan, 2014). Bayesian classifier presents a very credible approach that is used to consider and assess set of learning algorithms. It is considered to be very strong when dealing with data with lots of inconsistencies and it executes clear probabilities for hypotheses. Unlike other probability algorithms, Bayes' classifiers can be set up properly in a controlled learning environment. It adopts the strategy whereby the inclusion or exclusion of a specific attribute is separated from the inclusion or exclusion of any other attribute when classifying the variables into their respective

classes (Nikam, 2015). Another name for this approach is idiot's Bayes, simple Bayes or independence Bayes.

According to Archana and Elangovan, (2014), the Bayes' classifier is used because of the following factors.

- Ease of construction
- No need to approximate any difficult iterative parameters.
- Easy applicability to massive data sets.
- It is stress-free to explain,

These make workers with no skill in classifier technology work with and appreciate how the algorithm works and why it is making those classifications.

**2.1.2.2 Estimation**

Estimation works with constantly cherished results. When the algorithm is fed with some input data, estimation would be able to show a variable for some new constant variable such as salary, stature or others. Estimation is usually used to complete classification duties (Radhakrishnan et al 2013). Sample works of estimation According to Anand et al (2014) include:

- Guessing the total of kids in a house using the mothers' education level as input
- Guessing overall household income using the number of vehicles in the family as input data
- Guessing the cost of a piece of a real estate using distance of that land from a major business district as input data.

#### 2.1.2.3 Prediction

Prediction and classification have some similarities. The difference is that, in prediction the rows are grouped in such a way that emphasis is placed on upcoming forecast behaviour or figure. The wait and see approach is used to check the accuracy of the model. It is because of extra concerns about the temporal pattern of the input values to the marked values that is why prediction is seen to be different from classification and estimation (Radhakrishnan et al 2013 and Anand et al, 2014 ).Prediction examples are:

- Forecasting the amount of cash to be transmitted should a credit card holder accept balance transfer offer.
- Forecasting the number of customers likely to attrite in the next few months.

#### 2.1.2.4 Affinity Grouping or Association Rules

Affinity grouping has as its main purpose, to identify related variables or materials. A perfect example is discovering the materials that are usually bought together in electronic shopping cart in e-commerce site. Affinity grouping is said to be the core of *market basket analysis*. In a shopping mall this DM tool can be used to put in place a scheme that would enhance the organization and display of stuffs. This usually brings stuffs bought together into one place for easy accessibility (Radhakrishnan et al 2013).

According to Perner and Fiss (2002), Association rule is also very useful in like manner. It gives deeper understanding by identifying links between various variables in data with no semantic reliance like customer behaviour. For example if a bank executive discovers that most customers

who are able to pay back their car loan also go in for house loan may add the two in one package just to create the sense of belongingness to the bank.

Given a set of records, where each record is a set of ordinary articles, an association rule can be expressed in the form of M N, where M and N are groups of the articles. The natural implication of this expression is that records in the data store which contain M also contain N (Anand et al, 2014).

**2.1.2.5 Clustering**

In clustering, Radhakrishnan et al (2013) indicate that variables of diverse characteristics are grouped into subcategories or clusters based on their similarities. Unlike classification, there are no already stated labels or sample data for which the algorithm would be guided in clustering. The rows are simply clustered depending on their similarities. In this case a room is left for the data miner (person) to assign any meaning to the subsequent clusters. Variables that are grouped together in one cluster should be very alike and variables from different clusters should be very unalike. After grouping the variables, the various clusters could be tagged with class labels and assign the records in the database to their respective classes. This can be used for data classification (Perner and Fiss, 2002).

According to Anand et al (2014), cluster analysis could be considered as independent data mining method that could be used to analyze data set or organize the data set that could be used by other data mining techniques (Anand et al, 2014). Density-based clusters, contiguous clusters, Center-based clusters, Well-separated cluster and Conceptual clusters or shared property are the common forms of clusters.

#### 2.1.2.6 Description and visualization

In the analysis of huge data set, figures or numbers are hard to concentrate on. Using graphics to portray the meaning of these numbers or figures would go a long way to help administrators and improve comprehension. Clustering is usually done in numbers and the dendrogram uses these numbers to plot graphs and make pictorial representations of the various groups and subgroups (Perner and Fiss, 2002). Data visualization is identified as a serious form of descriptive data mining (Anand et al, 2014).

Patil and Bhoite, (2015) cited Han and Kamber that data mining consists of a number of techniques that assist with extracting information from existing data and transforming them into significant trends and rules to improve your understanding and the usefulness of the data.

## 2.3 IMPORTANCE AND APPLICATION OF DATA MINING

Bhambri (2011) writes that DM assists in finding solutions to business problems by discovering correlations, associations and patterns that are unknown in the business data kept in the data stores.

In general terms according to Patil and Bhoite (2015), with the application of data mining to scrutinize patterns and trends in huge data that are generated every day, bank executives and administrators can forecast (predict), how customers may behave with the introduction of new products and services, new interest rates, those with high tendency of defaulting loans, and how to institute a more gainful customer relationship schemes, all these with the highest possible precision. The most popular data mining techniques include clustering and classification. The former is a technique that is used to predict the tendency of a data instance to belong to a group whiles the latter is a technique that is used to group similar and dissimilar variables into their respective groups (Patil and Bhoite, 2015).

There are numerous data requirements of any industry of which the banks are not excluded. For any industry to be in operation in today's competitive market place there is the need for the executives to uniquely identify each and every customer. The profile, transaction behavior and desired services are among the data every industry needs to collect. Indisputably such data when used with DM would allow the banks to cross sell their products, distinguish between profitable and non-profitable customers, and avoid defaults and customer attrition. Chopra, et al (2011)

## 2.4 HOW DATA MINING CAN ASSIST IN THE BANKING SECTOR

The following section discusses the various aspects where DM is applied in the banking sector. The sections are grouped into risk management, marketing, customer relationship management, software support, fraud detection, and present industry status.

### 2.4.1 Finding a good prospect

Knowing more about your prospects is very necessary because campaign messages or advertisements must be directed to the right audience (Linoff & Berry, 2011) and searching through all bank's data by hand to find a good prospect is impractical (Finlay, 2014). According to Radhakrishnan et al (2013), a good prospect can be defined as simply a person who would perhaps show that he is concerned with becoming a client. In other word anybody who signifies that he is interested in the company's products and is willing to do business with the company is termed as a good prospect. In determining who is a good prospect, they write that DM is used in establishing the criterion to be used in detecting who is a good prospect. The medium and rules established are used to target people with those characteristics. Companies use numerous ways to identify prospects. An example is the use of surveys in organizations where customer data is difficult to maintain like the print media but in companies where customer data is readily available DM can be used to find people who have similar characteristics with existing customers.

Radhakrishnan et al (2013) gave example of one nationwide publication that found out statistics that were used to identify good prospects.

### 2.4.2 Customer Relationship Management

The discovery capability of DM helps in finding new customers, separating the customers by grouping them and discovering the most appropriate ways to relate with them (Kadiyala & Srivastava, 2011).

In today's modern times, according to Bhambri (2011), customers are considered to be the most important assets of any company. The reason is simply because there are many more companies which are offering the same products and services in a competitive style. Customers can decide to switch to companies which are ready to satisfy their needs and provide the needed items at the right time, quantity and price. To maintain customers and acquire more in the future, customer satisfaction should not be taken for granted (Bhambri, 2011). With this the concept of productorientation where products are designed and developed with no specific customer in mind has paved way for customer-orientation where products are designed and developed for specific customers. Now the main objective is to touch the heart of the customer and create the sense of belongingness for the company. Interestingly, organizations are running and maintaining huge databases that store billions of records about their customers. With this huge data available, Data mining possesses the ability to categorize customers into their respective groups and then all actions necessary can be performed on them. Data mining is applicable in the three steps of the customer relationship cycle: attracting the Customer, instituting measures to grow the profits of the customer and running programs to keep the Customer. Doug Alexander also includes that DM predict the number of customers that are not satisfied with the current business process and have high tendency of swapping the company for its competitor.

According to Radhakrishnan et al (2013), the main focus of CRM is to increase the income gained from existing customers by allowing the companies to cross-sell and up-sell their products. DM is employed to discover the kind of products to offer, whom to offer the products to and at what time to offer the products. They cited Charles Schwab an Investment Company which applied data mining on customer historical data to enhance its corporate policies.

### 2.4.3 Marketing

Because the products to be marketed are tailored to the preferences of segmented customers, target marketing increases the certainty that a customer would certainly react to the campaign (Kadiyala & Srivastava, 2011). The banking sector according to Bhambri, (2011), likewise other companies is faced with stiff competition in the market places. In Ghana there are places with only one bank. The competition becomes stiffer when there exists about four or more banks in one district capital. Know Your Customer (KYC) is a popular word these days. Financial institutions are instituting powerful marketing strategies to win more floating customers and others from their competitors. On the other hand customer attrition is considered one of the problems facing the financial institutions and the cost it takes to sell to a new customer is estimated to be six times more than that of selling to existing customers (O'Brien, 2003). These attest to the fact that existing customers are the most valuable assets of any organization. To determine which customer is a good prospect for a product, how they would accept new interest rates or otherwise, which customers would patronize new product offers and determining risky customers to grant loans to, among others are areas where data mining can assist with tremendous benefits (Bhambri, 2011). By labeling the various customers into their respective groups, data mining can enhance the response rates in direct mail campaigns and this would go a long way to increase income. Also the classification done through DM can be used to target new customers that match such characteristics.

Shaw et al (2001) admit that making marketing decision like promotions, distribution channel and advertising media using customary segmentation style is ineffective and usually yields poor response rate and is very expensive. In recent times products are designed and marketed with the individual tastes of the customers in mind not to meet the general homogenous characteristics of all customers (Shaw et al, 2001). With this DM is more than capable of using the rich customer transaction data stored in huge databases to perform Database marketing.

Smirnov-M (2007) published an article titled "Data Mining and Marketing" on the website cited below and he said, to profile customers, database of customers or surveys/interviews results can be used. The profiling information can be in the form of buying habits, interests, desires, and more personal records like age, gender, marital status and annual earnings. Information like this can be simply examined using DM software (www.estard.com).

Database marketing methods are used by Banks to pinpoint the most loyal customers. With this data Banks can aim at customers for loan promotions, to estimate the cost of retaining a customer and to follow direct marketing campaigns (www.estard.com). In fact many companies (like telephone companies) use data mining to examine customers' calling patterns. In effect, they are able to suggest the most suitable plan for each new client. Again Chitra and Subashini (2013) writes that analysts in Banks can examine previous trends, estimate the current demand and be able to predict the behaviour of customers in terms of products and services to gain competitive advantage.

### 2.4.4 Risk Management

According to Foust and Pressman (2008) there are generally three main risks that exist in every bank. These are credit risk, market risk and operational risk. The one that affects banks massively is the credit risk. The process of managing credit risk typically involves identifying the possible

risk, measuring the degree of this risk, determining the suitable remedy and the practical execution of risk models (Van Gestel & Baesens, 2008). There are two types of credit risk analysis - traditional and advanced method (Dzelihodzic and Donko, 2013). The traditional method has proven to be ineffective in analyzing credit for new loan products (Abrahams & Zhang, 2009). The advanced method makes use of neural networks, K-nearest neighbor, decision tree analysis and support vector machine in analyzing credit risk (Laha, 2007, Wang et al, 2005, Dasarathy, 1991, and Quinlan, 1986). One common thing among all these methods is the use of data. Dzelihodzic and Donko (2013) writes that lately, a lot of researches has been done to implement data mining and machine learning tools in analyzing credit scoring.

DM techniques allow banks to differentiate between borrowers who can pay back loans on time from those who may not. DM also equips banks to predict when a borrower may be at default, and determining the credit risk (Chitra and Subashini, 2013). Risk management is a major problem of any business organization. In commercial lending, risk assessment is generally the process of trying to measure the risk of loss to be incurred by the lender while planning to lend money to a customer. Data mining technique can assist in labeling groups of customers as Risky or Less Risky where the former group has high rate of defaulting and latter group is found to be safe. By leveraging Data mining technique Bank administrators examine the characteristics of customers (Bhambri, 2011 and Davis, 2001).

### 2.4.5 Fraud Detection

Kadiyala and Srivastava (2011) termed this as deviation from the norm. They say Forensic analysis can detect delinquent characteristics and use them to categorize customers' behaviour into delinquency or exception to the detected pattern. This is very good in detecting fraud. Fraud is one of the major concerns raised when it comes to financial institutions. Both banks and their customers

are susceptible to fraud and for that reason they should always be on alert. For the two parties to be involved in a secure transaction there should be a system to detect fraudulent patterns in customer data. Data mining techniques have the ability of detecting such fraudulent patterns in customer data and predicting the behaviour in advance (Bhambri, 2011).

### 2.4.6 Present industry status

Banks worldwide have started reaping the benefits of data mining. Bhambri (2011) cites several authors when giving examples of banks that are benefiting from data mining. In his work, a bank in New York called Chase Manhattan Bank used DM techniques to examine customer data and by doing that came out with a strategy to solve its problem of customer attrition. Again Fleet Bank in Boston used DM to find the best nominees to offer mutual funds (Bhasin, 2006). ICICI, IDBI, Citi bank, HDFC and PNB in India have started benefiting from the use of DM. Specifically Citi bank has profiled customers. Also DM is being used by Hutchison Max in Mumbai (Saxena, 2000). Fingerhut Corporation used DM to identify that, customers who used a particular zip code spent high on gold per order than the other products. With data mining, Fingerhut realized that many people from around that zip code were Hispanic and so they quickly reviewed their catalogs and added more jewelry made from gold. This catalog was to be marketed to Hispanic customers only (Kadiyala & Srivastava, 2011).

### 2.4.7 Customer Churn/Attrition

Customer churn is referred to as the situation where customers change to the products and services of an opponent (Kadiyala & Srivastava, 2011).

Customer churn concerns every company and it is mostly a major problem for big companies.

This is usually the case because as the customer base increases, the attention given to individual customers may no longer exist. Dissatisfied customers are most like to churn. Radhakrishnan et al (2013) indicates that customer churn is a major area where DM is applied. Anil and Ravi (2008) came up with a collective DM system that incorporated majority voting and using a number of DM algorithms to solve the problem of customer attrition by predicting credit card customer churn in banks. Kadiyala and Srivastava (2011), states that the profiles of churned customers can be analyzed to identify future customers who are likely to churn.

### 2.4.8 Pattern Management

According to Kadiyala & Srivastava (2011), with the outbreak of electronic commerce and online stores, there tend to be a very huge data generation as compared with the traditional brick and mortar store. Because of the explosion of larger data, industries involved have come out with a solution to the data storage problem that is pattern management which is parallel to database management system. With this, Pattern Query Language (PQL) was developed which is like Structured Query Language (SQL). The PQL does the mining of relationships from the huge data and store them in the pattern base after which the rest are discarded. New identified patterns which are identified as a deviation from the already stored patterns are added to the pattern base (Kadiyala & Srivastava, 2011). DM plays very significant role in the pattern discovery.

## 2.5 EMPIRICAL EVIDENCE

This section discusses what other researchers have done with respect to enhancing direct marketing and loan application assessment by using data mining in the form of case studies. The first subsection deals with direct marketing and the second subsection deals with loan application assessment which is also referred to as credit risk assessment by other researcher.

### 2.5.1 Enhancing direct marketing using data mining: Case studies

#### 2.5.1.1 Case one

Moro et al (2012) conducted a similar research with the aim of using data mining to enhance direct marketing. The dataset they used was from a Portuguese bank. The dataset was compiled between May 2008 and November 2010. It contained 79354 instances and 59 attributes. The research was guided using CRISP-DM. After going through the data preparation phase of the CRISP-DM, the number of instances decreased to 45211 with 5289 successful ones, the rest being unsuccessful. The software they used is rminer package and they modeled support vector machine (SVM). 20 – fold cross validation and confusion matrix were used. ROC curve was used and with the nature of their work they also used Lift cumulative cure. The ROC value of their work is 0.938. The Lift cumulative curve led to the AUC (Area Under Curve) value of

0.887. At the system modeling and system evaluation phases, the SVM was modeled against NB (Naïve Bayes) and DT (Decision Tree). The results obtained after calculating confusion matrix, the ROC and Lift curves of the SVM, DT and NB are tabulated below in table 2.3 and table 2.4.

They identified that the best time do to marketing is the last month in every quarter(April, August, December). They also identified that prospects who were successfully contacted have a higher chance of doing business with the bank in future.

| **Predicted Observed** | **Unsuccessful** | **Success** | **Correctly Classified** |
|---|---|---|---|
| **Unsuccessful** | 225016 | 41144 | 83.60% |
| **Success** | 3287 | 31973 | 90.68% |

**Table 2. 3: Confusion matrix of the best classifier**

Source: Moro et al (2012)

| Model | NB | DT | SVM |
|---|---|---|---|
| ROC | 0.870 | 0.868 | 0.938 |
| Lift | 0.827 | 0.790 | 0.887 |

**Table 2. 4: ROC and Lift values of NB, DT, SVM**

Moro et al (2012)

#### 2.5.1.2 Case two

Elsalamony (2014) concludes that business and marketing decisions are very essential to keep a good rapport with very profitable customers. He writes that for a business to succeed and survive, there is a necessity to put in place measures to enhance marketing strategies and customer care and with this data mining can assist with tremendous benefits. He evaluated and compared the performance of four different classification models which are Naïve Bayes (TAN), Logistic regression (LR), Multilayer perception neural network (MLPNN) and C5.0 Decision tree, that were applied on bank's direct marketing dataset. SPSS Clementine data mining workbench was the data mining software used. The dataset used consisted 45211 instances and 17 attributes. He divided the dataset into 70% training set and 30% testing set. Accuracy, sensitivity and specificity were used to determine the performance of the models. The results indicated that C5.0 performed a little bit better than MLPNN, LR and TAN with the results tabulated in table 2.5. It was also identified that the attribute "Duration" contributed very well in the development of C5.0, MLPNN and LR and also the attribute "Age" in TAN.

| Model | Partition | Accuracy | Sensitivity | Specificity |
|---|---|---|---|---|
| MLPNN | Training | 90.92% | 65.66% | 93.28% |
| | Testing | 90.49% | 62.20% | 93.12% |
| TAN | Training | 89.16% | 55.87% | 91.97% |
| | Testing | 88.75% | 52.19% | 91.73% |
| LR | Training | 90.09% | 64.83% | 91.76% |
| | Testing | 90.43% | 65.53% | 92.16% |

| | | | | |
|---|---|---|---|---|
| **C5.0** | **Training** | 93.23% | 76.75% | 94.92% |
| | **Testing** | 90.09% | 59.06% | 93.23% |

**Table 2. 5: The Confusion Matrices of TAN, MLPNN, LR and C5.0 DT Models for Learning and Testing Subsets**

Source: Elsalamony (2014)

#### 2.5.1.3 Case three

Sing'oei & Wang (2013) came out with a data mining structure to enhance direct marketing. Their dataset contained 45212 records and 17 attributes. They chose decision tree algorithm and IBM SPSS Modeler 14.2 to perform their experiments. They conducted experiments using 10 – fold cross validation. Also they used Lift curve as a performance evaluation measure. The model they used made 42212 correct classifications representing (93.37%) and 2999 representing (6.63%) wrong classifications. They demonstrated really that data mining can serve as an efficient tool that can be used to enhance marketing campaigns of banks.

### 2.5.2 Enhancing loan application using data mining: Case studies

#### 2.5.2.1 Case one

Sousa and Figueiredo (2014) conducted a research to use data mining to analyze credits. They used Credit Unions as case study. The dataset they used consisted of 39 attributes and 321 instances. They modeled Artificial Neural Network and Decision tree that implements C4.5 algorithm. They also used Weka data mining software to conduct all the experiments to find out which model will perform better. They used 10 – fold cross validation and confusion matrix to evaluate and tabulate the performance of the two models (NN and DT). With respect to the DT, 302 (94.08%) instances were correctly classified. on the negative side, 19 (5.92%) instances were wrongly classified. Again with respect to Neural Network, 294 (91.59%) instances were correctly classified and 27

(8.41%) instances were wrongly classified. Comparatively, it was realized that the DT implementing C4.5 algorithm performed better that the ANN.

#### 2.5.2.2 Case two

Rohit and Supreethi (2014) proposed a method to be used for classification. The method was used to provide loans to customers who are verified with some parameters that relate to the loan application. Such parameters include lending rate, amount of loan, type of property, credit history of the customer, income and loan term. Their method centered on developing a Meta classifier that is based on a combination of a number of classifiers. The Meta Classifier is made up of k-Nearest Neighbor, Naïve Bayes Classifier and Fuzzy Set. The Classifier was made to use a voting algorithm so as to attain a better classification value and to increase the accuracy of the Meta Classifier while deciding on the final decision in the process of Loan approval.

The outputs of some classifiers are taken by the voting algorithms as input. Also the voting algorithms choose a class which is selected by majority of the classifiers as output. Each output of the classifier is passed as input to the Meta Classifier. Finally the Meta Classifier works on the input and produces the end result. Majority voting is determined when two or three of the classifiers approve a class for a test set. Thus, when two of the classifiers predict that the loan can be approved for the applicant, then the Majority Voting of the result is taken as approved.

Figure 2.2 is the diagrammatic form of the processing of the Loan Application using the different Classifiers

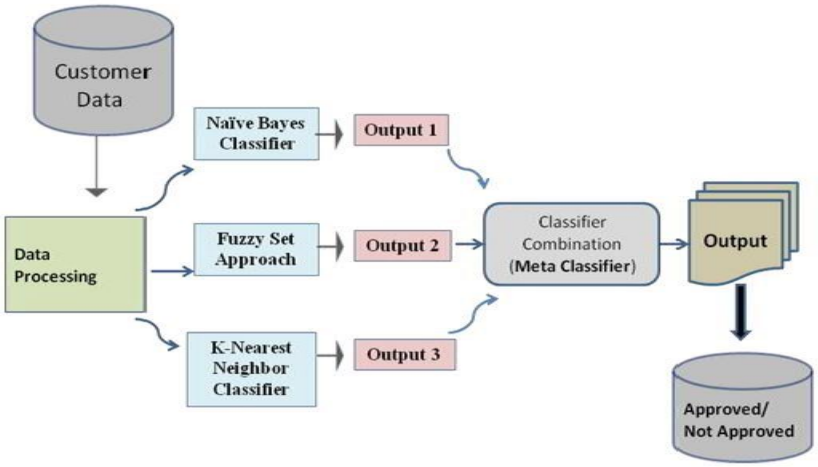



**Figure 2. 2: The Processing of the Loan Application using different Classifiers**

Source: Rohit and Supreethi (2014)

## 2.5 ETHICAL AND PRIVACY ISSUES IN DATA MINING

Robinson (2015), in his published paper concludes that with the day to day worldwide improvements concerning the state of internet, it is virtually impossible to get a complete guide to handle the misuse of information. A proper way to look at this is a mixture of real parameters. Thus there are types that apply to consumers and others that apply to companies. Customers should be mindful of who they give their information to. Ethical guidelines and legislative protections are suggested to reduce the fears about data mining but still allow companies to reap the maximum benefits. The benefits include better target products, cheap marketing and higher customer satisfaction (Robinson, 2015). He cites the code of ethics of three major American marketing associations – American Marketing Association (AMA), Business Marketing Association (BMA) and the Direct Marketing Association (DMA).

Islam and Brankovic (2003) argue that data miners need not to get full access to the data before they can develop classification algorithms. In other words, they claim that the data to be used for data mining should not necessarily have to be the same as that in the source (data warehouse,

database or files). They proposed the addition of noise to the data (attribute-wise) in order to distort the meaning of individual record but data miners would be able to detect the relationship in the data. Oliveira and Zaiane (2010) are of similar view but they proposed the use of geometric data transformation methods that distort private data that are stored in numbers.

With these methods data miners would be able to use the converted data for its intended purpose but the privacy and security of customers would be protected and ensured.

## 2.6 SUMMARY OF THIS CHAPTER

This chapter dealt with the definition of data mining and the tools used in data mining. It discussed what other researchers are discovering in terms of how data mining could be applied in the banking sector. In specifics, the chapter discussed data mining and data mining techniques, processes involved in data mining, importance and application of data mining, how data mining can assist in marketing in bank, the empirical evidence of applying data mining to enhance direct marketing and loan application assessment and the ethical and privacy in data mining. The next chapter would discuss the methodology adopted for this study.

# CHAPTER THREE

# METHODOLOGY

## 3.0 INTRODUCTION

This chapter discusses the research methods used for this work. According to Rajasekar et al (2013) research methods are the numerous ways, schemes and algorithms that are used in research. It consists of all the methods that are employed by a researcher in a process of conducting a research study (Rajasekar et al, 2013). Research methods assist the researcher to select the best ways of gathering samples, collecting data and discovering answers to a question. Rajasekar et al (2013) explain that scientific research methods specifically try to look for justification using data at hand (measurements and observation) not simply by using thinking only. In other words conclusions are made based on discovery after executing the methods and analyzing the data. Again, they describe research methodology as the process involved in solving a question.

For simplicity in reading, the various headings are outlined as:

- RESEARCH DESIGN
    - Descriptive research
      design ○ Open-ended
      interviewing ○ Case study
- THE CASE (RURAL BANK)
- DESCRIPTION OF THE DATA SET
- MODEL EVALUATION METHODS
    - Cross validation

- Confusion matrix ○ Receiver Operating Characteristics (ROC) curve



- WEKA (Waikato Environment for Knowledge Analysis ) SOFTWARE
- THE CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING (CRISP-DM)
  - Business Understanding ○ Data Understanding ○ Data Preparation ○ System Modeling ○ System Evaluation ○ Deployment
- PRIVACY AND CONFIDENTIALITY OF CUSTOMER DATA

## 3.1 RESEARCH DESIGN

The nature of the study required the need to use both numbers and texts. The data to be used consisted of both numerical (age, salary, income, loan amount) and textual (name, addresses, gender) data. Therefore it became necessary to use both quantitative and qualitative research methods.

To be specific the research design used in this study was the mixed methods approach. This type of research method is very good when there is the need to adopt both quantitative and qualitative approaches (Creswell, 2003) or capture and work with numerical and textual data (Williams, 2007). This approach offers the researcher extensive understanding of the problem understudy which is difficult when only one of them is used.

### 3.1.1 Descriptive research design

In order to accurately describe the problem understudy, the research environment, data set used in the study, algorithms and other materials, the descriptive research approach was used. This approach was very helpful in most of the phases of the CRISP-DM methodology.

According to Shields & Rangarajan (2013) descriptive research is adopted when the researcher wants to explain the characteristics of the problem to be studied. This design allows the researcher to probe and find answers to problems that involve who, what, when, where, and how these problems are related to certain research problems (Shields & Rangarajan, 2013). The descriptive research style in its simplest form tries to study the case, problem or phenomenon in its natural form (Williams, 2007). The main focus is to create an oral picture of the case, phenomenon or participant understudy (Hayes and Lemon, 1990).

### 3.1.2 Open ended interviewing and the interview checklist

According to Polkinghorne (2005) and the Merriam Webster dictionary definition, the process of conducting interview involves asking questions by the interviewer and answering question by the interviewee for which the goal is to collect detailed facts or information. Interview was chosen because the researcher can get a clearer view of what is happening (Weiss, 1994). Also, interview was used because the researcher has explicit ability to tune the questions asked so as to get the right information needed for the research and follow up questions could be asked (Emans, 1986).

Series of interviews were conducted with the bank authorities to identify the problem understudy because there was little information known to the researcher. Interviewing is a method that is used to gather data and also to advance one's knowledge from persons in a field about a topic (Kvale, 1996).

The interview checklist was not designed with pre-specified questions. It was designed to contain about five open-ended questions which were developed as the authorities answered the questions. It became prudent to use this technique because most of the data needed by the researcher to fully identify the problem understudy were not documented and for this reason officers in the field of marketing and loan application were interviewed.

It must be stated that the findings that were obtained after the implementation of the interview checklist were not analyzed and interpreted as the research but they form the basis for the research. It was very helpful during the first phase (Business Understanding) of the CRISP-DM.

### 3.1.3 Case study

Considering the benefits of case study and the nature of this research, it was a good idea to use it. In literature case study allows the examination of one or more entities over a period of time. Again, it allows the drawing and usage of deep information about the case understudy using different types of tools to collect the data (Creswell, 2003).

## 3.2 THE CASE (RURAL BANK)

According to Hayes and Lemon (1990) dealing with one case may not necessarily mean the researcher is dealing with one person. It can range from a whole family to one organization. Case studies usually work with a number of persons.

The name of the case bank is Yaa Asantewaa Rural Bank ltd. which is located at Ejisu in the Ashanti Region of Ghana. The Bank became part of the rural bank community on 14th April, 2010 under the companies' code 1963 (Act 179) as a company with limited liability to provide banking, financial and other services to the rural folks and to its clients in other areas.

Yaa Asantewaa Rural Bank received its license on May, 2012 and started operating on October, 2012. It is owned by 100 percent local Ghanaians especially those in the Ejisu community.

The vision statement of the Bank is "To provide the best and innovative rural banking services in Ghana". Again the Bank has as its mission statement "To be a rural focused, customer-oriented, providing its customers with innovative and complementary range of savings products which meet customers' specific needs, with emphasis on courteous service and effective and efficient delivery".

Yaa Asantewaa rural bank ltd was chosen for this study because of its closeness and availability to the study. It is the only bank with its headquarters close to the research area. Introductory letter was obtained from the department which the research report would be submitted and administrative permission was obtained from the bank. In order to limit the cost of the study in terms of time and finances, this bank was chosen.

### 3.3 DESCRIPTION OF THE DATA SET

The data set to be used for this study is in two folds (survey.csv and credit_risk.arff). Survey.csv contains survey data that were collected by the case bank. These data were taken to know how people in the Ejisu locality would respond when approached by the Bank's mobile bankers. One thousand different prospective customers were surveyed resulting in the survey data set containing one thousand (1000) records and eleven (11) attributes (10 categorical, 1 numeric) which were later increased to twelve (12) by adding the class attribute. No single prospective customer was contacted more than once. This data set was fed into the model to identify people who responded or otherwise to buy a loan product. The data set was encoded to ensure customer privacy. For example the age attribute was transformed into ranges. The income was changed to low, medium and high. Table 3.1 refers to the attributes, types and the allowed values in each column.

| No. | Attribute | Type | Categorical Values |
|---|---|---|---|
| 1 | Name | Categorical | NA |
| 2 | gender | Categorical | F-female; M-male |
| 3 | age | Categorical | 20-30; 31-40; 41-50; 51-60; 61-70; 71-100 |
| 4 | Marital status | Categorical | S-single; M-married |
| 5 | Occupation | Categorical | Trader; Carpenter; Driver; Tailor; Farmer; Pastor; Nurse; Mechanic; Teacher; Seamstress |
| 6 | Children | Numeric | 1; 2; 3; 4; 5; 6; 7; 8 |
| 7 | Income | Categorical | Low; Medium; High |
| 8 | Education | Categorical | Primary; Secondary: Tertiary |
| 9 | House | Categorical | N-no; Y-yes |
| 10 | Contact | Categorical | First; Second; third; Fourth |
| 11 | Residence | Categorical | NA |

| | | | |
|---|---|---|---|
| 12 | Class | Categorical | N-no; Y-yes |

**Table 3. 1: Attributes of survey.csv data set used for direct marketing**

Source: Yaa Asantewaa Rural Bank Limited

The second data set (credit_risk.arff) contains data that would be analyzed to determine the credit worthiness of the customers. The data set was developed by Professor Dr. Hans Hofmann from University of Hamburg and was taken from UCI Machine Learning Repository – Center for Machine Learning and Intelligent Systems (Lichman, 2013). Credibility, integrity and excellent standard of the data are highly assured. With the help of the loan officer and the head of information technology unit of the bank, the data set was modified to suit our local environment. This data set contained 20 attributes (7 – numerical, 13 - categorical) and 1000 records. Later the class attribute was added to make 21 attributes. The data set was also encoded to ensure kanonymity. Table 3.2 refers to the attribute, type and allowed values for each column. Extensive description of the data sets is made available in the Appendix C.

| **No.** | **Attribute** | **Type** | **Categorical Values** |
|---|---|---|---|
| 1 | Current_balance | Categorical | A11; A12; A13; A14 |
| 2 | Age_of_account_(months) | Numeric | 4 to 74 Months |
| 3 | Credit_history | Categorical | A30; A31; A32; A33; A34 |
| 4 | Purpose_of_loan | Categorical | A40; A41; A42; A43; A44; A45; A46; A47; A48; A49; A410 |
| 5 | Claimed_amount | Numeric | 250 to 18424 Cedis |
| 6 | Savings_account | Categorical | A61; A62; A63; A64; A65 |

| | | | |
|---|---|---|---|
| 7 | Yrs_of_current_employment | Categorical | A71; A72; A73; A74; A75 |
| 8 | Installment_rate | Numeric | 1 to 4 |
| 9 | Marital_status/sex | Categorical | A91; A92; A93; A94; A95 |
| 10 | Other debtors/guarantors | Categorical | A101; A102; A103 |
| 11 | Yrs_in_current_residence | Numeric | 1 to 4 |
| 12 | property | Categorical | A121; A122; A123; A24 |
| 13 | age | Numeric | 19 to 75 |
| 14 | Installment_plans | Categorical | A141; A142; A143 |
| 15 | shelter | Categorical | A151; A152; A153 |
| 16 | Pending_loan_in_this_bank | Numeric | 1 to 4 |
| 17 | occupation | Categorical | A171; A172; A173; A174 |
| 18 | Number_of_dependants | Numeric | 1 to 2 |
| 19 | telephone | Categorical | A191; A192 |
| 20 | Local_or_foreign_worker | Categorical | A201; A202 |
| 21 | Class | Categorical | Y - yes N - no |

**Table 3. 2: Attributes of credti_risk data set used to determine the credit worthiness of customers.**

## 3.4 MODEL EVALUATION METHODS

This section presents the evaluation methods that were employed to evaluate the performance of the models generated to identify prospects to direct marketing efforts to and also to determine the credit worthiness of loan applicants.

### 3.4.1 Cross validation

Validation is simply a method used to assess how good the performance of a model is. There are many forms of validation which are holdout method, k-fold cross validation and leave-one-out validation (www.cs.cmu.edu). K-fold Cross Validation usually referred to as 10-fold cross validation is the most common form of model validation method and is widely known among DM and machine learning folks (Refaeilzadeh, et al, 2009). This form of validation is mostly used when the information available for the DM project is limited and there is the need for an unbiased statistics about the performance of the model (Sing'oei & Wang, 2013).

The data set to be fed into the model was partitioned into k subsets and a holdout scheme is applied k times. This is good because during the validation process, single subset of the k subset is used as a test set and the other k-1 subsets are summed together as training set. After this, the average error is calculated for all the k times that the evaluation process was ran. K is an integer.

### 3.4.2 Confusion matrix

One of the popular performance evaluation methods in data mining, machine learning artificial, intelligence and statistics is confusion matrix. In order to quantify the performance of a problem that contains two classes, confusion matrix is usually used (Gupta et al, 2012). Most Data Miners use this method to evaluate their classification models, popular metrics are based on the confusion

matrix (Kohavi and Provost, 1998). Confusion matrix comprises of evidence about actual and predicted classification performed by the classification model.

| | | PREDICTED | |
|---|---|---|---|
| | | Positive | Negative |
| ACTUAL | Positive | True Positive (TP) | False Positive (FP) |
| | Negative | True Negative (TN) | False Negative (FN) |

*TP* = the number of correct classifications that an instance is Positive.

*TN* = the number of incorrect classifications that an instance is Negative.

*FP* = the number of incorrect classifications that an instance is positive.

*FN* = the number of correct classifications that an instance is Negative.

The values obtained for *TP, TN, FP and FN* is further used to calculate Accuracy, Sensitivity and Specificity. These three statistical measures according to Kohavi and Provost (1998), are termed as confusion matrix and the later reveals facts about actual and predicted classifications that are done by the classification algorithm (Elsalamony, 2014).

### 3.4.3 Receiver Operating Characteristics (ROC)

In DM, statistics and machine learning a common method for evaluating the performance of a classifier with binary class attribute is ROC curve (Jiménez-Valverde, 2012). The ROC graph displays the tradeoff between sensitivity and specificity (Fawcett, 2006). Fawcett (2006) writes that, in ROC graphs, algorithms with arbitrary performance displays a straight sloping line from

(0, 0) to (1, 1) and the sloping line is usually referred to as the ROC baseline. Receiver Operating Characteristics curve displays the performance measure for a single algorithm. A performance value of 0.5 means classification by random and a value of 1 means perfect classification. From equations 5 and 6 in chapter four, Sensitivity and True Positive Rate (TPR) are alike likewise specificity and False Positive Rate (FPR) are the same (Zhu et al, 2010). The ROC space (Zhu et al, 2010) is made up of all possible permutations of TPR and FPR (sensitivity and specificity). A point in the ROC graph is determined by TPR and FPR and the point's position shows a tradeoff between sensitivity and specificity. In other words the position of a point shows whether the classifier performs good or bad as indicated in figure 3.1. Hypothetically, the coordinate (0, 1) is the preferred coordinate for a point plotted by both TPR and FPR. In simple terms it could be described as lying on the upper left corner of the space (Zhu et al, 2010). This theoretical point shows that the classifier has 0% specificity and 100% sensitivity and means perfect classification (Zhu et al, 2010). Classification with 50% specificity and 50% sensitivity will display a sloping line with coordinates (0, 0) and (1, 1). A point plotted below the sloping line shows that the classifier performs badly.

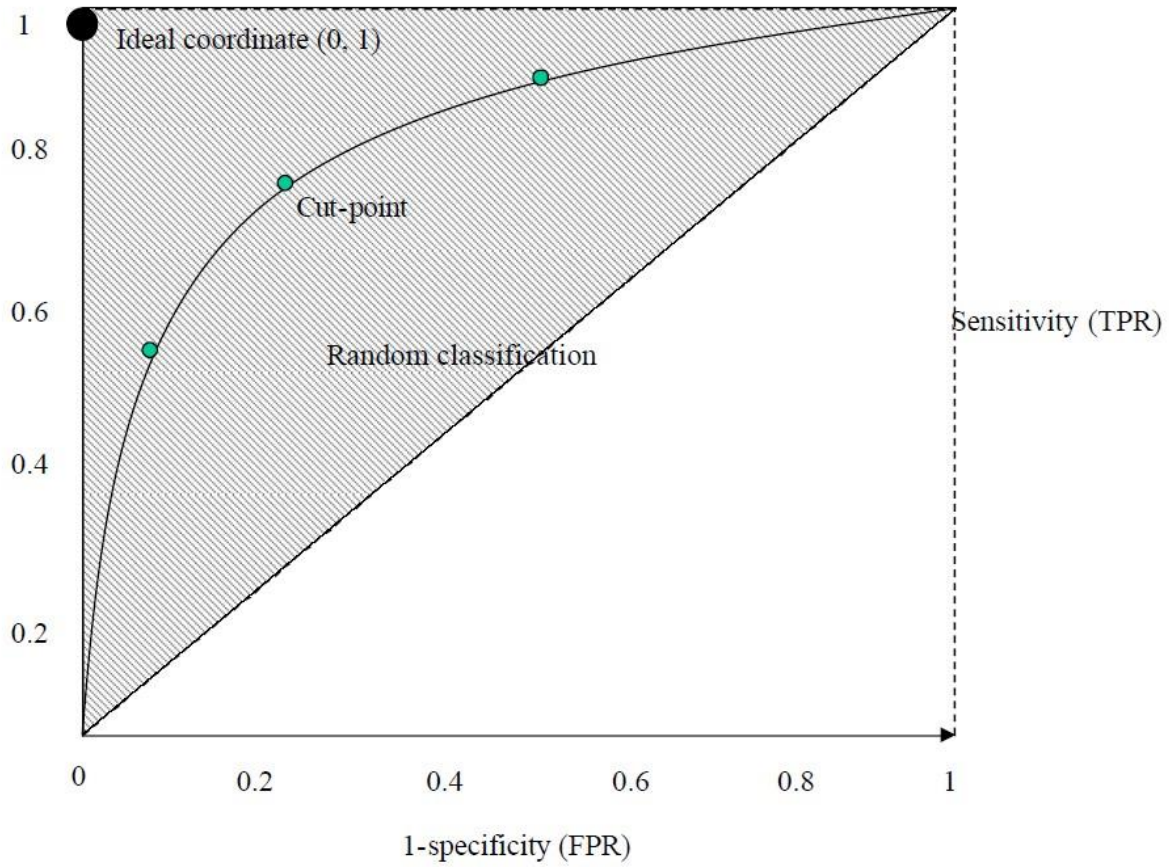



**Figure 3. 1: ROC space: shaded portion shows better classification performance Source: Zhu et al (2010).**

### 3.5 WEKA (Waikato Environment for Knowledge Analysis) SOFTWARE

All data mining tools employ logical and systematic schemes or algorithms (logistic regressions, and decision trees). To reduce the workload of amateurs in the field of data mining, experts and programmers have designed many programs that use the various DM algorithms. The main aim is to save time and prevent the situation of developing the algorithms from scratch. Examples include IBM Intelligent Miner, Knowledge STUDIO, Weka, SPSS Clementine and SAS Enterprise Miner (Chiu and Tavella, 2008).

Weka was chosen for this study because it is an open source machine learning suite and possesses a remarkable graphical and visualization capability that allows the presentation of findings in a more comprehensive manner (Abbas, 2015). Again Weka provides extensive algorithms for data mining. Its popularity stems on the fact that, it has enormous tools for preprocessing, classification, regression, clustering, association rules and visualization (Frank et al,
2004 and www.cs.waikato.ax.nz/ml/weka).

## 3.6 REASONS FOR THE CHOSEN ALGORITHMS

### 3.6.1 Decision Tree

Rokach (2005) writes that, Decision tree is clear and understandable. When it is condensed, it becomes simple to trail. It can also work with both categorical and numeric data. Representation done by decision tree is deep enough to represent any discrete classifier. Again decision tree can work with a dataset with data errors and missing entries. Decision trees are among the top 10 most popular algorithms (Rokach, 2005, Gupta, 2015 and www.saedsayad.com)

### 3.6.2 Random Forest

According to Montillo (2009), when it comes to accuracy Random forest has almost the same value as SVM and NN but Breiman (2001) writes that RF has better performance than SVM and NN. However, RF is easy to interpret. Again RF welcomes larger numbers of predictors. It is faster to learn from training data. As will be discussed in chapter four, RF has few parameters to be supplied before running. Lastly cross-validation is considered to be unnecessary since RF creates unbiased estimate of the generalization error internally (Montillo, 2009). RF addresses the problem of overfitting (Friedman et al, 2001)

### 3.7 THE CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING (CRISP-DM)

The cross industry standard process for data mining (CRISP-DM) usually shortened as CRISPDM is a framework for guiding and recording data mining tasks. It is a model that consists of phases that are followed to solve data mining problems (Bruno and Rafael, 2010).

The (CRISP-DM) is a popular methodology adopted by many Data Miners because when followed critically there is a higher chance in succeeding in the DM project (Chapman et al., 2000).

IBM SPSS Modeler CRISP-DM guide (2011) describes the model in two different forms – as a methodology and process model. The methodology basically describes the normal stages of a project, the works to do in every stage and an explanation of the links between the various works. As a process model, the CRISP-DM gives a general outline of the data mining life cycle.

There are six phases in the data mining life cycle. There are arrows that link the various phases with no exact order. The arrows show that, the various phases of the DM life cycle are interdependent. This means that the arrows depict reliance of the phases. For some projects to be successful there is the need to be switching between two or more phases before completion; others may not necessarily need that. The outermost circle shows data mining itself is revolving.

The popularity of CRISP-DM is as a result of it not designed, owned and sold by one particular organization. Again when followed successfully all DM models produced or projects done with the CRISP-DM become stress-free (Leventhal, 2010).

The CRISP-DM is easy to use. It is the most widely used model that experts and non-experts use in solving data mining question. Polls were conducted and as at September 18, 2016, 189 participants had responded. 96 (51%) of the respondents reported using CRISP-DM (PiatetskyShapiro, 2016).

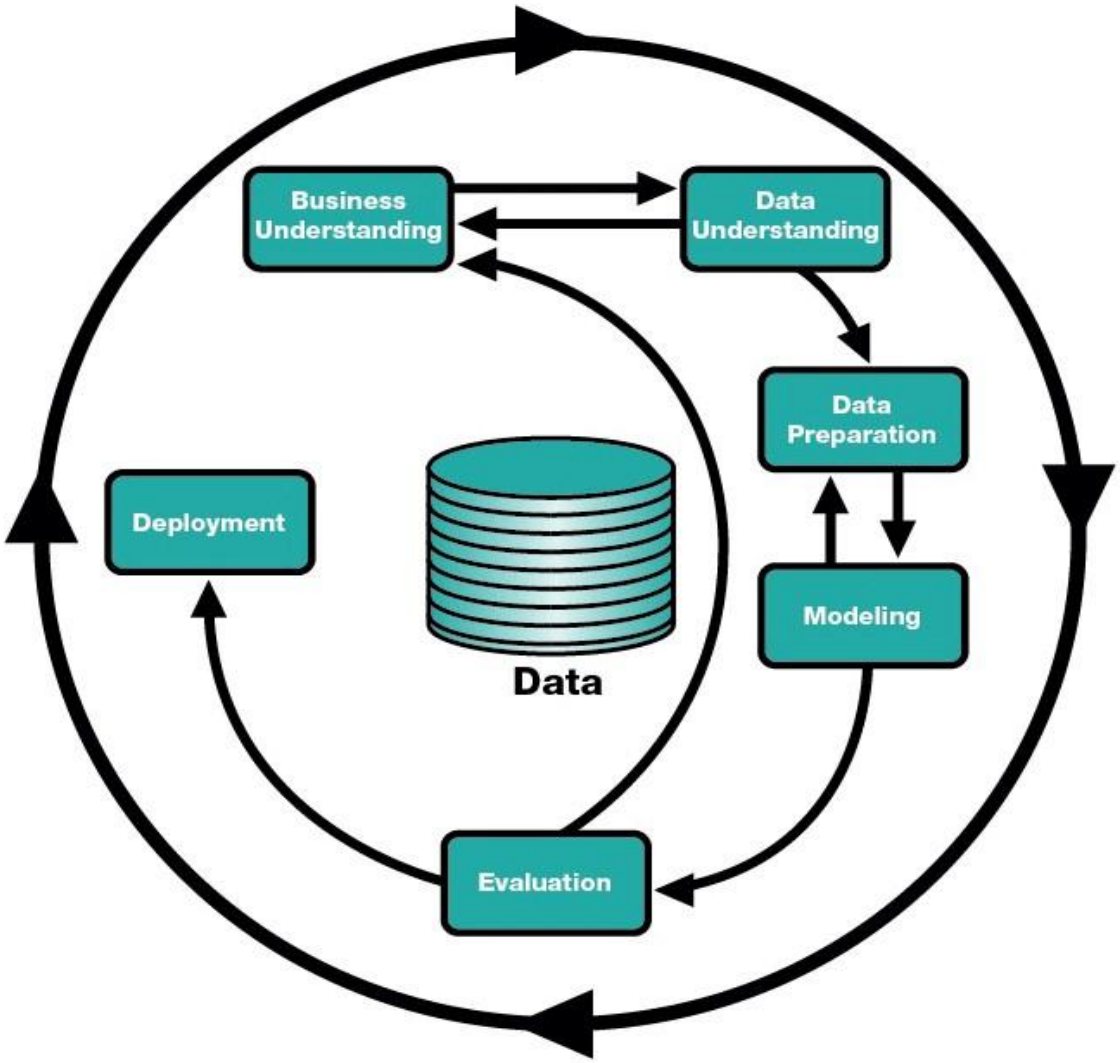



**Figure 3. 2: Cross Industry Standard Process for Data Mining CRISP-DM image source:** www.sv-europe.com

### 3.7.1 Business Understanding

This phase is also known as Problem understanding and entails the processes used to comprehend the opportunities and business purposes of the company (Bruno and Rafael, 2010).

According to the 2011 version of the IBM CRISP-MD guide, this phase consists of a number of tasks that must be completed before proceeding to the next phase.

After conducting series of interviews with the bank authorities, it was realized that the bank spends much money on trying to win more customers to expand its customer base and also sell its products to existing customers. The loan officer happens to be the head of the marketing team. They develop weekly "root plan" that tells them what to do, the route to pass, persons to contact and the amount of time to spend on each person. According to the loan officer this approach is expensive since most of the people they contact turn down the offer to apply for loan. To clarify this numerically, when about ten (10) customers are contacted, about five (5) representing 50 percent or less accept the offer to apply for loan. In other words, about 50 percent or more turn down the offer. Out of the 50 percent or less that accept to deal with the bank, there are some that are considered to be risky in terms of credit scoring. This means that, they have the tendency of not being able to pay the loan with the agreed interest rate and time.

Therefore the business objective here is to identify the 50 percent or less customers who are likely to accept the offer in advance during the preparation of the "root plan", so that they would be contacted leaving the other customers in order to save the Bank some money. Again out of this number another business objective is to identify the number who would be able to pay back the loan or may not.

The first DM goal of this research is to identify the number of people likely to respond when an offer is made to them to buy a loan. The second data mining objective is to identify the characteristics of customers who are of the tendency to default a loan and use it to prevent future defaults.

The models were designed and simulated using Weka software and ran on an intel core i3 processor laptop using Windows 7 operating System.

### 3.7.2 Data Understanding

This is the next phase of the CRISP-DM methodology. The data to be used in the DM project must be taken and described. The quality must be verified (Bruno and Rafael, 2010).

According to the CRISP-DM guide (2011), it is very important to study the data to be used for the DM project in order to avoid future problems especially during the data filtering phase. This is mainly the purpose of data understanding phase of the CRISP-DM model. In most projects existing corporate data such as operational data, survey data etc. is enough to complete the DM project. But there are times additional data may be purchased from external companies.

The data sets to be used for this study are in two different forms. The first was a survey data that was taken by the Bank's mobile bankers to discover the level of acceptance the local people have for the Bank. In other words, the Bank wanted to know how many people would respond when contacted to sell a loan product. The data were entered into Microsoft Excel Spreadsheet and since Weka does not support the workbook format, it became necessary to convert it to the .CSV (Comma delimited) file. All the data are categorical (string) except one which is numeric. A mechanism was used to code the entries – M, F, M, S stand for Male, Female (Gender), Married and Single (Marital status) respectively.

The data set (credit_risk.arff) is of good quality. There are no missing values, typographical errors, incorrect inputs and coding errors. But the data set (survey.csv) had some misspelt entries.

### 3.7.3 Data Preparation

Taking into consideration the organizational policies governing the data and the privacy and confidentiality associated with data mining, the names and addresses (location and residence) of the participants were not included in the data set. Again in order to protect the actual earnings of

the participants, the inputs for the income attribute are grouped into low, medium and high. Low means amount less than 500 cedis, medium mean amount between 500 and 1000 cedis and high means amount of 1000 or higher. These two methods are termed as anonymity and transformation (Oliveira and Zaiane, 2010 and Ketel and Homaifar, 2005).

There were some misspelt entries identified in the data understanding phase and were looked at closely. Example is the spelling of teacher as "taecher", carpenter as "capenter" and trader as "trade" in the occupation column. These data errors would have affected the model in a significant way and were taken care of.

This phase is also known as data filtering and involves the process of organizing the data for mining. The following steps are completed in order to filter the data to be used in the DM process. These are combining data sets, choosing part of the data as subgroup, combining rows, developing new columns, arranging the data to be used in the modeling, taking care of problematic figures (blank or missing) and dividing into training and test data sets

### 3.7.4 System Modeling

This phase of the CRISP-DM employs DM models to simulate the data mining project (Bruno and Rafael, 2010). The algorithms to be used for this study are the J48 Decision Tree and Random forest algorithms. Weka provides enormous algorithms that can be used under the DT classifiers. Typical examples of such algorithms are ADTree, BFTree, DecisionStump, FT, Id3 and more. The j48 classifier implements the C4.5 form of Decision Tree algorithm for developing classifiers (Zamani S. and Mogaddam, 2015). It is actually an advanced version of the C4.5 and it was designed to solve the problems of C4.5 and Id3 classifiers (Cufoglu et al, 2009).

According to the Crisp-DM guide, at this stage a number of classifiers can be modeled and tested to find the most suitable classifier for the data mining project. In the light of this RandomForest classifier was added to the J48 classifier to solve the credit worthiness of loan applicants' problem. These two classifiers are among the top 10 most popular decision tree classifiers (Gupta et al, 2012).

### 3.7.5 System Evaluation

This is where the Cross validation, Confusion matrix and ROC curve would be calculated, tabulated and graphed. The values attained from the confusion matrix would be used to calculate the accuracy, sensitivity and specificity of the models.

### 3.7.6 Deployment

The deployment of the model is the exclusive decision to be made by the administrators of the bank (main manager, branch manager, loan officer and Information Technology head). However, there is a high certainty that this study will enhance the weekly "root" plan development and also identify customers who are classified to have high credit risk or otherwise.

## 3.8 PRIVACY AND CONFIDENTIALITY OF CUSTOMER DATA

In view of legality, government and other policies governing customer and other data of the bank, this study was conducted not to include any private or confidential data of any customer. Names and addresses were excluded from the data set as suggested by Oliveira and Zaiane (2010). In order to consider the research to be ethical, anonymity and confidentiality were ensured as well as administrative permission was obtained. Anonymity is defined by Sweeney (2002) as method used in protecting data so that no K individual among a list or data set released would be identified in N-K colleagues whose names or data are also part of the list or data set. Permission was obtained

from the main manager, branch manager, loan officer and the information technology head before any information were released.

### 3.9 SUMMARY OF THE CHAPTER

This chapter discussed the methodology used for conducting this study. The chosen research design, the case bank, the data set to be used for the study, model evaluation techniques and the Weka software were all discussed and explained. The CRISP-DM methodology was also explained in general terms and in the light of how it was used to solicit information and guide this study. Because participants (customers) were used in this study privacy, confidentiality and ethical issues were also discussed.

# CHAPTER FOUR

# SYSTEM MODELING AND PERFORMANCE EVALUATION

### 4.0 INTRODUCTION

This chapter discusses the modeling processes of the classification algorithms that employ J48 decision tree and random forest classifiers. A number of model evaluation techniques were used here in order to measure the performance of the classifiers. All the modeling and the evaluation processes are done using the Waikato Environment Knowledge Analysis software (Weka). The study was conducted in order to design data mining models to identify prospects so that people

who match the characteristics would be targeted during loan marketing and also credit worthiness of such customers who apply for the loans would be predicted. The data sets used for this study are in two folds. The first one is used by the J48 model to detect prospects to direct marketing campaigns to. The main objective of this detection is to limit the scope of marketing so as to save the bank some cedis. It consisted of 11 attributes (10 categorical, 1 numeric) but the class attribute was added to make 12. Only 10 attributes were used by the classifier. The second data was analyzed by the model to detect the credit worthiness of customers who apply for loans. This data set contained 20 attributes, 7 were numerical and 13 were categorical. The class attribute was added and the number of attributes increased to 21. The experiments were executed using 10-folds cross validation which means that each data set was divided into 10 equal subsets and the model was ran ten times. As discussed in the chapter three, in every run nine subsets are put together and used as the training set whiles the remaining one is used as a test set. The confusion matrix was used to validate the model after which accuracy, specificity and sensitivity were calculated to evaluate the performance of the model. Receiver operating characteristics (ROC) was generated to pictorially evaluate the performance of the model. On the part of the random forest (RF), the out-of-bag error rates were used to evaluate the performance of the model and also the RF was tested against a single decision tree model.

## 4.1 DECISION TREE MODELING

This subsection deals with how the J48 decision tree classifier was developed and how the information gain was calculated. Two trees were generated, thus the pruned and unpruned tress.

### 4.1.1 J48 Decision Tree Induction and Classification

Decision tree (DT) is one of the mostly adopted DM techniques that is influenced by transforming the data set or information space into a structure that can be used to predict and classify new

records. The pruning path helps in achieving this (Abbas, 2015). Despite the development of many decision tree classifiers so that the small group of attributes needed to build the classifier can be obtained, it is difficult to find the best tree structure for any given data set (Patil et al, 2012 and Sun et al, 2008). All DTs that employ C4.5 algorithm are based on ID3 that was developed by Quinlan John Ross. ID3 uses a core that operates by using *Entropy* and *Information Gain* (www.saedsayad.com) but according to Linoff and Berry (2004), these two terms describe decision tree algorithms.

### 4.1.2 Entropy Reduction or Information Gain

According to Linoff and Berry (2004), Information gain defines a brilliant notion for ensuring purity in leaves. If a leaf is totally pure, then it means that all the classes in the leaf can be easily described and understood (they all belong to one class). If a leaf is considered as highly impure, it means that all classes in the leaf are complex and difficult to understand (Linoff and Berry, 2004). According to Fuchs and Peres (1996), Information theory a field in computer science has come up with a way of determining purity that is referred to as entropy. Entropy may be discussed as the amount of Y/N (yes or no) questions that should be answered to define the form of a system. If there exist eight (8) states of a system, it will take $\log_2(8)$ or three bits to number all of them or to uniquely distinguish one (Fuchs and Peres, 1996). Extra information which will decrease the entropy by reducing the amount of questions that needs to be answered in order to define the form of the system is termed as information gain (Fuchs and Peres, 1996). In this view, information gain can be said to mean the same as entropy reduction (Linoff and Berry, 2004).

A test must be made so as to choose the most qualifying attribute as the root node when generating the decision tree classifier and the kind of test is very important if the decision tree to be constructed will be simple (Quinlan, 1986). Seat-of-the-pants induction is the method that was used by the

initial ID algorithms which performed inductions very well. But Peter Gacs made a recommendation that was later incorporated into ID3. Because of this recommendation ID3 implemented information base technique that is influenced by two assumptions (Quinlan, 1986). If *D* is considered as a set of *m* and *t* objects with *M* and *T* classes respectively then these are the assumptions:

1. A correct DT constructed for *D* will categorize the same objects as part of *D*. A random object that is chosen by probability will belong to class M with *m/ (m + t)* and also to class T with *t/ (m + t).*
2. If a DT is adopted to group a set of items, the outcome will be a class. Therefore a DT can be considered as a message source of '*M'* or '*T'*, with the information required to produce the message stated by

$$\boldsymbol{I\,(m,t) = -\frac{m}{m+t}\,log_2\,\frac{m}{m+t} \quad \frac{t}{m+t} - log_2\frac{t}{m+t}} \qquad (\mathbf{1})$$

If attribute *P* contains items ($P_1$, $P_2$, …, $P_n$) and it is implemented as the root node of the DT, it will divide *D* into ($D_1$, $D_2$, …, $D_n$) where $D_1$ holds all the objects in *D* which have $P_1$ of *P*. Let $D_i$ contain $m_i$ items of class *M* and $t_i$ items of class *T* the required information for the subtree for $D_i$ is $I(m_i, t_i)$. Then the required information for the tree which has P as root node will be calculated as weighted average

$$\boldsymbol{E^{(P)} = \sum_{i=1}^{n} \frac{mi + ti}{m + t}\, I_{(mi,\, ti)}} \qquad (\mathbf{2})$$

where the value for the ith section is the part of items in D that exists for $D_i$. Therefore by branching on attribute P the information gained is calculated as

$$\boldsymbol{Gain\,(P) = I(m,t) - E(P)} \qquad (\mathbf{3})$$

Catlett (1986) reports that, at this point a popular thing to do is to select the attribute with the highest gains of information as a branch. ID3 analyzes all the contesting attributes and selects *P* to get the best out of *Gain(P)* and constructs the tree. The process is repeated to generate decision trees for the remaining subsets $D_1$, $D_2$, …, $D_n$. The entropy will be 1 (maximum) when the items have the same amount of yes and no classes. But when it happens that there exist unequal amount of yes and no classes the entropy lies between 0 and 1 depending on the skewness of the distribution (Quinlan, 1986).

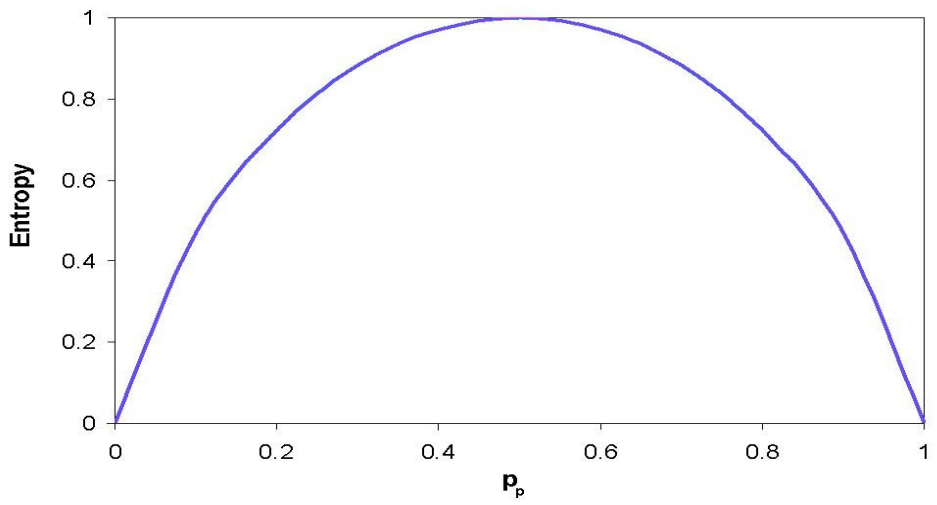



**Figure 4. 1: Entropy graph for binary classification. Source: dms.irb.hr**

Figure 4.1 illustrates the type of entropy function that is dependent on binary classification where the number of yes items varies from 0 to 1.

In Weka environment, there is a function called InfoGainAttributeEval which calculates the value of an attribute by computing the information gain. The function places emphasis on the class using formula (4) and the output is used to generate the decision tree. The results of calculating the information gain is tabulated below in table 4.1 after using the function InfoGainAttributeEval.

$$\boldsymbol{InfoGain(Class, Attribute) = H(Class) - H(Class \mid Attribute)} \qquad \textbf{(4)}$$

| NO. | ATTRIBUTE | INFORMATION GAIN |
|---|---|---|
| 1 | Contact | 0.595014 |
| 2 | Occupation | 0.183485 |
| 3 | Education | 0.120942 |
| 4 | Income | 0.074047 |
| 5 | Age | 0.047463 |
| 6 | Children | 0.012925 |
| 7 | Gender | 0.010837 |
| 8 | House | 0.010819 |
| 9 | Marital status | 0.000172 |

**Table 4. 1: Information Gain ration of attributes in descending order.**

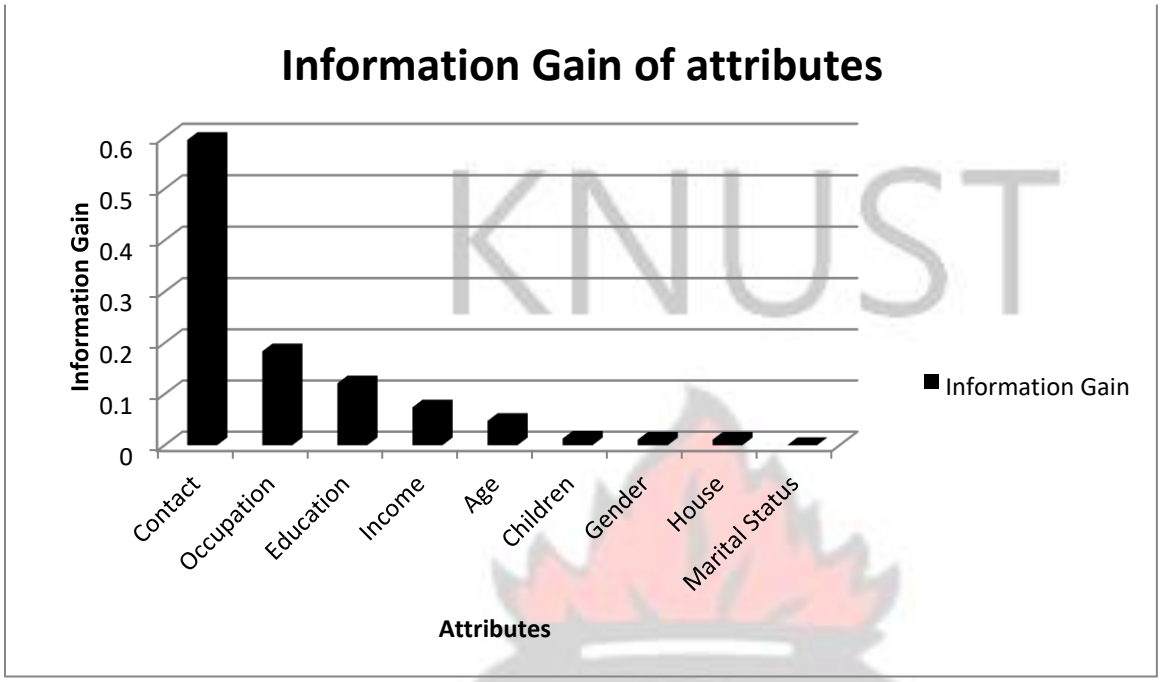



**Figure 4. 2: Histogram of Information Gain of attributes.**

The attribute with the highest information gain will be in the root node Catlett (1986). Clearly Contact with the information gain of 0.595014 will be in the root node. After this, the algorithm used the test figures to construct the decision tree. The first was unpruned decision tree and the second was pruned. For complete decision rules and decision trees for both pruned and unpruned trees refer to appendix A.

### 4.1.3 Unpruned Tree for J48 Classifier

The unpruned tree had 115 leaves, the tree size was 148 and it took virtually 0 seconds to build. The J48 decision tree classifier successfully classified 937 out of 1000 instances representing 93.7 percent as correct. Again 63 instances representing 6.3 percent were incorrectly classified. Viera & Garrett, (2005) make the assertion that, in research, there are some findings that depend on some level of individual subjective interpretation by viewers. Examples include radiographic explanation. Researches that take into account the agreement between two or more viewers must

include statistics that cater for the fact that, viewers will at times agree or disagree just by accident or chance. Researchers mostly use the Kappa statistics for calculation of the level of agreement between two or more viewers. The Kappa statistics is used to measure the agreement of prediction made concerning the true class. A kappa value of 0 means agreement by chance whereas kappa value of 1 means perfect agreement. Any kappa value between 0.81 and 0.99 is considered as "almost perfect agreement" (Viera & Garrett, 2005). With this interpretation, the J48 classifier built to detect potential prospect of the bank produced a kappa statistic of 0.8643 which is a very good value as it is in the range of "almost perfect agreement". For the unpruned decision tree and its associated classification rules that were generated refer to Appendix A.

### 4.1.4 Statistics for unpruned tree

*Number of Leaves: 115*
*Size of the tree: 148*
*Time taken to build model: 0 seconds*
*=== Stratified cross-validation ===*
*=== Summary ===*

| | | |
|---|---|---|
| *Correctly Classified Instances* | *937* | *93.7 %* |
| *Incorrectly Classified Instances* | *63* | *6.3 %* |
| *Kappa statistic* | *0.8643* | |
| *Mean absolute error* | *0.074* | |
| *Root mean squared error* | *0.229* | |
| *Relative absolute error* | *15.8339 %* | |
| *Root relative squared error* | *47.3776 %* | |
| *Total Number of Instances* | *1000* | |

### 4.1.5 Pruned Tree for J48 Classifier

Decision trees are pruned so that the best tree can be generated. In other words, when there is a set of trees with size that ranges from an unpruned tree to that of a null tree, pruning is done to get the

right sized tree (Weiss and Indurkhya, 1994). Tree pruning is done basically to reduce or eradicate inconsistencies in data set (training set). A pruned tree is smaller in size and very easy to understand (Weiss and Indurkhya, 1994). The Merriam-Webster dictionary defines pruning among others as "to reduce by removing parts that are not necessary or wanted".

The decision tree developed initially was very complex and difficult to understand. So it became necessary to prune it so as to get the right tree size with minimal leaves and which will be less complex. After pruning the tree, the number of leaves decreased to 15, the tree size also decreased to 22. But this time, it took 0.02 seconds to build the classifier. The number of correctly classified records decreased to 925 instances representing 92.5 percent. The number of incorrectly classified instances rather increased to 75 which represent 7.5 percent. The Kappa statistics also decreased to 0.8378 which is also considered as "almost perfect agreement" by Viera & Garrett (2005). For the pruned decision tree and its associated classification rules that were generated refer to Appendix A.

### 4.1.6 Statistics for pruned tree

*Number of Leaves: 15*
*Size of the tree: 22*
*Time taken to build model: 0.02 seconds*
*=== Stratified cross-validation ===*
*=== Summary ===*

| | | |
|---|---|---|
| *Correctly Classified Instances* | *925* | *92.5 %* |
| *Incorrectly Classified Instances* | *75* | *7.5 %* |
| *Kappa statistic* | *0.8378* | |
| *Mean absolute error* | *0.1237* | |
| *Root mean squared error* | *0.2581* | |
| *Relative absolute error* | *26.4637 %* | |

*Root relative squared error* *53.3959 %*

*Total Number of Instances* *1000*

## 4.2 MODEL EVALUATION FOR J48 DECISION TREE

This section discusses the model performance evaluation methods that were employed in the study. The popular confusion matrix was used and also accuracy, specificity and sensitivity were calculated. Precision, Recall and F-Measure were generated by weka and were also tabulated in table 4.5. The last performance evaluation method that was used is the receiver operating characteristics (ROC) curve that was displayed graphically.

### 4.2.1 Confusion matrix for unpruned tree

Considering the values from table 4.2, 602 negative items were correctly classified as negative. 335 positive items are correctly classified as positive. On the other hand 37 positive items were wrongly classified as negative. The same applies to 26 negative items that were wrongly classified as positive.

| | | PREDICTED | |
|---|---|---|---|
| | | Negative (N) | Positive (Y) |
| ACTUAL | Negative (N) | (TN) = 602 | (FP) = 26 |
| | Positive (Y) | (FN) = 37 | (TP) = 335 |

**Table 4. 2: Confusion Matrix for unpruned J48 decision tree**

### 4.2.2 Confusion matrix for pruned tree

Also looking at the values in table 4.3 600 negative items were correctly classified as negative. 325 positive items are correctly classified as positive. On the negative side 47 positive items are wrongly classified as negative. The same applies to 28 negative items that are wrongly classified

as positive. These values are used to calculate the accuracy, specificity and sensitivity of the J48 decision tree model.

| | | PREDICTED | |
|---|---|---|---|
| | | Negative (N) | Positive (Y) |
| ACTUAL | Negative (N) | (TN) = 600 | (FP) = 28 |
| | Positive (Y) | (FN) = 47 | (TP) = 325 |

**Table 4. 3: Confusion Matrix for pruned J48 decision tree**

### 4.2.3 Accuracy, Specificity and Sensitivity for both Unpruned and Pruned trees

**Accuracy**

Accuracy of a classification algorithm is termed as the proportion of the number of instances that are classified correctly which is equal to *TP* plus *TN* divided by the total amount of instances in the information space *N* (Nisbet et al, 2009). Accuracy can be expressed mathematically as:

$$Accuracy = \frac{TP + TN}{N} \quad (5)$$

**Sensitivity**

Sensitivity points to the number of positive instances that are correctly classified and it is equal to *TP* divided by *TP* plus *TN* (Elsalamony, 2014 and Linoff & Berry, 2004).

$$Sensitivity = \frac{TP}{TP + FN} \quad (6)$$

**Specificity**

Specificity refers to the proportion of negative instances that are correctly classified and it is equal to the proportion of TN divided by TN plus FP (Nisbet et al, 2009 and Linoff & Berry, 2004).

$$Specificity = \frac{TN}{+ FP} \qquad (7)\ TN$$

| | Accuracy | Sensitivity | Specificity |
|---|---|---|---|
| **Unpruned** | 0.937 | 0.9005 | 0.9586 |
| **Pruned** | 0.925 | 0.8737 | 0.9554 |

**Table 4. 4: Accuracy, sensitivity and specificity of unpruned and pruned J48 decision tree**

### 4.2.4 Precision, Recall and F-Measure for both Unpruned and Pruned trees

Recall according to Powers (2011) is the same as sensitivity in psychology. It refers to part of real positive instances that are predicted correctly as positives. The values obtained in table 4.5 were calculated by Weka. It is very important if Precision - Recall curve model performance evaluation were to be calculated.

| | Precision | Recall | F-Measure | Class |
|---|---|---|---|---|
| **Unpruned** | 0.942 | 0.959 | 0.95 | No |
| | 0.928 | 0.901 | 0.914 | Yes |
| **Weighted Avg.** | 0.937 | 0.937 | 0.937 | |
| **Pruned** | 0.927 | 0.955 | 0.941 | No |
| | 0.921 | 0.874 | 0.897 | Yes |
| **Weighted Avg.** | 0.925 | 0.925 | 0.925 | |

**Table 4. 5: Precision, Recall and F-Measure of unpruned and pruned J48 decision tree**

The value obtained for the ROC is 0.9254 which means that the J48 classifier has 92.5 percent accuracy of correctly classifying future attributes. The ROC value for a perfect classifier is 1.


**Figure 4.3: ROC curve for J48 pruned decision tree classifier.**

## 4.3 RANDOM FOREST CONSTRUCTION

This section discusses the approach used in the development of the data mining algorithm that was used to determine the credit risk or credit worthiness of loan applicants in the bank. A number of evaluation methods were employed to test the reliability of the algorithm in future events of classifying unknown records.

Random forest was proposed by Breiman in 2001. RF is a concept of the normal technique of random decision forests (Ho, 1995 and 1998). It uses numerous learning algorithms (ensemble learning method) for classification and regression to achieve improved performance in prediction than it could have achieved using any other learning algorithm (Opitz & Maclin, 1999 and Polikar,

2006). RF functions by generating many decision trees when it is training the algorithm and produces the class that has the highest value (mode) of all the classes for classification or calculates mean value for regression. Random forest was built to address the problem of overfitting in decision trees when using the training data (Friedman et al, 2001).

### 4.3.1 Bagging in random forest

According to Breiman (1996), bagging is a term in machine learning which is used to refer to ensemble algorithms that are used to decrease variance and overfitting in classification and regression. In bagging sequential trees are generated and they do not depend on previous trees. Each tree is generated self-reliantly using bootstrap version of the data set. Finally the maximum value is used for prediction (Breiman, 2001).

When given training data $X = x_1, x_2, \ldots, x_n$ with responses $Y = y_1, y_2, \ldots, y_n$, bagging continuously selects a sample (B times) arbitrarily from and replaces it in the training set and develops trees using these samples:

For $b = 1, \ldots, B$:

1. Sample with replacement, $n$ training examples from $X$, $Y$; call these $X_b$, $Y_b$.
2. Train a decision or regression tree $f_b$ on $X_b$, $Y_b$.

After training the algorithm, predictions for unknown samples $x'$ could be made by calculating the average of the predictions from all the regression trees on $x'$ or selecting the highest value in the case of decision tree classification:

$$\bar{f} = \frac{1}{B}\sum_{b=1}^{B} f_b \quad (8) \quad (x')$$

Bagging is usually used in association with random feature selection. The two main reasons why bagging is used are because bagging appears to increase accuracy when it is used together with random feature and secondly bagging can be used to state a real-time approximation of the generalization error of the joined trees as well as the strength and the correlation. These approximations are done using out-of-bag estimation discussed in section 4.3.3 (Breiman, 2001)

### 4.3.2 Using random features

Breiman (2001) proposed randomness as an extra layer that is added to bagging. This randomness alters the way classification and regression trees are built. Under normal circumstance ordinary decision trees are built by splitting at each node based on the best split in the midst of all the attributes. Random forest differs in a way such that, splitting is done by basing on the best among a subgroup of predictors (trees) arbitrarily selected at that node. This method is considered to have better performance when related to some of the other classifiers such as SVM, ANN, other decision trees and it is not affected by overfitting (Breiman, 2001).

According to Liaw and Wiener (2002), random forest is very easy to use as it uses only two factors thus the amount of variables available in the arbitrary subset at each node and the number of trees to be constructed in the forest, and error in the values does not affect the performance. For more information on random forest refer to Leo Breiman (2001). Random Forests. Machine learning. 45, 5-32.

### 4.3.3 Out-of-bag estimates for error monitoring

Tibshirani (1996) and Wolpert and Macready (1999) suggested the use of out-of-bag as a way of finding generalization error. When bootstrap versions of the original training set are made, trees are developed from these bootstrap which are not pruned (Breiman, 2001). There is an empirical evidence to demonstrate that out-of-bag estimate is very correct just like running a test set which

has the same size as the training set. This means that there is no need for a separate test set of data (Breiman, 2001). One third of the records are left out of the training set in each bootstrap. Out-of-bag estimate is unbiased unlike cross-validation where there exist some form of error but the degree is unknown.

### 4.3.4 Random Forest in Weka

Random forest algorithm was executed to determine the credit worthiness of loan applicants. Weka has implemented random forest as a classifier with a number of parameters that are supplied before running the classifier. MaxDepth is the maximum depth of trees, numFeatures is the number of attributes to be used in random selection, numTrees is the number of trees to be generated and the seed is the random number seed. All these values are supplied as numbers (integers).

## 4.4 EVALUATION OF RANDOM FOREST CLASSIFIER

This section talks about the evaluation procedure for random forest classifier that was generated to determine the credit worthiness of loan applicants. Three different experiments were conducted. In order to generate a robust classifier, the random forest was evaluated using the following parameters (Sazonau, 2012). The parameters are number of trees in the forest, number of arbitrarily used features and using a decision tree.

### 4.4.1 Evaluating Random Forest using number of trees (*N*)

First of all, experiment number one was conducted and ran a hundred times. It was conducted to test the out-of-bag error rate against the number of trees. The number of trees that was generated in each run was changed. The first run generated one tree and the last run generated one hundred trees. However, the out-of-bag error was recorded in tens. 10-folds cross validation was used for this experiment and the averaged out-of-bag estimate for each run is graphed below in figure 4.4.

The result confirmed that the higher the number of trees the lesser the error rate (Sazonau, 2012).

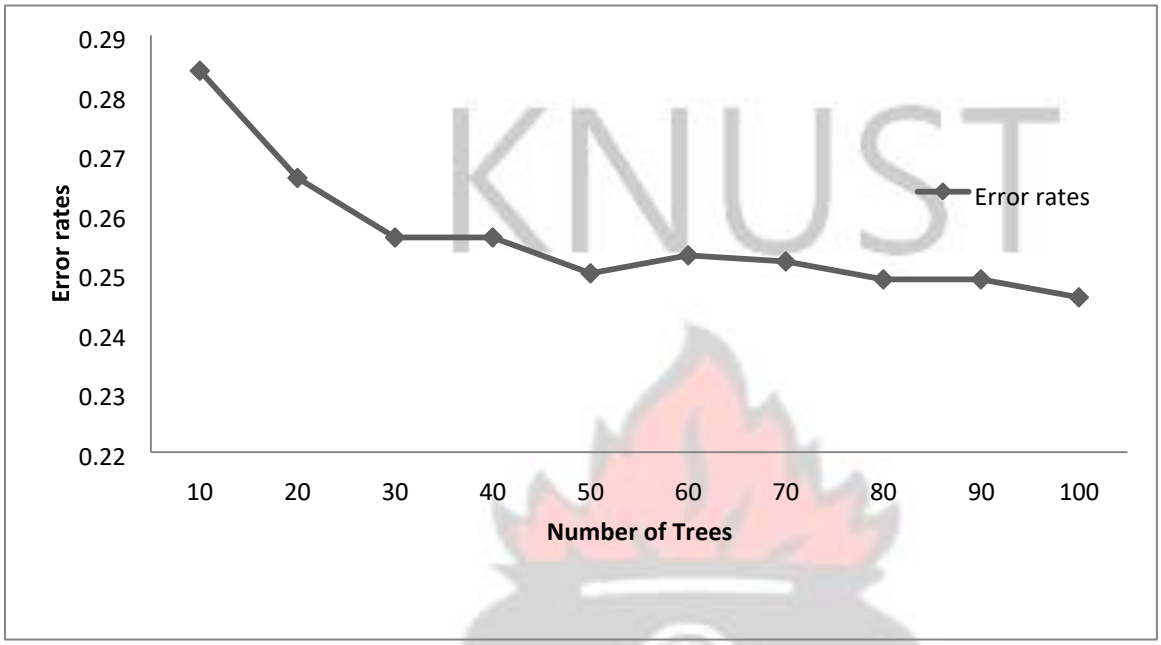



**Figure 4.4: The number of trees and their corresponding Out-of-bag error rates.**

### 4.4.2 Evaluating Random Forest using number of features (*NF*)

Experiment number two was conducted to test the error rate using the value supplied for the *numFeatures*. The experiment was run twenty times and in each run the *numFeatures* value was changed. The first run had a value of 1 while the last run had a value of 20. The number of trees for this experiment is 100 which is the default value in Weka. 10-folds cross validation was used and the averaged out-of-bag error rates are graphed below in figure 4.5.

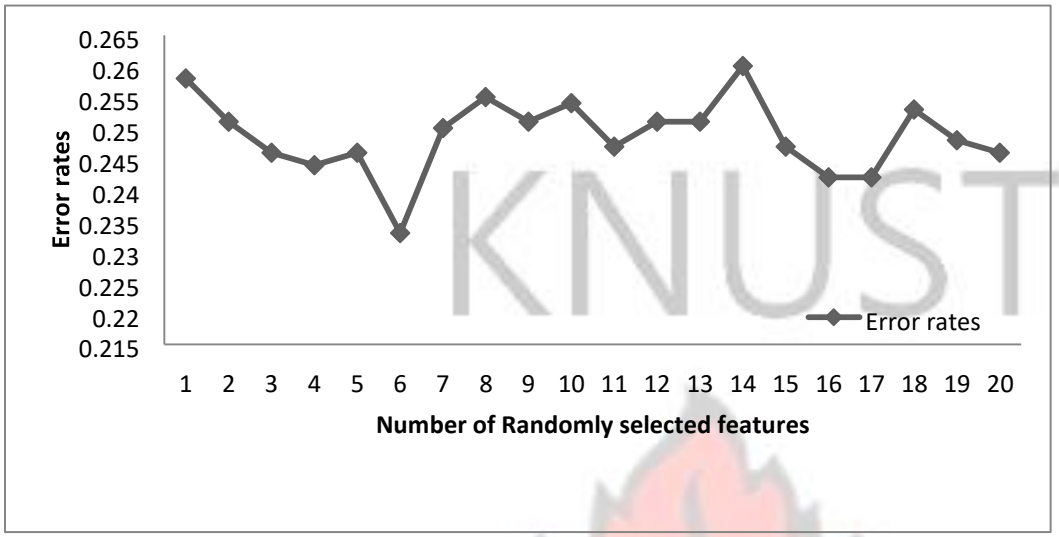



**Figure 4.5: The number of randomly selected features and their corresponding out-of-bag error rates.**

### 4.4.3 Evaluating Random Forest using a decision tree

This experiment was conducted to evaluate the performance of random forest against that of decision tree algorithm (J48 classifier). With this form of evaluation, there was the need to generate ROC curves which have averaged threshold for random forest and J48 classifier. Calculating the number of arbitrarily selected attributes *t* is equal to the square root of total number of attributes ($t = \sqrt{tn}$) (Sazonau, 2012). The random feature value obtained for the random forest and the tree is the same *t=4* however the number of trees for the random forest is 100. The 10-fold cross validation also applied to the two classifiers. The knowledge flow diagram that is used to generate the ROC curves is shown in figure 4.6. The threshold averaging scheme was implemented on the two ROC curves for the random forest and the J48 classifiers.

The resulted ROC graph is finally generated in figure 4.7.

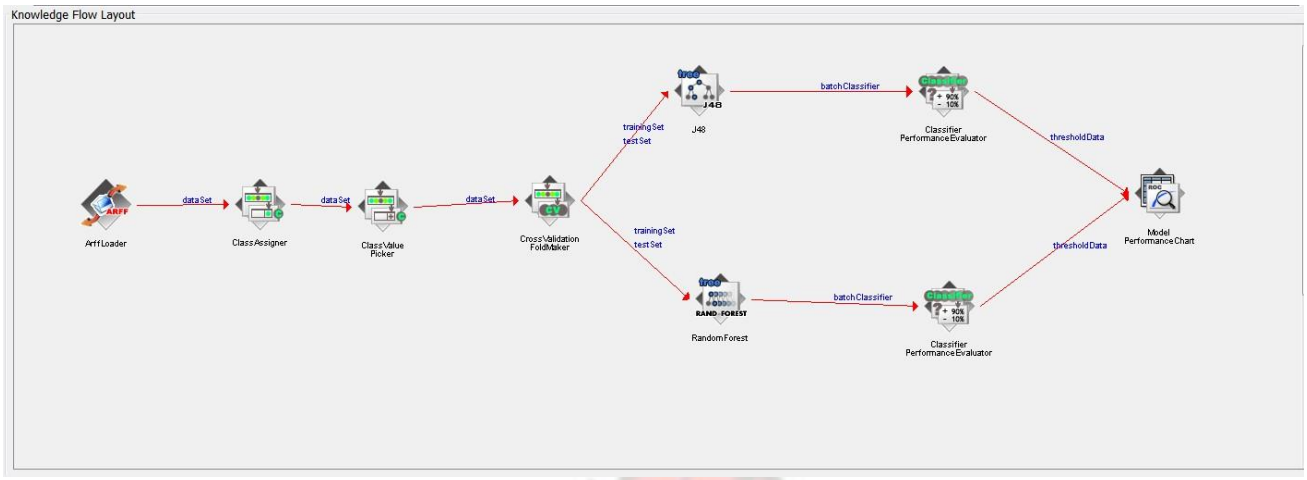



**Figure 4.6: The knowledge flow diagram for constructing the ROC curve for the random forest and the decision tree classifiers.**

The ROC graph indicates that the random forest does better classification than the decision tree classifier. The x (light line) represents the curve for the decision tree whiles the + (thick line) represents the curve for the random forest.

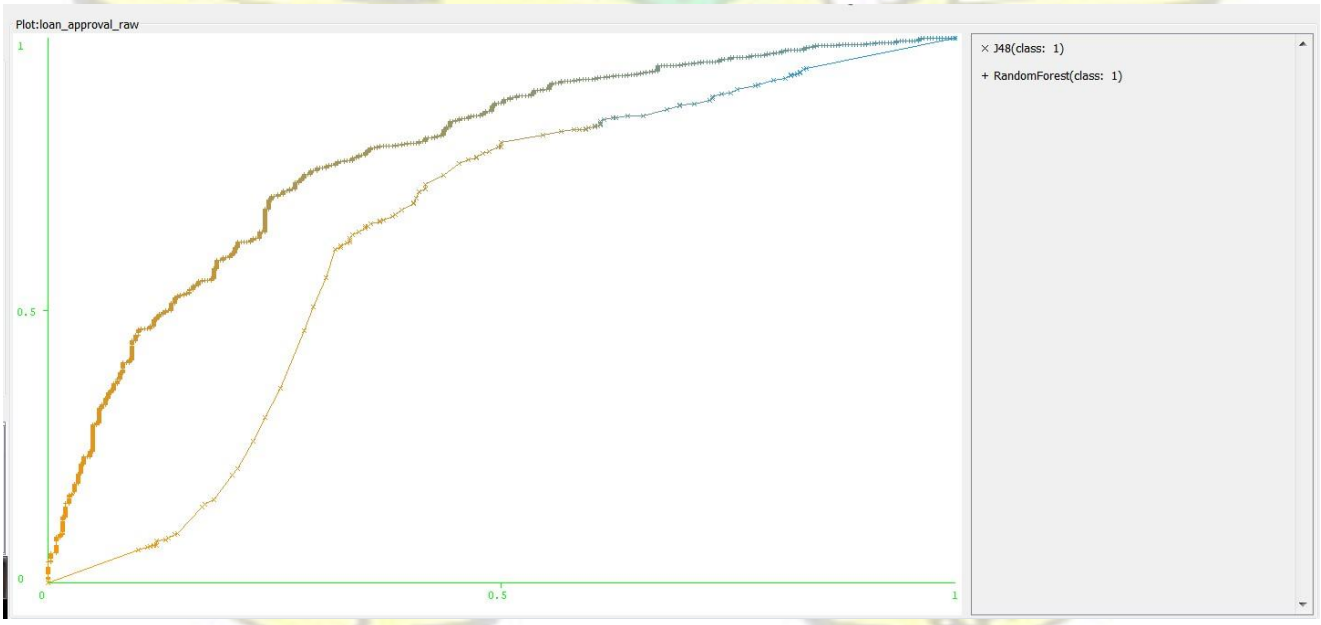



**Figure 4.7: ROC curves generated to compare the performance of random forest and decision tree classifiers.**

### 4.4.4 Statistics for random forest

Below are the statistics for the random forest classifier that was generated to determine the credit worthiness of loan applicants. Out of the numerous experiments conducted, it was realized that the model works best with the highest number of trees (100) (figure 4.4) and 6 number of random features (figure 4.5). Therefore these values were supplied as the parameters for the best random forest model and the results are shown below. The out-of-bag error estimate was 0.233. The RF made 764 instances of correct classification, 236 instances of incorrect classification and the kappa statistics of 0.3776. The kappa statistics value is considered to be in the range of 0.21 and 0.40 which means "fair agreement" (Viera & Garrett, 2005).

*=== Classifier model (full training set) ===*

*Random forest of 100 trees, each constructed while considering 6 random features.*

*Out of bag error: 0.233*

*Time taken to build model: 2.64 seconds*

*=== Stratified cross-validation ===*

*=== Summary ===*

| | | |
|---|---|---|
| *Correctly Classified Instances* | *764* | *76.4 %* |
| *Incorrectly Classified Instances* | *236* | *23.6 %* |
| *Kappa statistic* | *0.3776* | |
| *Mean absolute error* | *0.3353* | |
| *Root mean squared error* | *0.4039* | |
| *Relative absolute error* | *79.8094 %* | |
| *Root relative squared error* | *88.1301 %* | |
| *Total Number of Instances* | *1000* | |

## 4.5 SUMMARY OF CHAPTER FOUR

In this chapter, the processes involved in generating the two data mining classifiers (J48 decision tree classifier and the random forest classifier) as well as the performance evaluation procedures were discussed. For the decision tree algorithm entropy and information gain were discussed. Pruned and unpruned decision trees were generated and their statistics were discussed. The confusion matrix was tabulated for the pruned and unpruned tree as well as the accuracy, sensitivity and specificity were calculated. To be able to generate the ROC curve, the values generated by Weka for precision, recall and f-measure for both pruned and unpruned tree were tabulated and were eventually used to plot the ROC curve. Again the random forest classifier was explained and two of the most important concepts (bagging and out-of-bag estimate) were also explained. The random forest classifier was evaluated using the out-of-bag error rates against the number of tree and number of random features to be selected and also the random forest against a decision tree. The next chapter is results and analyses.

# CHAPTER FIVE

# RESULTS AND ANALYSES

## 5.0 INTRODUCTION

In model explanation, Du (2006) writes that it is appropriate to use visualization tools to explain the model. By this means, the model would be translated into a more comprehensive form for users. This chapter discusses the results that were obtained after the implementation of the J48 decision tree classifier ran on the dataset to determine prospects so that marketing efforts would be directed to them and Random Forest to determine the credit worthiness of loan applicants. Two datasets were used for this study and 10 – folds cross validation was used to run the models. The J48 was used because of its popularity and ease of usage (Abbas, 2015) and the random forest was used because of the high performance it promises (Breiman, 2001). The section is divided into four main sections (5.1, 5.2, 5.3 and 5.4). Section 5.3 is the summary of the main findings and section 5.4 is the summary of this chapter. Section 5.1 deals with the J48 decision tree classifier in relation to research question one and section 5.2 also deals with the random forest classifier in relation to research question two. After the implementation of the models and conducting the experiments, these are the results.

## 5.1 J48 DECISION TREE CLASSIFIER

The J48 DT classifier was used to detect prospective customers so that marketing campaigns would be directed to them which is in line with research question one. In chapter four, after the development of the J48 decision tree classifier, a number of evaluation techniques were implemented. Confusion matrix was used after which accuracy, sensitivity and specificity were calculated. Voznika and Viana claims that accuracy, specificity and sensitivity are measures that

determine the robustness and usefulness of classification algorithms. The J48 decision tree classifier that was developed to identify prospective customers yielded accuracy values of 93.7% for unpruned and 92.5% for pruned. This means that the model has about 92.5% accuracy when asked to make a classification or prediction based on an unknown data set using pruned decision tree and 93.7% using unpruned decision tree. These accuracy values are very good as Witten and Frank (2005) estimates high accuracy to be in the range of 0.75 and 1 representing 75% and 100% respectively.

### 5.1.1 Attribute contribution for J48 decision tree classifier

According to the J48 classifier, three attributes that contributed massively in the classification are *Contact*, *Education* and *Age* with the result tabulated below in descending order in table 5.1. As a result of this, the results analysis and discussion would be made around the three most contributing attributes.

| **Attribute** | **Percentage** | **Attribute** | **Percentage** |
|---|---|---|---|
| Contact | 100% | Children | 0% |
| Education | 100% | Occupation | 0% |
| Age | 90% | Gender | 0% |
| House | 0% | Marital status | 0% |
| Income | 0% | NA | NA |

**Table 5.1: Attributes and their percentage of contribution to the J48 decision tree classifier**

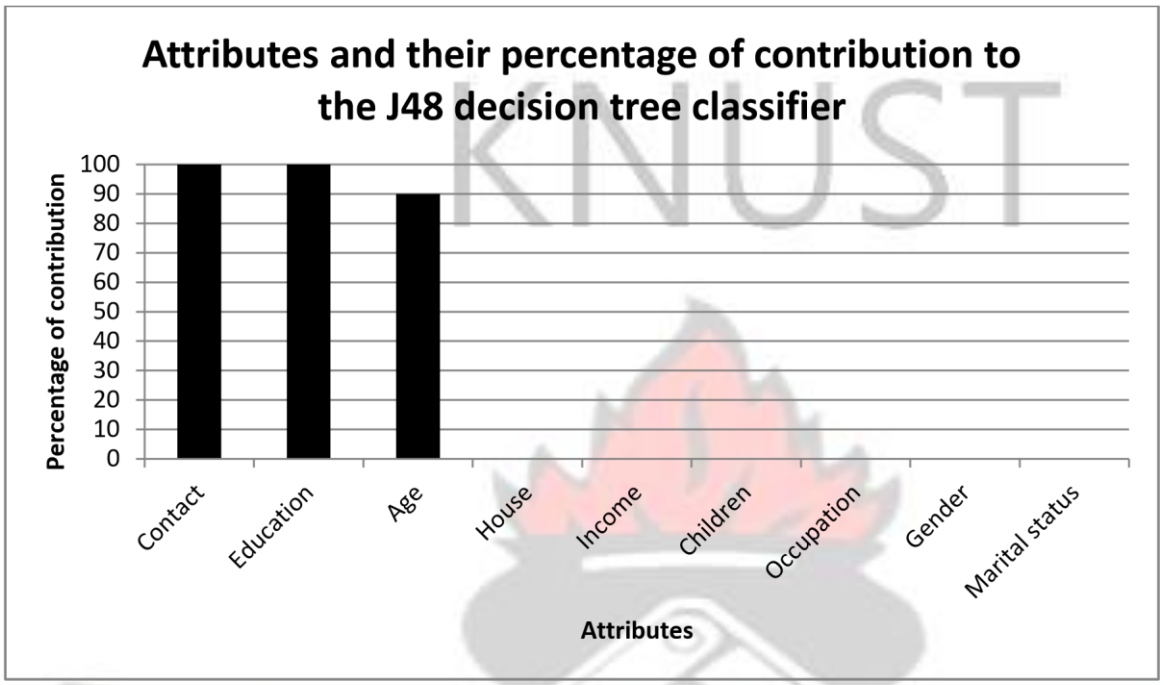



**Figure 5. 1: Attributes and their percentage of contribution to the J48 decision tree classifier**

#### 5.1.1.1 Contact Attribute

The prospects were approached on four different weeks in one month – first (red), second (cyan), third (blue), and fourth (deep dark green) week. It was discovered that most of the prospects were contacted in the third week (blue) with the total value of 341 contacts. These are summarized in figure 5.2.

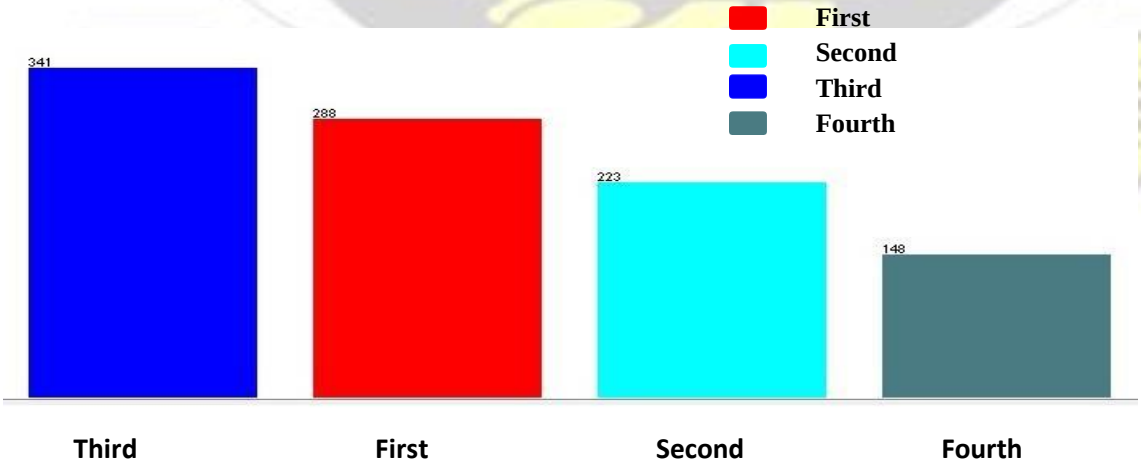


**Figure 5. 2: Periods that prospective customers were contacted.**

Figure 5.3 indicates that majority of the prospects who were contacted in the third week responded to purchase the loan product. Out of the 341 contacts made, 322 responded and 19 did not. This represents 94.43 percent for positive response and 5.57 percent negative response. Comparing the findings in figure 5.3, it could be analyzed that prospects who are contacted in third weeks are more likely to respond to the offer. The next highest contacts were made in the first week which were 288 contacts and for this, 15 responded and 273 did not respond indicating 5.21 percent positive responses and 94.79 percent negative responses respectively. The second week contacts recorded 24 (10.76%) positive responses and 199 (89.24%) negative responses. Likewise the fourth week contact also recorded 11 (7.43%) positive responses and 137 (92.57%) negative responses. Therefore it is better for the bank to market its loan products in third weeks.

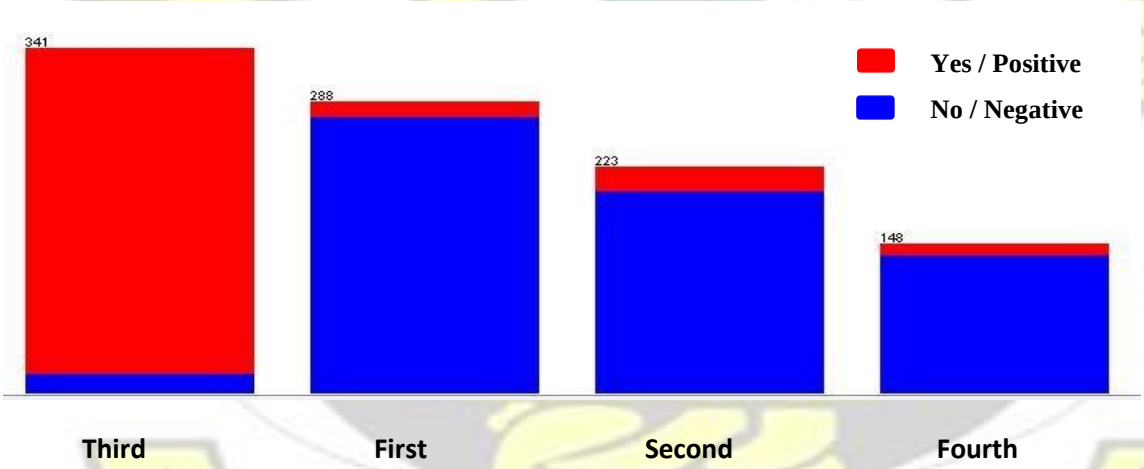



**Figure 5.3: Number of prospective customers who responded or otherwise to the offer with respect to the contact times.**

### 5.1.1.2 Education Attribute

All the prospects contacted were grouped into three different categories for the education attribute – Primary (blue), Secondary (red) and Tertiary (cyan). The results indicated that most of the people (448) contacted had secondary school as their highest educational level at the time the data was

being collected. The rest are having 284 and 268 people having primary and tertiary as their highest educational level respectively as indicated in figure 5.4.

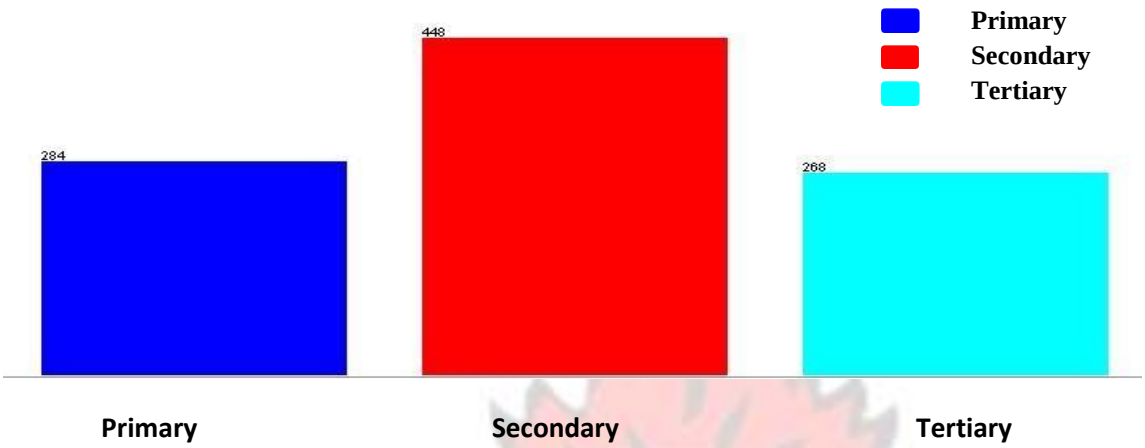



**Figure 5.4: The highest educational level of prospective customers.**

Figure 5.5 indicates majority of the people who responded to buy the loan product that was offered had tertiary education. Out of the 268 prospective customers contacted, 186 responded positively whiles 82 responded negatively indicating 69.40 percent and 30.60 percent positive and negative responses respectively. On the other hand the group that had primary as their highest level of education recorded 88 (30.90%) positive responses and 196 (69.10%) negative responses. For the 448 prospects who had secondary as their highest education, 98 (21.87%) responded positively and 350 (78.13%). Although the number of contacts made for secondary education prospects is very high, when compared with tertiary education in terms of positive responses, tertiary is significantly higher than secondary with 69.40 percent against 21.87 percent. This means that people with tertiary education has a higher possibility of responding to a loan offer and that the bank can focus on this group of people for loan marketing.

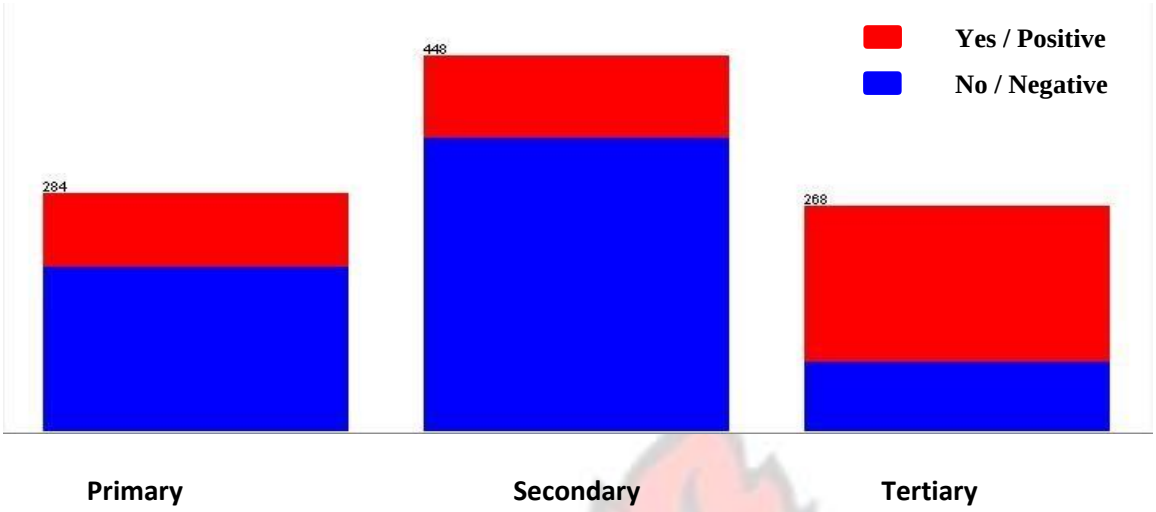



**Figure 5.5: Number of prospective customers who responded or otherwise to the offer with respect to their educational level.**

### 5.1.1.3 Age Attribute

Under the age attribute there were six categories. These were 20-30 (blue), 31-40 (red), 41-50 (cyan), 51-60 (deep green black), 61-70 (pink), and 71-100 (green). As depicted by figure 5.6 the classifier identified the age group with the highest contacts as 41-50 (cyan) with total number of 370 contacts followed by 31-40 with total number of 300 contacts and 51-60 with 150 contacts.

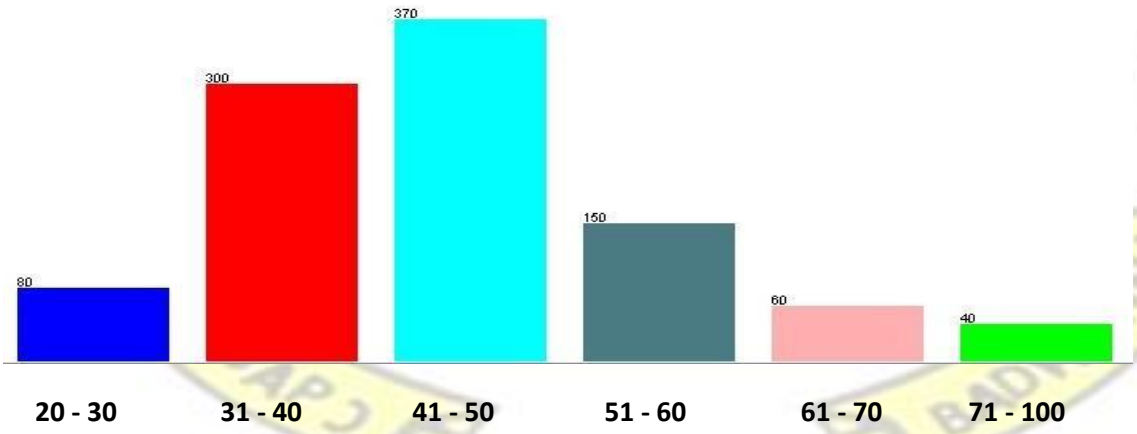


**Figure 5.6: Age group of prospective customers**

In figure 5.7, it was realized that the age group that responded massively to the offer was 31-40 with 150 (50%) positive and 150 (50%) negative responses. The second group that also responded quiet significantly with 150 contacts is 51-60, which recorded 72 (48%) positive and

78 (52%) negative responses. Again the third group that also responded quite well is 41-50 with 91 (24.60%) positive and 279 (75.40%) negative responses. Therefore it can be analyzed that most of the prospective customers who responded to the offer are in the age group of 31-40 followed by 51-60 (50% and 48% respectively) and that the bank could target this group for loan marketing and make it very attractive to the local people.

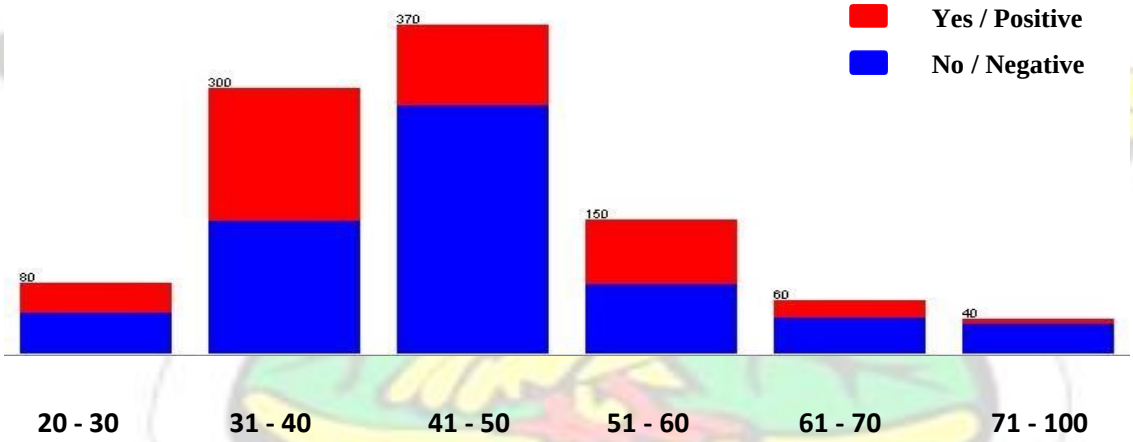



**Figure 5. 7: Number of prospective customers who responded or otherwise to the offer with respect to age group.**

### 5.1.1.4 Class Attribute

Finally figure 5.8 depicts the overall findings of the classifier taking into consideration all the attributes. The main objective for implementing the J48 decision tree algorithm was to determine the number of prospective customers who are likely to respond when approached with an offer to buy a loan product from the bank and then their characteristics would be used in direct marketing. Out of the 1000 records used by the classifier 372 prospects representing 37.2 % responded

positively while 628 prospects representing 62.8 % responded negatively. Therefore it can be analyzed and concluded at this point that, the J48 decision tree classifier has successfully classified the instances and that data mining can assist in direct marketing.

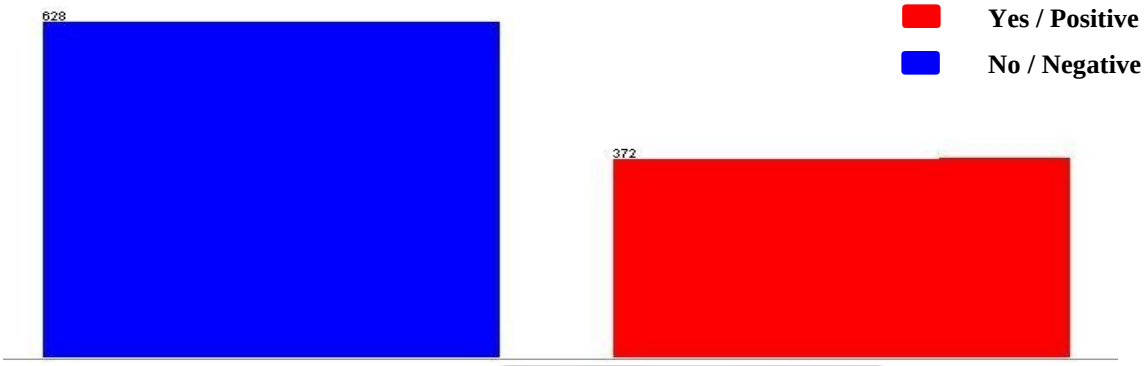



**Figure 5.8: Number of prospective customers who responded and otherwise to the offer.**

## 5.2 RANDOM FOREST CLASSIFIER

The random forest classifier was generated to determine the credit worthiness of loan applicants which is in line with research question two. The dataset used consisted of 1000 records and 20 attributes excluding the class attribute. With the RF classifier a number of experiments were conducted. The RF classifier correctly classified 774 representing 77.4 percent and wrongly classified 226 instances representing 22.6 percent. According to Witten and Frank (2005) estimates of high accurate model, the RF can be considered as very accurate since it is above the 75 percent lower bound.

### 5.2.1 Analyzing Random Forest using number of trees (*N*)

The experiment conducted for this evaluation used out-of-bag error rates against the number of trees generated. It was realized that the higher the number of tress the lower the error rate. This indicates that the higher the number of trees generated the better the RF classifier and the result confirms the findings of Sazonau (2012). The graph in figure 4.4 indicates a drastic reduction of

error rate in the first 30 generated trees. Above the first 30 trees generated, the difference in error rates is low with *N+1* slightly better than *N*

### 5.2.2 Analyzing Random Forest using number of features (*NF*)

This experiment was conducted to compare the error rate against the number of features that are arbitrarily selected as indicated in figure 4.5. The results indicate that there is no specific trend in random forest performance with respect to the *NF*. The last one can be described as the persistent compromise by raising the strength and firmness of a single tree from one side and resulting growth of the relationship between learners from the other, affecting the growth in generalization error upper bound (Sazonau, 2012).

### 5.2.3 Analyzing Random Forest using a decision tree

The experiment produced a resulting graph that confirms a general report that random forest performs better than one tree in a ROC space (Sazonau, 2012). The RF ROC curve indicated as series of connected + signs is a little bit thicker and higher than that of the decision tree as indicated in figure 4.7 as series of * signs. Again, because the RF ROC curve is thicker than that of the decision tree, the RF is considered to have high confidence interval than the decision tree.

### 5.2.4 Analyzing Random Forest using the statistics

A similar research was conducted by Zamani and Mogaddam (2015) in Mellat Bank of Iran. They used J48 decision tree classifier in Weka to detect credit risk of customers in the bank on the same data set. They also used 10-folds cross validation. Their findings are a little different from the findings in this study. Their model correctly classified 711 (71.1%) instances as against 764 (76.4%) instances in this study. Also their model incorrectly classified 289 (28.9%) instances as against 236 (23.6%) in this study. They had a kappa statistic of 0.259 as against 0.3776 in this

study. This indicates that Random forest model performs a little better in classification than J48 decision tree models (Breiman, 2001).

### 5.2.5 Attribute contribution for Random Forest

According to the RF classifier, three attributes that contributed very well in the classification are *Current_balance, Age_of_the_account_(months)* and *Credit_history* with the result tabulated below in table 5.2 and graphed in figure 5.9. As a result of this, the results analyses and discussion for the RF model would be made around the three most contributing attributes. For complete attribute contribution chart refer to Appendix B.

| No | Attribute | % | No | Attribute | % |
|---|---|---|---|---|---|
| 1 | current_balance | 100 | 11 | yrs_in_current_residence | 0 |
| 2 | age_of_account_(months) | 100 | 12 | property | 0 |
| 3 | credit_history | 100 | 13 | age | 20 |
| 4 | purpose_of_loan | 0 | 14 | installment_plans | 0 |
| 5 | claimed_amount | 20 | 15 | shelter | 40 |
| 6 | savings_account | 50 | 16 | pending_loans_in_this_bank | 0 |
| 7 | yrs_of_current_employment | 0 | 17 | occupation | 0 |
| 8 | installment_rate | 0 | 18 | number_of_dependants | 0 |
| 9 | marital_status/sex | 0 | 19 | telephone | 0 |
| 10 | other_debtors/guarantors | 0 | 20 | local_or_foreign_worker | 10 |

**Table 5.2: Attributes and their percentage of contribution to the random forest classifier**

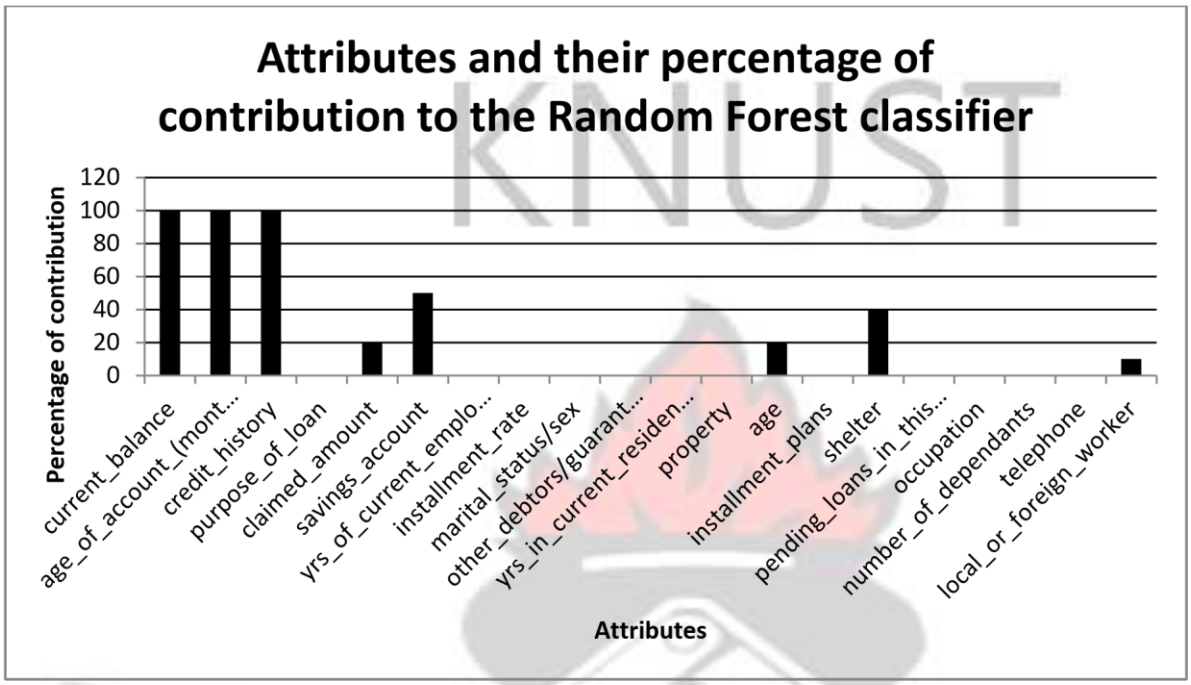



**Figure 5.9: Attributes and their percentage of contribution to the random forest classifier**

**5.2.5.1 Current Balance**

This attribute recorded the amount of cedis loan applicants have in their accounts at the time of loan application. Items under the current balance attribute were grouped into four categories in Ghana cedis. These are A11 (blue), A12 (red), A13 (cyan) and A14 (deep dark green). A11 are instances that are less than 0, A12 are instances that range between 0 and 200, A13 is greater than or equal to 200 and A14 means no checking account. It was identified by the model that A14 recorded the highest contribution value of 394 applications. The second highest is A11 which recorded a value of 274 followed by A12 and A13 which recorded 269 and 64 values respectively. The graph is in figure 5.10.

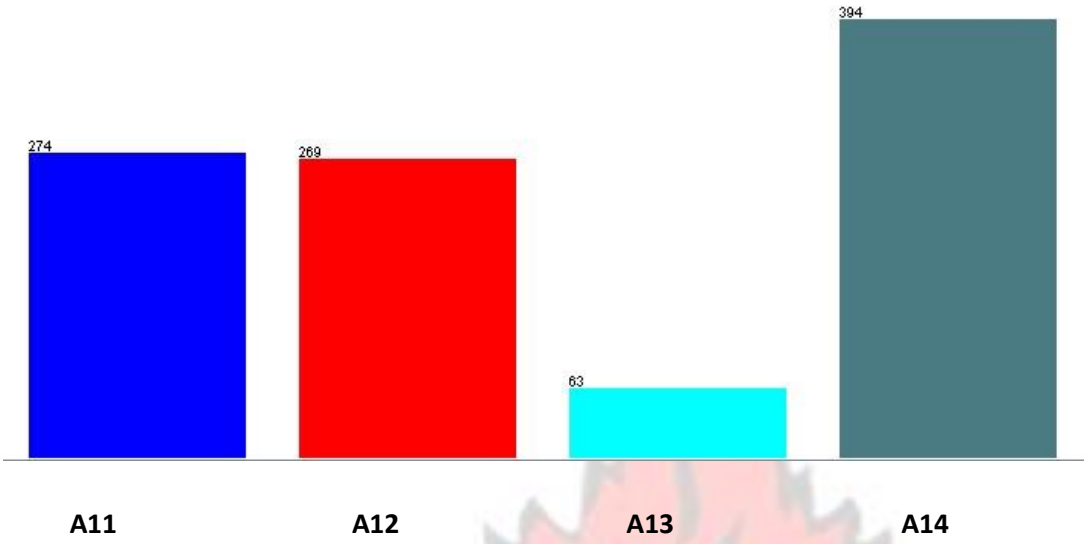



**Figure 5.10: Number of customers that are grouped according to account balance**

Figure 5.11 shows that among the four groups, A11 and A12 which mean customers with account balance less than 0 and between 0 and 200 cedis are very risky (bad) for granting loan to. A11 recorded 139 (50.73%) good and 135 (49.27%) bad loans. A12 also recorded 164 (60.97%) good and 105 (39.03%) bad loans. On the positive side, it was identified that customers with no accounts (A14) are less risky (good) for granting loans to with 348 (88.32%) good and 46 (11.68%) bad loans. This can be analyzed that, loan applicants in the category of A11 and A12 should be scrutinized critically before granting loans to them.

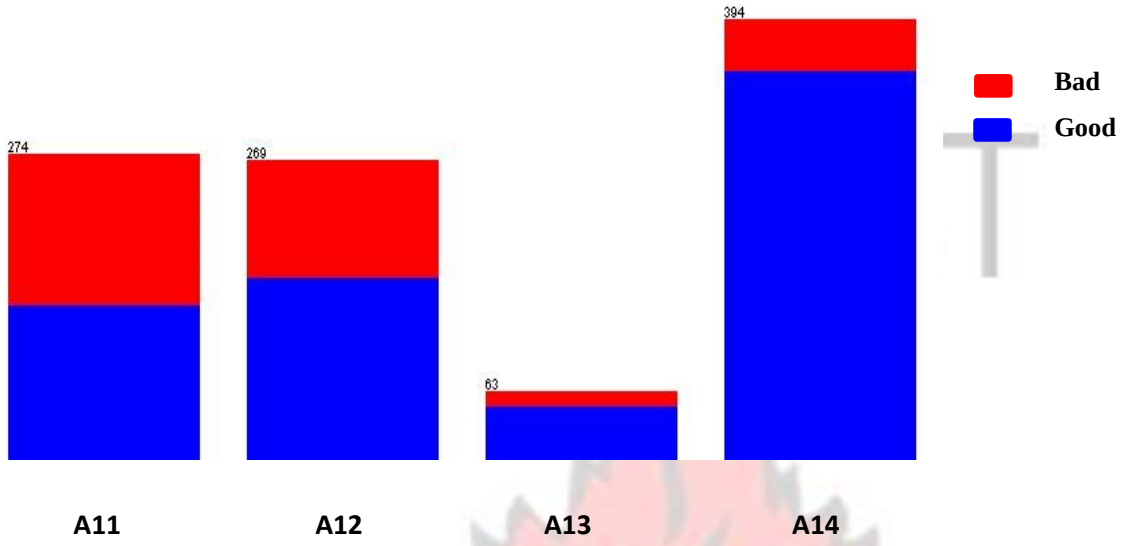



**Figure 5.11: Number of customers that are grouped according to account balance and whether good or bad**

### 5.2.5.2 Age of the Account in months

Items under this attribute represent how old the accounts of customers are in months. Because the items under this attribute are numeric (integer), the minimum value was 4 (months), the maximum value is 72 (months), the mean is 20.9 (months) and standard deviation is 12.059 (months). It was realized that the highest number (265) of applicants had their account active from 8 to 12 months followed by 186 applicants having their account active for 21 to 25 months. The last group of 151 applicants had their account active between 16 and 21. The other groups of applicants are less than 100. Figure 5.12 summarizes this information.

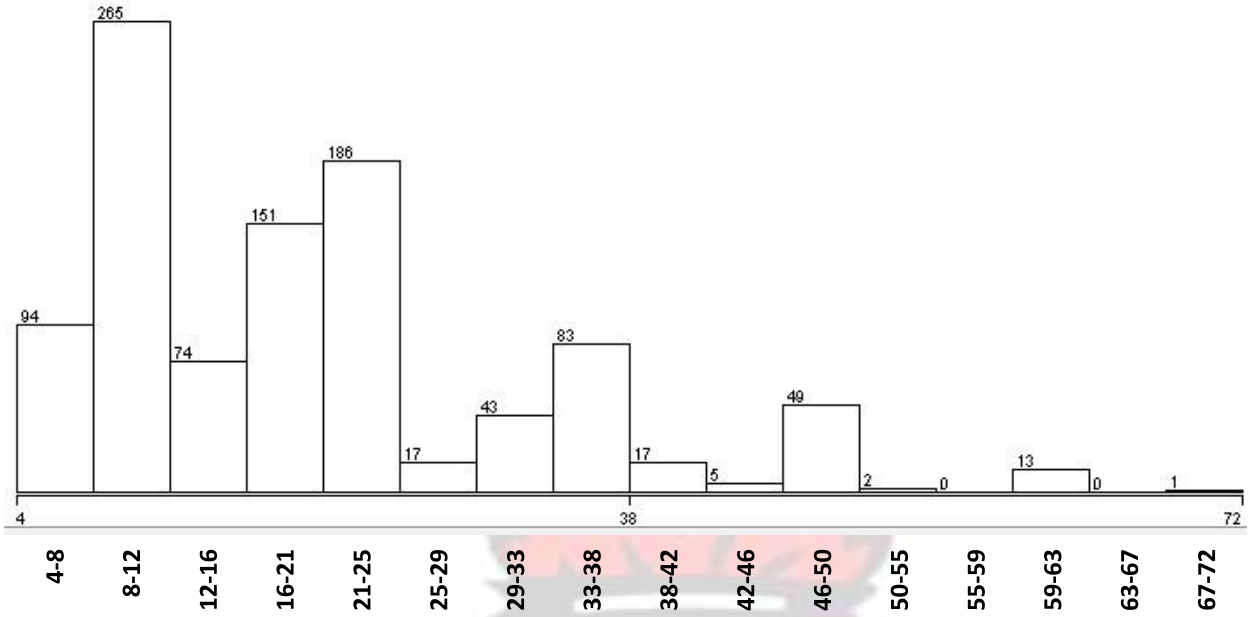



**Figure 5.12: Age of loan applicants' account in months**

Figure 5.13 show the summary of how old loan applicants' accounts are in the light of good or bad loan application. The loan applicants who had their account between 46-50 months old recorded 58% more bad loans. The second highest is 33-38 months which recorded 55% bad loans. Again 21-25 months old and 8-12 months old recorded 28.89% and 25% bad loans respectively. This means that 46-50 months old group is more risky followed by, 33-38, 21-25 and 8-12.

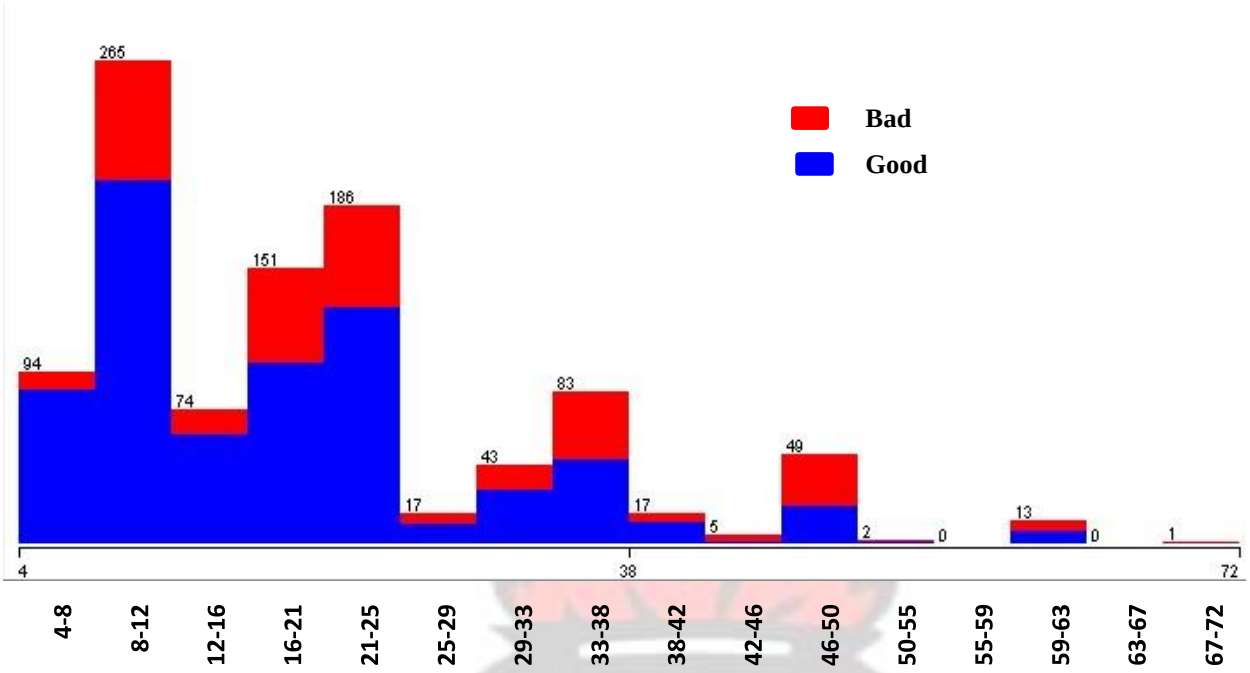



**Figure 5.13: Age of loan applicants' account in months and whether good or bad**

### 5.2.5.3 Credit history

Items under this attribute are categorical. They represent the credit history of loan applicants. There are five categories – A30 (blue, no credit taken or all credits paid back duly), A31 (red, all credits at this bank paid back duly), A32 (cyan, existing credit paid back duly till now), A33 (deep dark green, delay in paying off in the past) and A34 (pink, critical account/ other credits existing but not at this bank). It was realized that 530 (53%) of the loan applicants had A32 (existing credit paid back duly till now) and the second highest group is A34 with 293 (29.3%) applicants. This information is summarized in figure 5.14. The next figure 5.15 describes that out of the 530 applicants about 70 percent were found good and about 30 percent were considered bad loan applicants.

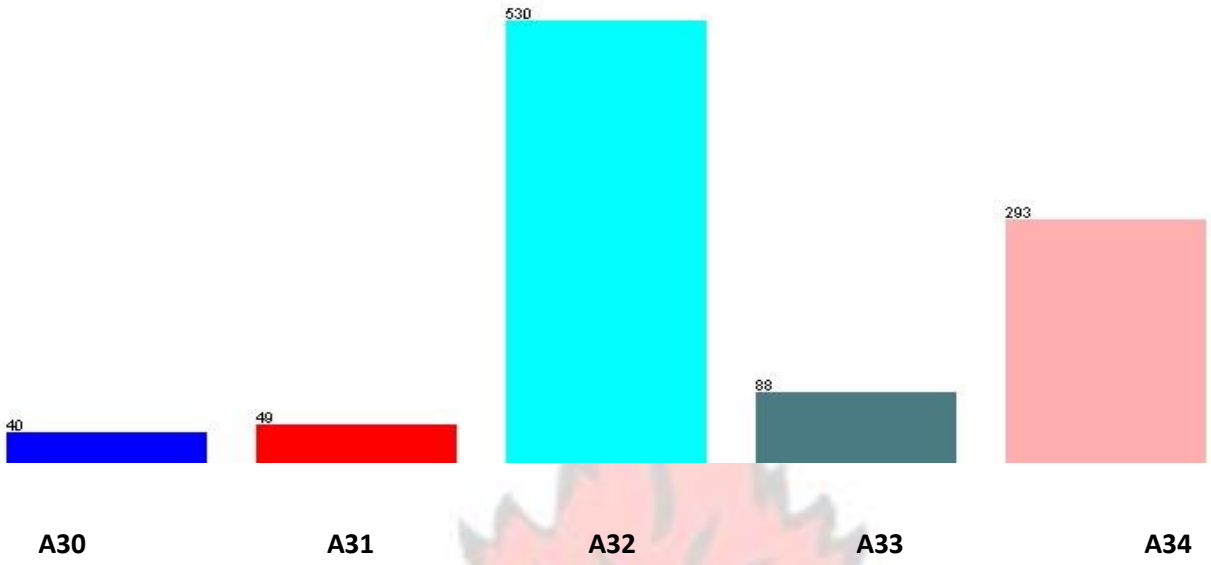



**Figure 5.14: Credit history of loan applicants**

Figure 5.15 shows very fascinating results. Although A30 and A31 recorded the least contribution values with 40 and 49 respectively as against A32 (530) and A34 (293), the percentage of bad loan in these groups are significantly noticeable with 25 (62.5%) bad loans for A30 and 28 (57.14%) for A31.

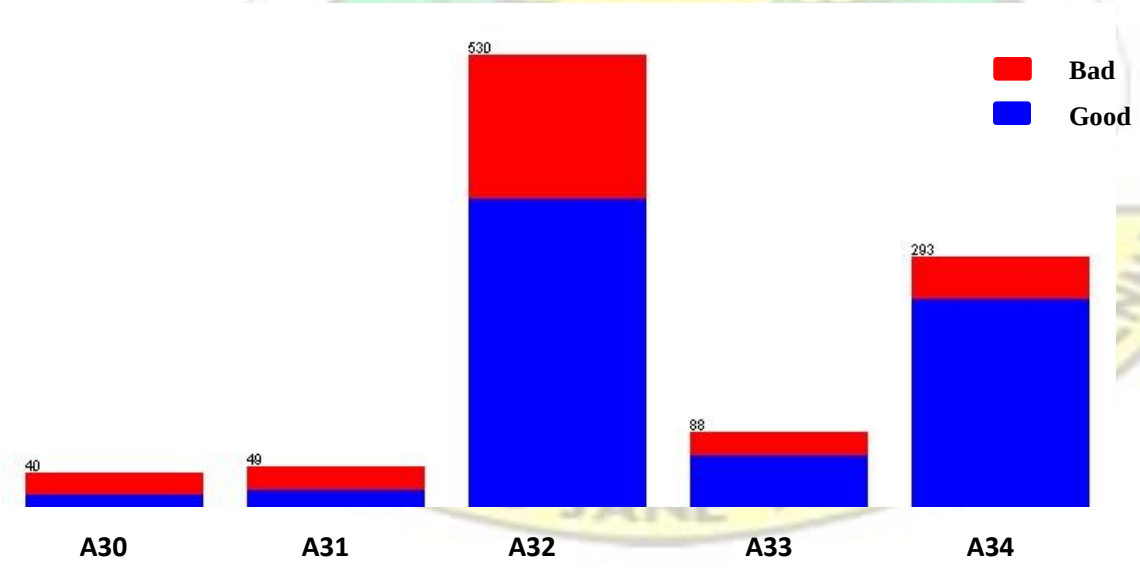


**Figure 5.15: Credit history of customers and whether good or bad**

### 5.2.5.4 Class attribute

Finally figure 5.16 shows the total findings of the classifier taking into consideration all the attributes. The main objective of using the RF algorithm was to determine the credit worthiness of loan applicants who apply for loans at the bank. The data set used for this study had two classes (Good and Bad). Out of the 1000 instances used by the classifier 700 (70%) were classified as good (not risky) to be given the loan. On the other hand 300 (30%) of the applicants were classified as bad (risky) and not to be given the loan. Therefore it can be analyzed and concluded at this point that, the RF has successfully classified the instances.

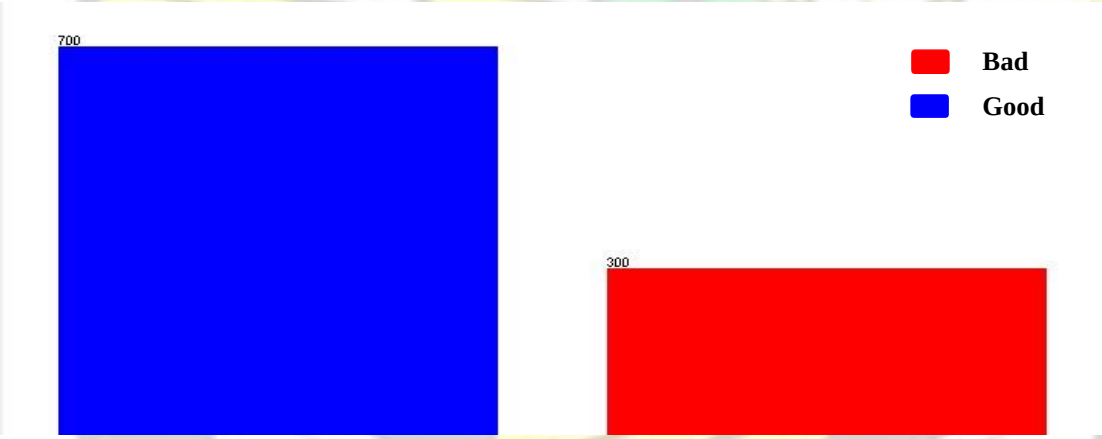



**Figure 5.16: Number of loan applications classified as good or bad**

## 5.3 SUMMARY OF MAIN FINDINGS

The following section presents the summary of the main findings of this work. For easy readability it is presented in bulleted form.

- The J48 decision tree classifier identified that 322 (94.42%) prospects contacted in the third week responded positively as against 19 (5.58%) who responded negatively. In the first week 15 (5.21%) prospects responded positively and 273(94.79%) responded negatively. The second week recorded 24 (10.76%) positive and 199 (89.24%) negative responses. The fourth week also recorded 11 (7.43%) positive and 137 (92.57%) negative responses.
- Again prospects with tertiary education responded very well with 186 (69.40%) positive and 82 (30.60%) negative responses. This is against primary education with 88 (30.99%) positive and 196 (69.01%) negative responses as well as secondary education with 98 (21.87%) positive and 350 (78.13%) negative responses.
- The age group that patronized the loan product mostly is 31-40 with 150 (50%) positive and 150 (50%) negative responses. The next interesting finding is that, 51-60 had 72 (48%) positive and 78 (52%) negative responses. 41-50 age group recorded 91 (24.60%) positive and 279 (75.40%) negative responses.
- In totality, 372 prospects representing 37.2 % responded positively while 628 prospects representing 62.8 % responded negatively.
- The random forest classifier also identified that A14 thus loan applicants with no account are very good for granting loans to as out of 394 applicants 348 (88.32%) are good and 46 (11.62%) are bad. This is followed by A13 with 49 (77.78%) good and 14 (22.22) bad application. Again A12 recorded 164 (60.97) good and 105 (39.03%) bad application.

The last one is A11 which recorded 139 (50.73%) good and 135 (49.27%) bad application.

- The loan applicants who had their account between 46-50 months old recorded 58% bad loans. The second highest is 33-38 months which recorded 55% bad loans. Again 16-21, 21-25 and 8-12 months old recorded 33.33%, 28.89% and 25% bad loans respectively.
- Under the credit history of loan applicants, although A30 and A31 recorded the least contribution values with 40 and 49 respectively as against A32 (530) and A34 (293), the percentage of bad loan in these groups are significantly higher with 25 (62.5%) bad loans for A30 and 28 (57.14%) for A31.
- In all 700 (70%) were classified as good (not risky) to be granted the loan and 300 (30%) of the applicants were classified as bad (risky) and not to be given the loan.

## 5.4 SUMMARY OF CHAPTER FIVE

The discussion of this chapter was about the two classifiers that were generated in chapter four. In relation to research question one, the J48 DT classifier was modeled to find out how data mining can be used to detect probable customers so that marketing campaigns can be directed to them. Again in relation to research question two the random forest classifier was also modeled to analyze the credit worthiness of loan applicants. The next chapter is about the summary, conclusion, recommendation and future works.

# CHAPTER SIX

# CONCLUSION AND RECOMMENDATIONS

This chapter summarizes the entire findings of the study. It is divided into three main sections. Again it provides the conclusion, recommendations made for the case bank, and the future works.

## 6.1 SUMMARY

The main objectives of this work were first, to determine how DM model could be used to detect probable customers who may be interested in a loan product so that the scope of marketing could be limited and administrative focus enhanced and second, to detect the credit worthiness or assess the credit risk of loan applicants. Upon careful consideration of the literature, data mining seems to have enormous benefits in the application of marketing (Kadiyala & Srivastava, 2011) and credit risk analysis (Davis, 2001). Out of this, two data mining algorithms (J48 decision tree and random forest) were implemented on two different data sets. In order to check the performance of the decision tree model the confusion matrix was used and then accuracy, specificity and sensitivity were calculated. ROC curve was also generated to depict the robustness and usefulness of the decision tree model in a graphical form. Also the random forest was evaluated using the number of trees generated, number of features arbitrarily selected and against a single decision tree. The two models yielded remarkable results when it comes to correctly classifying instances with accuracy of 92.5% (pruned DT), 93.7% (unpruned DT) and
77.4% (random forest). Open source data mining software called Weka from university of Waikato was used to simulate all the experiments.

## 6.2 CONCLUSION

With fields like the banking industry, data are being generated massively on regular basis. So there is the need to turn these data into useful information. Employing human expert to analyze the huge data stored in databases and data warehouses is nearly impossible. With this, fields like data mining, machine learning and artificial intelligence allow the development or generation of well defined, robust and simple to use models to sort through and analyze huge data that are impossible for the human brain (Doug Alexander). This study has demonstrated with practical methods and experiments that data mining can be used to assist in marketing and analyze credit risk assessment. Typically it was realized in chapter five that out of the 341 contacts made in the third week, 322 responded and 19 did not respond as against 288 contacts made in first week where 15 responded and 273 did not respond. Again, it was identified that out of 394 loan application submitted by customers with no checking account, 88.32 percent were found to be good (not risky) to be granted the loan. Therefore bank managers, bank administrators, loan officers and Information technology officers in Yaa Asantewaa Rural bank and other banks in Ghana can apply these data mining models to detect probable prospects to channel their resources and products to so that administrative focus can be enhance and expensive resources can be used gainfully. Also they can use the RF model to automate the process of granting loans so as to limit or possibly completely desist from granting loans to risky customers. By adopting this way it is expected that Banks in Ghana would operate very well to serve its customers and will be able to make the expected returns.

## 6.3 RECOMMENDATION AND FUTURE WORK

In light of the general nature of data mining and the findings of this study, the following recommendations are made to help bank administrators (Managers, Loan officers) to fully leverage the benefits of data mining.

- The marketing team should market loans in third weeks rather than any other weeks.
- Prospects with tertiary as their educational level should be targeted more because they are very promising for loan marketing.
- The age group 31-40 is more likely to respond when approached to buy a loan product.
- The loan officer can approve loans of applicants with no checking account but applicants with account balance less than 0 must be critically scrutinized.
- Applicants with account age between 46-50 and 33-38 are very risky for granting loans to.
- Data mining requires huge data to be analyzed for interesting pattern and so data should be collected on regular basis by the bank's mobile bankers from the local people so that new and helpful discoveries could be made to enhance the process of marketing the bank's products.
- The J48 DT classifier yielded very intriguing results in that maximum number of prospects contacted in the third week responded as compared to the other weeks. With this it is recommended that the bank authorities should further investigate if there could be some factors influencing the massive third week purchases. That could be used to gain competitive advantage.

As identified by Foust and Pressman (2008) there are generally three main risks that exist in every bank. These are credit risk, Market risk and operational risk. This study has demonstrated that DM can be used to limit or possibly absolutely extinguish the effect of market risk and credit risk. Therefore future work may include the following:

- To find ways to leverage data mining to reduce operational risk. By this banks may be expected to reap the maximum benefits from their operations and be able to serve their customers well.
- A very big problem is that banks are particularly sensitive about their data. This prevented some data from being included in this study. It is believed that if bank authorities are assured of good data protection and customer privacy then very important data could be released for inclusion in similar studies. With this future work should be, how to find ways of ensuring proper security, confidentiality and privacy protection so that customer data, transactional data and some other proprietary data could be protected and at the same time could be mined for maximum benefits.



## REFERENCES

Abbas S. (2015), Deposit subscribe Prediction using Data Mining Techniques based Real Marketing Dataset, International Journal of Computer Applications (0975 – 8887) Volume 110 – No. 3, January 2015 (Accessed Monday, October 26, 2015, 8:21:10 AM)

Abrahams, C. & Zhang,M.(2009). *Credit risk assessment: the new lending system for borrowers, lenders, and investors.* New Jersey: John Wiley & Sons, Inc.

Anand V., S., Bhujade V., Bhagat P., Khaparde A. (2014), A Review Paper on Various Data Mining Techniques, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 4, April, 98 – 101, available online at www.ijarcsse.com (accessed Wednesday, August 26, 2015, 11:43:48 AM)

Anil K., D., & Ravi, V. (2008). Predicting credit card customer churn in banks using data mining. International Journal of Data Analysis Techniques and Strategies, 1(1), 4-28

Archana S., Elangovan K. (2014), Survey of Classification Techniques in Data Mining International Journal of Computer Science and Mobile Applications, Vol.2 Issue 2, February, 65-71, available online at www.ijcsma.com (accessed Monday, October 12, 2015, 9:31:29 AM)

Bhambri V. (2011), Application of Data Mining in Banking Sector, International Journal of Computer Science and Technology Vol. 2, Issue 2, June 2011, 199 – 202, available online at www. ijcst.com (Accessed Wednesday, October 14, 2015, 12:41:15 PM)

Bhasin, M. L. (2006). Data mining: a competitive tool in the banking and retail industries. Banking and finance, 588.

Boateng, F. G., Nortey, S., Asamanin Barnie, J., Dwumah, P., Acheampong, M., & AckomSampene, E. (2016) Collapsing Microfinance Institutions in Ghana: An Account of How Four
Expanded and Imploded in the Ashanti Region.*International Journal of African Development*, *3*(2), 5.

Breiman L., Friedman J., Olshen R., and Stone C. (1984). Classification and Regression Trees, Belmont, California: Wadsworth Int. Group

Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), 123-140.

Breiman, L. (2001). Random forest. Machine learning, 45(1), 5-32

Bruno C. da Rocha and Rafael T. de Sousa Junior (2010), Identifying Bank Frauds Using CRISP-DM and Decision Trees, International journal of computer science & information Technology (IJCSIT) Vol.2, No.5, October 2010, 162 – 169, (accessed online Monday, November 09, 2015, 4:18:57 PM)

Catlett, J. (1986). *Induction using the shafer representation* (Technical report). Basser Department of Computer Science, University of Sydney, Australia.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000), *CRISP-DM 1.0 - Step-by-step data mining guide*, CRISP-DM Consortium.

Chitra, K., & Subashini, B. (2013). Data Mining Techniques and its Applications in Banking Sector, International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 8, August 2013

Chiu S., and Tavella D. (2008), *Data Mining and Market Intelligence For Optimal Marketing Returns*, Oxford,UK Elsevier Inc.

Chopra, B., Bhambri, V., Krishan, B.(2011), Implementation of Data Mining Techniques for Strategic CRM Issues, Int. J. Comp. Tech. Appl., Vol 2 (4), 879-883 Available online at www.ijcta.com (Accessed Wednesday, August 26, 2015, 10:29:24 PM)

Coppock, D., S. (2002). *Why Lift? - Data Modeling and Mining*, Information Management Online (June).

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

Creswell J., W. (2003), Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Thousand Oaks, CA: SAGE Publications. (Accessed Wednesday, November 04, 2015, 2:52:08 AM)

Creswell W. J. (2003). RESEARCH DESIGN Qualitative, Quantitative. and Mixed Methods Approaches. SECOND EDITION, **SAGE** Publications International Educational and Professional Publisher.

Cufoglu, A., Lohi, M., & Madani, K. (2009). A comparative study of selected classifiers with classification accuracy in user profiling. In Computer Science and Information Engineering, 2009 WRI World Congress on (Vol. 3, pp. 708-712). IEEE.

Dasarathy, B.V. (1991). *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*, ed. IEEE Computer Society Press

Davis, C. (2001). QUADSTONE Insures Liverpool Victoria CRM Strategy With Scalable Technology.

Dick, D. L. (2013). Bankruptcy's Corporate Tax Loophole. *Fordham L. Rev.*,*82*, 2273.

Doug Alexander, "Data Mining", http://www.laits.utexas.edu/~anorman/BUS.FOR/course.mat /Alex/, (accessed Thursday, October 22, 2015, 8:05:15 PM)

Du, X. (2006). Data Mining Analysis and Modeling for Marketing Based on Attributes of Customer Relationship

Dzelihodzic, A. D. N. A. N., & Donko, D. Z. E. N. A. N. A. (2013). Data Mining Techniques for Credit Risk Assessment Task. Recent Advances in Computer Science and Applications, 6-8.

Elsalamony, H. A. (2014). Bank Direct Marketing Analysis of Data Mining Techniques. Network, 5, 0. *International Journal of Computer Applications (0975 – 8887) Volume 85 – No 7, January 2014*

Emans, B. (1986). Interviewen; theorie, techniek en training. Groningen: Wolters – Noordhoff.

Fawcett T. (2006). An introduction to ROC analysis. Pattern Recognition Letters. 2006; 27(8): 861–874.

Fawcett, T. (2005). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861874.

Finlay, S. (2014). Predictive Analytics, Data Mining and Big Data: Myths, Misconceptions and Methods. Palgrave Macmillan.

Foust, D. and Pressman, A. (2008.), *"Credit Scores: Not-So-Magic Numbers,"* Business Week, February 7, 2008.

Frank, E., Hall, M., Trigg, L., Holmes, G., & Witten, I. H. (2004). Data mining in bioinformatics using Weka. Bioinformatics, 20(15), 2479-2481.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer, Berlin: Springer series in statistics.

Fuchs, C. A., and Peres, A. (1996). Quantum-state disturbance versus information gain: Uncertainty relations for quantum information. Physical Review A, 53(4), 2038.

Gattari, P. (2012). Role of Patent Law in Incentivizing Green Technology, The.*Nw. J. Tech. & Intell. Prop.*, *11*, vii.

Gupta, D. L., Malviya, A. K., & Singh, S. (2012). Performance analysis of classification tree learning algorithms. IJCA) International Journal of Computer Applications, 55(6).

Han and Kamber, Morgan Kaufmann Publishers is an imprint of Elsevier. 225Wyman Street,Waltham, MA 02451, USA 2012 by Elsevier Inc. All rights reserved.

Hayes, N., & Lemon, N. (1990). Stimulating positive cultures: the Concorde Informatics case. *Leadership & organization development journal*, *11*(7), 17-21.

Ho, T. K. (1995, August). Random decision forests. In Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on (Vol. 1, pp. 278-282). IEEE.

Ho, T. K. (1998). The random subspace method for constructing decision forests. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 20(8), 832-844.

http:// www.cs.waikato.ax.nz/ml/weka

http://www.saedsayad.com/decision_tree.htm (accessed Friday, October 09, 2015 09:41:36 AM)

https://www.cs.cmu.edu/~schneide/tut5/node42.html

https://www.cs.cmu.edu/~schneide/tut5/node42.html

IBM SPSS Modeler CRISP-DM Guide (2011), available online at https://the-modeling-agency.com/crisp-dm.pdf (Accessed Thursday, November 12, 2015, 8:20:24 AM)

Islam, M. Z., & Brankovic, L. (2003). Noise addition for protecting privacy in data mining. In Proceedings of The 6th Engineering Mathematics and Applications Conference (EMAC2003), Sydney (pp. 85-90).

Jiménez-Valverde, A. (2012). Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modeling. Global Ecology and Biogeography, 21(4), 498-507.

Kadiyala, S., S. & Srivastava (2011). Data Mining For Customer Relationship Management, International Business & Economics Research Journal, Volume1, Number 6

Ketel M. and Homaifar A. (2005), Privacy-Preserving Mining by Rotational Data Transformation, 43rd ACM Southeast Conference, March 18-20, 2005, Kennesaw, GA, USA. Copyright 2005 ACM 1-59593-059-0/05/0003 (Accessed online Wednesday, October 28, 2015)

Kiron, D., Shockley, R., Kruschwitz, N., Finch, G., and Haydock, M. (2012). Analytics: The Widening Divide. MIT Sloan Management Review, 53(2), 1-22.

Kohavi, R. & Provost, F. (1998). Glossary of Terms. *Machine Learning, 30*(2-3), 271-274.

Kvale, D. (1996). *Interviews.* London: SAGE Publications.

Laha, A.(2007). Building contextual classifiers by integrating fuzzy rule based classification technique and k-nn method for credit scoring" Advanced Engineering Informatics archive Volume 21 Issue 3, Pages 281-291, July, 2007

Leventhal, B. (2010). An introduction to data mining and other techniques for advanced analytics. Journal of Direct, Data and Digital Marketing Practice, 12(2), 137-153.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Linoff, G. S., & Berry, M. J. (2004). Data mining techniques: for marketing, sales, and customer relationship management. John Wiley & Sons.

Linoff, G. S., & Berry, M. J. (2011). Data mining techniques: for marketing, sales, and customer relationship management. John Wiley & Sons.

Montillo, A., A. "Random forests," *Guest lecture: Statistical Foundations of Data Analysis*. New York: Springer-Verlag, Apr. 2009.

Moro, S., Laureano, R., & Cortez, P. (2012). Enhancing bank direct marketing through data mining. In *Proceedings of the 41th European Marketing Academy Conference (EMAC)*. European Marketing Academy.

Neelamegam S., Ramaraj E. (2013), Classification algorithm in Data mining: An Overview, International Journal of P2P Network Trends and Technology (IJPTT) – Volume 4 Issue 8- Sep 2013, 639 – 374, available online at http://www.ijpttjournal.org (Accessed Wednesday, October 07, 2015, 12:34:58 PM)

Nikam s., s. (2015), A Comparative Study of Classification Techniques in Data Mining Algorithms, Oriental Journal of Computer Science & Technology, Vol. 8, No. (1), 13 – 19, available online www.computerscijournal.org (Accessed Monday, October 26, 2015, 1:31 PM)

Nisbet, R., Miner, G., & Elder IV, J. (2009). *Handbook of statistical analysis and data mining applications*. Academic Press.

O'brien j. A. (2003), "*Introduction To Information Systems: essentials for the e-business enterprise. Mcgraw-hill, inc.*". Pp 219 – 220

Oliveira S., R., M., Zaiane O., R. (2010), Privacy Preserving Clustering by Data Transformation, Journal of Information and Data Management, Vol. 1, No. 1, February 2010, Pages 37 - 51. (Accessed online Wednesday, October 28, 2015, 10:33:47 PM)

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. Journal of Artificial Intelligence Research, 169-198.

Patil, N., Lathi, R., & Chitre, V. (2012). Comparison of C5. 0 & CART classification algorithms using pruning technique. International Journal of Engineering Research and Technology (Vol. 1, No. 4 (June-2012)). ESRSA Publications.

Patil, S., Bhoite, D. S. (2015), Applying Data Mining Tool (Weka) to Study Consumer Acceptance of E-Banking, International Journal of Scientific Research, vol 4, 517 – 520 (Accessed Monday, September 28, 2015, 11 hours ago, 10:41:11 AM)

Perner P. and Fiss G. (2002), Intelligent E-Marketing with Web Mining, Personalization and User adapted Interfaces, Institute of Computer Vision and Applied Computer Science (Accessed Thursday, October 22, 2015, 8:30:26 PM)

Piatetsky-Shapiro G. (2016). KDnuggets Methodology Poll (http://www.kdnuggets.com/ polls/2002/methodology.htm) Accessed Sunday, September 18, 2016

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and systems magazine*, *6*(3), 21-45.

Polkinghorne, D. E. (2005). Language and meaning: Data Collection in qualitative research. Journal of Counselling Psychology, 52, 137 – 145.

Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

Quinlan, J.R.(1986). Induction of Decision Trees, Machine Learning, vol. 1, pp. 81-106, 1986.

Radhakrishnan B., Shineraj G., Anver Muhammed K.M. (2013), Application of Data Mining In Marketing, International Journal of Computer Science and Network, Volume 2, Issue 5, October 2013, 41 – 46, available online at www.ijcsn.org

Rajagopal S. (2011), Customer Data Clustering Using Data Mining Technique, International Journal of Database Management Systems ( IJDMS ) Vol.3, No.4, November 2011, 1 – 11, (Accessed Thursday, October 22, 2015, 8:24:16 PM)

Rajasekar S., Philominathan P., Chinnathambi V. (2013), Research Methodology, available online at rajasekar@cnld.bdu.ac.in (accessed Today, November 02, 2015, 12:39:08 AM)

Raju, P. S., Bai, V. R., Chaitanya, G. K.(2014), Data mining: Techniques for Enhancing Customer Relationship Management in Banking and Retail Industries, International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 2, Issue 1, January 2014. Available online at www.ijircce.com (Accessed Friday, September 04, 2015, 4:07:29 PM)

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. In Encyclopedia of database systems (pp. 532-538). Springer US.

Robinson S., C.(2015), The Good, the Bad, and the Ugly: Applying Rawlsian Ethics In Data Mining Marketing, Journal of Media Ethics, 30:19–30, 2015, (accessed Thursday, October 22, 2015, 10:12:47 PM)

Rokach L, Maimon O. Decision trees. The Data Mining and Knowledge Discovery Handbook. 2005; 165–192.York.

Sakshi and Khare S. (2015), A Comparative Analysis of Classification Techniques on Categorical Data in Data Mining, International Journal on Recent and Innovation Trends in Computing and Communication Volume: 3 Issue: 8 - August 2015, 5142 – 5147 available online at http://www.ijritcc.org (Accessed Wednesday, October 07, 2015, 10:39:44 AM)

Saxena, M. (2000). Vice-President (Retail Marketing). *HDFC Bank, Press Statement, Reported by Leena Baliga in India Express Newspaper on 9th Nov*.

Sazonau, V. (2012). Implementation and Evaluation of a Random Forest Machine Learning Algorithm. University of Manchenster, 9.

Shaw J. M., Subramaniam C., Tan G. W., and Welge E., M. (2001), Knowledge management and data mining for marketing, Elsevier Decision Support Systems 31 (2001), 127–137, available online at www.elsevier.comrlocaterdsw

(Accessed Thursday, October 22, 2015, 8:11:12 PM)

Shields, P. M., & Rangarajan, N. (2013). A playbook for research methods: Integrating conceptual frameworks and project management. New Forums Press.

Sing'oei, L., & Wang, J. (2013). Data Mining Framework for Direct Marketing: A Case Study of Bank Marketing. J. Comput. Sci. Issues, 10(2), 198-203.

Singh A., Kaur A., Kaur J., Singh R., Raheja S. (2015), Pattern Analysis On Banking Dataset, International Journal of Scientific & Technology Research Volume 4, Issue 06, June 2015, 252-256, available online at www.ijstr.org (accessed Tuesday, September 29, 2015, 12:00:20 PM)

Smirnov-M H. (2007), Data Mining and Marketing http://www.estard.com/data_mining_marketing/data_mining_campaign.asp (accessed Thursday, October 22, 2015, 8:02:19 PM)

Sousa, M. D. M., & Figueiredo, R. S. (2014). Credit analysis using data mining: application in the case of a credit union. JISTEM-Journal of Information Systems and Technology Management, 11(2), 379-396.

Sun, H., Huai, J., Liu, Y., & Buyya, R. (2008). RCT: A distributed tree for supporting efficient range and multi-attribute queries in grid computing. *Future Generation Computer Systems*, *24*(7), 631-643.

Sweeney, L. (2002). Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(05), 571-588.

Tibshirani, R. (1996). Bias, Variance, and Prediction Error for Classification Rules, Technical Report, Statistics Department, University of Toronto.

Van Gestel, T. & Baesens, B. (2008). *Credit Risk Management:Basic Concepts: Financial Risk Components, Rating Analysis, Models, Economic and Regulatory Capital: Basic Concepts: Financial Risk Components, Rating Analysis, Models, Economic and Regulatory*

*Capital*. Oxford University Press

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. Fam Med, 37(5), 360-363.

Voznika F., Viana L., Data Mining Classification Courses.Cs.Washington.Edu/Courses/ Csep521/07wi/.../**Leonardo_Fabricio**.Pdf (Accessed Tuesday, October 06, 2015, 11:35:45 AM)

Wang,Y., Wang,S. & Lai,S.S.(2005) *A New Fuzzy* Support Vector Machine to Evaluate Credit Risk, IEEE Trans. Fuzzy Systems, vol. 13, no. 6, pp. 820-831, Dec. 2005.

Weiss, R., S. (1994). Learning from strangers, the art and method of qualitative interview studies. New York, NY: Free Pr.

Weiss, S. M., & Indurkhya, N. (1994). Decision tree pruning: biased or optimal?. In AAAI (pp. 626-632).

Williams C. (2007), Research Methods, Journal of Business & Economic Research, Volume 5, Number 3, 65 – 73, (accessed online Wednesday, November 04, 2015, 3:01:11 AM) Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Wolpert, D. H., & Macready, W. G. (1999). An efficient method to estimate bagging's generalization error. *Machine Learning*, *35*(1), 41-55.

Zaïane O. R. (1999), CMPUT690 Principles of Knowledge Discovery in Databases, University of Alberta https://webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/notes/Chapter1/ (accessed Monday, October 05, 2015, 2:20PM)

Zamani S. and Mogaddam A. (2015), Natural Customer Ranking of Banks in Terms of Credit Risk by Using Data Mining: A Case Study: Branches of Mellat Bank of Iran, Journal UMP Social Sciences and Technology Management Vol. 3, Issue. 2,2015. (Accessed Monday, October 26, 2015, 8:35:45 AM)

Zhaohan Y. (2009). Optimization techniques in data mining with applications to biomedical and psychophysiological data sets. University of Iowa Iowa Research Online Available Online: http://ir.uiowa.edu/etd/274 (accessed Monday, October 12, 2015, 9:24:21 AM)

Zhu, W.,Zeng, N., and Wang, N. (2010). Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS® Implementations. NESUG proceedings: health care and life sciences, Baltimore, Maryland, 1-9.

# APPENDIX A

## COMPLETE CLASSIFICATION RULES FOR J48 DECISION TREES

### PRUNED J48 DECISION TREE CLASSIFICATION RULES

contact = third: y (341.0/19.0) contact

= first

| marital status = s

| | gender = m

| | | house = n

| | | | age = 20-30

| | | | | children <= 1

| | | | | | education = primary: n (2.0)

| | | | | | education = secondary: y (2.0)

| | | | | | education = tertiary: n (0.0)

| | | | | children > 1: y (4.0)

| | | | age = 31-40: y (5.0/1.0)

| | | | age = 41-50: y (0.0)

| | | | age = 51-60: n (5.0)

| | | | age = 61-70: y (0.0)

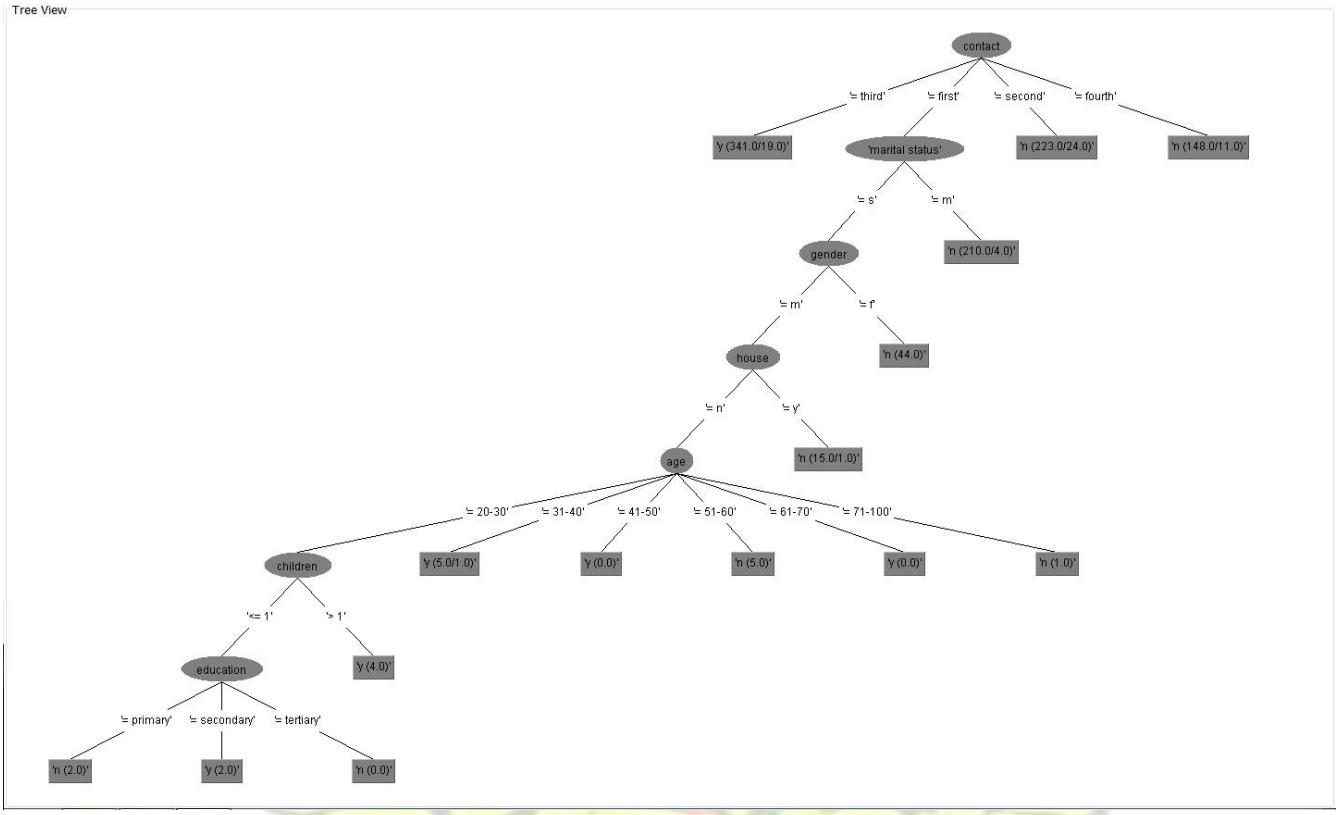
| | | | age = 71-100: n (1.0)

| | | house = y: n (15.0/1.0)

| | gender = f: n (44.0) | marital status = m: n (210.0/4.0) contact = second: n (223.0/24.0) contact = fourth: n (148.0/11.0)

**J48 PRUNED DECISION**

**TREE**




**UNPRUNED J48 DECISION TREE CLASSIFICATION RULES**

contact = third

| gender = m

| | education = primary

| | | age = 20-30: n (1.0)

| | | age = 31-40: y (2.0)

| | | age = 41-50: y (4.0)

| | | age = 51-60: y (2.0)

| | | age = 61-70: y (11.0)

| | | age = 71-100: y (0.0)

| | education = secondary

| | | age = 20-30: n (6.0/2.0)

| | | age = 31-40

| | | | occupation = trader |

| | | | marital status = s

| | | | | | income = low: n (3.0)

| | | | | | income = medium: n (7.0/3.0)

| | | | | | income = high: y (7.0/2.0) |

| | | | marital status = m: y (6.0)

| | | | occupation = carpenter: y (4.0)

| | | | occupation = driver

| | | | | income = low: n (0.0)

| | | | | income = medium: y (5.0/2.0)

| | | | | income = high: n (3.0/1.0)

| | | | occupation = tailor: y (2.0)

| | | | occupation = farmer: y (0.0)

| | | | occupation = pastor: y (0.0)

| | | | occupation = nurse: y (0.0)

| | | | occupation = mechanic: y (1.0)

| | | | occupation = teacher: y (0.0)

| | | | occupation = seamstress: y (0.0)

| | | age = 41-50: y (5.0)

| | | age = 51-60: y (4.0)

| | | age = 61-70: y (4.0)

| | | age = 71-100: y (2.0)

| | education = tertiary: y (51.0)

| gender = f

| | occupation = trader: y (55.0)

| | occupation = carpenter: y (0.0)

| | occupation = driver: y (0.0)

| | occupation = tailor: y (0.0)

| | occupation = farmer: y (0.0)

| | occupation = pastor

| | | income = low: y (2.0)

| | | income = medium: y (2.0)

| | | income = high: n (1.0)

| | occupation = nurse: y (76.0)

| | occupation = mechanic: y (0.0)

| | occupation = teacher: y (59.0) | |

occupation = seamstress: y (16.0)

contact = first

| marital status = s

| | gender = m

| | | house = n

| | | | age = 20-30

| | | | | children <= 1

| | | | | | education = primary: n (2.0)

| | | | | | education = secondary: y (2.0)

| | | | | | education = tertiary: n (0.0)

| | | | | children > 1: y (4.0)

| | | | age = 31-40: y (5.0/1.0)

| | | | age = 41-50: y (0.0)

| | | | age = 51-60: n (5.0)

| | | | age = 61-70: y (0.0)

| | | | age = 71-100: n (1.0)

| | | house = y: n (15.0/1.0)

| | gender = f: n (44.0)

| marital status = m

| | age = 20-30

| | | house = n: n (4.0)

| | | house = y: y (2.0)

| | age = 31-40: n (41.0/1.0)

| | age = 41-50: n (109.0)

| | age = 51-60: n (30.0/1.0)

| | age = 61-70: n (19.0)

| | age = 71-100: n (5.0)

contact = second |

gender = m

| | age = 20-30

| | | occupation = trader: n (0.0)

| | | occupation = carpenter: n (2.0)

| | | occupation = driver: y (3.0/1.0)

| | | occupation = tailor: y (8.0/3.0)

| | | occupation = farmer: n (2.0)

| | | occupation = pastor: n (0.0)

| | | occupation = nurse: n (0.0)

| | | occupation = mechanic: n (0.0)

| | | occupation = teacher: n (0.0)

| | | occupation = seamstress: n (0.0)

| | age = 31-40

| | | occupation = trader: n (3.0)

| | | occupation = carpenter: n (12.0) | | | occupation = driver: y (5.0/2.0)

| | | occupation = tailor: y (6.0/1.0)

| | | occupation = farmer: n (6.0/2.0)

| | | occupation = pastor: n (0.0)

| | | occupation = nurse: n (0.0)

| | | occupation = mechanic

| | | | income = low: n (11.0)

| | | | income = medium: y (2.0)

| | | | income = high: n (3.0)

| | | occupation = teacher: n (0.0)

| | | occupation = seamstress: n (0.0)

| | age = 41-50: n (50.0/1.0)

| | age = 51-60: n (12.0/1.0)

| | age = 61-70

| | | children <= 4: n (2.0)

| | | children > 4: y (3.0)

| | age = 71-100: n (3.0)

| gender = f: n (90.0)

contact = fourth | gender

= m

| | marital status = s

| | | income = low

| | | | occupation = trader: n (5.0/1.0)

| | | | occupation = carpenter: n (0.0)

| | | | occupation = driver

| | | | | children <= 1: n (2.0)

| | | | | children > 1: y (2.0)

| | | | occupation = tailor: y (1.0)

| | | | occupation = farmer: n (0.0)

| | | | occupation = pastor: n (0.0)

| | | | occupation = nurse: n (0.0)

| | | | occupation = mechanic: y (1.0)

| | | | occupation = teacher: n (0.0)

| | | | occupation = seamstress: n (0.0)

| | | income = medium: n (6.0)

| | | income = high: n (0.0)

| | marital status = m

| | | age = 20-30: n (2.0)

| | | age = 31-40

| | | | occupation = trader: n (1.0)

| | | | occupation = carpenter: n (11.0)

| | | | occupation = driver

| | | | | children <= 2: y (3.0)

| | | | | children > 2: n (6.0/1.0)

| | | | occupation = tailor: n (0.0)

| | | | occupation = farmer: y (1.0)

| | | | occupation = pastor: n (3.0)

| | | | occupation = nurse: n (0.0)

| | | | occupation = mechanic: n (5.0)

| | | | occupation = teacher: n (0.0)

| | | | occupation = seamstress: n (0.0)

| | | age = 41-50: n (24.0)

| | | age = 51-60: n (8.0)

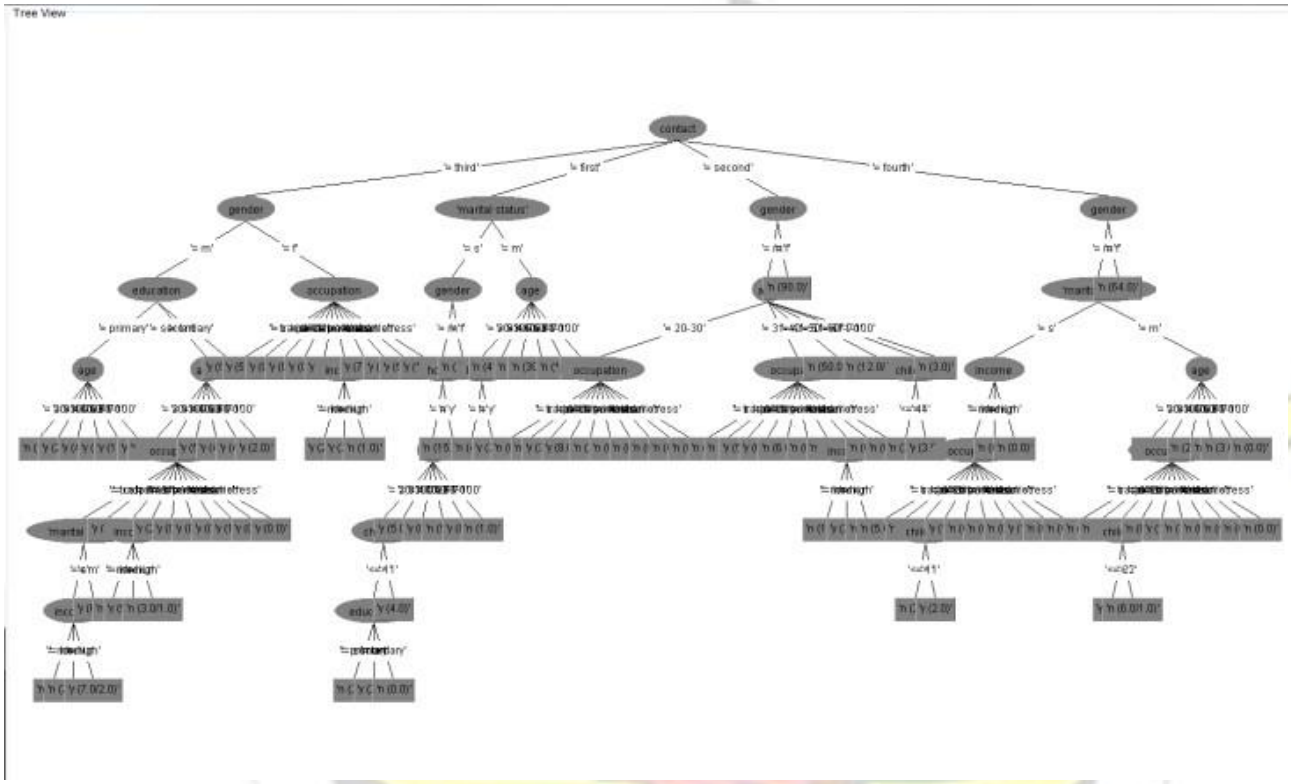
| | | age = 61-70: n (3.0/1.0)

| | | age = 71-100: n (0.0)

| gender = f: n (64.0)



**J48 UNPRUNED DECISION TREE**

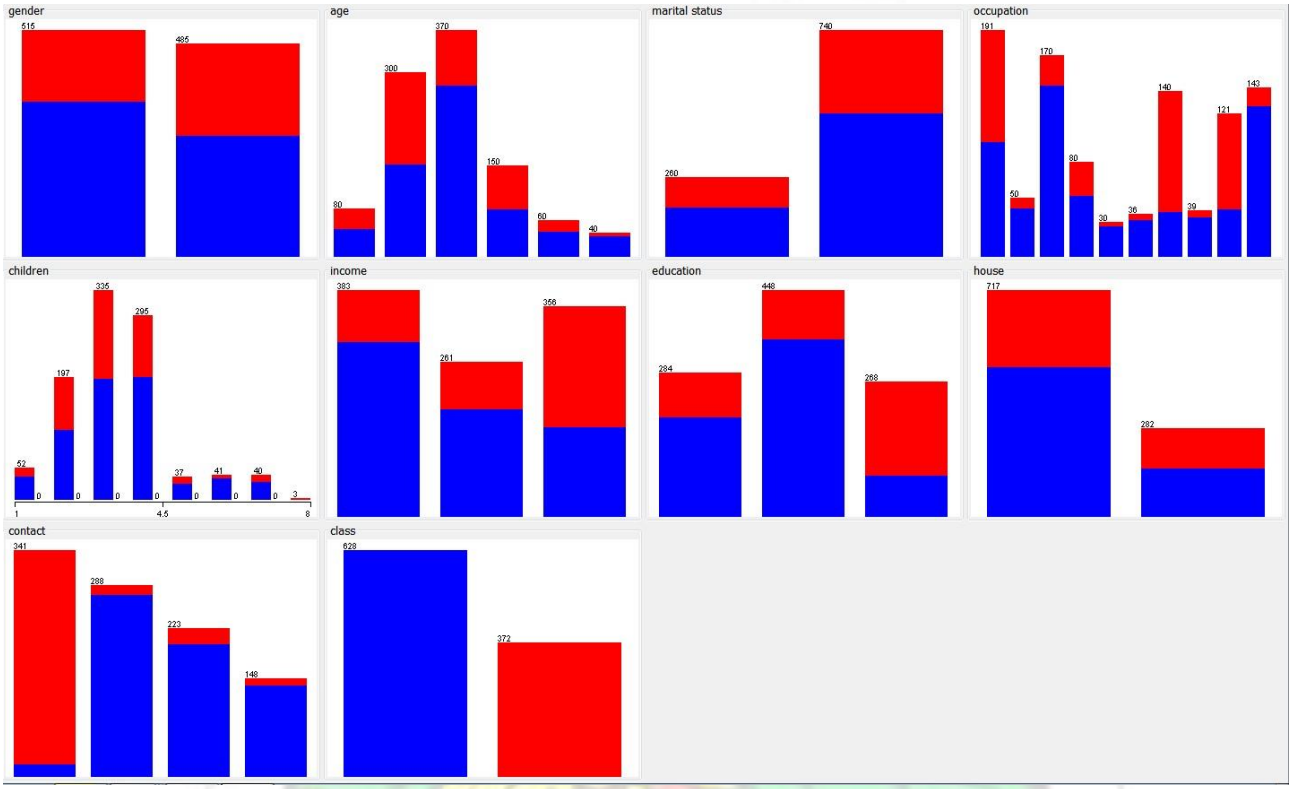


# APPENDIX B

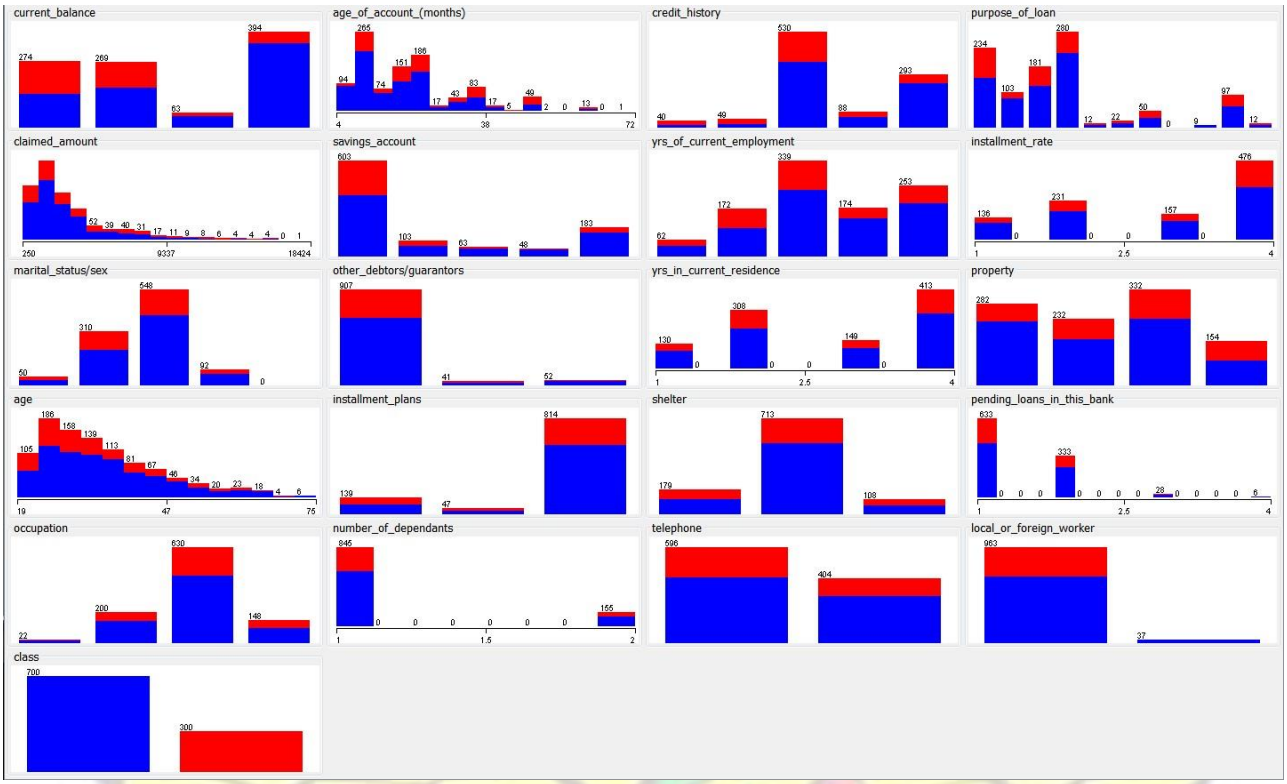
## COMPLETE VISUAL REPRESENTATION OF FINDINGS



## J48 DECISION TREE CLASSIFIER



## RANDOM FOREST CLASSIFIER




# APPENDIX C

## COMPLETE DATA DESCRIPTION FOR THE CREDIT RISK ANALYSIS MODEL

## DATASET

Description of the German credit dataset.

1. Title: German Credit data

2. Source Information



Professor Dr. Hans Hofmann
Institut f"ur Statistik und "Okonometrie
Universit"at Hamburg
FB Wirtschaftswissenschaften
Von-Melle-Park 5
2000 Hamburg 13

3. Number of Instances: 1000

Two datasets are provided. the original dataset, in the form provided by Prof. Hofmann, contains categorical/symbolic attributes and is in the file "german.data".

For algorithms that need numerical attributes, Strathclyde University produced the file "german.data-numeric". This file has been edited and several indicator variables added to make it suitable for algorithms which cannot cope with categorical variables. Several attributes that are ordered categorical (such as attribute 17) have been coded as integer. This was the form used by StatLog.

6. Number of Attributes german: 20 (7 numerical, 13 categorical)
Number of Attributes german.numer: 24 (24 numerical)

7. Attribute description for german

Attribute 1: (qualitative)
Status of existing checking account
A11 : ... < 0 DM
A12 : 0 <= ... < 200 DM
A13 : ... >= 200 DM /
salary assignments for at least 1 year
A14 : no checking account

Attribute 2: (numerical)
Duration in month

Attribute 3: (qualitative)
Credit history

A30 : no credits taken/
all credits paid back duly
A31 : all credits at this bank paid back duly
A32 : existing credits paid back duly till now
A33 : delay in paying off in the past
A34 : critical account/
other credits existing (not at this bank)

Attribute 4: (qualitative)
Purpose
A40 : car (new)
A41 : car (used)
A42 : furniture/equipment
A43 : radio/television
A44 : domestic appliances
A45 : repairs
A46 : education
A47 : (vacation - does not exist?)
A48 : retraining
A49 : business
A410 : others

Attribute 5: (numerical)
Credit amount

Attibute 6: (qualitative)
Savings account/bonds
A61 : ... < 100 DM
A62 : 100 <= ... < 500 DM
A63 : 500 <= ... < 1000 DM
A64 : .. >= 1000 DM
A65 : unknown/ no savings account

Attribute 7: (qualitative)
Present employment since
A71 : unemployed
A72 : ... < 1 year
A73 : 1 <= ... < 4 years
A74 : 4 <= ... < 7 years
A75 : .. >= 7 years
Attribute 8: (numerical)
Installment rate in percentage of disposable income

Attribute 9: (qualitative)
Personal status and sex
A91 : male : divorced/separated
A92 : female : divorced/separated/married
A93 : male : single
A94 : male : married/widowed
A95 : female : single

Attribute 10: (qualitative)
Other debtors / guarantors
A101 : none
A102 : co-applicant
A103 : guarantor

Attribute 11: (numerical)
Present residence since

Attribute 12: (qualitative)
Property
A121 : real estate
A122 : if not A121 : building society savings agreement/
life insurance
A123 : if not A121/A122 : car or other, not in attribute 6
A124 : unknown / no property

Attribute 13: (numerical)
Age in years

Attribute 14: (qualitative)
Other installment plans
A141 : bank
A142 : stores
A143 : none

Attribute 15: (qualitative)
Housing
A151 : rent
A152 : own
A153 : for free

Attribute 16: (numerical)
Number of existing credits at this bank
Attribute 17: (qualitative)

Job
A171 : unemployed/ unskilled - non-resident
A172 : unskilled - resident
A173 : skilled employee / official
A174 : management/ self-employed/
highly qualified employee/ officer

Attribute 18: (numerical)
Number of people being liable to provide maintenance for

Attribute 19: (qualitative)
Telephone
A191 : none
A192 : yes, registered under the customers name

Attribute 20: (qualitative)
foreign worker A201 : yes
A202 : no

8. Cost Matrix

This dataset requires use of a cost matrix (see below)

| | 1 | 2 |
|---|---|---|
| 1 | 0 | 1 |
| 2 | 5 | 0 |

(1 = Good, 2 = Bad)

the rows represent the actual classification and the columns the predicted classification.

It is worse to class a customer as good when they are bad (5), than it is to class a customer as bad when they are good (1).