Using Data Mining Techniques for Credit Risk Scoring

By: Adila Abdelraheem Mohamed Hamad
B.Sc in Computer Science, University of Khartoum, 2006
Faculty of Mathematical Sciences

A thesis submitted to the Khartoum University in partial fulfillment for the degree of Master of Computer Science

Supervisor: Dr. Mohammed Abdallah

Computer Science Department
Faculty of Mathematical Science
December -2013
ABSTRACT:

Data Mining is powerful, it has many uses, in the banking sector in Sudan there's a lack of application of data mining tools. In this thesis data mining techniques used in Credit risk scoring are reviewed and analyzed, a commercial Sudanese bank example has been studied and used to create a Credit risk scoring model 'with WEKA' using two functions, and identify which of the function is the most suitable to be used.
ACKNOWLEDGMENT

My gratitude is owed to my colleagues past and present for their help and support in producing this Thesis, my supervisor Dr. Mohamed Abdelrahman, and my family for their patience and support during my whole master's and the writing of this thesis and reviews whom valuable comments helped improve it.

Adila Abdelraheem Mohamed Hamad
# TABLE OF CONTENTS

Table of contents ...................................................... 1  
Figure List ...................................................................... 2  
Table List ........................................................................ 2  

1 Introduction  
1.1 Banking History .................................................... 3  
  1.1.1 History of Credit Scoring ...................................... 3  
1.2 Problem Statement ................................................ 4  
1.3 Objectives of the Study ............................................ 5  
1.4 Scope of Work ........................................................ 5  
1.5 Significance of the Study ......................................... 5  
1.6 Structure of the Thesis ........................................... 5  

2 Literature Review  
2.1 Data Mining 'DM' .................................................. 7  
2.2 Credit Risk Scoring ............................................... 8  
  2.2.1 Expert Systems ................................................. 9  
  2.2.2 Linear Discriminant Analysis "LDA" ..................... 10  
  2.2.3 Logistic Regression "LR" .................................... 11  
  2.2.4 Decision Trees "DT" .......................................... 11  
  2.2.5 Artificial Neural Networks (ANNs) ..................... 13  
  2.2.6 Support Vector Machines "SVM" ......................... 14  
  2.2.6 Genetic Programming or genetic algorithms ........ 15  
2.3 Literature Review summary ................................... 17  

3 Methodology  
3.1 Methodology ....................................................... 18  
3.2 Research Procedure .............................................. 19  
  3.2.1 Introduction/ problem statement ......................... 20  
  3.2.2 Literature review .............................................. 20  
  3.2.3 Study bank's risk scoring operation ................... 20  
  3.2.4 Preprocess and clean dataset ............................ 20  
  3.2.5 Convert excel to .csv ....................................... 20  
  3.2.6 Classify with Weka ......................................... 20  
  3.2.6.1 Classify with SVMs .................................... 21
CHAPTER 1
INTRODUCTION

This chapter introduces the current research with the background of the problem described first. After that the statement of the problem, objectives, scope and importance of the study are described respectively.

1.1 Banking History

Freixas & Rochet (2008) suggest in their book Microeconomics of banking, a simple operational definition of a bank which is: “A bank is an institution whose current operations consist in granting loans and receiving deposits from the public” [1].

In February 1960, the Bank of Sudan began acting as the central bank of Sudan, issuing currency, assisting the development of banks, providing loans, maintaining financial equilibrium, and advising the government.

There were originally five major commercial banks (Bank of Khartoum, An Nilein Bank, Sudan Commercial Bank, the People's Cooperative Bank, and the Unity Bank) but the number subsequently grew. The public was dissatisfied with the commercial banks, however, because they were reluctant to lend capital for longterm development projects. Since the government decreed the 1970 Nationalization of Banks Act, all domestic banks have been controlled by the Bank of Sudan.

In addition, the government established numerous specialized banks, such as the Agricultural Bank of Sudan (1959) to promote agricultural ventures, the Industrial Bank of Sudan (1961) to promote private industry, the Sudanese Estates Bank (1966) to provide housing loans, and the Sudanese Savings Bank established to make small loans particularly in the rural areas [18].

1.1.1 History of Credit Scoring

"David Durand (1941) was the first to recognize that one could use the statistics techniques to discriminate between good and bad loans, decisions on whether to give loans or send merchandise had been made judgmentally by credit analysts for many
years. However, these credit analysts were being drafted into military service and there was a severe shortage of people with this expertise. So the firms got the analysts to write down the rules of thumb they used to decide to whom to give loans, these rules were then used by non-experts to help make credit decisions which formed one of the first examples of expert systems.

The automation of the lending decision started after the arrival of credit cards in the 1960's. When the lending organizations used credit scoring they found that it also was a much better predictor than any judgmental scheme and default rates would drop by 50% or more.

In the 1980s the success of credit scoring meant that banks started using scoring for their other products like personal loans, while in the last few years scoring has been used for home loans and small business loans.

Advances in computing allowed other techniques to be tried to build scorecards. In the 1980s logistic regression and linear analysis were introduced. Recently, artificial intelligence techniques have been piloted [4].

1.2 Problem Statement

Credit risk scoring is one of the problems that face the banking sector, and it can be defined as the method used to classify borrowers into good or bad classes, in other words score credit users according to their historical data into different classes, using certain techniques.

In Sudanese banks there's no clear model used in scoring borrowers, which caused some banks to fail, in this research one of the Sudanese banks is under study to produce a model that is suitable to be used for two of the most popular funding formulas used in it.

Credit risk scoring is mainly a classification problem, which has been treated with different approaches, at the beginning there was the Expert systems which is decided by the expert's subjective opinion, and then it evolved to the use of traditional statistical approaches such as linear discriminant analysis, logistic regression, and decision trees, it
has also been processed through modern artificial intelligence techniques such as: support vector machines and artificial neural network and genetic programming.

1.3 Objectives of the Study

- Study different credit risk scoring approaches
- Define the merits and limitations and of each approach
- Choose an appropriate approach for credit scoring to be used in Sudanese banks
- To develop an automated credit scoring model that is suitable for Sudanese banks

1.4 Scope of Work

In this thesis a commercial bank’s data will be used to create and test credit risk model for two of its popular funding formulas "Murabaha & Musharaka" and that will be through refining data attributes and applying different models to the data sets and then choosing the most accurate to use.

1.5 Significance of the Study

Sudan banks today are offering more loan options to its clients and with no clear method or model to scoring the risk of each borrowing operation, which can cause catastrophic results to any bank.

Credit risk scoring offer a clear and straight forward way to classify each borrower according to their historical data, which will help achieve robust efficient banking services.

1.6 Structure of the Thesis

Chapter one includes an introduction to banking history in Sudan, loans history in general and credit risk scoring different approaches, the scope of work and the structure of the thesis.

Chapter two of this thesis will introduce data mining as a means to solve classification problems, then will cover the different approaches used in credit risk scoring, and define their strengths and limitations.
Chapter three of this thesis will present methodology to be adopted in continuing this research. Research procedures, operational framework, at the end a case study of the "bank risk analysis department" is explained along with the different financing formulas used in Sudanese banks.

Also in chapter three the use of WEKA is illustrated as it will be used as the analysis tool to provide the classification needed in the credit scoring model, the results, screen shots, and discussion.

Chapter four concludes this thesis and provides recommendations for further works.
CHAPTER 2

LITERATURE REVIEW

This chapter describes the preliminary concepts and presents current approaches for credit risk scoring. The chapter begins with the definition and some background of data mining in prediction, as well as, the explanation of credit risk and each of its approaches strengths and limitations.

2.1 Data Mining 'DM'

Knowledge is the most valuable asset, it helps in taking good decisions and actions, prevent catastrophic losses and keep businesses ahead. Data on the other hand is the base for knowledge, from analysing data one can extract useful information and discover interesting patterns that can help in making different important decisions.

The vast amount of data and the need for powerful data analysis tools has been described as a "data rich but information poor situation" [8].

What is Data mining? "DM 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" [8]

Data mining is the concept of answering questions that you’ve not asked yet, by reaching deep into databases. The tasks performed can be classified into two categories: descriptive and predictive data mining; descriptive data mining provides information to understand what is happening inside the data without a predetermined idea, on the other hand, predictive data mining allows the user to submit records with unknown field values, and the system will guess the unknown values based on previous patterns discovered form the database, the DM models can be categorized according to the tasks they perform: Classification and Prediction, Clustering, Association Rules Classification and prediction is a predictive model, but clustering and association rules are descriptive models [16].
The type of DM task to use depends on the problem to be solved, and on the type of patterns and data to be mined. In addition, depending on the specific problem, techniques from various fields can be used, including machine learning, artificial intelligence, natural language processing, pattern recognition and visualization [3].

To build predictive models that predict trends or behaviours based on the analysis of historical and structured data, one needs to construct and use a model to assess the class of an unlabeled example, or to assess the value or value ranges of an attribute that a given example is likely to have [3, 16].

Classification is one of the major types of prediction problems, classification is finding a pattern that describes the group to which an item belongs. That is, the model is derived based on the analysis of a set of training data (i.e., data objects for which the class labels are known). The model is used to predict the class label of objects for which the class label is unknown; classification is used to predict categorical (discrete, unordered) labels. To represent the derived model various forms can be used like classification rules “if-then”, decision trees, mathematical formulae and neural networks [3, 8, 16].

Classification may need to be preceded by relevance analysis, which attempts to identify attributes that are significantly relevant to the classification process. Such attributes will be selected for the classification process. Other attributes, which are irrelevant, can then be excluded from consideration.

One of the fundamentals for credit scoring is selecting appropriate and more predictive variables, variable selection is the process of selecting the best predictive subset of variables from the original set of variables in a dataset [8].

2.2 Credit Risk Scoring

Risk is defined as an element or factor indicating a certain danger that will affect the ability to reach one objective, in financial terms, risk is an unknown component of the future value of a financial asset. Banking risks can be split in three main categories: credit risk, interest rate risk, and liquidity risk. Each type of risk is reported in the bank's balance sheet such as:
- Credit risk: Credit risk occurs when a borrower is unable to pay back the loan he has been granted.

- Liquidity risk: Liquidity risk occurs when a bank is facing unexpected cash withdrawals on deposit accounts. A massive number of withdrawals would lead to a liquidity shortage for the bank.

- Interest rate risk: Interest rate risk occurs when maturity transformation happens. For instance, deposit is a short term interest product whereas loans are long term interest products.

The liquidity risk and interest rate risk is out of scope in this thesis.

In order to minimize credit risk, financial institutions must establish a sound risk management of retail and corporate lending, financial institutions can minimize credit risk in five different ways: through accurate loans pricing, credit rationing, use of collateral, loan diversification and more recently through asset securitization and the use of credit derivatives.

In retail banking, credit risk can be assessed in two different ways, by using qualitative or quantitative methods. Ideally, the bank relies on qualitative methods to decide whether to grant a loan or not. The most well-known and also recognized quantitative method is credit scoring, which is the main focus of this thesis [1].

Credit risk scoring is mainly a classification problem, which has been treated through different approaches; at the beginning there was the Expert systems which is decided by the expert's subjective opinion, and then it evolved to the use of traditional statistical approaches such as linear discriminant analysis, logistic regression, and decision trees, also it has been processed through modern artificial intelligence techniques such as support vector machines and artificial neural network and genetic programming, each of this approaches will be described in details.

2.2.1 Expert Systems

In Expert system the decision is left to the local officer or relationship manager expertise, using their subjective judgment and weighting of certain key factors. One of the most common expert systems in the five C's of credit:
I. Character: reputation & willingness to repay and repayment history; as the character of the borrower and his reputation in society and in market defines his ability to repay.

II. Capital: how much is been asked for, is the amount to be borrowed is sensible and matches the borrowers business size.

III. Capacity: the ability to repay, how much free income do they have?, is the size of business is significant in comparison to the loan.

IV. Collateral: if a default occurs the bank claims a collateral pledge by borrower, the greater the point of this claim and higher the market value the lower the exposure risk of the loan.

V. Cycle (or economic conditions): the state of the business cycle, what are the conditions in the market? Depending on the borrowers' business type and the current conditions of its relevant market defines their ability to repay the loan.

Problems: consistency: what are the important common factors to analyze across different types of borrowers?, subjectivity: what are the optimal weights apply to the factors chosen(result in different standard applied by credit officers) [2,4].

2.2.2 Linear Discriminant Analysis "LDA"

In LDA Customers are divided into two (or more) groups that are supposed to have different means, such groups are described based on past data "ones who defaulted and ones who did not", so that new data can be assigned to one of them. Discriminant Analysis is a descriptive tool: no model for the process leading to default is present, the assumption of an equal variance/covariance matrix across groups is not necessary, although it helps keeping things simple, probability estimates require “multi-normality”, which is seldom supported by empirical data.[1]

Z-score model is a linear probability model example that can explain how the scoring models work; the best fitting scoring model for commercial loans took this form:

\[ Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1.0X_5 \]

Where:

\[ X_1 = \text{working capital/total assets ratio;} \]
\( X_2 = \text{retained earnings/total assets ratio}; \)

\( X_3 = \text{earnings before interest and taxes/total assets ratio}; \)

\( X_4 = \text{market value of equity/book value of total liabilities ratio}; \)

\( X_5 = \text{sales/total assets ratio}. \)

The resulting a Z score below a critical value they would be classified as "bad" and the loan would be refused [2].

The LDA is very easy to implement and interpret, and outperform neural networks with large data samples, but the problem with the linear discriminant analysis is it requires strong statistical assumptions that could be unavailable and unrealistic; the logistic regression solves that with equally acceptable results [1, 9].

1.2.3 Logistic Regression "LR"

This type of modelling approach is based on the concept that each single attribute should be tested before inclusion in the model. Logistic regression (LR) can be spitted in several categories: binomial logistic regression, multinomial logistic regression, ordinal logistic regression, and so on.

The general formula is as follows:

\[
\text{Logit}(p_i) = B_0 + B_1 X_1 + \ldots + B_k X_k
\]

Where:

\( p = \text{probability of default based on the characteristics given} \)

\( X_1, \ldots, X_k = \text{Independent variables} \)

\( B_0 = \text{Constant} \)

\( B_1, \ldots, B_k = \text{Coefficients} \)

\[
\text{Logit} = \log \left( \frac{p(\text{default})}{p(\text{non-default})} \right) \]

LR doesn't require statistical assumptions like LDA, it also achieved better accuracy than most other models, a difficulty with logistic regression was that one has to use maximum likelihood to estimate the independent variables \( X_i \), also the independent variables must be linearly related to the logit of the dependent variable [4,9].

2.2.4 Decision Trees "DT"
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 [8].

Decision tree induction algorithms have been used for classification in many application areas such as medicine, manufacturing and production, financial analysis, astronomy, and molecular biology [8].

ID3, C4.5, and CART are algorithms that adopt a greedy non-backtracking approach in which decision trees are constructed in a top-down recursive divide-and-conquer manner, which starts with a training set of tuples and their associated class labels, the training set is recursively partitioned into smaller subsets as the tree is being built [8].

The C4.5 is one of the algorithm used in the credit risk scoring problem, decision tree classifiers have good accuracy; however, successful use may depend on the data at hand, are easy to understand and doesn't require any domain knowledge and applicable whatever the nature of response and explanatory variables, the disadvantages of DT: greater demand for computational resources and structure of DT depends on the observed data; a small change alters the structure of tree, and the accuracy of DTs is not stable and is easily affected by noisy data and the redundancy of the data attributes [9].

For conventional statistical classification techniques, an underlying probability model must be assumed in order to calculate the posterior probability upon which the classification decision is made, the more recently developed data mining techniques has three models based on artificial intelligence techniques that have been applied since the beginning 90's. These include neural networks (BP), Support vector machines (SVM) and genetic programming (GP) [15].

These models can perform the classification task without the limitations with the statistical traditional approaches, also the artificial intelligence methods also achieved better performance than traditional statistical methods, since they are highly capable of extracting meaning from imprecise data and detecting trends that are too complex to be discovered by either humans or other conventional techniques [5, 13].
2.2.4 Artificial Neural Networks (ANNs)

Are mathematical representations, inspired by the neuron node functioning of the human brain. They are composed by a number of simple processors (neurons) working in parallel, without any centralized control. The neurons are arranged in a particular structure which is usually organized in layers. ANNs can be used to simulate the non-linear relationship in complicated data through a system of weighted connections that determines the information flow through the network [7, 15].

The input of each neuron is the weighted sum of the network inputs, and the output of the neuron is a sigmoidal function value based on its inputs. This network is trained by supervised learning, through an iterative process on a sample of a known class, in credit risk scoring problem feedforward or multilayer perceptron networks with the backpropagation algorithm are used [8].

ANNs are used broadly in classification, clustering and optimization problems, as they are able to recognize the complex and non-linear patterns between input and output variables, in credit scoring which they can predict the credit worthiness of a new applicants and classify current borrowers into good or bad classes. [6,10].

They have four major components: inputs, weights, neurons and outputs, nodes shown in Figure 2.1.

Figure 2.1
Neural network models are more accurate, adaptive and robust, compared to linear discriminant analysis, logistic regression and decision trees [5, 6].

Major Limitations of using neural networks is their lack of transparency; the internal structure of the network is hidden "black box nature" which leads to a lack of accountability, and it doesn't reveal anything about the intermediate steps that lead to the final output, also there is an over learning risk, that is: the networks learns the sample but looses the ability to generalize, For this reason a portion of the sample is kept for testing purposes "to make sure the same input always have the returns the same output" [2, 8,15].

2.2.5 Support Vector Machines "SVM"

Support Vector Machines is the state-of-the-art technology based on statistical learning, it's a method for the classification of both linear and nonlinear data, SVM were first suggested by Vapnik in 1992 and have recently been used in a range of problems including pattern recognition, bioinformatics, and text categorization [5, 8].

SVMs are a set of data-driven, supervised learning methods that do not require specific assumptions on the underlying data generating process [14].

SVM is an algorithm that works as follows: It uses a nonlinear mapping to transform the original training data into a higher dimension. Within this new dimension, it searches for the linear optimal separating hyperplane (a "decision boundary" separating the tuples of one class from another). With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyperplane figure 2. The SVM finds this hyperplane using support vectors ("essential" training tuples) and margins (defined by the support vectors) [8], in the case of credit scoring, SVM is used to classify the applicants usually based on non-linear input variables [9].

The complexity of the learned classifier is characterized by the number of support vectors "no. of training tuples" rather than the dimensionality of the data. Hence, SVMs tend to
be less prone to overfitting than some other methods, they lie closest to the decision boundary (Maximum Margin Hyperplane MMH) [8].

Credit scoring models based on SVM, using parameters selection algorithms, have good accuracy and consume less computational time compared with traditional classification methods, and their classification accuracy is very similar to neural networks [5,14].

Figure 2.2

2.2.6 Genetic Programming or genetic algorithms

GP is based on the Darwinian principle of natural selection and evolution. GP makes use of the idea of survival of the fittest by progressively accepting better solutions to the problem. It is inspired by biological processes of inheritance, mutation, natural selection, and the genetic crossover that occurs when parents mate to produce offspring [15].

A Genetic Algorithm (GA) is a data driven, non-parametric heuristic search process, where the training algorithm can be chosen to optimize a wide range of objective functions. Because the training algorithm is guided only by the performance of
competing solutions, GAs have the potential to generate models that outperform other approaches [17].

Genetic learning starts as follows. An initial population is created consisting of randomly generated rules. Each rule can be represented by a string of bits. As a simple example, suppose that samples in a given training set are described by two Boolean attributes, $A_1$ and $A_2$, and that there are two classes, $C_1$ and $C_2$. The rule “if $A_1$ and not $A_2$ then $C_2$” can be encoded as the bit string “100,” where the two leftmost bits represent attributes $A_1$ and $A_2$, respectively, and the rightmost bit represents the class. Similarly, the rule “if not $A_1$ and not $A_2$ then $C_1$” can be encoded as “001.” If an attribute has $k$ values, where $k > 2$, then $k$ bits may be used to encode the attribute’s values. Classes can be encoded in a similar fashion [8].

The goal of the GA is to combine together and mutate different solutions so that over time fitter (better) solutions evolve, new populations are created over a number of generations (iterations) through the application of the genetic operators: selection, cross over and mutation, in selection: From the existing population, a number of strings are selected for breeding, with selection favouring those strings that represent the best solutions found to date. In crossover, substrings from pairs of rules are swapped to form new pairs of rules. In mutation, randomly selected bits in a rule’s string are inverted, the process of generating new populations based on prior populations of rules continues until a population, $P$, evolves where each rule in $P$ satisfies a pre-specified fitness threshold [8, 17].

One of the advantages of genetic programming is that it works better in large collection of data, speed and efficiency increase is parallel to data size [11].

A generalization, besides the approaches described above, there are two types of classification techniques used in credit scoring and they can be divided into:

I. single classifiers:
   Like decision trees and support vector machines.

II. hybrid systems:
Are systems which combine more than one good approach to overcome their weaknesses to some extent [10].

2.3 Literature Review summary

This chapter discussed the data mining role and its use in solving many problems, one of these problems is credit risk scoring, then credit risk scoring and its different approaches are discussed in detail.
CHAPTER 3

METHODOLOGY

This chapter presents methodology to be adopted in continuing this research. Research procedures, operational framework, assumptions and limitations and research schedule are included.

3.1 Methodology:

The Waikato Environment for Knowledge Analysis (WEKA), is a popular machine learning and data mining tool developed by the university of Waikato, and is available free under the GNU general public license [18]. WEKA is used in this thesis to classify the data of current borrowers into good or bad loans. Banks data is saved in an excel file that holds all the borrowers detail for their current loans. The data will be cleaned and only the attributes needed in classifying them to produce a credit risk model will be used, the file will be converted from the excel "xlsx" file into a comma separated value file "csv" that can be used with WEKA.

Only two types of loans will be modeled: Murabaha and partnership, as they are the most commonly used, the classifying techniques used are the most popular [3,5,12]: logistic regression and SVM, the results of both classifiers are compared to decide which is the most suitable to use in commercial Sudanese banks.
3.2 Research Procedure

Start

- Introduction/problem statement
  - Chapter 1

- Literature review
- Study bank's risk scoring operation
  - Chapter 2

- Preprocess and clean dataset
  - Chapter 3

- Convert excel to .csv

- Classify with Weka

- Classify with SVM
- Classify using ANNs

- Compare results

- Discrimination of the dissertation results

End
  - Chapter 4
3.2.1 Introduction/ problem statement:

An introduction to banking to banking and loans history, the credit risk scoring problem in Sudanese banks are discussed in chapter one of this thesis.

3.2.2 Literature review:

The literature review covering the concept of data mining and its use in classification, credit risk scoring as a classification problem and the different approaches used in modeling, varying from traditional statistical models into AI based models; are elaborated in chapter 2.

3.2.3 Study bank's risk scoring operation:

The case study adopted by this dissertation is based on "bank's" risk analysis unit operations, is also explained in chapter 2 along with the different funding formulas that are popular and used in Sudanese banks.

3.2.4 Preprocess and clean dataset.

The data set used is an excel file with attributes written in Arabic, and it contains data covering all types of loans formula, the cleaning and preprocessing needed was to extract the data used in Murabaha and partnership, translation attribute name, and fill null values in the data.

3.2.5 Convert excel to .csv:

The data is now in .xlsx file type, conversion to .csv is done by saving as the new file type, a straight forward operation, or it can be done online.

3.2.6 Classify with Weka:

Choosing the WEKA explorer, open the .csv file and list all the attributes in the preprocess table, choose the attribute that will be used in this case:
## Table 3.1: Preprocess attributes

<table>
<thead>
<tr>
<th>Attribute #</th>
<th>Attribute name</th>
<th>Attribute type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Funding formula</td>
<td>Numerical</td>
</tr>
<tr>
<td>2</td>
<td>Funding amount</td>
<td>Numerical</td>
</tr>
<tr>
<td>3</td>
<td>Used amount</td>
<td>Numerical</td>
</tr>
<tr>
<td>4</td>
<td>Profit margin</td>
<td>Numerical</td>
</tr>
<tr>
<td>5</td>
<td>Current balance</td>
<td>Numerical</td>
</tr>
<tr>
<td>6</td>
<td>Total balance</td>
<td>Numerical</td>
</tr>
<tr>
<td>7</td>
<td>Value of collateral</td>
<td>Numerical</td>
</tr>
<tr>
<td>8</td>
<td>duration of loan in months</td>
<td>Numerical</td>
</tr>
<tr>
<td>9</td>
<td>Class</td>
<td>Nominal</td>
</tr>
</tbody>
</table>

In both of the funding formulas the same model "both have the same attributes and date" will be built using both classification techniques, in the classification tab.

The date format in the data is not compatible to WEKA, to solve this problem a new column is added to the date that counts the number of months between the issue date and due date to have the duration of the loan in months, now all the needed attributes are in numerical format which is well accepted by WEKA.

The last attribute "class" is the key of the classifying, it defines the relevant type of each credit by applying the if statement. "if( total balance is greater than funding amount then good, otherwise its bad )"

### 3.2.6.1 Classify with SVMs:

Under the classify tab choose the classifier under function name SMO which refers to the Support vector machines, button choice is set to: cross-folds =10, and the classifying attribute being "Class".
Figure 3.1

Time taken to build model: 14.26 seconds

--- Stratified cross-validation ---
--- Summary ---
Correctly Classified Instances 1483 98.2119%
Incorrectly Classified Instances 27 1.7881%
Kappa statistic 0.9275
Mean absolute error 0.0179
Root mean squared error 0.1337
Relative absolute error 6.9394%
Root relative squared error 37.25%
Total Number of Instances 1510
Ignored Class Unknown Instances 5

--- Detailed Accuracy By Class ---

<table>
<thead>
<tr>
<th>Class</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>ROC Area</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>good</td>
<td>0.990</td>
<td>0.109</td>
<td>0.981</td>
<td>0.998</td>
<td>0.99</td>
<td>0.935</td>
<td>good</td>
</tr>
<tr>
<td>bad</td>
<td>0.891</td>
<td>0.002</td>
<td>0.95</td>
<td>0.931</td>
<td>0.938</td>
<td>0.945</td>
<td>bad</td>
</tr>
</tbody>
</table>

Weighted Avg. 0.982 0.093 0.982 0.982 0.982 0.937

--- Confusion Matrix ---

a  b <-- classified as
1279 2 | a = good
25 204 | b = bad

Figure 3.1 shows:

- Time taken to build SMO model = 14.26 seconds
- the SMO classification accuracy: 98.2119%
- confusion Matrix: first row means: 1279 records were correctly classified as good loans, and 2 were incorrectly classified as bad loans; the second row means only 25 records were incorrectly classified as good loans, and 204 were correctly classified as bad loans.
3.2.6.2 Classify with logistic regression:

Under the classify tab choose the classifier under function name logistic, which is the logistic regression, also the button choice is set to: cross-folds = 10, and the classifying attribute being "Class".

Figure 3.2

<table>
<thead>
<tr>
<th>Time taken to build model: 197.47 seconds</th>
</tr>
</thead>
<tbody>
<tr>
<td>--- Stratified cross-validation ---</td>
</tr>
<tr>
<td>--- Summary ---</td>
</tr>
<tr>
<td>Correctly Classified Instances 1369 90.6623 %</td>
</tr>
<tr>
<td>Incorrectly Classified Instances 141 9.3377 %</td>
</tr>
<tr>
<td>Kappa statistic 0.6312</td>
</tr>
<tr>
<td>Mean absolute error 0.0928</td>
</tr>
<tr>
<td>Root mean squared error 0.3015</td>
</tr>
<tr>
<td>Relative absolute error 36.0261 %</td>
</tr>
<tr>
<td>Root relative squared error 84.0504 %</td>
</tr>
<tr>
<td>Total Number of Instances 1510</td>
</tr>
<tr>
<td>Ignored Class Unknown Instances 5</td>
</tr>
</tbody>
</table>

--- Detailed Accuracy By Class ---

<table>
<thead>
<tr>
<th>Class</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>ROC Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.946</td>
<td>0.326</td>
<td>0.942</td>
<td>0.946</td>
<td>0.945</td>
<td>0.656</td>
</tr>
<tr>
<td>b</td>
<td>0.672</td>
<td>0.052</td>
<td>0.7</td>
<td>0.672</td>
<td>0.686</td>
<td>0.561</td>
</tr>
</tbody>
</table>

Weighted Avg. 0.907 0.286 0.905 0.907 0.906 0.857

--- Confusion Matrix ---

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th></th>
<th>---- classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td>1215</td>
<td>66</td>
<td></td>
<td>a = good</td>
</tr>
<tr>
<td>75</td>
<td>154</td>
<td></td>
<td>b = bad</td>
</tr>
</tbody>
</table>

Figure 3.2 shows:

- Time taken to build Logistic model = 197.47 seconds.
- the Logistic classification accuracy: 90.6623%.
- confusion Matrix first row means: 1215 records were correctly classified as good loans, and 66 were incorrectly classified as bad loans, the second row means: only 75 records were incorrectly classified as good loans, and 154 were correctly classified as bad loans.
3.2.7 Compare results:

Looking at the previous two figures 3.1, 3.2, the results clearly show that the SVM produced, faster building time, more accurate results, while logistic regression, while the margin between the two results is big, the SVM confusion matrix showed more accurate results in discriminating between good and bad loans.

3.2.8 Discrimination and dissertation results:

The Support Vector Machines SVM show more accurate results which means that its preferred to be used in the Sudanese banks.

3.3 Case Study

The bank presents multiple funding formulas to its clients, they all in the form provisions; with different format and names, the client is required to meet each format conditions to be eligible for funding.

There are two types of clients: individual clients and institution or cooperation clients, both types of clients must present a number of documents to proceed with the funding request "application" to the investment office at the nearest bank's branch, after the application documents are completed, they are sent to the bank's head quarters risk analysis unit, where they will be analyzed and reviewed.

In the case of individual clients the risk analysis unit analyzes every application through:

I. 6 month bank statement prior to application date of client's active account, and the analysis of every transaction [deposits, withdrawals...].

II. Collateral provided by the client that must be sufficient to cover the finance cost.

III. The down payment by the client must be more than the minimum required by the transaction, (e.g. 40% in the case of small cars).

IV. Making sure that the client doesn't have outstanding internal commitments to the bank.

V. Concentration: making sure that the client fund doesn't exceed half of the bank's capital.

VI. Default: making sure that the client is not a defaulter with its previous commitments to the bank.
VII. Making sure that the client doesn't have outstanding commitments with other banks through the central bank governing body.

In the case of institution or cooperation clients; the risk analysis unit analyzes the financial position of cooperation's budget for three consecutive years that's legally audited and stamped from the tax bureau.

After the risk analysis is done, the analysis result is sent to the investment unit at the bank's headquarters where they make the decision based upon it, then the feedback is returned to their relevant branch, then the funding operations are executed accordingly.

a. Bank financing formulas applied by Islamic banks vary in their term and conditions, here's a brief definition for the most popular formulas used by Sudanese banks [13]:

II. Murabaha: one of the religiously approved sales, that sells the product a customer "client" want to buy through the bank "e.g. a car" with extra profit added to the original price, under a number of conditions "active account, sufficient collateral”.

III. Alsallam: is a type of selling that pays the price for a certain product that will be delivered later "e.g. crops not yet harvested".

1. Musharaka: funding consists of a bank becoming a partner in a clients business or investment with a share in the capital and known percentage of the profit "under clear conditions".

2. Mudaraba: is a contract between two parties, requires that one party "in this case the bank”, fund the other part in investing it in the market, with fixed percentage for each in the profit under certain conditions, and the loss is on the bank.

3. Ijara: is a rent contract for a certain product, with a relaxed margin for giving?, for a specified fee and time.

4. Istisnaa: is a contract with a factory on manufacturing a specific product that a bank will purchase.

Table 3.2: shows the data attribute included in the excel file that contain current borrowers and their loan description.

<table>
<thead>
<tr>
<th>Attribute #</th>
<th>Attribute name</th>
<th>Attribute type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Description</td>
<td>Type</td>
</tr>
<tr>
<td>---</td>
<td>--------------------------</td>
<td>------------</td>
</tr>
<tr>
<td>1</td>
<td>Branch code</td>
<td>Numerical</td>
</tr>
<tr>
<td>2</td>
<td>Funding sector</td>
<td>Categorical</td>
</tr>
<tr>
<td>3</td>
<td>Funding formula</td>
<td>Numerical</td>
</tr>
<tr>
<td>4</td>
<td>Funding amount</td>
<td>Numerical</td>
</tr>
<tr>
<td>5</td>
<td>Used amount</td>
<td>Numerical</td>
</tr>
<tr>
<td>6</td>
<td>Issue date</td>
<td>Date</td>
</tr>
<tr>
<td>7</td>
<td>Profit margin</td>
<td>Numerical</td>
</tr>
<tr>
<td>8</td>
<td>Due date</td>
<td>Date</td>
</tr>
<tr>
<td>9</td>
<td>Current balance</td>
<td>Numerical</td>
</tr>
<tr>
<td>10</td>
<td>Total balance</td>
<td>Numerical</td>
</tr>
<tr>
<td>11</td>
<td>Type of collateral</td>
<td>Categorical</td>
</tr>
<tr>
<td>12</td>
<td>Value of collateral</td>
<td>Numerical</td>
</tr>
<tr>
<td>13</td>
<td>Currency type</td>
<td>Numerical</td>
</tr>
</tbody>
</table>
4.1 Conclusion

In this thesis different approaches to credit risk scoring are reviewed and analyzed, the case of a commercial Sudanese bank has been studied and its financing data were used in building a Credit risk scoring model, to build the model WEKA data mining tool has been used, a couple of classification functions SMO and Logistic were applied to the same preprocessed data to compare the accuracy of each in classifying the financing data, the result of applying the two functions clearly identified SMO 'Support Vector Machines' as the better functions in terms of accuracy and time to build model.

The result of the comparison proves that new AI methods are far better in classification than traditional methods, as Support Vector machines have the upper hand in this particular case of a commercial bank.
Reference

[1] IMPLEMENTING A CREDIT RISK MANAGEMENT SYSTEM BASED ON INNOVATIVE SCORING TECHNIQUES, by Linda Delaire A thesis submitted to The University of Birmingham for the degree of DOCTOR OF PHILOSOPHY University of Commerce and Social Science, Department of Economics, The University of Birmingham, February 2012


[8] Data Mining Concepts and Techniques, Third Edition, Jiawei Han, University of Illinois at Urbana–Champaign, Micheline Kamber, Jian Pei, Simon Fraser University

[9] IEEE Credit Scoring Using Data Mining Techniques with Particular Reference to Sudanese Banks, Eiman Kambal, Izzeldein Osman, Methag Taha, Noon Mohammed, Sara Mohammed, University of Khartoum, Khartoum, Sudan

[10] Journal of AI and Data mining, Credit scoring in banks and financial institutions via data mining techniques: A literature review, S. M. Sadatrasoul1*,
M.R. Gholamian1, M. Siamal, Z. Hajmoohammadi2, 1. Department of Industrial engineering, Iran University of Science and technology, Tehran, Iran, 2. Department of Computer Science, Amirkabir University of technology, Tehran, Iran, Received 22 November 2012; accepted 11 March 2013
[11] Evaluation of Data Mining Classification Techniques and Performances to Banking Customers Credit Scoring, 5th symposium on advances in science and technology, SASTech, Iran, Mashhad May 2011
[12] Journal of the Operational Research Society (2013) 64, 1384–1399, A literature review on the application of evolutionary computing to credit scoring, Al Marque’s, V García, JS Sa’nchez, Universitat Jaume I, Castello de la Plana, Spain
[15] A Comparison Study of Credit Scoring Models, Defu Zhang, Hongyi Huang, Qingshan Chen1, Yi Jiang, Department of Computer Science, Xiamen University, China