**International Journal of Computing and Digital Systems**
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 10, No.1 (Oct.-2021)

http://dx.doi.org/10.12785/ijcds/100193

# Boosting Algorithms to Analyse Firm's Performance Based on Return on Equity: An Explanatory Study

**Elsayed A. Elamir**[1]

[1] *Department of Management & Marketing, University of Bahrain, Kingdom of Bahrain*

*Received 08 Jul. 2020, Revised 22 Jul. 2021, Accepted 05 Aug. 2021, Published 28 Oct. 2021*

**Abstract:** This study aims to use the boosting algorithms especially gradient boosting and its extension extreme gradient boosting in predicting firm performance in terms of return on equity that could be considered as a measure of profitability and to use the partial dependent plot and local interpretable model-agnostic explanation techniques to explain the model and its prediction. The models are evaluated using R-squared, root mean square error, and mean absolute error. The global interpretations in terms of partial dependent plot and local interpretations in terms of local interpretable model-agnostic explanations are performed to interpret the prediction for any individual or group of cases. The results show that the extreme gradient boosting is improving the model by about 39% for training set and about 4% for testing set in terms of R-squared. Interesting results are given by the partial dependent and local model-agnostic explanation plots where they are suggesting that the total assets, the total liability and the board size have the most effect on the predicting and interpreting return on equity. By taking over-fitting in consideration the gradient boosting model is a better choice than extreme gradient boosting. The important scores suggest that the total liability, beta coefficients and the total assets have the most effect on return on equity.

**Keywords:** Business Analytics, Financial Management, Global Model, Gradient Boosting, Machine Learning, R-squared.

## 1. Introduction

In recent years, there are a lot of concerns to have accurate and reliable prediction of earnings, growth and firm performance especially with the huge advancements in machine learning algorithms and the availability of large data. These predictions not only measure financial performance of a firm but also helping financial and operation managers with investment, production and financing decision making and outside investors to understand the performance of the firm; see, [1], [2], [3], [4], [5], [6] and [7].

Return on equity (ROE) is calculated as a net profit after tax over the total shareholder's equity. This quantity measures the shareholders rate of return on their investment in the firm. In other words, it can be considered as a profitability quantity which is used to evaluate the effectiveness of the firm in creating profits that are the rights of capital owners. With this respect, the banks must hold capital to prohibit bank failure and to deal with the capital requirements put by the regulatory authorities. On the other hand, the banks do not need to keep too much capital because this will reduce the returns on equity holders; see, [8], [9], [10], and [11].

Multiple linear regression (MLR) is a well-known model that uses one equation to build model over the complete data space. When the assumptions of multiple linear regression model do not satisfy such as nonlinearity and the existence of interactions, the estimates and predictions are under severe limitations; see, [12] and [13].

Reference [14] developed gradient boosting machine method (Gboost) that included both regression and classification problems. The basic principles are that if a loss function (such as squared error) and a weak learner (such as regression trees) are given, the method looks for an additive model that minimize the loss function. The method is guessing the best value for the response (such as the average). The gradient (residual) is computed, then the model is fit to the gradients to minimize loss function. The present model is summed to old one, and the steps continue untill a specified criterion (number of iterations); see [13] and [15]. Extreme gradient boosting method (Xgboost) is an advanced application of Gboost to overcome overfitting in Gboost and introducing more accuracy and scalability over simple algorithms. Xgboost supports several types of objective functions including regression; see, [16].

*E-mail: shabib@uob.edu.bh*

The aims of this study are twofold. The first aim is to predict return on equity ("ROE") which depends on the logarithm of total assets ("logTOTA"), liability ("LIAB"), ratio of market to book value ("BOKV"), Beta coefficient ("SVOL"), firm's age ("AGEB"), cash availability ("NCSH") and board size ("BSZ") using the multiple linear regression, the gradient boosting and extreme gradient boosting methods. The root mean square error (RMSE) and R-squared are used to compare the model performance among all methods. The second aim is to use the partial dependent plot and local interpretable model-agnostic explanation techniques to explain the model and prediction in interpretable and faithful ways where the users need to be confident that the model will perform well on real datasets. According to ([17] p. 1) "Understanding the reasons behind predictions is, however, quite important in assessing trust, which is fundamental if one plans to act based on a prediction, or when choosing whether to deploy a new model. Such understanding also provides insights into the model, which can be used to transform an untrustworthy model or prediction into a trustworthy one". To satisfy these two aims, sixty-three banks are selected from the bourse in 8 "Middle East and North Africa" ("MENA") countries for the period 2009 to 2018 that has 630 observations. The data from 2009 to 2017 is used to build the training model. This training model is used to predict and interpret data in year 2018.

This study is organized as follows. The literature review is illustrated in Section 2. In section 3 the methodology is outlined. The results and discussion are presented in Section 4. The conclusion is given in Section 5.

## 2. Literature Review

Return on equity and return on assets are the most profitability measures vastly used as overall measures of corporate financial performance; see, [18]. Reference [10] mentioned that ROE could be consider as the most ratio an investor should take it into account. In fact, ROE clarifies the result of "structured financial ratio analysis (Du Pont analysis)"; see, [19], [9] and [20].

Reference [21] discussed the link between the performance indicators of market and accounting indicators using 40 public firms from Jodan using years 1984 to 1996. The results indicated that a significant positive relation between return on equity and the market price per share. Reference [11] examined the relation between ROE, ROA (return on assets) and ROI (return on investment) with Jordanian insurance public firms" share prices" from 2002 to 2007. They found that a positive relationship between ROE, ROA and ROI together with" share prices" while the results illustrated no relations between the ROE separately with market" share prices" for Jordanian insurance public firms. Reference [22] developed a predictive model using extreme gradient boosting to predict a bankruptcy with several economic measures and permits to expect a financial condition of a company. They applied extreme gradient boosting to Polish firms from 2007 to 2013 (bankrupt) and from 2000 to 2012 (operating) and concluded that the extreme gradient boosting produced significant increase in the prediction quality.

Reference [23] applied extreme gradient boosting to predict default of bank in banking sector in U.S. They used annual chains of 30 financial ratios for 156 commercial banks from years 2001 to 2015. The results indicated that the chance of bank financial distress is increased by pretax return on assets, lower values for kept gains to mean equity and total risk-based capital ratio. Reference [24] applied extreme gradient boosting to characterize a group of main indicators that could assist in forecasting and prohibiting failure of banks in the Eurozone. They used 25 annual ratio series from commercial banks of 2006-2016 period. They built classification model to decide about the key variables that causes bank defaults. The results indicated that the bank managers should follow the most important variables in the study, such as net loan to total asset and equity to liability and taking early action rather than waiting for government action.

Reference [25] offered a reference to investors and creditors for taking investment decision by making financial prediction of operating revenue, earning per share, cash flaw and net working capital using multivariate adaptive regression spline and queen genetic algorithm-support vector machine to make. Reference [26] examined the ability of machine learning algorithms to improve the prediction of the sign of earnings changes and its usefulness for return forecasting. They concluded that there were 62.3 percent prediction accuracy using stepwise logit regression and 76.8 percent out of sample accuracy using random forest method while elastic net method performed similarly to stepwise logit method.

Reference [27] applied extreme gradient boosting method to forecast systemic banking crises. The results showed that the extreme gradient boosting outperformed the existing methods in terms of the predictive power. They considered" being the demand for deposits"," the level of domestic credit" and" banking assets" are the most important variables to explain the causes that produce systemic bank crises. Reference [28] studied the automatically detect fraudulent claims and group them into several fraud types. The results showed that a high-performance gain obtained by Xgboost in revealing and grouping fraudulent claims compared to other machine learning algorithms.

This study extends the previous studies to MENA countries where it makes the following contributions.

This study could be considered the first in the MENA area that apply the gradient boosting approaches to predict banks' performance based on return on equity. Second, characterizes four main variables, namely, the total assets, the total liability to total assets and the stock volatility and the board size that may help the bank manager to anticipate and increase financial stability for the bank. Third, fill the gap between the prediction and the interpretations or knowing reasons behind this prediction on the global and local levels.

## 3. Methodology

### A. Study variables

The data is chosen from 8 MENA nations, namely, "Egypt", "Jordan", "Qatar", "Oman", "Saudi Arabia", "Kuwait", "Bahrain" and "United Arab Emirates" from "2009 to 2018" across 63 banks in all countries that gives a total of 630 observations. Since there is homogeneity between these nations in terms of cultures and assets, they are chosen in the sample. The internet sites of the enrolled banks in the stock market are utilized to assemble the financial information. To satisfy study aim, eight variables are chosen. Return on equity ("ROE") as response variable and seven predictor variables, logarithm of total assets ("logTOTA"), liability ("total liabilities to total assets- LIAB"), market to book ratio ("bank book value to its market value - BOKV"), Beta coefficient ("stock volatility with respect to the market - SVOL"), bank's age ("AGEB"), cash availability ("net cash of the bank - NCSH") and board size ("BSZ"). These variables can be defined as: ROE: return on equity is computed by dividing net income by equity of shareholders and may be considered as a measure of how effectively firm manager is using assets to make profit, TOTA: the total assets are the total of fixed and current assets in the firm balance sheet, LIAB: Liability is computed as assets minus equity of firm, BOKV: market to book ratio is computed as market cap to common shareholder equity, SVOL: Beta coefficient computed as the covariance between security returns and the market returns over the variance of the market return for a specified time, AGEB: bank age since it is established. NCSH: cash availability is the difference between firm total cash and total liabilities, BSZ: board size is the number of people in the board that includes executive and non-executive directors.

### B. Feature selection

To identify the key features necessary to predict the return on equity ("ROE"), Boruta algorithm is used to characterize and test important variables that are statistically significance. Boruta is a feature standard information system to judge which of the variables or features are statistically significant and which are not. Boruta uses what is called shadow features which are reduplicates of original variables but with random selected values so that the distribution of them stays the same yet their importance is sponge out; see, [29] and [30]. The variable will be retained if its distribution above the distribution of shadow features.

TABLE I. THE IMPORTANCE AND FEATURE SELECTIONS RESULTS USING 100 RUN OF BORUTA ALGORITHM

| Variables | Mean Imp | Median Imp | Min Imp | Max Imp | Decision |
|---|---|---|---|---|---|
| logTOTA | 16.84 | 16.91 | 14.54 | 19.01 | Confirmed |
| LIAB | 12.99 | 13.10 | 11.50 | 15.38 | Confirmed |
| BOKV | 7.21 | 07.51 | 04.35 | 09.93 | Confirmed |
| SVOL | 5.38 | 05.66 | 03.19 | 06.87 | Confirmed |
| AGEB | 7.81 | 08.18 | 06.64 | 08.64 | Confirmed |
| NCSH | 6.14 | 06.01 | 04.51 | 08.44 | Confirmed |
| BSZ | 8.15 | 08.10 | 05.58 | 10.06 | Confirmed |

(*) Important: Imp

Table I gives the results of 100 run of Boruta algorithm to select the feature in ROE model. The method had confirmed or retained all the variables. Moreover, Figure 1 displays boxplots to illustrate the distribution of feature importance over Boruta run. The green color indicates the confirmed or retained variables while the blue color reflects the distribution of the importance of worst, average and best shadow in each iteration. From Table I, it can be noted that all the independent variables are confirmed to be selected in our models because they are significant at the level of 0.01. In other words, all five independent variables namely logTOTA, LIAB, BOKV, SVOL, AGEB, NCSH and BSZ, are helping in predicting ROE.

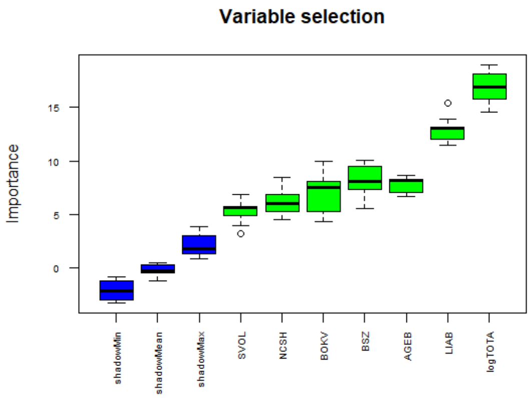


Figure 1. the boxplot of feature selections based on 100 runs of Burota algorithm

*C. Analysis methods*

The multiple linear regression, gradient boosting and extreme gradient boosting are explained briefly.

*1) Multiple linear regression*

The minimization of sum squares errors between original values, $y_i$, and estimated values, $\hat{y}_i$ is considered as the objective of multiple linear regression

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where

$$y_i = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \varepsilon_i$$

$\beta_i$ are the parameters, $y_i$ the response, $x_i$ the predictors and $\varepsilon_i$ are the errors; see, for example, [13] Kuhn and Johnson (2013).

*2) Gradient boosting*

The gradient boosting methods are proposed by [14]. Following the procedures in [15], [31], [27] A and [32], for a given training data $D\{x_i, y_i,\ i = 1, \ldots, N\}$ the gradient boosting is an ensemble of $K$ classification and regression trees (CART), the model is trained by the minimization of the objective function

$$Obj(\theta) = L(\theta) + \Omega(\theta) = \sum_{i} l(y_i, \hat{y}_i)) + \sum_{k=1}^{K} \Omega(f_k)$$

Where $\theta$ are the parameters, $\Omega$ is a regularization term and $L$ is the loss function, $K$ is the numbers of decision trees, $\hat{y}_i = \sum_k f_k(x_i)$, $f_k \in F$ are $K$ decision trees and $f_K(x) = w_{q(x)}$ where $q(x)$ is the function that gives the independent path in the structure of the tree. The regularization term can be expressed as

$$\Omega(f) = \gamma T + 0.5\lambda \sum_{t=1}^{T} w_t^2$$

Where $\lambda$ and $\gamma$ are the parameters of the regularization part and $T$ is the number of leaves.

In regression task a classic loss function is the squared error loss ($L_2$) as

$$L(y, f) = 0.5(y - f)^2$$

The derivative is the error $y - f$ that imply that the gradient is just the residuals; see [13].

According to [31], the training algorithms using Xgboost can be summed up as follows.

1. For every predictor: order the values and obtain the best dividing value (min. RMSE).
2. Select the predictor with the best dividing value that optimizes the training target.
3. Continue dividing until obtain the specified maximum depth of the tree.
4. Specify the prediction value to the leave and prune it.
5. Iterate these procedures in a collective way until the fixed number of trees is obtained.

Therefore, the prediction $y$ at step $t$ is

$$\hat{y}_i^{(t)} = \sum_{k=1}^{K} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)$$

The simplified objective function can be written as

$$Obj(\theta)^{(t)} = \sum_{j=1}^{T} \left[ \left( \sum_{i \in I_j} g_i \right) w_j + 0.5 \left( \sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T$$

Where $g_i = \partial_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$ and $h_i = \partial^2_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$ are the first and second derivative for loss function and $I_j = \{i | q(x_i) = j\}$ is the type set of leaf $j$, $w_j$ the optimal leaf weight for a known structure $q(x)$. For more details, see, [31]. For most practical tasks, there is an empirical evidence that "the simple additive models corresponding to the first term of the analysis of variance decomposition provide good results" ([33; [34]).

The hyperparameters in Gboost are: trees number (optimal number of trees that reduce the loss function using cross validation), trees depth (split numbers in each tree that control the complexity of boosted ensemble), rate of learning (how the tree proceeds down the gradient descent called shrinkage) and subsampling (control use a fraction of the available training observations).

According to [35], Xgboost has regularization unlike Gboost that helps in reducing overfitting where it is known as “regularized boosting” technique. The hyperparameters are eta (controls the learning rate), gamma (minimum loss reduction to do more partition on a leaf node of the tree), max depth (tree depth), min child weight (minimum number of values required in each terminal node and subsample (percent of training phase to sample for each tree).

### *D. Interpretations*

There are two types of interpretations: (a) global meaning: assist to recognize the inputs and the whole model relationship with the response variable. The most known ways are variable important measures and partial dependent plot and (b) local meaning: help to understand the predicted values for specified row(s) of data.

Variable importance reflects the overall contribution of each predictor variable to the forecasting of a machine learning model where they compute the value of a variable that has relationship with the dependent as emulated to other variables used in the model; see, [13] Kuhn and Johson (2013). Partial probability plot is introduced by [14] Friedman (2001) to interpret the dependency of several input features to the predictions by plotting the effect of changing a specific input feature over its marginal distribution on the predicted values with holding other variables fixed; see, [12] Hastie et al. (2008).

Local Interpretable Model-agnostic Explanations (LIME) is a graph method that aids in explanations of individual predictions. The idea of LIME it is likely to fit a simple model around a single value that will imitate how the global model pursues at that locality where it assumes that “every complex model is linear on a local scale”. Then, the simple model could be used to interpret the forecasting of the more complex model locally; see [17] Ribeiro et al. (2016).

## 4. THE RESULTS

The ROE-model could be written as

$$ROE = f(logTOTA, LIAB, BOKV, SVOL, AGEB, NCSH, BSZ)$$

In case of linear relationship

$$ROE = \beta_0 + \beta_1 logTOTA + \beta_2 LIAB + \beta_3 BOKV + \beta_4 SVOL + \beta_5 AGEB + \beta_6 NCSH + \beta_7 BSZ + \varepsilon_i$$

Where $\beta_0$ , $\beta_1$ , $\beta_2$ , $\beta_3$ , $\beta_4$ , $\beta_5$ , $\beta_6$ and $\beta_7$ are the parameters in the model, $\varepsilon_i$ are the errors.

Data from year 2008 to year 2016 (training set) is used to predict ROE in year 2017 (testing set). The training set is composed of 567 values and the testing set is composed of 63 values. In the case of training and testing sets, root mean square error (RMSE), R-squared and mean absolute error (MAE) are used to assess the overall model. Note that the less values for RMSE and MAE mean better model performance while large value for R-squared means better model performance.

References [36], CARET package ([37], [38]), gbm package ([39]), xgboost package ([25]) and lime package ([40] are used to analyse the data and to obtain all the results in this study. References [41] and [42] indicated that recurring $k$-fold cross validation can be applied to increase the accuracy of the estimates while preserving less bias. All methods in this study are used the function train in Caret package with 5 repeats of 10-fold cross validation; see, [13].

### *A. Multiple linear regression (MLR)*

Table II illustrates the results of multiple linear regression analysis. Since “p-value” for “F-statistics” is zero, the regression model is significance at all common levels (0.01, 0.05 and 0.10). From column “p-value” it can note that the variables “logTOTA”, “LIAB”, “SVOL”, “AGEB” and “BSZ” are significance at 0.05 level of significant. The variables “logTOTA”, “LIAB” and “BSZ” have a positive effect on “ROE” while “SVOL” and “AGEB” have a negative effect on ROE. The variables “BOKV” and “NCSH” are not significant.

TABLE II. MLR ANALYSIS FOR ROE MODEL

| Term | Coeff. | Std. Error | t-stat | p-value | F-stat |
|---|---|---|---|---|---|
| Intercept | -7.773 | 3.667 | -2.120 | 0.034* | F=10.67 |
| logTOTA | 3.532 | 0.656 | 5.380 | 0.000*** | p-val=0 |
| LIAB | 1.025 | 0.204 | 5.018 | 0.000*** | $R^2 = .12$ |
| BOKV | -2.159 | 1.395 | -1.548 | 0.122 | |
| SVOL | -1.396 | 0.679 | -2.056 | 0.040* | |
| AGEB | -0.104 | 0.037 | -2.803 | 0.005** | |
| NCSH | 0.0001 | 0.0001 | 0.633 | 0.527 | |
| BSZ | 0.6075 | 0.2278 | 2.666 | 0.008** | |

Note that “Coeff: ceoefficients, Std: standard, stat: statistics, val: value, (***) significance at 0.001, (**) significance at 0.01, (*) significance at 0.05 and (.) significance at 0.10”

Table III shows the variable importance for ROE model for all methods. FOR MLR method logTOTA is at the top of important metric. The score started to decrease with LIAB, AGEB, BSZ, SVOL and NCSH. Consequently, logTOTA and LIAB have the most influence on the prediction of ROE while the less importance variables are BOKV and NCSH.

TABLE III. MLR, GBOOST AND XGBOOST IMPORTANT SCORES FOR EACH VARIABLE

| MLR | | Gboost | | Xgboost | |
|---|---|---|---|---|---|
| Variable | Score | Variable | Score | Variable | Score |
| logTOTA | **5.38** | LIAB | **0.379** | LIAB | **0.466** |
| LIAB | **5.02** | SVOL | **0.252** | logTOTA | 0.143 |
| AGEB | 2.80 | logTOTA | 0.113 | SVOL | 0.136 |
| BSZ | 2.67 | BOKV | 0.093 | AGEB | 0.086 |
| SVOL | 2.05 | AGEB | 0.080 | BOKV | 0.075 |
| BOKV | 1.55 | NCSH | 0.052 | NCSH | 0.052 |
| NSCH | 0.63 | BSZ | 0.029 | BSZ | 0.040 |

Table IV shows the performance results for multiple linear regression. The RMSE is 11.087 for training set and it is 7.736 for testing set. With respect to R-squared, it is 11.8% for training set and increases to 21.7% for testing set. For MAE, it is 6.735 for training set and decreases to 5.666 for testing set.

TABLE IV. PERFORMANCE METRICS FOR ROE MODEL USING MLR, GBOOST AND XGBOOST METHODS

| | Data | |
|---|---|---|
| | Training | Testing |
| | MLR | |
| RMSE | 11.087 | 7.736 |
| R-square | 0.118 | 0.217 |
| MAE | 6.735 | 5.666 |
| | Gboost | |
| RMSE | 8.001 | 7.477 |
| R-square | 0.585 | 0.304 |
| MAE | 4.768 | 5.150 |
| | Xgboost | |
| RMSE | 1.782 | 7.033 |
| R-square | 0.979 | 0.381 |
| MAE | 1.287 | 5.027 |

*B. Gradient boost (Gboost)*

Figure 2 displays the results of the tuning parameters for Gboost algorithm over 567 cases and 7 features. The best model parameters are selected based on the minimum value for RMSE. The last values that utilized for the ROE model are the number of trees is 100, the interaction depth is 3, the shrinkage is 0.01 and the minimum observation in node is 10.

Table 2 shows the variable importance for ROE model using Gboost method. LIAB, logTOTA and SVOL are the most variables affecting the prediction of ROE. The importance scores start decreasing with AGEB, BOKV, NCSH and BSZ. The performance results for Gboost are shown in Table 3.

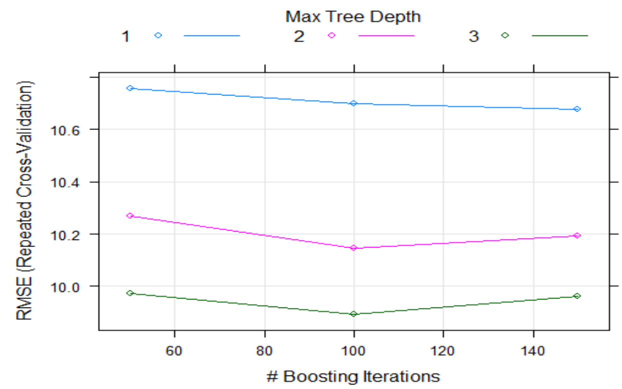


Figure 2. Tune length for ROE model using Gboost method

The RMSE is 8.001 for training set while it decreases to 7.477 for testing set. R-squared is 58.5% for training set and decreases to 30.4% for testing set. MAE is 4.768 for training set and increases to 5.150 for testing set.

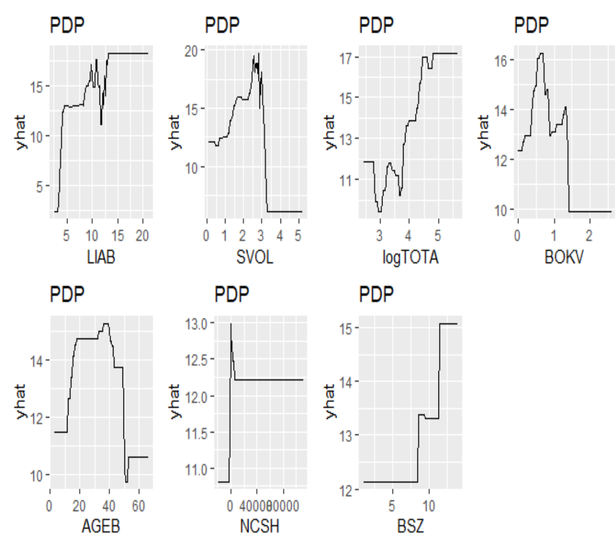


Figure 3. Partial dependent plot (PDP) for ROE model using Gboost method

Figure 3 displays the partial dependent plot for ROE model using Gboost method. This plot shows the changes in the average predicted value of ROE with a given feature while holding other variables constant. The advantage of this plot is that it reflects the changes in the predicted value across the whole range of the independent variables. For example, it can divide range of AGEB variable to two intervals, from 0 to about 40 where there is trend up in the average of the predicted value while for more than 40 there is trend down in the average of predicted value. On average, it can see that the variables logTOTA, LIAB, NCSH and BSZ will increase the average of the predicted value. In addition, on average the variables SVOL, AGEB and BOKV will decrease the average of the predicted value.

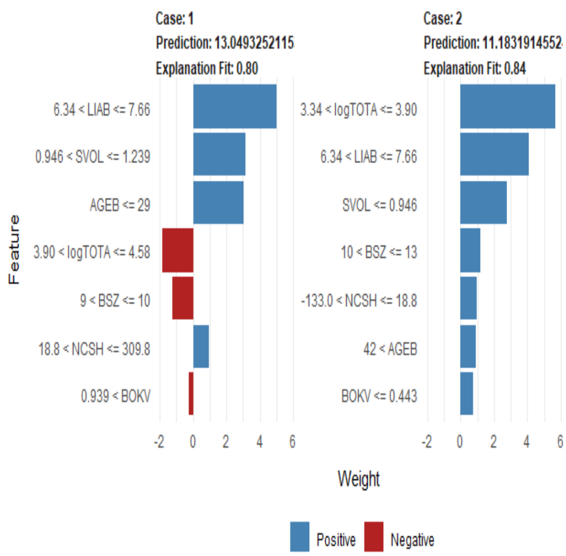


Figure 4. LIME plot for ROE model using Gboost method

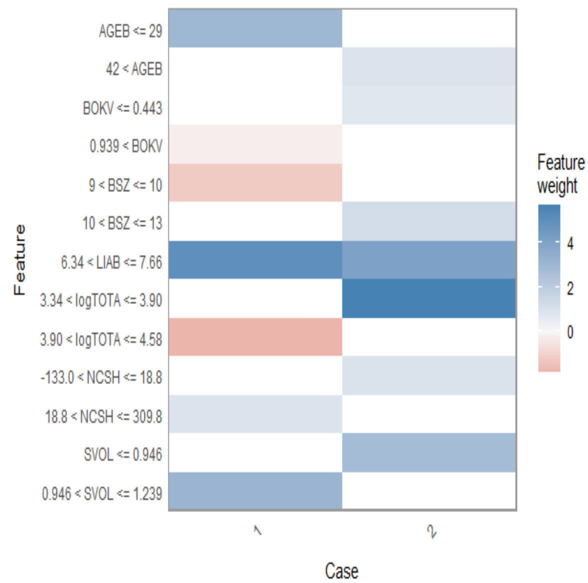


Figure 5. Heatmap for ROE model using Gboost method

Figure 4 displays the LIME plot for cases 1 and 2 (first two years) using Gboost method. This figure shows an individual plot for the first two cases from testing data and provide the predicted value for each case based on the 7 features that interpret the linear model in the local region for this observation and whether the feature brings an increase (positive) or decrease (negative) in average predicted value for ROE. It also gives the model fit for each case to show how well the model illustrates the local region. Therefore, it can be inferred that case 1 has a good explanation fit about 80% with a positive support for 6.34<LIAB<=7.66, SVOL>0.946, AGEB<=29 and 18.8<NCSH<=309.8 and negative support from 3.9<logTOTA<=4.58 and 9<BSZ<=10. For case 2, it has high explanation fit about 84% with positive support from all features especially 3.34<logTOTA<=3.9, 6.34<LIAB<=7.66, SVOL<=0.946, 10<BSZ<=13, -133<NCSH<=18.8, AGEB>42 and BOKV<=0.433.

Figure 5 displays the heatmap for ROE model using Gboost method. This plot is showing the selection of different variables across all the observations that influence each case. In other words, it is very useful plot to find out common features which affect all observations. For example, in case 1, AGE <= 29 and 0.946<SVOL<= 1.239 has the same positive effect on the average predicted value. Similarly, 9<BSZ<=10 and 3.90<logTOTA<=4.58 has the same negative effect on the average predicted value.

### *C. Extreme gradient boosting (Xgboot)*

Figure 6 displays the results over the tuning parameters for 567 cases and 7 features. The best model parameters are selected based on the less value for RMSE. The last values are the number of rounds=150, the maximum depth=3, eta=0.3, gamma=0, the column sample by tree = 0.8, the minimum child weight=1 and sub-sample=1.

Table 2 shows the variable importance for ROE model using Xgboost. The LIAB, logTOTA and SVOL are the most variables affecting the prediction of ROE. The importance scores start to decrease with AGEB, BOKV, NCSH and BSZ. Table 3 shows the performance results for ROE model using Xgboost. The RMSE is 1.782 for training set and it increases to 7.033 for testing set. R-squared is 97.9% for training set and decreases to 38.1% for testing set. MAE is 1.2877 for training set and increases to 5.027 for testing set.

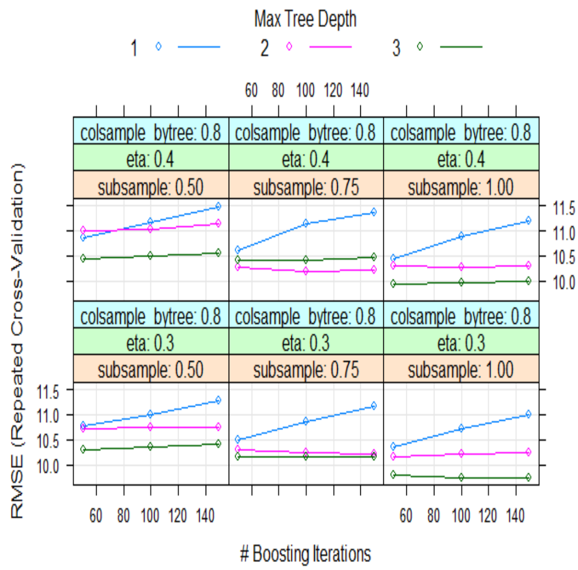


Figure 6. Tune length for Xgboost

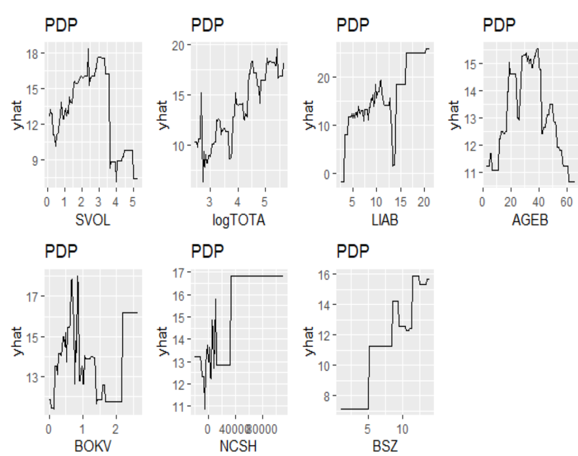


Figure 7. Partial dependent plot (PDP) for ROE model using Xgboost method

Figure 7 displays the partial dependent plot for ROE model. This plot shows the changes in the average predicted value of ROE with a given feature while holding other variables constant. The advantage of this plot is that it reflects the changes in the predicted value across the whole range of the independent variables. For example, it can divide range of AGEB variable to two intervals, from 0 to about 35, there is trend up in the average of predicted value while for more than 35 there is trend down in the average of predicted value. On average, it can see that the variables logTOTA, LIAB, NCSH and BSZ will increase the average of the response variable. In addition, on average the variables SVOL, AGEB and BOKV will decrease the average of the response variable.

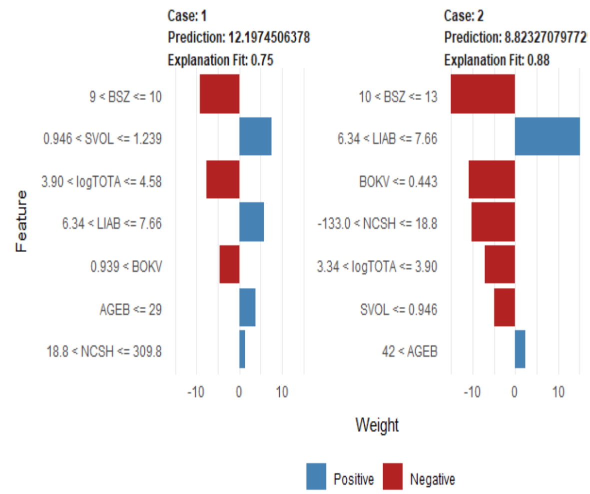


Figure 8. LIME plot for ROE model using Xgboost method

Figure 8 displays the LIME plot for cases 1 and 2 (first two years). This figure shows an individual plot for the first two cases from testing data and provide the predicted value for each case based on the 7 features that clarify the linear model in the local region for this observation and whether the feature brings an increase (positive) or decrease (negative) in average predicted value for ROE. It also gives the model fit for each case to see how well the model interprets the local region. Therefore, it can be inferred that the case 1 has a good explanation fit about 75% with negative support from 9<BSZ<=10, 3.90<logTOTA<=4.58, and BOKV>0.939 and positive support from 0.946<SVOL<=1.239, 6.34<LIAB<=7.66 , AGEB<=29 and 18.8<NCSH<=309.8. Similarly, the case 2 has a high explanation fit about 88% with positive support from 6.34<LIAB<=7.66, and AGEB>42 and negative support in 10<BSZ<=13, BOKV<=0.433, -133<NCSH<=18.8, 3.34<logTOTA<=3.9 and SVOL<=0.946.

Figure 9 displays the heatmap for ROE model. This plot is showing the selection of different variables across all the values that influence each case. In other words, it is very useful plot to find out common features which affect all observations. For example, in case 1, AGE <= 29, 6.34<LIAB<= 7.66 and 0.946<SVOL<= 1.239 has the same positive effect on the average predicted value. Similarly, BOKV<= 0.443 and -1.33<NCSH<=18.8 has the same negative effect on the average predicted value.

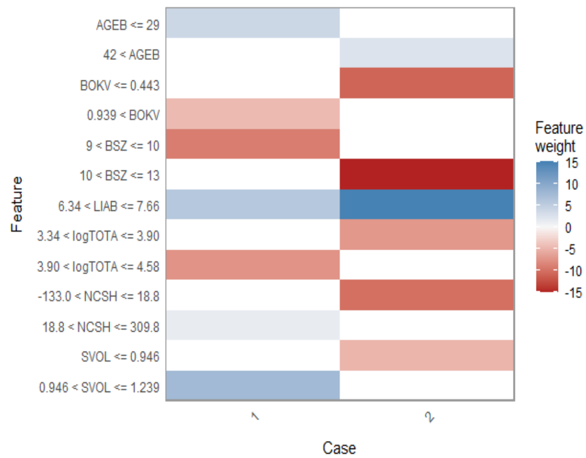


Figure 9. Heatmap for ROE model using Xgboost method

### D. Discussion

The performance metric results are collected in Table 3. In terms of RMSE and training set the preferable method is Xgboost 1.782, pursued by Gboost 8.001 and regression 11.087. This means that Xgboost improves the results by at least 6.219 (8.001-1.782). With respect to testing set, the preferable method is Xgboost 7.033, followed by Gboost 7.477 and regression 7.736. This means that Xgboost improves the results by 0.444 (7.477-7.033).
Regarding R-squared and training set, the preferable method is Xgboost 97.9%, followed by Gboost 58.5% and regression 11.8%. This means that the Xgboost improves the results by at least 39.4% (97.9%-58.5%). With respect to testing data, the best method is Xgboost 35.1%, followed by Gboost 30.4% and regression 21.7%. This means that the Xgboost improves the results by 4.4% (35.1%-30.7%). In the same way, it can rank models with respect to MAE. Prior studies such as De Graph (2017) who used support vector machine and fuzzy fingerprint to predict financial performance of firm using return on equity. The accuracy for the support vector machine and fuzzy fingerprint were 70.8% and 83.3%, respectively, while our results for the gradient boosting and extreme gradient boosting in terms of R-squared were 58.5% and 97.9%, respectively. The Xgboost achieves accuracy superior to other methods. In addition, the results of this study are consisted with the results of Balakrishnan et al. (2010) who considered the total assets and mark to book ratio as important variables in building predictive models.
By looking at the error Table IV, there is suspect for over-fitting where Xgboost has test error (7) much higher than the training error (1.78). To solve this, it might consider Gboost model is a better choice where it has the test error 7.5 which is very close to the training error 8.

## 5. Conclusion

Three methods of machine learning, namely, the multiple linear regression, the gradient boosting and the extreme gradient boosting are investigated to be used in predicting and interpreting firm performance based on return on equity in MENA countries. These methods are utilized data from 8 countries in MENA area. Sample of 63 banks is selected to give a total of 630 cases over 10 years period. To validate the models, the data are divided into the training data that included first 9 years to train the models and the testing data that included the last year to test the models.
The results suggested that the extreme gradient boosting method is outperformed the multiple linear regression and gradient boosting methods. In case of using extreme gradient boosting method, the model performance is improved by at least 6.219 for training data and 0.444 for testing data in terms of RMSE. In terms of R-squared, the extreme gradient boosting showed improving in the model performance by about 39% for training set and about 4% for testing set.
Although the extreme gradient boosting illustrated better performance in terms of RMSE, there is a big difference between training and testing errors that may indicate over-fitting. Since Gboost has small difference between training and testing errors, it is selected as a better model to overcome over-fitting as suggested by one of the reviewers.
The importance scores for the gradient boosting methods illustrated that the most important variables in predicting and interpreting return on equity are the total liabilities to total assets, the total assets and the stock volatility with respect to the market. The partial dependent and local interpretable model-agnostic explanation plots showed that the total assets, the total liabilities to total assets and beta coefficients had the most stability in predicting and interpreting return on equity over the whole range of the data.
Since this study characterized a set of main variables that have the most importance scores, this may benefit the mangers of banks in MENA countries by keeping a close watch on these relevant variables to help in increasing the market financial stability for the banks. For example, the total liability to total assets (leverage) reflects how a bank is financially stable. The higher the ratio, the higher the degree of leverage and, therefore, the higher the risk to invest in that bank. This study could be extended to other sectors such as service sector.

## ACKNOWLEDGEMENT

The authors thank the editor and five anonymous reviewers for insightful comments on the paper that improved the quality and presentation of the manuscript.


## REFERENCES


[1] C.K Delen, and A. Uyar, "Measuring firm performance using financial ratios: A decision tree approach," *Expert System Applications*, Vol. 40. pp. 3970-3983, 2013.

[2] K. Bochkay and B. Levine, "Using MD&A to improve earnings forecasts," *Journal of Accounting, Auditing & Finance*, Vol.15, pp.1–25, 2017.

[3] E. Onder, E., and A.T. Altintas, "Financial performance evaluation of Turkish construction companies in Istanbul Stock Exchange (BIST)," *International Journal of Academic Research in Accounting, Finance and Management Sciences*, Vol. 7, pp. 108–113, 2017.

[4] G. Mousa and E. Elamir, E, "Determinants of forward-looking disclosure: evidence from Bahraini capital market,". *Afro-Asian Journal of Finance and Accounting*, Vol. 8, pp. 1-19, 2018.

[5] E. A. Elamir, "Modeling and predicting earnings per share via regression tree approaches in banking sector: Middle East and North African countries case," Investment Management and Financial Innovations, 17(2), pp. 51-68, 2020.

[6] M. S. Mahmoud, M. Eltoweissy, D. B. Rawat, and W. M. El-Medany, "Introduction to Special Issue on: Web of Things and Big Data," International. Journal of Computing and Digital Systems, vol. 5, 2016.

[7] Z.A. Hamza and M. Hammad, "Testing approaches for web and mobile aplications: An overview," International. Journal of Computing and Digital Systems, vol. 9, pp 1-8, 2020.

[8] G.R. Muehlhauser, "Putting performance measures to wor,". *Journal of Applied Corporate Finance*, Vol. 8, pp. 47-54, 1995.

[9] C. Correia, D. Flynn, E. Uliana, and M. Wormald, "*Financial management*," 5th Edition. Cape Town: Juta, 2003.

[10] A. Monteiro, "A quick guide to financial ratios," *The Citizen*, Money web Business Insert, 6 May:3, 2006.

[11] M.A. Kabajeh, S.M. Al Nu'Aimat and F.N. Dahmash, "The relationship between the ROA, ROE and ROI ratios with Jordanian insurance public companies market share prices," *International Journal of Humanities and Social Science*. http://www.ijhssnet.com/journals/Vol_2_No_11_June_2012/12.pdf. Accessed on April 23rd 2015. pp. 115-120, 2012.

[12] T. Hastie, R. Tibshirani and J. Friedman, "The Elements of statistical learning: Data mining, inference and prediction," Springer, 2 ed, 2008

[13] M. Kuhn, and K. Johnson, "*Applied predictive modelling,*" New York: Springer, 2013.

[14] J. Friedman, "Greedy function approximation: A Gradient boosting machin," *Annals of Statistics*, Vol. 29, pp. 1189–1232, 2001.

[15] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontier in Neurorobotics*, Vol. 7, pp.21-42, 2013.

[16] Y.C. Chang, K.H. Chang and G.J. Wu "Application of extreme gradient boosting trees in the construction of credit risk assessment models for financial institutions," *Applied Soft computing*, Vol. 73, pp. 914-920, 2018.

[17] M.T. Ribeiro, S. Singh and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier," In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). ACM, New York, NY, USA, pp. 1135-1144, 2016.

[18] A. Rappaport, "*Creating shareholder value*," New York: The Free Press, 1986.

[19] J.D. Stowe, T.R. Robinson, J.E. Pinto, and D.W. McLeavy, D.W. "*Analysis of equity investments: Valuation,*" Baltimore: AIMR, 2002.

[20] C. Firer, S.A. Ross, R.W. Westerfield, and B.D. Jordan, "*Fundamentals of corporate finance,*" 3rd South African edition. New York: McGraw-Hill, 2004.

[21] M. AL Khalaileh, "The Relationship between Accounting Performance Indexes and Market Performance Indexes, An Applied Study on Listed Corporations at Amman Security Exchange," *Administrative Sciences Studies Magazine*, Jordan University, Amman, Vol. 1, pp 1-10, 2001.

[22] M. Zieba, K. Tomczak and J. Tomczak, "Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction," *Expert Systems with Applications*, Vol. 58, pp. 93-101, 2016.

[23] P. Carmona, F. Climent, and A. Momparler, "Predicting failure in the U.S. banking sector: An extreme gradient boosting approach," *International Review of Economics and Finance*, 61, pp. 304-323, 2019.

[24] F. Climent, A. Momparler, and P. Carmona, "Anticipating bank distress in the Eurozone: An extreme gradient boosting approach," *Journal of Business Research*, 101, pp. 885-896, 2019.

[25] Y.J. Chen, J. Lin, Y. Chen and J. Wu, "Financial forecasting with multivariate adaptive regression splines and queen genetic algorithm-support vector regression," *IEEE Access*, Vol. 7, pp. 112931-112938, 2019.

[26] J. Hunt, J. Myers and L. Myers, "Improving earnings predictions with machine learning," https://zicklin.baruch.cuny.edu/wp-content/uploads/sites/10/2019/12/Improving-Earnings-Predictions-with-Machine-Learning-Hunt-Myers-Myers.pdf., 2019.

[27] D. Alaminos, M. Fernandez-Gamez, C.J. Santos, and J. Campos-Soria, "Predicting systemic banking crises using extreme gradient boosting," *Journal of Scientific & Industrial Research*, Vol. 78, pp. 571-575, 2019.

[28] N. Dhieb, H. Ghazzai, H. Besbes and Y. Massoud, "Extreme gradient boosting machine learning algorithm for safe auto insurance operations," IEEE International Conference on Vehicular Electronics and Safety (ICVES), Cairo, Egypt, pp. 1-5, 2019.

[29] T. Chen , T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell., I. Cano I, T. Zhou, M.. Li , J. Xie., M. Lin., Y. Geng and Y. Li, "Xgboost: extreme gradient boosting, R package," version 0.90.0.2. https://CRAN.R-project.org/package=xgboost, 2019.

[30] B. Greenwell, B. Boehmke, J. Cunningham and GBM Developers "gbm: Generalized Boosted Regression Models," R package version 2.1.5. https://CRAN.R-project.org/package=gb, 2019.

[31] I.B. Mustapha, and F. Saeed, "Bioactive molecule prediction using extreme gradient boosting," *Molecules*, 21, pp. 1-11, 2016.

[32] R. De Graaff, "Sentiment analysis of annual reports as a financial performance indicator," Master of Science in Business Information Systems, Eindhoven University of Technology, Holland, 2017.

[33] R. Schapire, "The boosting approach to machine learning:an overview," *Nonlinear Estimate Classification Lecture Notes Statistics*, Vol. 171, pp. 49–171, 2002.

[34] J. Wenxin, "On weak base hypotheses and their implications for boosting regression and classification," *Annal Statistics*, *Vol.* 30, pp. 51–73.Available on line at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.134.9366, 2002

[35] T. Chen and T. He, "Xgboost: extreme gradient boosting," R package version 0.3.0. technical repor, 2015.

[36] . R Core Team, "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/, 2017.

[37] M. Kuhn, "Caret package. *Journal of Statistical Software*," Vol. 28, pp.1-26, 2008.

[38] M. Kuhn, "caret: Classification and Regression Training," R package version 6.0-86. https://CRAN.R-project.org/package=caret, 2020.

[39] R. Balakrishnan, X. Qui, and SP. rinivasan, "On the predictive ability of narrative disclosures in annual reports." *European Journal of Operational Research*, 202, pp. 789-801, 2010.

[40] T. Pedersen and M. Benesty, "Lime: Local Interpretable Model-Agnostic Explanations," R package version 0.5.1. https://CRAN.R-project.org/package=lime., 2019.

[41] A. Molinaro, "Prediction Error Estimation: A Comparison of Resampling Methods," *Bioinformatics*, **21**(15), 3301–3307

[42] Kim JH (2009). "Estimating Classification Error Rate: Repeated Cross–Validation, Repeated Hold–Out and Bootstrap." *Computational Statistics & Data Analysis*, **53**(11), 3735–3745, 2009.

**Elsayed A.H. Elamir** is an associate professor at the University of Bahrain, Management & Marketing department. He obtained his Ph.D. from Durham University, UK. His research interest includes: Business statistics, data mining, quantitative modeling and machine learning. Worked as a Research Fellow at University of Southampton.