Research Article

Predicting the Performance of Rural Banks in Ghana Using Machine Learning Approach

Emmanuel Awoin, Peter Appiahene, Frank Gyasi, and Abdulai Sabtiwu

University of Energy and Natural Resources, Sunyani, Ghana

Correspondence should be addressed to Emmanuel Awoin; emmanuelawoinpass@gmail.com

Received 2 September 2019; Revised 4 January 2020; Accepted 13 January 2020; Published 19 February 2020

Copyright © 2020 Emmanuel Awoin et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The idea of rural banks was introduced as a result of limited commercial bank branches in rural areas to mobilize their resources for rural development. It is also believed that financial institutions such as rural banks are powerful tools for mitigating poverty. Nevertheless, some of these banks are rather increasing the burden of people through illegal activities and mismanagement of resources. Assessing banks’ performance using a set of financial ratios has been an interesting and challenging problem for many researchers and practitioners. Identification of factors that can accurately predict a firm’s performance is of great interest to any decision-maker. The study used ARB’s financial ratios as its independent variables to assess the performance of rural banks and later used random forest algorithm to identify the variables with the most relevance to the model. A dataset was obtained from the various banks. This study used three decision tree algorithms, namely, C5.0, C4.5, and CART, to build the various decision tree predictive models. The result of the study suggested that the C5.0 algorithm gave an accuracy of 100%, followed by the CART algorithm with an accuracy of 84.6% and, finally, the C4.5 algorithm with an accuracy of 83.34 on average. The study, therefore, recommended the usage of the C5.0 predictive model in predicting the financial performance of rural banks in Ghana.

1. Introduction

In order to fast track the development of rural areas in Ghana, rural banks were introduced. The Association of Rural Banks (ARB) described some of the roles of rural banks as follows: cultivating the habits of savings among rural inhabitants, mobilizing resources locked up in the rural areas into the banking systems to facilitate development, and identifying viable industries in their respective areas for investment and development [1]. It is believed that financial institutions are powerful tools to mitigate poverty [2], but some of them are rather increasing the burden of people through illegal activities and mismanagement of resources [3]. The collapse of most of these institutions can be seen as a result of them not being able to evaluate and predict their financial standings in the years ahead [4]. Therefore, the main objective of this project is to develop and propose a predictive model capable of predicting the financial standing of financial institutions as well as identifying the most influencing financial ratios, using rural banks in Ghana as a case study area and decision tree algorithms (C5.0, C4.5, and CART).

The remainder of this paper is organized as follows: the next section (Section 2) provides a literature review; Section 3 presents the methodology developed and followed in this study and documents its findings; Section 4 summarizes and concludes the paper.

2. Review of the Literature

It is self-evident and based on empirical observation that most financial institutions in Ghana over the years have failed in their operations because they were unable to predict their financial status with respect to their progress or failure. An institution like Diamond Microfinance Limited (DKM) company, which started its operations in 2005, had its operations suspended by the Central Bank of Ghana after it violated the banking Act. Upon the analyses made by Citi
Advances in Fuzzy Systems

Business News, the company failed to hold sufficient assets to meet its liabilities to depositors [5]. Other financial institutions that have suffered a similar collapse are UT Bank and Capital Bank. In August 2017, the bank of Ghana revoked their license describing them as “deeply insolvent” [4]. Because of these cases, many people feel reluctant to invest in some of these financial institutions. A lot of studies have been made to reduce this issue through predictive models using decision tree algorithms and the financial ratios of these institutions.

According to [6], the financial ratios have proven to be the most accurate method for analyzing financial reports which have a high accuracy to treat the points of weakness, effectively and efficiently. Using financial ratios to evaluate a firm’s performance is not an emerging field of study, but rather this has been in the system for quite a long time. Financial ratios derived from financial statements can be used to predict stock price trends in emerging markets [7]. A simple literature search can find literally thousands of publications on topics related to financial ratios. The underlying studies often differentiate themselves from the rest by developing and using different independent variables (financial ratios) and/or employing different statistical or machine learning-based analysis techniques such as decision tree, neutral network, to mention but a few [8].

Machine learning is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence [9]. Machine learning makes use of the construction of algorithms that can learn from and make predictions from data [10]. A research conducted by [11] indicated that recent years have seen a boost in using a predictive classification in medical diagnosis to perform research in this field, and the majority of these papers focus their goal on improving accuracy. Their results, using age, weight, waist, hip, and height as indicators, explained that after a successful experiment with four different machine learning algorithms to make the predictions on type II diabetes, ID3 had an accuracy of 78.57, Naïve Bayes had 79.89, AdaBoost had 84.19, and random forest had the highest accuracy of 85%.

Authors of [12], in their paper, sought to address the difference between deep neural network- (DNN-) based computer vision and human vision with respect to image production/capturing. It indicated that it is easy to produce images that are completely unrecognizable to humans, but that state-of-the-art DNNs believe to recognize objects with over 99% confidence. Given the near-human ability of DNNs to classify visual objects, questions arise as to what differences remain between computer and human vision. A study conducted by [13] revealed a major difference between DNN and human vision. Changing an image, originally correctly classified (e.g., as a lion), in a way imperceptible to human eyes, can cause a DNN to label the image as something else entirely.

Moreover, [14] used multiple regression to evaluate the performance of students using English, Mathematics, Chemistry, and Physics to predict their grade point average. They used performance measures such as Root Mean Square Error in measuring their model’s efficiency. At the end of their analysis, they had a Root Mean Square Error of 0.342. Another AI research work performed by [15] used financial ratios together with the various popular machine learning algorithms to predict the bankruptcy possibility of some companies in China. After the analysis of their results, random forest and decision tree had the highest accuracy, that is, 95% and 94%, respectively.

The prediction of corporate bankruptcies is an important and widely studied topic since it can have a significant impact on bank lending decisions and profitability. Atiya [16], performed an empirical experiment with a total of 37 ratios which is composed of financial and other nonfinancial ratios and used principle component analysis (PCA) to extract suitable variables. The decision tree (DT) classification methods (C5.0, CART, and CHAID) and logistic regression (LR) techniques were used to implement the financial distress prediction model. Finally, the experiments acquired a satisfying result, which testifies to the possibility and validity of their proposed methods for the financial distress prediction of listed companies that were used as test subjects. Ocal et al. [17] used C5.0 and CHAID decision tree algorithms to estimate the financial failure and/or success of a given manufacturing company, and 35 financial ratios were used as independent variables calculated on the grounds of both company’s annual financial statements and notes from 2007 to 2013. According to the results, the model’s classification accuracy was 90.97% for the training set and 87.5% for the testing set. Thus, the classifications made by the C5.0 algorithm can be considered successful.

Kim and Kang [18] tried to strengthen the predictive accuracy by combining ensemble with neural networks, but other studies have made it obvious that decision trees combined with financial ratios analysis might predict the financial distress of companies more accurately (Kim and Kang, 2010). Apart from the use of decision tree in making predictions, one interesting technique in the event of prediction is neural networks [19]. But this research paper will restrict itself to the use of the decision tree algorithm as the decision tree algorithm is one of the most important classification measures in data mining [20].

3. Methodology

The main purpose of this project is to develop a predictive framework capable of classifying rural banks’ financial status in Ghana. This could not have been accomplished without the necessary steps of data acquisition, data preprocessing, feature selection, and classification. An overview of the framework is shown in Figure 1.

3.1. Study Area. This study used rural banks in Ghana that are registered under the Association of Rural Banks in Ghana as its case study. According to the Bank of Ghana, as of August 2018, the number of rural banks that are legally known is one hundred and forty-five (145) [1]. This research project used the financial ratings of rural banks in Ghana as
collated by the ARB Apex Bank for the performance of rural banks in the various quarters.

3.2. Data Collection. The data were collected from the ARB Apex Bank, Sunyani branch, and it was within the following quarters.

(i) The second quarter ended in June 2014.
(ii) The fourth quarter ended December 2014.
(iii) The fourth quarter ended in December 2015.
(v) The second quarter ended in June 2018.

3.3. Dependent Variable. Our dependent variable is the status of the financial institution’s end of the financial period quarter as to whether it was strong, satisfactory, fair, marginal, or unsatisfactory. If a financial period is marked as strong, satisfactory, fair, marginal, or unsatisfactory, it will continue to remain so until the next financial period proves otherwise. At the end of the preprocessing process, we were left with six hundred and fifty-seven (657) Data Manipulation Units (DMUs), with ninety-nine (99) being strong, three hundred and thirty (330) being satisfactory, one hundred sixty-four (164) being fair, and sixty-four (64) being marginal. Table 1 shows the rating range and interpretation of each financial status.

3.4. Independent Variables. The independent variables or predictors used in the research are the various financial ratios or parameters that are being used by the ARB Apex Bank to assess the performance of their member rural banks. They are in four broad categories, namely, capital, asset including asset quality and asset utilization, earnings/profitability, and liquidity. The broad themes have been detailed into sixteen (16) parameters for the assessment in Table 2.

3.5. Adopted Decision Tree Algorithms. C5.0 was developed by Ross Quinlan in 1994. It works by splitting the sample based on the field that provides the maximum information gain. Each subsample defined by the first split is then split again, usually based on a different field. And the process repeats until the subsamples cannot be split any further. The process repeats until the subsamples cannot be split any further. C5.0 R package version 0.1.2 was developed by Max et al. [21].

\[
H(T) = I_E(p_1, p_2, \ldots, p_J) = - \sum_{i=1}^{J} p_i \log_2 p_i
\]

\[
IG(T,a) = H(T) - \sum_{i=1}^{J} p_i \sum_{a=1}^{J} p(a) - \log_2 p(a)\]

\[
H(T|a) = \sum_{i=1}^{J} p_i \log_2 p_i - \sum_{a}^{J} p(a) \sum_{i}^{J} \log_2 p_i|a). \tag{1}
\]

3.6. Classification and Regression Tree (CART). The Classification and Regression Tree was developed by Breiman et al. [22]. It is a binary decision tree, constructed by splitting the node into two child nodes repeatedly. Classification and Regression Tree (CART) algorithm is a classification algorithm for building a decision tree based on Gini’s impurity index as a splitting criterion. CART is a binary tree built by splitting a node into two child nodes repeatedly. It uses the R package “rpart” version 4.1-13, developed by Therneau et al. [23].

\[
IG(p) = \sum_{i=1}^{J} p_i \sum_{k=i}^{J} p_k = \sum_{i=1}^{J} p_i = (1 - p_i) = \sum_{i=1}^{J} (p_i - p_i^2)
\]

\[
= \sum_{i=1}^{J} p_i - \sum_{i=1}^{J} p_i^2 = 1 - \sum_{i=2}^{J} p_i^2. \tag{2}
\]
C4.5 builds decision trees from a set of training data and the training data is a set \( S = \{ s_1, s_2, \ldots \} \) of already classified samples. The attribute with the highest normalized information gain is chosen to make the decision. C4.5 goes back through the tree once it is created and attempts to remove the branches that do not help by replacing them with leaf nodes. It uses the RWeka package, version 0.4–39, developed by Hornik et al. [24], and contains the J48 function for executing C4.5 models.

3.7. Selecting the Most Important Variables with the Random Forest Algorithm. Random forest algorithm is very popular for feature selection in data science because of its tree-based strategy that is naturally ranked by how good they add up to the purity of the nodes. Nodes that contribute greatly to the purity of the tree are known or seen as the most important ones. Reference [25] proposed a technique for feature selection using the mean decrease accuracy (MDA) and mean decrease gini (MDG) in random forest. This technique scores the MDA and MDG of variables, sum up the scores, rank the total scores of the variables in decreasing order, and runs the random forest algorithm again with the top 50% variables.

In this paper, we used a similar technique like the one proposed by [25]. We focused on the MDG and set our threshold to 0.10; that is, all the variables that gave MDG less than 0.10 were eliminated and considered less relevant. After identifying the most relevant variables, a new dataset was created for comparative purposes.

3.8. The Proposed Model. The proposed DT model for classifying rural banks’ financial status is shown in Figure 1.

3.9. Discussions and Findings

(i) Initially, all sixteen (16) predictors were used to build the random forest model, and upon the summary of the model, it was observed that thirteen (13) of the predictors (i.e., the predictors with asterisks (*) in Table 3) had an MDG greater than or equal to 0.10, making them the most relevant for the model, whereas the other three were indicated to be less relevant.

(ii) The model with all the predictors came out with an error rate of 10.1% and an execution time of 2.11 seconds, whilst the model with the thirteen predictors came out with an error rate of 11.05% and an execution time of 1.75 seconds.

(iii) Upon building the C5.0 models using both the dataset with all the predictors and the one with the eleven predictors, the models came out with the same results. This can be seen in Figure 2 as they both yielded the same confusion matrix and statistics. The models showed a 100% accuracy and also showed a 100% accuracy when the test data was used for prediction (i.e., all 53 banks with fair status, all 14 banks with marginal status, all 34 banks with strong status).

(iv) Upon building the C4.5 model with the training dataset containing all the predictors, the model came out with a 94.12% accuracy and also showed 82.83% when the test data was used for prediction (i.e., only 39 out of 53 banks with fair status, only 10 out of 14 banks with marginal status, only 89 out of 97 banks with satisfactory status, and only 26 out...
Table 3: Relevant (*) and less relevant variables after building the MLR model.

<table>
<thead>
<tr>
<th>Independent variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paid-up capital</td>
</tr>
<tr>
<td>Capital adequacy ratio (CAR)*</td>
</tr>
<tr>
<td>Loss/NPL ratio*</td>
</tr>
<tr>
<td>Other assets to total assets ratio*</td>
</tr>
<tr>
<td>Property, plant, and equipment to total assets ratio*</td>
</tr>
<tr>
<td>Return on earning assets*</td>
</tr>
<tr>
<td>Costs to income ratio*</td>
</tr>
<tr>
<td>Past due ratio*</td>
</tr>
<tr>
<td>Liquid assets to total assets ratio*</td>
</tr>
<tr>
<td>Return on assets (ROA)*</td>
</tr>
<tr>
<td>Average primary reserve ratio for the quarter*</td>
</tr>
<tr>
<td>Nonperforming loans (NPL) ratio*</td>
</tr>
<tr>
<td>Advances to loanable funds ratio</td>
</tr>
<tr>
<td>Return on equity (ROE)*</td>
</tr>
<tr>
<td>Average secondary reserve ratio for the quarter*</td>
</tr>
</tbody>
</table>

Table 4: Statistical information on the various adapted DT algorithms.

<table>
<thead>
<tr>
<th>Decision tree algorithm</th>
<th>Accuracy</th>
<th>P value</th>
<th>Kappa</th>
<th>No information rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset with all initial predictors</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C5.0</td>
<td>1</td>
<td>2.2e − 16</td>
<td>1</td>
<td>0.4899</td>
</tr>
<tr>
<td>C4.5</td>
<td>0.9412</td>
<td>2.2e − 16</td>
<td>0.9094</td>
<td>0.5076</td>
</tr>
<tr>
<td>CART</td>
<td>0.8787</td>
<td>2.2e − 16</td>
<td>0.8078</td>
<td>0.4949</td>
</tr>
<tr>
<td>Dataset with the 13 predictors</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C5.0</td>
<td>1</td>
<td>2.2e − 16</td>
<td>1</td>
<td>0.4646</td>
</tr>
<tr>
<td>C4.5</td>
<td>0.9387</td>
<td>2.2e − 16</td>
<td>0.9038</td>
<td>0.4646</td>
</tr>
<tr>
<td>CART</td>
<td>0.8131</td>
<td>2.2e − 16</td>
<td>0.7048</td>
<td>0.4949</td>
</tr>
</tbody>
</table>
of 34 banks with strong status were predicted correctly). Using 10-fold cross validation, the model gave a Kappa value of 0.6908. More details about the C4.5 models with all sixteen predictors can be seen in Table 3.

(v) Upon building the C4.5 model with the training dataset, containing the 13 predictors, the model came out with a 93.87% accuracy and also showed an 83.84% accuracy when the test data was used for prediction (i.e., only 40 out of 53 banks with fair status, only 12 out of 14 banks with marginal status, only 81 out of 97 banks with satisfactory status, and only 33 out of 34 banks with strong status were predicted correctly). More details about the C4.5 models with the eleven predictors can be seen in Table 3.

(vi) Upon building the CART model with the training dataset containing all the predictors, the model came out with an 87.87% accuracy and also showed 87.88% when the test data was used for prediction (i.e., only 46 out of 57 banks with fair status, only 15 out of 18 banks with marginal status, only 94 out of 98 banks with satisfactory status, and only 19 out of 24 banks with strong status were predicted correctly). More details about this model can be seen in Table 3. Cross-validating this model, the CV model gave the best CP of 0 but came out with an accuracy less than the initial CP of 0.01 (this can be seen from the graph at the left side of Figure 2), so the final model was based on the default CP.

(vii) Upon building the CART model with the training dataset containing the eleven predictors, the model came out with an 81.31% accuracy when the test data was used for prediction (i.e., only 37 out of 57 banks with fair status, only 19 out of 18 banks with marginal status, only 91 out of 98 banks with satisfactory status, and only 14 out of 24 banks with strong status were predicted correctly). Cross-validating this model, the CV model gave the best CP of 0 but came out with an accuracy less than the initial CP of 0.01 (this can be seen from the graph at the right side of Figure 2), so the final model was based on the default CP.

(viii) The random forest model with all the initial predictors had an error rate negligently less than that of the one with the thirteen predictors. However, the second model attained a much better execution time as a result of the trade-off. Moreover, with the most relevant financial ratios unravelled by the random forest model, rural banks in Ghana can focus more on getting better results on those ratios so as to attain a better financial standing.

(ix) Figure 3 is a chart showing the accuracies obtained from all the models.


