International Journal of Computer (IJC)

ISSN 2307-4523 (Print & Online)



http://ijcjournal.org/

# A Hybrid Machine Learning Approach for Credit Scoring Using PCA and Logistic Regression

Sylvester Walusala W[a]*, Dr. Richard Rimiru[b], Dr. Calvin Otieno[c]

[a]*School of computing Jomo Kenyatta University of Agriculture and Technology Nairobi*

[b,c]*Senior Lecturer, School of computing Jomo Kenyatta University of Agriculture and Technology Nairobi*

[a]*Email: sylvesterww@gmail.com*

[b]*Email: rimirurm@gmail.com*

[c]*Email: otienocalvins22@gmail.com*

**Abstract**

Credit scoring is one mechanism used by lenders to evaluate risk before extending credit to credit applicants. The method helps distinguish credit worthiness of good credit applicants from the bad credit applicants. Credit scoring involves a set of decision models and with their underlying techniques helps aid lenders in issuing of consumer credit. Logistic regression (LR) is an adjustment of linear regression with flexibility on its preposition of data and is also able to handle qualitative indicators. The major shortcoming of Logistic regression model is the inability to deal with cooperative (over fitting) effect of the variables. PCA is a feature extraction model that is used to filter out irrelevant un-needed features and hence, it lowers model training time and costs and also increases model performance. This study evaluates the shortcomings of simple models and proposes to develop an efficient and robust machine learning technique combining Logistic and PCA models to evaluate firms in the deposit taking SACCO sector. To achieve this, experimental methodology is adopted. The proposed hybrid model will be two staged. First stage will be to transform the original variables to get new uncorrelated variables. This will be done using Principal Component Analysis (PCA). Stage two is the use of LR on the principal component values to compute the credit scores. Inferences and conclusions were made based on the analysis of the collected data using Matlab.

***Keywords:*** Credit scoring; Machine learning; PCA; Multinomial logistic regression.

---

* Corresponding author.

## 1. Introduction

Credit risk is an important and highly studied topic in finance accounting and banking [44]. Many financial institutions take credit risk to be the most critical and difficult risk to manage and analyze. Growth in information technology has led to lowering the costs of managing analyzing and acquiring data, hence able to build a more efficient and robust technique for credit risk management [26]. Credit scoring is one mechanism used by lenders to evaluate risk before extending credit to credit applicants. The method helps distinguish credit worthiness of good credit applicants from the bad credit applicants. Credit scoring involves a set of decision models and with their underlying techniques helps aid lenders in issuing of consumer credit. Ming-Chang [45] states that bankruptcy in companies lead to economic damages for creditors, management, investors and employees together with social cost hence need for credit scoring in finance.

Reference [22] recognized the need to differentiate between good and bad loan by measuring applicants characteristics. He stated that after the measurement credit analysts then advise financial companies on whether to lend or not. The arrival of credit cards in the 1960s increased need and expansion of credit scoring. Credit scoring is believed to be a better means of evaluating a creditworthy borrower as compared to the traditional methods of risk assessment. With enhancement in technology over the years many banks are adopting credit scoring models for evaluating of borrower. Credit scoring simplifies the task of estimating the probability of default and calculates the loss given default from a range of complicated possible scenarios [95]

Reference [37] define machine learning as the process by which a computer learns by being exposed to data, generally by using an algorithm that optimizes some mathematical function of that data. They state that there are two fundamental categories of machine learning i.e. supervised learning and unsupervised learning. Supervised learning models are in the business of prediction while unsupervised learning focuses on understanding the structure behind a data set. The main characteristic of supervised machine learning algorithm is that it consists of target or outcome variable. The target variable being used to predict other features from a given set of variables i.e. independent variables. In essences, using the target variable, a model is generated that maps the input to a desired output. In this regard the training process generally continues until the function achieves the required level of accuracy. Supervised machine learning techniques are achieved using classification and regression techniques and algorithms or approaches that range from non-linear regression, generalized linear regression methods, linear discriminant analysis, Support Vector Machines (SVMs) decision trees, ensemble methods among others. For the unsupervised machine learning, there is no specific target or outcome variable to help predict or develop an estimate. The algorithm is mainly used for clustering or segmenting entities into various groups for specific intervention. Some major methods of unsupervised machine learning algorithms are Apriori and K-means algorithms [62]

Reference [62] defines machine learning as a computer science branch with capacity to learn from a given data set. He suggests that machine learning models can be used in different fields. These include credit scoring, fraud detection, internet search, stock trading, and medical design among other areas. Christopher [15] observed that Machine learning is related to the fields of Pattern Recognition, Statistics and Data Mining while also it emerges as a subfield of computer science and gives special attention to the algorithmic part of the knowledge extraction

process. Reference [19] states that the objective of machine learning is to help solve real world challenges with the use of a model that provide a good data estimation and approximation. Machine learning models have been used to predict stock price movements over a long period of time. Mehak and his colleagues [43] used Machine Learning Techniques to predict stock market predict performance of Karachi Stock Exchange (KSE). The results of the study confirmed that machine learning models are capable of predicting the stock market performance. The Karachi Stock Market with KSE-100 index was consistent with the behavior that could be predicted using machine learning methods.

Reference [23] Used SVM machine learning in health informatics. They observed that Machine learning technique has played a key role in epidemic forecasting for the past years. They state that the selected studies showed that by using the machine learning technique, the challenges faced in detection of digital disease can be overcome. They however propose that real-time monitoring and forecasting is the next big frontier for the researchers in the field of machine learning. Jung and his colleagues [30] demonstrated use of machine-learning application in determining significant features with risk of Radiation Pneumonitis (RP) patients. They used classification experiments with the chosen features using kernel SVM supervised machine learning model.

Reference [46] carried out a survey of Machine Learning Applications for Energy-Efficient Resource Management in Cloud Computing Environments. They observed that there are various examples of machine learning based solutions to several resource management issues in the cloud. They note that it although those solutions were not entirely designed for the purpose of energy efficiency; some of them generally reduced energy consumption as a side effect. Melmet [46] in their survey list several machine learning methods that have helped in energy resource management: They state that demand forecasting is one of the key problems in data center management, and ability to have good forecasting techniques can lead to optimal allocation strategies which will minimize energy consumption. Through accurate estimation according to [46] can allow cloud providers to avail intelligent resource management structures which rely on scheduling of tasks and consolidations to turn on the least number of equipment in the data centers, hence reducing energy consumption. Researchers have also found ways on how to reduce cooling costs by concentrating on the thermal distribution that arise as a result of different workload placements. On the scale of individual equipment and machines, machine learning has helped save energy by optimally configuring CPU usage and frequency.

Machine Learning algorithms have also been used for betterment in Education systems. This has been done through predictive models built to predict the retention of new students in Science & engineering course while classifying them into three categories based upon overall GPA [28]. The classes were based on three categories i.e. at-risk, Intermediate, Advanced. The results showed that students under at-risk category were more likely to drop their studies [28]. It thus means that predictive analytics along with machine Learning can be a great tool and beneficial in exploring opportunities challenges and improvise the hidden loop holes in education system (40)

### *1.1. Feature selection*

Reference [60] define the term feature selection to refer to algorithms that output a subset of the input feature

set. They state that one factor that plagues classification algorithms is the quality of the data. Irrelevant or redundant information or the data is noisy and unreliable then knowledge discovery using training will become more difficult Immaterial of whether a learner attempts to select features itself or ignores the noise, feature selection prior to learning can be beneficial [60]. Reducing the dimensionality of the data reduces the size of the hypothesis space and allows algorithm to operate faster and more effectively. In some cases it can improve accuracy on classification. Ramlee and his colleagues [61] observed that dimension reduction is more concerned with eliminating the redundant data with the main purpose being to improve the performance of the process. Redundancy increases the relation among the features and will make features to be strongly depended on each other. Dimension reduction improves the verification process when selecting the features because after the unimportant features are removed the classification process becomes easier. Novakovic and his colleagues [54] observed that feature selection reduces the dimensionality of feature space, removes irrelevant, redundant or noisy data. It brings the immediate effects for application: speeding up a data mining algorithm, improving the data quality and thereof the performance of data mining, and increasing the comprehensibility of the mining results. Feature selection as a process of selecting necessary variables of a subset for the purposes a learning model [36]. The following are the advantages of feature selection: Increases accuracy of the developed model, helps reduce on resource use during data collection, removes redundant unnecessary variables, and improves data quality, Reduces dimensionality of feature space increasing speed of the model while lowering storage space and enhances data understanding enabling understanding of the processes that led to generation of the data i.e. enhances ability for one to clearly visualize and abstract the data [36]. Feature selection is a critical and important task for data preprocessing. This has high considerable ability to improve and enhance the prediction accuracy and performance of financial crises. Feature selection helps process and choose a minimum subset of S features from the original set of T dataset features (S<T) such that the dimensionality is optimally lowered. This is critical since there are no generally agreed upon variables (i.e. financial ratios) to be the most representative features in bankruptcy prediction and credit scoring activities, The collected variables are first examined to evaluate their necessity, importance and explanatory power in the chosen dataset. This helps filter out noise features reducing dimensionality while enhancing performance of the classifiers

According to [36] variables for the purposes of feature selection can be classified into three:

- Relevant- Which are variables that will strongly influence the output of the model hence cannot be assumed or left out
- Irrelevant- Variables that have no influence on the output of the model and these values are generated at random
- Redundant- Where variables take the role of another variable in the model i.e. lead to model redundancy.

The main objectives of feature selection are to obtain feature space with:

- Low dimensionality
- Retention of sufficient information
- Removing of noisy or unwanted features/variables

- Comparability of the features

Principal component analysis (PCA) is a method which enables to explain the variance –covariance structure of a set of p-variables through a few linear combinations of these variables (components) [48]. PCA seeks to maximize the variance of the linear combination of these variables. PCA is a technique applied to data that are not groups of observations and fragmentation of variables. PCA aims to reduce the dimension of a dataset with the advantage of simplicity. Steps of computing principal component analysis involves: selecting initial analysis variables based on the research problem, Selecting covariance matrix or correlation matrix based on the characteristics of the initial variables and then computing the characteristic roots of the covariance or the correlation matrix [13].

The main applications of factor analysis techniques are: to reduce the number of variables and to detect structure in the relationships between variables, i.e. to classify variables. Main goal of a factor analysis is to explain the covariance relationships among the variables in terms of some unobservable and non-measurable elements [63]. Very often, the unobserved factors are more interesting than mainly the observed quantitative measures in respective data themselves. Due to this feature, factor analysis is able to be widely used in many fields such as marketing, psychology, finance, economics, political sciences among others [12]. Factor analysis is a method used to reduce the information that have a similar pattern of response so as to get a small set of variables from a large set of variables. This ensures that the relationship and pattern of the data can be easily interpreted. They explain that there are two main types of factor analysis defined as Exploratory Factor Analysis (EFA), and Confirmatory Factor Analysis (CFA) whereby EFA works by determining the relationship between variables and factor while, CFA measures the relationship among the many factors [12].

Linear Discriminant Analysis (LDA) is a linear classical dimension reduction method that is designed to optimally cluster different classes of data providing for a low dimensional subspace [61]. LDA is involved the measurement of between-class scatter and within-class scatter so as to quantify the quality of the cluster in a given sample data. The method thus detects the significant features from the classes, and then transforms it to new space data consisting of only most necessary features

Stepwise regression selects out only the significant curves from available data on the basis of variance contribution computation and collinearity verification, hence enabling the established fitting model to be optimal [71]. The aim of stepwise regression method is to maximize the estimation power by use of the minimum number of independent variables. Stepwise regression combines both the forward and backward selection technique which involves an automatic process of dropping and selecting of independent variables [71]. Regina and his colleagues [62] states that the growing varieties, volumes and velocity of amounts data due to the growth of technology in particular and the availability of cheaper data sharing and storage facilities enhanced by availability of cheaper and more so more powerful computational tools have led to a new frontier in the data science field. This thus has enhanced active ongoing research within the fields of machine learning i.e. the process of building analytical models using methodologies for machine to learn from data. The algorithms extract knowledge and pattern from a set of data. They propose development of hybrid machine learning system that is incorporating the most important features that determine credit worthiness.

Cristián and his colleagues [16] observed it's possible to improve traditional credit scoring models by the use of more sophisticated techniques. Finding a model to replace human assessment in credit default is of major interest for risk management researchers [73]. PCA is a feature extraction model that is used to filter out irrelevant un-needed features and hence, it lowers model training time and costs and also increases model performance [67].

Logistic regression (LR) is an adjustment of linear regression with flexibility on its preposition of data and is also able to handle qualitative indicators. Logistic regression unlike in LDA has the capacity to predict default chance of a loan applicant and identify the other variables related to applicant. LR is the best when a decision maker wants to be sure a client will pay a loan [52]. According to [33] LR is likely a better performing model when used on smaller number of parameters. Generating an accurate, as well as explanatory; classifying model is an increasingly important issue in credit risk management [44]. Marcos & Reginaldo [39] propose that future studies on credit scoring should strive to use hybrid models, combining different techniques to improve classification and predictive performance. Employing feature reduction models will enhance accuracy and improve on results . Pejic and his colleagues [47] states that the amount of data is constantly increasing, which thus creates a lot of challenges in successful decision making for managers. Growth of both valuable as well as invaluable data in databases has led to a need for the use of different techniques which can help in extracting, finding and analyzing data important for decision making.

The major shortcoming of Logistic regression model is the inability to deal with cooperative (over fitting) effect of the variables [6]. This lowers accuracy of the model. Any slight increase in the accuracy levels of an existing model is of great importance. This study proposes to develop a hybrid machine learning model combining Logistic and PCA models to evaluate firms in the deposit taking SACCO sector. A hybrid model combines advantages of the two models creating a robust model [35]. Each algorithm will have its own advantages and disadvantages hence developing prediction model for a classification problem cannot be obtained effectively and efficiently by using of only any one of the available classification algorithms. They propose that Instead of relying on a single classifier, we can combine two or more models which enhance accuracy of classification [18]

### *1.2. Specific objectives*

- To evaluate and discriminate the most suitable variables for developing the model.
- To assess the ability of the simple existing models to discriminate between good, neutral and bad score.
- To evaluate the ability of the developed hybrid model to discriminate between good, neutral and bad score

### *1.3. Research questions*

- Can the model discriminate the most suitable variables?
- Can the simple existing models discriminate among good, neutral and bad scores?
- Can the developed hybrid model discriminate among good, neutral and bad scores?

## 2. Related studies

There has been increased demand on credit loans and the limited number of credit experts due to need to minimize costs on paying the experts. Lenders hence require efficient and accurate automated credit scoring model. Due to significant number of customer portfolios, need for enhancement in credit scoring system which will lead to loss reduction, future savings, faster processing, and a closer behavioral study on the existing customers there has been to numerous attempts to develop models [103]. Aduda and his colleagues [4] reinforces on these benefits of credit scoring which include accuracy in the decision making process. They state that accuracy is gained as a result of reduction in adverse selection cases due to making better assessments. They further state credit scoring decreases information asymmetries between borrowers and lenders in financial markets and also allow lenders to more accurately evaluate risks and improve portfolio quality. Credit scoring plays a vital role in economic growth by helping expand access to credit markets, lowering the price of credit and reducing delinquencies

Machine learning to be the process by which a computer learns by being exposed to large data, generally by using an algorithm that optimizes some mathematical function of that given data [37]. There are two main fundamental categories of machine learning i.e. supervised learning and unsupervised learning. Supervised learning models are in the business of prediction while unsupervised learning focuses on understanding the structure behind a data set [37]. Domingos [21] defines machine learning as a computer science branch with capacity to learn from a given data set. He suggests that machine learning models can be used in different fields. These include credit scoring, fraud detection, internet search, stock trading, and medical design among other areas. Machine learning is related to the fields of Pattern Recognition, Statistics and Data Mining while also it emerges as a subfield of computer science and gives special attention to the algorithmic part of the knowledge extraction process [15]. The objective of machine learning is to help solve real world challenges with the use of a model that provide a good data estimation and approximation [19].

Reference [51] define a credit-scoring model as a formula that puts weight on different characteristics of a borrower, lender and loan. Several methods have been used for scoring. Examples include the traditional models such as logistic regression and probit, artificial intelligence algorithms and data mining approaches. The early studies of bankruptcy mainly used some accounting ratios where a comparison of accounting ratios of failed firms and of the non-failed firms was done. These studies used time series analysis [14].

Research and studies on credit scoring models is wide spread: Reference [58] on logistic regression, Reference [72] on nearest space method, Reference [38] on hybrid models, Reference [40] on Artificial neural network, Reference [17] on genetic algorithmic, Reference [56] on decision trees among others. According to [13] Artificial Intelligence models have made significant contribution to information systems. These models include neural networks, genetic programming decision trees among others. Neural networks are non-linear, algorithmic procedure using interconnected nodes with ability to transform inputs into desired outputs. Comparison between different statistical credit scoring approaches demonstrates that advanced techniques, such neural networks and genetic programming perform better than more conventional techniques, such as discriminant analysis and logistic regression, in terms of their higher predictive ability [2]. Results of some studies show that the

predictive capabilities of both approaches were sufficiently similar to make it difficult to distinguish between them. Every methodology is likely to give its own result and no methodology gives an optimal result hence difficult for decision makers to use only one model in decision making due to many qualitative considerations [2]. Probit analysis as a technique that finds coefficient values, which results in the probability of a unit value of a dichotomous coefficient value. The probit model transforms linear combination of the independent variables into their cumulative probability values from a normal distribution. ANN has more advantages in credit scoring compared to other statistical methods [40]. Reference [39] adopted a case study methodology in evaluating credit risk for a Credit Union. Using data mining techniques, decision trees and ANN they developed credit risk models. The models were able to produce better results than previous models. However they recommend that future studies need to combine different techniques to improve predictive and classification power. While studying computational time reduction for credit scoring [29] concluded that neural networks and evolutionary models are likely to have very high costs since certain number of parameters need optimization so as to achieve accuracy rates shown by support vector machine based on stratified sampling. The neural network models support both classification and regression algorithms hence, are appropriate for studying the classification problems [62].

Genetic algorithm (GA) emulates natural selection as proposed by Charles Darwin evolution theory. It is a search model that ensembles process of natural evolution. Reference [2] analyzed literature review of credit scoring models using 214 articles. They found out that advanced techniques such as genetic programming performed better. Reference [17] carried out a study on GA for credit scoring system maintenance function. With the help of experts in the field of credit scoring a model was developed to try and learn a function with a given set of dataset of 25 models developed earlier. GA was found to have a better performance compared to traditional techniques. However David states that GA may not analyze all classes of data, the model is dependent on data and business objective; require good processing power and getting an optimal combination of parameter data is difficult for a given problem. Huang and his colleagues [30] reviews GA model that uses appropriate fitness evaluation function with consideration values of both good classified customers and negatively classified customers in a set. The results showed high accuracy rate for the positively classified good customers, but poor performance on negatively classified customers. Reference [1] analyzed capacity of genetic programming model in the Egyptian market. The model proved better than probit model. Decision trees use branching mechanism to explain all the possible outcomes in a scenario. Each tree branch represents a possible occurrence or decision. Reference [11] carried out a quantitative survey on 342 online surveys and 31 face to face Chinese and Australian firms. The study aimed to determine business collaboration using Decision trees. They conclude that decision trees are a good supplement in analyzing collaborating strategies. Reference [66] studied ability of decision trees and support vector machines (SVM) in detecting credit card fraud. Using a population of 978 fraudulent records and 22 million normal records 3 stratified samples were developed. They found that decision trees outperformed SVM.

Reference [36] was the first to develop linear discriminant analysis (LDA) model. According to [50], the most widely used statistical method for credit scoring is LDA. Mircea and his colleagues [46] used discriminant analysis to develop a credit scoring model. They found out that discriminant analysis represented an effective method of extracting relevant information. It has however been criticized for its linear relationships requirement

between dependent variables and independent variables and its assumptions that the input variables have to adhere to normal distribution which does not always hold. Discriminant analysis is based on an assumption that different classes of data will be generated by using different Gaussian distributions. The main types of discriminant algorithms used for classification being the linear and the quadratic discriminant [62].

Logistic regression (LR) is an adjustment of linear regression with flexibility on its preposition of data and is also able to handle qualitative indicators. Logistic regression unlike in LDA has the capacity to predict default chance of a loan applicant and identify the other variables related to applicant. LR is the best when a decision maker wants to be sure a client will pay a loan [52]. LR is likely a better performing model when used on smaller number of parameters. Ricardas and Vytautas [64] used discriminant analysis, LR and ANN methods classification of Lithuanian companies into 2 groups of default and non-default. The three methods were found to be relevant methods of classification of bank clients however LR model was the best with 97% accuracy.

Reference [8] used support vector machine (SVM) to develop a credit scoring model for a micro finance institution. Using a sample data of 157 instances consisting of customer information they developed a model using SVM and did a comparison of its performance with other models. Percentage of accurately predicted data of the default and the non-default cases and the inaccurately classified data were monitored using five methods i.e. (LR), Quadratic discriminant Analysis (QDA), (LDA), Back Propagation neural network (BP) method and SVM. SVM outperformed other models with accuracy of 85.36%, LR 75.19%, QDA 71.37%, LDA 75.50% and BP 80.73%. Dynamic models are better and more flexible way for modeling and forecasting consumer risk.

Many studies done show the advantages and limitations of credit scoring models. neural networks are reliable tools of predicting the determinants of relationship quality; robust to handle missing or inaccurate data i.e. associative ability; the model automatically handle variable interaction; their accuracy can be assessed using statistical methods like mean squared error; they can easily be updated over time and provide high levels of accuracy of 96.9% [12.78] However neural network models also have some technical limitations: No method devised yet to determine the significance of independent inputs in a neural network directly; learning process in ANN can be time consuming: difficult to state the results in simple precise analytical model statement and if environment changes, the network must be reconstructed

SAS Institute [65]) outlines the following as the main advantages and disadvantages of Bayes statistics: Provides a convenient background setting for a range of other models, e.g. hierarchical models and problems on missing data; provides interpretable answers; obeys likely hood principle when having two distinct likelihood scenarios; provides a natural and a principled way of combining prior information and data; it provides results that is conditional on data and exact without reliance on asymptotic approximation. However the Bayes model has the below observed limitations: does not tell us how to select prior, can produce posterior distribution results that are likely to be heavily influenced by priors, and involves high computational cost especially where large parameters might be needed. Advantages of decision trees as follows; accuracy of classification, effectiveness, straightforward and intuitive explanation, easy interpretation of results and easy interpretation of results. Disadvantages of decision trees include; need for discrete values in the algorithm, high data required and over sensitivity to training and costly.

Advantages expert systems in credit scoring to be; the systems can take qualitative variables, the system does not assume any statistical distribution of data, the systems builds a number of if – then which can be applied to new cases of classification decisions [27]. The disadvantages of expert systems as observed by [27] are; programming intuition of an expert system is very difficult, inability of systems to adapt to inductive learning when using if then rules in changing scenarios, inability of the systems to work with incomplete data, costly system to set up

Machine learning models are prone to over-fitting challenges which leads to bias [21]. This happens mainly on the basis of data being used for modeling. It's possible to improve traditional credit scoring models by the use of more sophisticated techniques [18]. Anomaly checks should be done on data to eliminate outlier variables before fitting in a model. They separately used PCA and Logistic regression in developing a credit scoring model. They observed that for improved performance of models, data needed cleaning and enhancement [42]. Using a single model may not consider all characteristics of data [34]. Even with the use of cross-validation to reduce over fitting, machine learning models may still produce results that may be difficult to interpret, understand and defend [20]. Any slight improvement may lead to a tremendous increase in lending using the score. According to [44] hybridization will increase the level of accuracy even by slight margins. They state that there is no universal credit scoring model that can be accepted in all circumstances hence there is need to continue improving on existing models.

Over-fitting is a critical issue development of credit model development [27]. Many machine learning models face the challenge of over-fitting [57]. Even in scenarios where there are very many relevant features used there will be a problem of over-fitting. These high dimension variables, according to [57] becomes a challenge to understand hence cannot be used to develop a good classifier. Reference [57] thus proposes use of lower effective dimensions or adopting algorithms that can help in lowering the variable dimensions before developing a learning model. Reference [91] states that main challenges of traditional machine learning models is over-fitting of data. They suggest correction of this can be through use of more advanced models like complex modeling techniques such as neural networks, random forests and support vector machines. They however observed these models will lead to marginal increase in accuracy and will to loss of interpretability. They thus suggest that for better improved classifier model there is need to improve on data quality through minimizing of outliers.

Seyed and his colleagues [67] observed that PCA is a feature extraction model that is used to filter out irrelevant un-needed features and hence, it lowers model training time and costs and also increases model performance. Mehdi K & Reference [44] observed that generating an accurate, as well as explanatory; classifying model is an increasingly important issue in credit risk management. Reference [39] propose that future studies on credit scoring should strive to use hybrid models, combining different techniques to improve classification and predictive performance. the major shortcoming of Logistic regression model is the inability to deal with cooperative (over fitting) effect of the variables. This lowers accuracy of the model [6].

## 3. Research design

The research is to develop an efficient and robust technique for credit risk management using PCA and logistic regression for evaluating SACCOs. To achieve this, experimental methodology is adopted. Experiments have shown to be a robust research methodology in mature sciences including applied and behavioral sciences [70]. Financial data from SASRA for the 5 years under study (2009-2014) was used.

Reference [3] ratio analysis can be used to measure performance of an organization. He states that ratio analysis is a good tool of financial analysis, which can be used as a predictive tool in measuring organization or business performance. Ratio analysis is used in aiding the decision making process, on the basis of the relationship between information presented in financial statements with the aim of determining value and evaluating risk. Main ratios used for credit scoring for corporate include liquidity, profitability, financial leverage, repayment capability and efficiency [10,35,101]. Historical financial ratios can conclusively be used to develop a credit predicting model [31].

Fundamental analysis involves analyzing the key fundamentals of a company's financial condition, operating results and its common share stock's underlying characteristics. This involves analyzing and reviewing financial statements for the financial ratios among other things [14]. The financial ratios can help us learn the company's strengths and weaknesses, identify underlying trends and developments in the company or organization, evaluate operating efficiencies, and gain a general overview and understanding of a business' nature and future operating characteristics. Reference [10] state that financial ratios has been used extensively in financial analysis due to its efficiency through an easy developing, understanding procedure and low computing equipment requirements.

Reference [53] states that financial ratios is one of the most reliable and exceptional ways to analyze a company's management of their daily operation activities, performance and management to the management of liabilities, assets and equity. The selection and review of the most appropriate and necessary ratios that closely represent a company's performance is important and critical. Choosing incorrect ratios as variable parameter may lead to poor quality and undermine the accuracy of the model predicting or estimating a firm's business performance. Financial ratio that measure profitability, liquidity and solvency are seen to be the most reliable indicators and are significant and critical predictors of insolvency. Thus financial ratios do give a lot of benefits to a company or any business entity. Financial modeling using financial ratios is a reliable tool to predict business performance either in the current existing environments or the future environments [53]

## 4. Data analysis

The data analysis is two staged. First stage will be to transform the original variables to get new uncorrelated variables. This will be done using Principal Component Analysis (PCA). An In-depth assessment and analysis of the variables helps in including important variables and excluding others that are not relevant. It has the advantage of providing more precise and succinct credit management models, reducing time of execution and improving decision accuracy. The analysis of outlier or discrepant, cases may be important to creating a new classification or finding undesirable patterns [50]. Stage two is use of LR on the principal component values to compute the credit scores. LR is applied to get specific values of Y as follows:

$Y_i$=1 if $Y^*_I \leq 0.29$ for low scores

$Y_i$=2 if $0.29 < Y^*_I \leq 0.95$ for neutral scores

$Y_i$=3 if $Y^*_I > 0.95$ for high scores

Where $Y_i$ are the independent ordinal values while $Y^*_I$ are the scales for credit scaling model. Inferences and conclusions were made based on the analysis of the collected data using Matlab.

## 5. Discussion

From our data analysis the following results are produced based on the models used

### *5.1. Hybrid model based on PCA and Multinomial Logistic Regression*

**Table 1:** Sacco classifications under hybrid PCA-MLR model

| | Low scores | Neutral scores | High scores |
|---|---|---|---|
| Correctly classified Saccos | 4 | 5 | 8 |
| Misclassified Saccos | 1 | 2 | 0 |
| TOTAL | 5 | 7 | 8 |

From our hybrid model results show that 4 Saccos under the low scores are correctly classified and 1 is misclassified as in the neutral category. This represents 80% correctly classified and 20% misclassified.

For the neutral scores 5 Saccos are correctly classified while 2 are classified as low score Saccos. This represents 71.43% correctly classified and 28.57% wrongly classified as of low scores Under the high scores all the 8 Saccos are correctly classified to be of high credit scores representing 100% correct classification.

On overall the model is able to correctly classify 17 Saccos and misclassify 3 Saccos. This represents 85% accuracy and 15% misclassification.

### *5.2. Multiple Regression model-PCA*

**Table 2:** Sacco classifications PCA-Multiple regression model

| | Low scores | Neutral scores | High scores |
|---|---|---|---|
| Correctly classified Saccos | 1 | 4 | 5 |
| Misclassified Saccos | 4 | 3 | 3 |
| TOTAL | 5 | 7 | 8 |

From our PCA-multiple regression model, results show that 1 Saccos under the low scores is correctly classified and 4 is misclassified. This represents 20% correctly classified and 80% misclassified. For the neutral scores 4 Saccos are correctly classified while 3 are miss-classified. This represents 57.14% correctly classified and 42.86% wrongly classified Under the high scores all the 5 Saccos are correctly classified to be of high credit scores while 3 are misclassified. This represents 62.5% correct classification and 37.5% misclassification. On overall the model is able to correctly classify 11 Saccos and misclassify 9 Saccos. This represents 45% accuracy and 55% misclassification.

### *5.3. Stepwise fit-Multinomial logistic Regression model*

This model produces all scores being negative values. This model may not be fit for this set of data because of lack of interpretability of output

### *5.4. Stepwise fit-Multiple regression*

**Table 3:** Sacco classifications under stepwise fit-multiple regression model

| | Low scores | Neutral scores | High scores |
|---|---|---|---|
| Correctly classified Saccos | 8 | 22 | 28 |
| Misclassified Saccos | 17 | 13 | 12 |
| TOTAL | 25 | 35 | 40 |

From the stepwise fit-multiple regression model results show that 8 Saccos under the low scores are correctly classified and 17 misclassified. This represents 32% correctly classified and 68% misclassified. For the neutral

scores 22 of the Saccos are correctly classified while 13 misclassified. This represents 62.86% correctly classified and 37.14% wrong classification. Under the high scores 28 Saccos are correctly classified while 12 are misclassified. This represents 62.5% correct classification and 37.5% misclassified. On overall the model is able to correctly classify 58 Saccos and misclassify 42 Saccos. This represents 58% accuracy and 42% misclassification.

### *5.5. Factor Analysis-Multinomial logistic Regression model*

This model produces a negative slope thus this model may not be fit for this set of data because of lack of interpretability of output

### *5.6. Factor Analysis-Multiple regression model*

**Table 4:** Sacco classifications under Factor analysis-multiple regression model

| | Low scores | Neutral scores | High scores |
|---|---|---|---|
| Correctly classified Saccos | 1 | 2 | 6 |
| Misclassified Saccos | 3 | 4 | 2 |
| TOTAL | 4 | 6 | 8 |

From the stepwise fit-multiple regression model results show that 1 Saccos under the low scores are correctly classified and 3 misclassified. This represents 25% correctly classified and 75% misclassified. For the neutral scores 2 of the Saccos are correctly classified while 4 misclassified. This represents 33.33% correctly classified and 66.67% wrong classification. Under the high scores 6 Saccos are correctly classified while 2 are misclassified. This represents 75% correct classification and 25% misclassified. On overall the model is able to correctly classify 9 Saccos and misclassify 9 Saccos. This represents 50% accuracy and 50% misclassification.

## 6. Conclusion

From the results in chapter four the developed models the summary of the results are as follows: the PCA-MR model is able to correctly classify 11 Saccos and misclassify 9 Saccos. This represents 45% accuracy and 55% misclassification. The FA-Multiple regression model is able to correctly classify 9 Saccos and misclassify 9 Saccos. This represents 50% accuracy and 50% misclassification. The stepwise fit –Multiple linear regression model is able to correctly classify 58 Saccos and misclassify 42 Saccos. This represents 58% accuracy and 42% misclassification.This Stepwise fit-Multinomial logistic Regression and Factor Analysis-Multinomial logistic Regression models produces all scores being negative values. This model may not be fit for this set of data

because of lack of interpretability of output.The PCA-Multinomial logistic Regression was able predict 80% accurately of the low credit SACCOs 20% was predicted as being in the moderate category. 71.47% of the moderate SACCOs were predicted accurately and 28.53% predicted to be of low credit score. The model predicted 100% accurate among the high score SACCOs. On average the PCA-Multinomial logistic Regression model achieved 85% accuracy and 15 % of the SACCOs were misclassified. The correlation matrix also shows only 2 variables strongly correlated implying over fitting (multicolinearity) has been eliminated. This model also out performs the models tasted in this study.From our analysis simple models may not accurately discriminate among good neutral and bad scores. The developed hybrid model discriminate among good, neutral and bad scores thus achieving our objectives.

**References**


[1] Abdou, HAH , 'Genetic programming for credit scoring : the case of Egyptian public sector banks' , Expert Systems with Applications, **36** (9) , pp. 11402-11417. 2009

[2] Abdou .H. and Pointon. J. “Credit scoring statistical techniques and evaluation criteria: A review of literature.” Intelligent systems in accounting, finance and management; 18 (2-3) pp 59-88. 2011

[3] Adedeji. Elijah. “A Tool for Measuring Organization Performance using Ratio Analysis” Research Journal of Finance and Accounting Vol.5, No.19. pp 16-22. 2014

[4] Aduda J, Peterson O. M. & Githinji M .W. (2012). “The Relationship between Credit Scoring Practices by Commercial Banks and Access to Credit by Small and Medium Enterprises in Kenya” International Journal of Humanities and Social Science Vol. 2 No. 9. pp 203-213. 2012

[5] Altman, E.I, Danovi, A. and Falini, A. “Z-Score Models’ Application to Italian Companies Subject to Extraordinary Administration”. Journal of Applied Finance, 23(1): pp. 128-137.2013

[6] Alireza H, Mohana O, Marthandan .G , Wan F, Wan Y , Sasan K “ Statistical and data mining methods in credit.” Proceedings of the Asia Pacific Conference on Business and Social Sciences 2015, Kuala Lumpur (in partnership with The Journal of Developing Areas). 2015 pp. 448-458

[7] Ayushi Sharma & Akshit Chopra “Artificial Neural Networks: Applications in management” Journal of Business and Management (IOSR-JBM) Volume 12, Issue 5 PP 32-40. 2013

[8] Asuri Venkata Madhavi & Radhamani .G, “ Improving the credit scoring model of microfinance institutions by support vector machines” International Journal of Research In Engineering and Technology: Volume: 03 pp: 29-33. 2014

[9] Beaver, William H. “Financial Ratios as Predictors of Failure,” Empirical Research in Accounting: Selected Studies, Vol. 4 pp 71 – 111. 1996

[10] Carlos.Cubaque-Zorro and Juan. C. Figueroa-García. "A fuzzy logic system for evaluating financial profit ratios," IEEE Conference on Norbert Wiener in the 21st Century (21CW), Boston, MA, 2014 pp. 1-7.

[11] Chao. S. and Yu Zhang. “Using decision tree in business collaborator” 8th SMEs in a global economy conference 2011 pp. 172-186.

[12] Chaiwut .K, W. Rueangsirarak and R. Chaisricharoen. "Factor analysis on student loan consideration in higher education level," International Conference on Digital Arts, Media and Technology (ICDAMT), 2017, pp. 296-301.

[13] Chen, X. Lu, and Z. Du(2014) "RBF neural network modeling based on PCA clustering analysis," in 2014 IEEE International Conference on Granular Computing (GrC), 2014 pp. 35–38.

[14] Chon Sern Tan, Chin Khian Yong and Yong Haur Tay. "Modeling financial ratios of Malaysian plantation stocks using Bayesian Networks," 2012 IEEE Conference on Sustainable Utilization and Development in Engineering and Technology (STUDENT), Kuala Lumpur, 2012 pp. 7-12.

[15] Christopher M Bishop et al. "Pattern recognition and machine learning". Springer, New York. 2006 pp 1-200

[16] Cristián B, Lyn C & Richard W (2015). "Improving credit scoring by differentiating defaulter behavior" Journal of the Operational Research Society Volume 66, pp 771–781 2015

[17] David J. Forgarty "Using genetic Algorithms for credit scoring systems maintenance functions" International Journal of Artificial intelligence and Applications Vol.3, No.6, November 2012

[18] Devi R. and R. M. Chezian. "A relative evaluation of the performance of ensemble learning in credit scoring," IEEE International Conference on Advances in Computer Applications (ICACA), Coimbatore. 2016 pp. 161-165.

[19] Dhage, S. N., Raina, C. K. "A review on Machine Learning Techniques". International Journal on Recent and Innovation Trends in Computing and Communication, vol. 4, no. 3, pp. 395-399. 2016

[20] Dinesh Bacham and Janet Zhao "Machine Learning: Challenges, Lessons, and Opportunities in Credit Risk Modeling" Moody's analytics risk perspectives/managing disruptions/Vol IX pp 1-5 2017

[21] Domingos Pedro(2012). "A few useful things to know about Machine learning" Communications of ACM 55.10 pg 78-87 2012

[22] D. Durand, "Risk Elements in Consumer Installment Financing," National Bureau of Economy Research, New York, 1941, pp. 189-201.

[23] Ekkarat. B and Khanita. D, "Digital disease detection: Application of machine learning in community health informatics," 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), 2016 pp. 1-5.

[24] Fang. K. & Huang H. "Variable Selection for Credit Risk Model Using Data Mining Technique" Journal of computer Vol 6 No 9 pp 1868-1874 2011

[25] Fisher .R. A. "The use of multiple measurements in taxonomic problems" pp 466-475, 1936

[26] Gabriela Mircea, Marilen Pirtea, Mihaela Neamtu and Sandra Băzăvan "Risk software application using a credit scoring model" International journal of applied mathematics and informatics Issue 1, Volume 6. Pp 1-8 2012

[27] Genriha. I and Voronova. I. "Methods for Evaluating the Creditworthiness of Borrowers" SCEE. Conference proceedings. RTU Publishing House. 2012 Vol.22 pp.42-49

[28] Halde, R.R "Application of Machine Learning algorithms for betterment in education system," International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), Pune, 2016 pp. 1110-1114.

[29] Hens .A.B and Tiwari .M. (2012). Computational time reduction for credit scoring: an integrated approach based on support vector machine and stratified sampling method" Expert systems with applications. Volume 39 Issue 8, Pp 6774-6781 2012

[30] Huang. X, Cai. W. Lin. X. and Zhong .H. "A genetic algorithm model for personal credit scoring in:

In: Proceedings of the International Conference on Computational Intelligence and Software Engineering, Wuhan, China, 2009 pp 1–4.

[31] Ivica P & Tamara K "The relative importance of financial and non- financial variables in predicting insolvency" Croatian Operational Research Review (CRORR), Vol. 4 pp 187-198 2013

[32] Jung. H. Oh, R. Al-Lozi and I. E. Naqa "Application of Machine Learning Techniques for Prediction of Radiation Pneumonitis in Lung Cancer Patients," International Conference on Machine Learning and Applications, Miami Beach, FL, 2009 pp. 478-483.

[33] Kalamkas N. & Gulna .B. "Algorithmic Scoring Models." Applied Mathematical Sciences, Vol. 7 pp. 12, 571 – 586. 2013

[34] Khashei, M.; Bijari, M.& Hejazi, S. "Combining seasonal ARIMA models with computational intelligence techniques for time series forecasting". Soft Comput. Vol 16, pp 1091–1105, 2012

[35] Khiem. Tran, T. Duong and Q. Ho, "Credit scoring model: A combination of genetic programming and deep learning," Future Technologies Conference (FTC), San Francisco, CA,2016,pp.145-149.

[36] Ladha. L. & Deepa. T. "Feature selection methods and algorithms" International Journal on Computer Science and Engineering (IJCSE) Vol. 3 No. 5 pp 1787-1797. 2011

[37] Lee M. & Evans M. "Learning by numbers" Predictions Technology Supplement Vol Autumn pp 1-2 2016.

[38] Lin. S. L. "A new two stage hybrid approach of credit risk in banking industry" Expert systems with applications, Vol 36(4), pp 33-41. 2009

[39] Marcos. M. and Reginaldo.S "Credit Analysis using data mining: Application in the case of Credit union" JISTEM, Brazil Vol. 11, No.2, pp. 379-396 2014

[40] Marques A. I, Garcia. V and Sanchez J.S. (2013) "A literature review on the application of evolutionary computing to credit scoring" Journal of operational research society. Vol 64, Issue 9 pp 1384–1399, 2013

[41] Mehmet. Demirci"A Survey of Machine Learning Applications for Energy-Efficient Resource Management in Cloud Computing Environments," IEEE 14th International Conference on Machine Learning and Applications (ICMLA), Miami, FL, 2015 pp. 1185-1190.

[42] Meera R. & Tulasi. B. "Credit scoring process using banking detailed data store" `International Journal of Applied Information Systems (IJAIS) –Foundation of Computer Science FCS, New York, USA Vol 8– No.6. pp 13-20 2015

[43] Mehak. Usmani, S. H. Adil, K. Raza and S. S. A. Ali. "Stock market prediction using machine learning techniques," 3rd International Conference on Computer and Information Sciences (ICCOINS), Kuala Lumpur, 2016, pp. 322-327.

[44] Mehdi K & Akram M. "A Soft Intelligent Risk Evaluation Model for Credit Scoring Classification" Int. J. Financial Stud. Vol 3, pp 411-422. 2015

[45] Ming-Chang Lee. "Enterprise Credit Risk Evaluation models: A Review of Current Research Trends" International Journal of Computer Applications, Vol 44. Pp 1-5 2012

[46] Mircea G., M.Pirtea, M.Neamiu & S. Bazavan. "Discriminant analysis in a credit scoring model" Recent Advances in Applied & Biomedical Informatics and Computational Engineering in Systems Applications. Pp 257-262 2011

[47] Pejic. Bach, J. Zoroja, B. Jaković and N. Šarlija "Selection of variables for credit risk data mining models: Preliminary research," 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, 2017, pp. 1367-1372.

[48] Muca M and Puka L. "R, Matlab and SPSS for factor analysis purposes: Some practical considerations and an" European Scientific Journal vol.10, No.3 pp 233-246 2014

[49] Mehdi Khashei & Akram Mirahmadi "A Soft Intelligent Risk Evaluation Model for Credit Scoring Classification" Int. J. Financial Stud. Vol 3, pp 411-422. 2015

[50] Marcos. M. &Reginaldo.S. "Credit analysis using data mining: Application in the case of credit union" Journal of Information Systems and Technology Management Vol. 11, No. 2, pp. 379-396. 2014

[51] Nanni L, Lumini A (2009). "An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring." Expert Systems with Applications Vol 36 pp 3028–3033. 2009

[52] Nataša .Š, Kristina Š & Silvija.V, "Logistic regression and multi-criteria decision making in credit scoring" Proceedings of the 10th International Symposium on Operational Research SOR ' 2009 pp 1-10

[53] Norazuaniza. M. Yunus and S. A. Malik "Development of financial model using financial ratios in predicting business performance of IBS Construction Company," International Conference on Statistics in Science, Business and Engineering (ICSSBE), Langkawi, 2012 pp. 1-6

[54] Novakovic J, P. Strbac, D. Bulatovic(2011) "Toward Optimal Feature Selection" Yugoslav Journal of Operations Research Vol 21, Number 1, pp 119-135. 2011

[55] Ouertani Nadia and Rangau L. Ureche "Corporate Default Analysis in Tunisia Using Credit Scoring Techniques." International Journal of Business, 15(2) pp 198-220. 2010

[56] Paleologo G, Elisseeff A and Antonini G." Subbaging for credit scoring models" European journal of operational research. Vol 201(2). 490-499. 2010

[57] Pedro Domingos. "A Few Useful Things to Know about Machine Learning" Commun. ACM Vol 55, pp 78–87 2012

[58] Psillaki. M, Tsolas, I & Margaritis D. "Evaluation of credit risk based on firm performance." European journal of operational research, 201(3), 873-88. 2010

[59] Radek Silhavy, Petr Silhavy& Zdenka Prokopova "Analysis and selection of a regression model for the Use Case Points method using a stepwise approach" The Journal of Systems and Software Vol 125 pp 1–14. 2017

[60] Raghavendra B and Simba B "Evaluation of Feature Selection Methods for Predictive Modeling Using Neural Networks in Credits Scoring" Int. J. Advanced Networking and Applications Volume:02, Issue: 03, Pages: 714-718 2010

[61] Ramlee R, Azah K. and Sharifah S "PCA and LDA as Dimension Reduction for Individuality of Handwriting in Writer Verification" 13th International Conference on Intelligent Systems Design and Applications (ISDA) IEEE.2013 pp 104-108

[62] Regina. E. Turkson, E. Y. Baagyere and G. E. Wenya "A machine learning approach for predicting bank credit worthiness," Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR), Lodz, 2016, pp. 1-7.

[63] Ricardas Mileris and Vytautas Boguslauskas. "Data Reduction Influence on the Accuracy of Credit

Risk Estimation Models" Inzinerine Ekonomika-Engineering Economics, Vol 21(1) pp 5-11, 2010

[64] Ricardas Mileris & Vytautas Boguslauskas "Credit Risk Estimation Model Development Process: Main Steps and Model." Inzinerine Ekonomika-Engineering Economics, Vol , 22(2), pp126-133. 2011

[65] http://support.sas.com/documentation/cdl/en/statugbayesian/61755/PDF/default/satugbayesian.pdf accessed 11th November 2015

[66] Sahin. Y. and Duman .E. (2011)"Detecting credit card fraud by decision trees and support vector machines" International Multiconference of Engineers and computer scientists, 2011. Pp 1-6

[67] Seyed S, Mohammad G, & Kamran S. "Combination of Feature Selection and Optimized Fuzzy Apriori Rules: The Case of Credit Scoring" The International Arab Journal of Information Technology, Vol. 12, No. 2 2015

[68] Steven, R: "Operational Risk" (10th Edition), Securities and Investment Institute, 24 Monument Street, London. 2006

[69] http://www.sasra.go.ke/index.php/welcome-to-sasra#.VfJ7c9Kqqko accessed on 11th May 2015

[70] Vinh Vo Xuan. "Using Accounting Ratios in Predicting Financial Distress: An Empirical Investigation in the Vietnam Stock Market" Journal of Economics and Development, Vol.17, No.1, pp. 41-49. 2015

[71] Xiao-Lin Li, Yu Zhong, "An overview of personal credit scoring: Techniques and future work" International Journal of Intelligence Science, Vol 2, pp 181-189. 2012

[72] Yair Levy and Timothy J. Ellis. "A Guide for Novice Researchers on Experimental and Quasi-Experimental Studies in Information Systems Research" Interdisciplinary Journal of Information, Knowledge, and Management Vol 6, pp 152-161 2011

[73] Yufeng et al "The porosity and permeability prediction methods for carbonate reservoirs with extremely limited logging data: Stepwise regression vs. N-way analysis of variance." Journal of Natural Gas Science and Engineering Vol 42 pp, 99-119 2017

[74] Zhou, L., Lai, K. K., & Yen, J. "Credit scoring models with AUC maximization based on weighted SVM." International journal of information technology & decision making. Volume 08, Issue 04 pp 5859-5865. 2009

[75] Zhou, X., Jiang, W. & Shi, Y. "Credit risk evaluation by using nearest subspace method." Procedia computer science, 1(1), 2443-2449. 2010

[76] Zoroja, J. Pejić B, M. & Ćurko, K. "Data mining applications framework for business organizations: Business functions approach" The Business Review Cambridge Vol 22(1), 119-126, 2014

[77] Zurada .J. & Kunene .N. K. "Comparison of the performance of computational intelligence methods for loan granting decisions" Proceedings of the 44th Hawaii International Conference on System Sciences IEEE 2011 pp 1-10