“Modeling and predicting earnings per share via regression tree approaches in banking sector: Middle East and North African countries case”

AUTHORS	Elsayed A. H. Elamir https://orcid.org/0000-0002-9430-072X http://www.researcherid.com/rid/K-3340-2018

ARTICLE INFO	Elsayed A. H. Elamir (2020). 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), 51-68. doi:10.21511/imfi.17(2).2020.05
DOI	http://dx.doi.org/10.21511/imfi.17(2).2020.05
RELEASED ON	Friday, 15 May 2020
RECEIVED ON	Friday, 20 March 2020
ACCEPTED ON	Tuesday, 05 May 2020

LICENSE	This work is licensed under a Creative Commons Attribution 4.0 International License

JOURNAL	"Investment Management and Financial Innovations"
ISSN PRINT	1810-4967
ISSN ONLINE	1812-9358
PUBLISHER	LLC “Consulting Publishing Company “Business Perspectives”
FOUNDER	LLC “Consulting Publishing Company “Business Perspectives”

NUMBER OF REFERENCES	NUMBER OF FIGURES	NUMBER OF TABLES
47	9	9

© The author(s) 2026. This publication is an open access article.

BUSINESS PERSPECTIVES

LLC "CPC "Business Perspectives"
Hryhorii Skovoroda lane, 10,
Sumy, 40022, Ukraine

www.businessperspectives.org

Elsayed A. H. Elamir (Kingdom of Bahrain)

MODELING AND PREDICTING EARNINGS PER SHARE VIA REGRESSION TREE APPROACHES IN BANKING SECTOR: MIDDLE EAST AND NORTH AFRICAN COUNTRIES CASE

Abstract

The regression tree approach is an effective and easy to interpret technique where it utilizes a recursive binary partitioning algorithm that divides the sample into partitioning variables with the strongest correlation to the response variable. Earnings per share can be considered as one of the main factors in making the investment decision. This study aims to build a predictive model for earnings per share in the context of the Middle East and North African countries (MENA) . The sample of the study consists of sixty-three banks, which were chosen from eight countries, with a total of six-hundred thirty observations. The simple regression, regression tree, and its pruned regression tree, conditional inference tree, and cubist regression are used to build the predictive model for earnings per share that depends on total assets, total liability, bank book value, stock volatility, age of the bank, and net cash. The results show that the cubist regression is outperforming other approaches where it improves root mean square error for the predictive model by approximately double in comparison with other methods. More interesting results are obtained from the important scores, where it shows that the total assets of the bank, bank book value, and total liability have the biggest impact on the prediction of earnings per share. Also, the cubist regression gives an improvement in R-squared over other methods by at least 30% and 23% using training and testing data, respectively.

Keywords earnings performance, forecasting, investment decision, machine learning, predictive model, risk management

JEL Classification C53, D22, F47, M10

Received on: 20th of March, 2020
Accepted on: 5th of May, 2020
Published on: 15th of May, 2020

© Elsayed A. H. Elamir, 2020

Elsayed A. H. Elamir, Ph.D., Associate Professor, Management & Marketing Department, College of Business, University of Bahrain, Kingdom of Bahrain.

This is an Open Access article, distributed under the terms of the Creative Commons Attribution 4.0 International license, which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

Conflict of interest statement:
Author(s) reported no conflict of interest

INTRODUCTION

Although the financial market analysis needs knowledge, perceptive insight, and experience, the automation techniques have been steadily used and growing because of the availability of huge financial data. There is much work growing in the fields of data mining, machine learning, and predictive models and their applications to business (Bose & Mahapatra, 2001; San Ong, Yichen, & The, 2010; Canhoto & Clear, 2020). Stock evaluation of a firm to buy or sell is a crucial decision to be taken by the investors, especially with the availability of large data. This decision is not easy to be made without the help of some modern models and determining the best model, which influences the investment decisions for a firm (McNichols, 2000; Goel & Gangolly, 2012; Onder & Altintas, 2017). Earnings per share (EPS) is considered an important profitability metric on financial statements for making the investment decision. It represents the returns delivered by the firm for each outstanding share of common stock. In finance and accounting literature,

many measures of financial performance were utilized, such as return on assets, return on equity, earnings per share, stock return, and others (Zhang, Cao, & Schniederjans, 2004). Qiu, Srinivasan, and Hu (2014) are given support for EPS that can help in making the investment decision in a reliable way better than other measures such as return on assets, especially in predictive models.

Linear regression (LR) is the simplest and popular model where there is one predictive equation holding over the complete data space. When there are many features, which have many nonlinear interactions, the linear regression prediction is subject to severe limitations, for example, high root means square error (Strobl, Malley, & Tutz, 2009; Kuhn & Johnson, 2013). Recently, machine learning techniques have become famous and broadly utilized models for nonparametric regression and classification in several scientific areas. Three popular approaches for regression based on machine learning are classification and regression trees (CART), conditional inference trees (CIT), and cubist regression trees (CRT). The main advantages of these approaches are that it is simple and easy to understand what variables contribute more to the prediction by looking at the tree, it can predict in case of missing some data, it can work when the actual regression surface is not smooth, and in many cases, it gives more prediction accuracy in comparison with linear regression (Breiman, Friedman, Olshen, & Stone, 1984; Quinlan, 1992, 1993; Hothorn, Hornik, & Zeileis, 2006; Han & Kwak, 2019; Kuhn & Johnson, 2013).

The literature is reviewed in Section 1. The methodology is described in Section 2. The data analysis results and discussions are presented in Section 3. Final section is devoted to conclusion.

1. LITERATURE REVIEW

Ou and Penman (1989) studied the two-step process to predict the sign of earnings changes. They used a stepwise logit regression model in estimating the historical relationship between observed financial ratio and sign of changes in future earnings. They obtained 78% accuracy of the sign of the changes for one year ahead earnings. In the out-of-sample prediction of the sign of the one year ahead earnings changes, they obtained approximately 60% accuracy.

Lawellen (2004) used regression models to predict aggregate stock returns using financial ratios such as dividend yield. They found that the predictive regressions are biased in the small sample, but the correction used by previous studies tends to improve predicting power substantially.

Bulgurcu (2012) used TOPSIS technique to analyze the financial performance of technology firms, which were registered in the Istanbul Stock Exchange. This study obtained performance scores by TOPSIS method to examine and assess the firms in terms of ten financial ratios.

Zekic-Susac, Sarlija, and Bensic (2004) compared neural network, logistic regression, and decision tree models on the Croatian dataset to characterize important features for small business credit scoring. They showed that the neural network models are better associated with data than logistic regression and decision tree models. They concluded that the neural network model extracted entrepreneur personal, business characteristics, and credit program characteristics as important features.

Tsai and Wang (2009) used a decision tree and artificial network models to predict stock prices on Taiwanese stock market data. They concluded that the F-score on trained stock exchange data was 77% using decision tree and artificial network, while the F-score was about 67% using a single algorithm.

Gepp, Kumar, and Bhattacharya (2010) studied discriminant, logit, and decision tree models to obtain accurate business failure prediction models in financial investment and lending sectors. In terms of predicting the failure or success of a business, they have concluded that the decision tree model could surpass the prediction technique of business failure as compared to logit and discriminant models.

Döpke, Fritsche, and Pierdzioch (2017) studied the usefulness of selected financial leading indi-

http://dx.doi.org/10.21511/imfi.17(2).2020.05

cators for forecasting recessions using a boosted regression tree method. Their results showed that short-term interest rate and the term spread are the most important indicators. Boosted regression trees helped them to find out the method in which the recession probability relies on the shares between the leading indicators. The spread term and the stock market gained importance, while the predictive power of the short term is declined.

Lin Yu-Cheng, Yu-Hsin Lu, Fang-Chi Lin, and Yi-Chen Lu (2017) applied a cubist regression tree model on data from Taiwanese companies to explain when and why auditors compromise their independence. They showed a positive relationship between auditor dependence and important clients in case of net losses in the current year as reported by clients. They also concluded that although the clients reported net losses in their financial statements, the auditors permitted more important clients to manage their discretionary accruals a bit upward.

Affes and Hentati-Kaffel (2019) studied bankruptcy forecasting using multivariate adaptive regression splines (MARS), classification and regression trees (CART), and hybrid models on US banks' data over a complete cycle for the market. They concluded that MARS provided better results than CART in terms of correct classification, hybrid method increased the correct classification in the training sample, and in general, nonparametric models (MARS, CART, hybrid) had given better results for bank failure forecasting than the logit model.

Carmona, Climent, and Momparler (2019) used extreme gradient boosting to forecast bank failure in the US banking sector. The data consisted of an annual series of 30 financial ratios for 156 national commercial banks from 2001 to 2015. They indicated that retained earnings, pre-tax return on assets, and total risk-based capital ratio are related to a higher risk of bank failure. The bank financial distress is increased by the exceedingly high yield on earning assets.

Bellotti, Brigo, Gambetti, and Vrins (2019) applied many regression and machine learning techniques on the database from a European Debt Collection Agency to predict recovery rates on non-performing loans. They found that the cubist regression, boosted trees, and random forest methods resulted in better than other approaches.

Chu, He, Hui, and Lehavy (2020) examined the managerial disclosure of modern products within the setting of the pressure between disclosure and managerial incentives. They developed a dictionary-based innovation disclosure measure obtained from the narratives in new product announcements. They found that a significant positive relationship between investor response and innovation disclosed up to two years' prediction can be obtained by the degree of innovation disclosed in new product announcements and the degree of innovation disclosure. The performance predictability is affected by managerial disclosure incentives.

Numerous earlier studies (Altman, Sabato, & Wilson, 2010; Altman, Iwanicz-Drozdowska, Laitinen, & Suvas, 2017; Appiah, Chizema, & Arthur, 2015) give evidence that the firm size plays an important role in making several choices within the firm and can impact on the productivity of the firm. Dias and Matias-Fonseca (2010) utilized 31 financial ratios to forecast corporate performance, including liability and others.

Different financial ratios are used in building predictive models to predict different outcomes such as corporate failure, bankruptcy, financial disasters, and financial performance of the firms. Appiah and Abor (2009) utilized many financial ratios, such as liability, liquidity, and profitability ratios, to construct their model. In Jordan, Alkhatib and Al-Horani (2012) utilized a set of 24 financial ratios to anticipate the financial distress of a sample of recorded companies. Kloptchenko, Eklund, Back, Karlsson, Vanharanta, and Visa (2002) utilized 7 ratios to forecast the financial performance of the firm. Balakrishnan, Qiu, and Srinivasan (2010) utilized firm measure, market-to-book ratio, and related ratios in their model.

This study is different from previous studies in many aspects. It can be considered one of the few studies dealing with the applications of the regression tree approaches in MENA countries. The high demand for investors and financial analysts in the financial markets, especially in MENA countries, to have expectation about the financial performance of firms and contributing

to the literature about building predictive models in MENA countries.

This main aim of this study is to build and predict earnings per share (EPS) based on the logarithm of bank total assets (logTOTA), total liabilities to total assets (LIAB), bank book value to its market value (BOKV), stock volatility with respect to the market (SVOL), age of the bank (AGEB), and net cash of the bank (NCSH) using classification and regression trees (CART), conditional inference trees (CIT), and cubist regression trees (CRT) approaches.

2. METHODOLOGY

2.1. Data collection and study variables

The banking sector in MENA countries is selected because the bank has the most assessed and reported capitalization in stock trades of these markets. In this study, the data are collected from eight MENA countries, namely, Egypt, Jordan, Qatar, Oman, Saudi Arabia, Kuwait, Bahrain, and United Arab Emirates from 2009 to 2018 through a sample of sixty-three banks in all of them, with a total of 630 observations. The predictive model that predicts EPS is built based on the data from 2009 to 2017 (training data) as year t and is called a training model. The data for 2018 (year $t+1$) is used as testing data to validate the training data or predict the model. Because of the homogeneity among these countries in terms of culture, conventions, and financial conditions, they are selected in the sample. The websites of the recorded banks in the bourse are used to collect the financial data. According to the aim of this study and previous arguments, seven variables are considered. Earnings per share (EPS) as dependent variable or measure of profitability and six independent variables, namely, total assets (TOTA), total liabilities to total assets (LIAB), bank book value to its market value (BOKV), stock volatility with respect to the market (SVOL), age of the bank (AGEB), and net cash of the bank (NCSH).

2.2. Predictive models

The linear regression, regression tree, conditional inference tree, and cubist regression are discussed briefly.

2.2.1. *Linear regression*

Classical linear regression aims to minimize the sum of square errors between actual values, y_i, and estimated values, $\hat{y}_i$, as

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

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

where β_i are the parameters, ε_i are the errors (Kuhn & Johnson, 2013).

2.2.2. *Regression tree*

Decision trees are one of the nonparametric predictive modeling approaches that are applied to classification and regression problems termed classification and regression tree (CART) analysis, first studied by Breiman, Friedman, Olshen, and Stone (1984). The regression tree uses distinct branches to go from values about features to conclude the target variable (leaves) using a set of if-then rules. The splitting points identify non-overlapping regions that have the most homogeneous responses to the target variable, and in each region, a simple model (such as the average) is fitted. The splitting corresponding to the tree deepness continues till a stopping criterion is reached. For prediction, the new data is divided following the trained split points (Breiman, 1996, 2001; Geurt, Ernst, & Wehenkel, 2006).

In case of regression, the model started with all observations, *D*, and searches each observation of each independent to locate the independent and divide the value, which divides the observations into two groups, say, D_1 and D_2, such that square sum of errors is minimized:

$$SSE = \sum_{i \in D_1} \left(y_i - \bar{y}_1 \right)^2 + \sum_{i \in D_2} \left(y_i - \bar{y}_2 \right)^2,$$

where $\bar{y}_1$ and $\bar{y}_2$ are the means of the training group outcomes inside sets D_1 and D_2, respectively. Besides, inside D_1 and D_2, this method finds out the predictor and divides value with the best minimizes *SSE*. The process continues within groups D_1 and D_2 until the sample numbers in the divisions will fall under a pre-specified value. Because this

http://dx.doi.org/10.21511/imfi.17(2).2020.05

method works recursively, this method is known as recursive partitioning (Hastie, Tibshirani, & Friedman, 2009; Kuhn & Johnson, 2013).

The tree could be large and overfit the training data. In this case, the tree should be trimmed to smaller depth (cost-complexity tuning) (Breiman, Friedman, Olshen, & Stone, 1984; Kuhan & Johnson, 2013). The aim is to obtain the actual tree size that has minimum "error rate". As suggested by Breiman, Friedman, Olshen, and Stone (1984), the "error rate" should be penalized based on the tree size as

$$SSE_{cp} = SSE + c_p \cdot (\#\text{final nodes}),$$

note that c_p is the complexity parameter that helps in finding the smallest trimmed tree, which has the least penalized "error rate" (Olson & Wu, 2020).

Although regression tree is interpretable to visualize the results, it has a natural way to perform variable selection, can handle missing data and is robust to outliers, has some weaknesses in terms of instability and model accuracy compared with methods, which depends on ensemble learning algorithms (Hastie, Tibshirani, & Friedman, 2009; Kuhn & Johnson, 2013).

2.2.3. Conditional inference tree

This approach is introduced by Hothorn, Hornik, and Zeileis (2006) to overcome the bias in the basic regression tree. They used significance test procedures to select variables that have many possible splits instead of selecting the variable based on maximizing cost measure. The permutation tests are used to compute multiple tests at each start of algorithm (chose feature – select split – repeat). The test is used to assess the difference between the averages of two sets created by the division, and a *p*-value can be evaluated for the test. A stopping point is used to decide whether more splits should be created, for example, one minus the *p*-value (Hothorn, Hornik, & Zeileis, 2006). Generally, the conditional inference trees used a feature selection scheme that is based on permutation significance tests that reduce the bias in the basic regression tree (Westfall & Young, 1993; Hothorn, Hornik, & Zeileis, 2006; Han & Kwak, 2019).

2.2.4. Cubist regression

Cubist regression is a rule-based regression described by Quinlan (1992). A model tree is generated from the training group, and the linear model is estimated and smoothed. The model tree is ceased into rules, and the pruned are applied. Quinlan (1992) described the equation for the smoothed model as

$$\hat{y}_p = \frac{n_k \hat{y}_k + c\hat{y}_p}{n_k + c},$$

$\hat{y}_p$ is the prediction for the parent model, $\hat{y}_k$ is the forecasting for the "child model", n_k is the sample size in the child model and c is a constant (Hastie & Pregibon, 1990).

To combine up the tree, the cubist is

$$\hat{y}_p = a\hat{y}_k + (1-a)\hat{y}_p,$$

where $a = \frac{Var(e_p) - Cov(e_k, e_p)}{Var(e_p - e_k)}$,

e_p is the model residuals for the parent model, e_k is the residuals for the child model, *Var* stands for variance, and *Cov* stands for covariance (for more details about cubist regression, see Quinlan, 1992, 1993; Kuhn & Johnson, 2013).

3. DATA ANALYSIS AND RESULTS

The general EPS model can be written as

$$EPS = f(logTOTA, LIAB, BOKV, SVOL, AGEB, NCSH).$$

This function could be linear as

$$EPS = \beta_0 + \beta_1 logTOTA + \beta_2 LIAB + \beta_3 BOKV + \\ + \beta_4 SVOL + \beta_5 AGEB + \beta_6 NCSH + \varepsilon_i,$$

where $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6$ are the parameters in the model, ε_i are the errors.

The predictive model that predicts EPS is built based on the data from 2009 to 2017 (training data) as year *t* and is called a training model. The data for

Table 1. Descriptive statistics for the study variables

Variables	Mean	Sd	Median	Skew	Kurtosis
EPS	0.932	1.793	0.175	3.277	14.477
logTOTA	3.996	0.790	3.905	0.226	–1.039
LIAB	7.915	2.380	7.624	0.968	2.734
BOKV	0.718	0.374	0.742	0.840	3.331
SVOL	1.366	0.701	1.223	1.690	4.425
AGEB	35.721	13.912	37.000	–0.234	–0.271
NCSH	507.555	5708.026	7.180	12.248	225.625

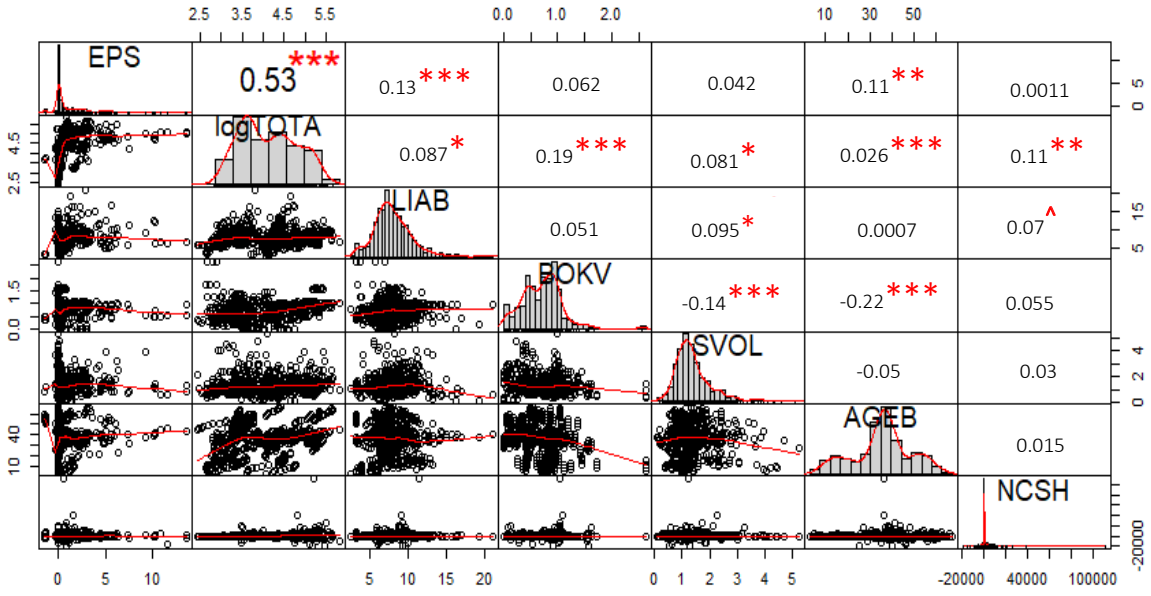
2018 as year $t + 1$ are used as testing data to validate the training model. In other words, the training model is built based on 9 years' data (2009–2017) to predict EPS in 2018. The training data from 2008 to 2017 consists of 566 observations, while the testing data for 2018 consists of 64 observations.

Different performance metrics are used to evaluate the model in case of using training and testing data such as root mean square error (RMSE), determination coefficients (R-squared), and mean absolute error (MAE). In RMSE and MAE, the less value, the better performance. In R-squared, the higher value, the better performance. Variable importance is a measure of the decrease in "squared error", where the advancement in "squared error" due to each independent is gathered inside every tree. The refinement values for every independent are then averaged toward the whole gathering to produce an aggregate importance value (Friedman, 2002; Ridgeway, 2007). The important variables that contribute to predictions of EPS are obtained for each method to reflect the rank or importance of the independent variables.

All the analysis in this study is done using R-software and CARET package (Kuhn, 2008; R Core Team, 2017).

3.1. Descriptive analysis

The descriptive statistics for the variables of the study are displayed in Table 1. It can be noted that the standard deviation (Sd) is high for *NSCH* variable that indicates high variability among banks with respect to this variable. The mean and median are almost equal for *logTOTA* variable. Where the measures of skewness and kurtosis for the



Note: (***) significance at 0.001, (**) significance at 0.01, (*) significance at 0.05, and (^) significance at 0.10.

Figure 1. The correlation matrix, histogram, and scatter plots for the study variables

http://dx.doi.org/10.21511/imfi.17(2).2020.05

Table 2. Linear regression analysis for EPS model

Term	Coefficients	Std. error	t-statistics	*p*-value
Intercept	–3.780	0.384	–9.84	0***
logTOTA	1.190	0.083	14.3	0***
LIAB	0.047	0.025	1.84	0.066^
BOKV	–0.234	0.176	–1.33	0.183
SVOL	–0.073	0.086	–0.852	0.394
AGEB	–0.004	0.005	–0.901	0.368
NCSH	–0.001	0.0001	–1.41	0.159
	F-statistics = 38.8, with *p*-value = 0			–

Note: (***) significance at 0.001, (**) significance at 0.01, (*) significance at 0.05, and (^) significance at 0.10.

Table 3. Linear regression variable importance scores and performance metrics for EPS model

	Predictor variables					
	logTOTA	**LIAB**	**NCSH**	**BOKV**	**AGEB**	**SVOL**
Score	100	7.394	4.165	3.578	0.359	0
Model performance						
	RMSE		R-squared		MAE	
Training data	1.409		0.294		0.837	
Testing data	2.183		0.315		1.216	

variables *EPS*, *logTOTA*, *SVOL*, *AGEB*, *NCSH* are far away from 0 and 3, respectively, this indicates that the distribution of these variables is mostly non-symmetric. While the measures of skewness and kurtosis for the variables *LIAB* and *BOKV* are near 0 and 3, respectively, this indicates that the distribution of these variables is nearly symmetric.

Figure 1 shows the correlation and significance of the study variables. It can note that *EPS* has a significant correlation with *logTOTA*, *LIAB*, *AGEB*, while it has no significant correlation with *BOKV*, *SVOL*, and *NCSH*. The highest correlation is between EPS and *logTOTA* and the lowest correlation between *AGEB* and *NCSH*.

3.2. Linear regression

The results of the linear regression analysis are given in Table 2. Where *p*-value for F-statistics is zero, the model is significant. From column *p*-value, it can be seen that the variables *logTOTA* and *LIAB* are significant at 0.001 and 0.10 levels of significance, respectively.

The results of linear regression performance metrics are given in Table 3. The RMSE is 1.409 for training data, while it is 2.183 for testing data. R-squared is about 29.4% for training data and goes up to about 31.5% for testing data. MAE is about 0.837 for training data and goes up to 1.216 for testing data.

Figure 2 and Table 3 show the linear regression variable importance for EPS model. *logTOTA* transpires to the top of important metrics. The important scores start receding with *LIAB*, *NCSH*, *BOKV*, and *AGEB*. Note that the variable *SVOL* has no importance score. Consequently, *logTOTA* has the biggest impact on EPS.

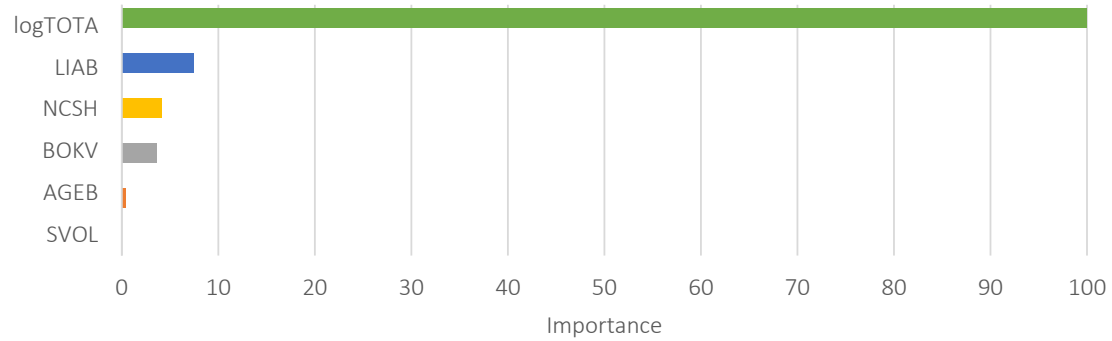


Figure 2. Linear regression variable importance scores for EPS model

3.3. Basic regression tree approach

Basic regression tree for EPS model with 13 terminal nodes is shown in Figure 3. The first decision node in Figure 3 is *logTOTA* that is most strongly associated with EPS. The left and right branches show that the best cut-off value equal to 0.45 is the best to reduce root mean square error. Then, the decision nodes are divided by the variable *BOKV* at cut-off value 0.98 and *logTOTA* at cut-off value 4.2 that are the best values to reduce the root mean square error. This process will be continued until the terminal nodes are obtained. The terminal nodes contain two values. The bottom one is the percentage of the data in this node that are used to compute the average of EPS (predicted value). For example, the training data have 566 values, and the predicted value 0.14 in the left node is the average of about 340 values (0.60·566) that fall in this branch. In other words, if *logTOTA* is less than 4.5, if *logTOTA* is less than 4.2, then the predicted EPS is 0.14. Another example is if *logTOTA* is more than or equal 4.5, *BOKV* is more than or equal 0.98, *SVOL* is more than or equal 0.84, then the predicted EPS is 5.3, using the average of about 11 EPS values in this branch (0.02·566).

The results of the basic regression tree performance metrics are given in Table 4. The RMSE is 0.978 for training data, while it is 2.263 for testing data. R-squared is about 66% for training data and goes down to about 25.3% for testing data. MAE is about 0.463 for training data and goes up to 1.176 for testing data.

Table 4. Basic regression tree variable importance scores and performance metrics for EPS

	Predictor variables					
	BOKV	**AGEB**	**LIAB**	**logTOTA**	**NCSH**	**SVOL**
Score	100	56.20	40.77	19.40	12.48	0
Model performance						
	RMSE		**R-squared**		**MAE**	
Training data	0.978		0.659		0.463	
Testing data	2.263		0.253		1.176	

Figure 4 and Table 4 show the basic regression tree variable importance for EPS model. *BOKV* and *AGEB* transpire to the top of important metrics, and important scores start receding with *LIAB*, *logTOTA*, and *NCSH*. Note that the variable *SVOL*

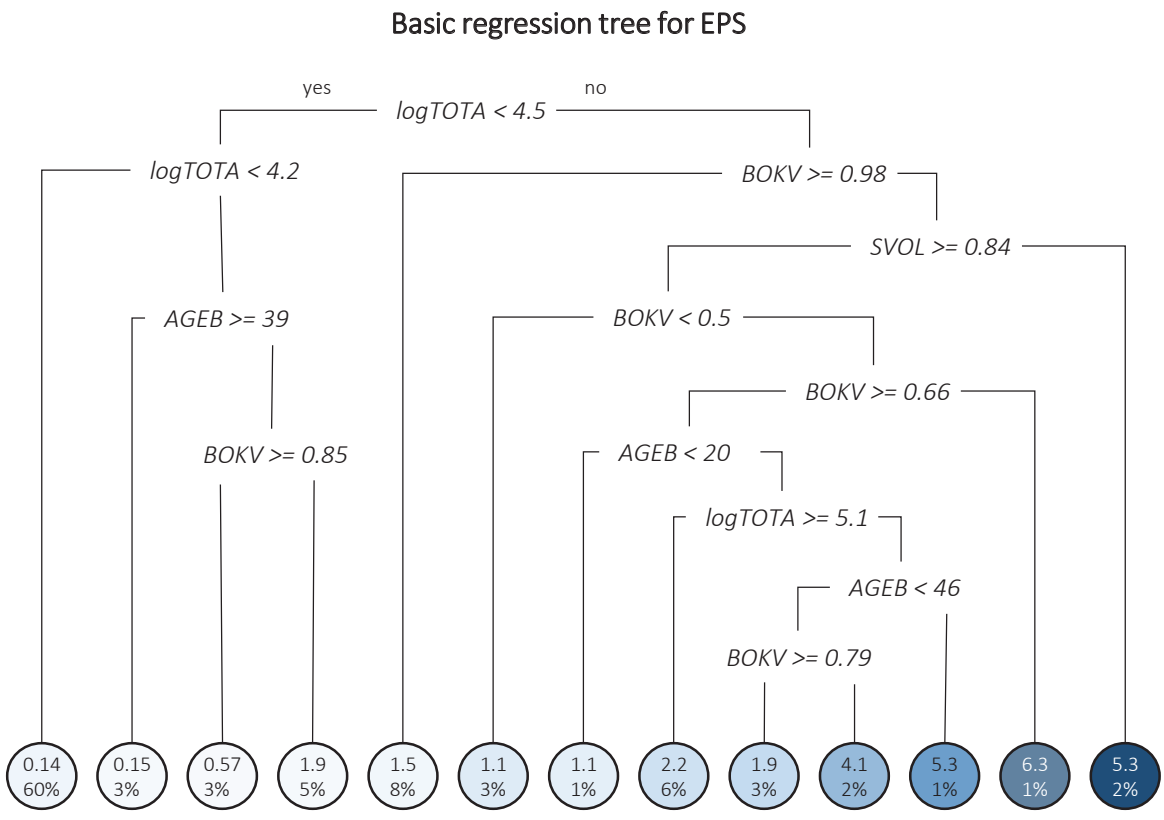


Figure 3. Basic regression tree for EPS model

http://dx.doi.org/10.21511/imfi.17(2).2020.05

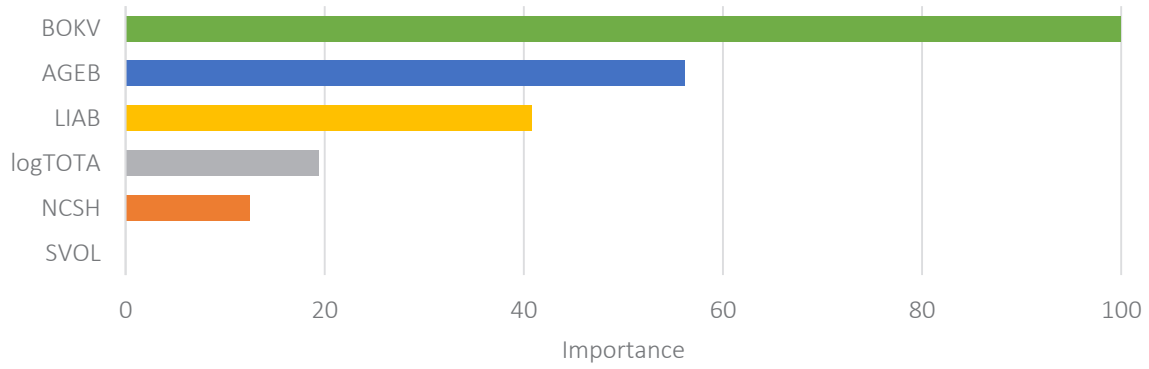


Figure 4. Basic regression tree variable importance scores for EPS model

has no importance score. Consequently, *BOKV*, *AGEB*, and *LIAB* have the biggest impact on EPS.

Figure 5 shows the pruned regression tree for EPS model. The number of tree nodes is 7, which is less than the basic regression tree. The first decision node in Figure 5 is *logTOTA* that is most strongly associated with EPS. The left and right branches show the best cut-off value equal to 0.45, which is the best value to reduce root mean square error. Then, the decision nodes are divided by the variables *BOKV* at cut-off value 0.98 and *logTOTA* at the cut-off value 4.2 that are the best values to reduce the root mean square error. This process will be continued until the terminal nodes are obtained. The terminal nodes contain two values. The bottom one is the percentage of the data in this node that used to compute the average of EPS (predicted value). For example, the training data is 566 values, and the predicted value 0.14 in the left node is the average of about 340 values (0.60*566) that fall in this branch. In other words, if *logTOTA* is less than 4.5, if *logTOTA* is less than 4.2, then the predicted EPS is 0.14. Another example is if *logTOTA* is more than or equal 4.5, *BOKV* is less than 0.98, then the predicted EPS is 1.5, using the average of about 45 (0.08*566) EPS values in this branch.

Table 5. Pruned regression tree variable importance scores and performance metric for EPS

	Predictor variables					
	BOKV	**logTOTA**	**AGEB**	**NCSH**	**LIAB**	**SVOL**
Score	100	72.7	49.4	23.4	22.1	0
Model performance						
	RMSE		**R-squared**		**MAE**	
Training data	1.058		0.602		0.535	
Testing data	2.321		0.217		1.283	

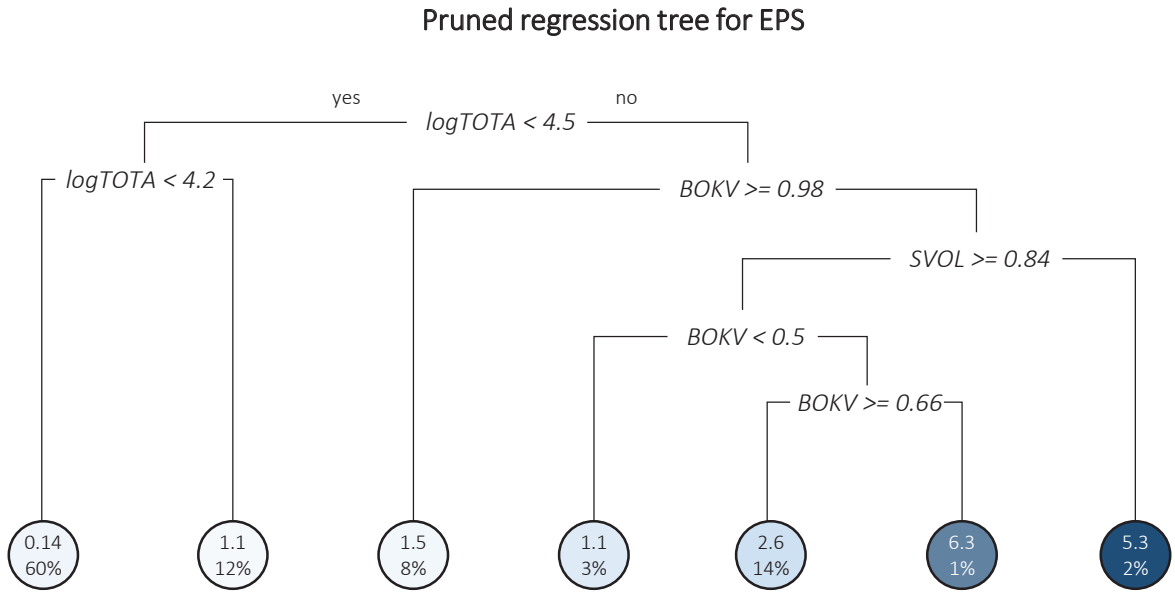


Figure 5. Pruned regression tree for EPS model

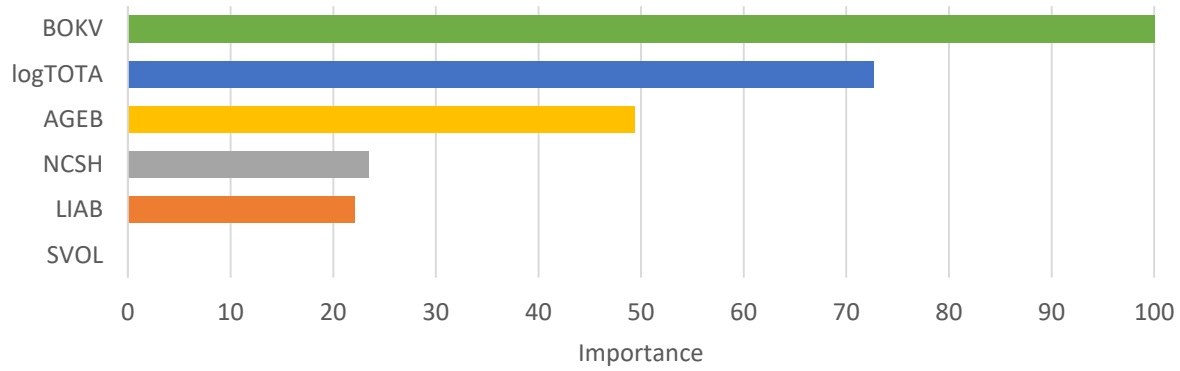


Figure 6. Pruned regression tree variable importance scores for EPS model

The results of the pruned regression tree performance metric are given in Table 5. The RMSE is 1.058 for training data while it goes up to 2.263 for testing data. R-squared is about 60.2% for training data and goes down to about 21.7% for testing data. MAE is about 0.535 for training data and goes up to 1.283 for testing data.

Figure 6 and Table 5 show the pruned regression tree variable importance for EPS model. *BOKV*, *logTOTA*, and *AGEB* transpire to the top of important metric, and important scores start receding with *NCSH* and *LIAB*. Note that the variable *SVOL* has no importance score. Consequently, *BOKV*, *logTOTA*, and *AGEB* have the biggest impact on EPS.

3.4. Conditional inference tree

Conditional inference tree is shown in Figure 7. The decision nodes are presented as circles with a number in each circle. The independent variable is divided twofold in each circle, with a *p*-value of the dependence test. The first decision node in Figure 7 is *logTOTA* that is most strongly associated with EPS that is measured by $p < 0.001$. The left branch shows the best cut-off value more than or equal 0.45, that is the best value to reduce root mean square error, and gives the predicted EPS of about 2.437, using 157 values. The decision node is divided by the variable l*ogTOTA* that is still strongly associated with EPS ($p < 0.001$) at cut-off

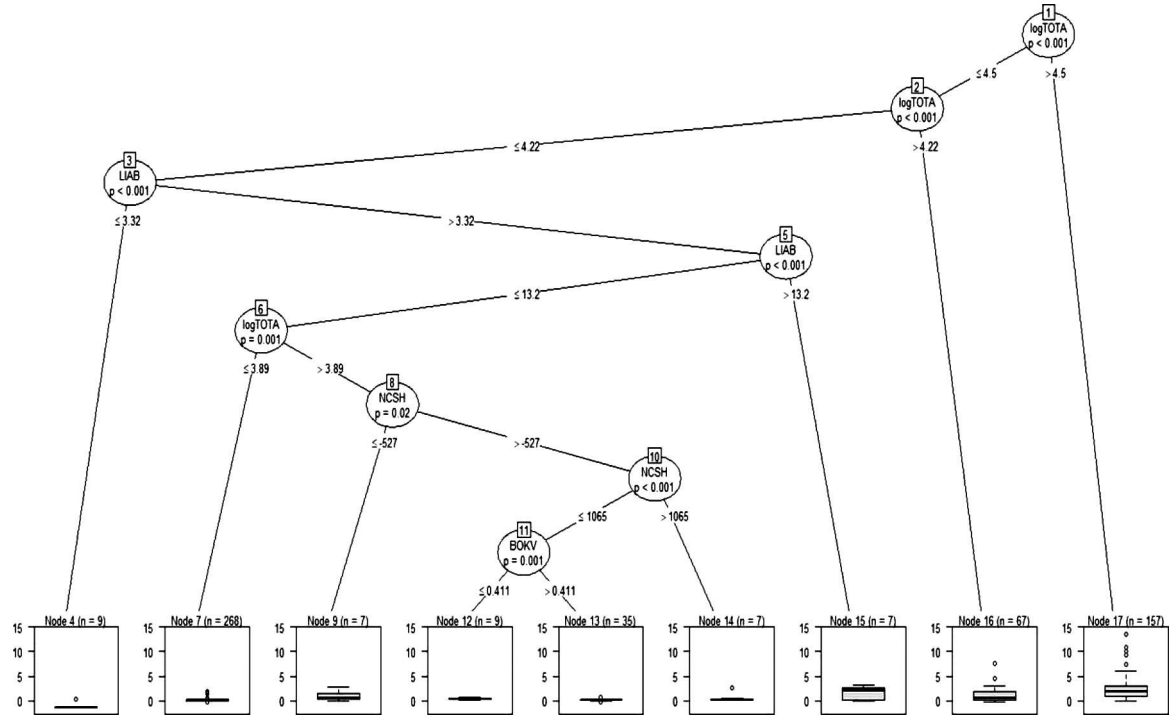


Figure 7. Conditional inference tree for EPS model

http://dx.doi.org/10.21511/imfi.17(2).2020.05

value 4.22. The left branch is a cut-off value less than or equal to 4.22, and the right branch gives the predicted EPS of 1.066, using 67 data. Then, the decision node is divided by the variable *LIAB* that is still strongly associated with EPS ($p < 0.001$) with a right branch at cut-off value 3.32, while the left branch gives the predicted EPS of 1.162, using 9 observations. This process will be continued until the terminal nodes are obtained. The results are displayed by boxplot because the response variable EPS is continuous.

Table 6. Conditional inference tree variable importance scores and performance metrics for EPS model

	Predictor variables					
	logTOTA	BOKV	LIAB	AGEB	SVOL	NCSH
Score	100	17.45	10.36	4.11	2.42	0
Model performance						
	RMSE		R-squared		MAE	
Training data	1.131		0.545		0.526	
Testing data	2.190		0.280		1.110	

The results of the conditional inference tree performance me tric are given in Table 6. The RMSE is 1.131 for training data, while it goes up to 2.190 for testing data. R-squared is about 54.5% for training data and goes down to about 28 % for testing data. MAE is about 0.526 for training data and goes up to 1.110 for testing data.

Figure 8 and Table 6 show the conditional inference tree variable importance for EPS model. *logTOTA* transpires to the top of important variables. The important scores start receding with *BOKV*, *LIAB*, *AGEB*, and *SVOL*. Note that the variable *NCSH* has no importance score. Consequently, *logTOTA* has the biggest impact on EPS.

3.5. Cubist regression

Cubist regression is an ensemble model that predicts using the linear regression models at the terminal node of the tree. The tree is decreased to a set of rules that are ways from the top to the bottom. Rules are dispensed through pruning or combined for simplification.

Table 7 shows the resampling results across tuning parameters for 566 samples and 6 predictors. The optimal model is chosen based on the smallest value for RMSE. The committees = 20 and neighbors = 5 are the final values that are used for the model.

Table 7. Cubist resampling results across tuning parameters for 566 samples and 6 predictors

Committees	Neighbors	RMSE	R-squared	MAE
1	0	1.36	0.427	0.587
1	5	1.31	0.465	0.557
1	9	1.32	0.455	0.575
10	0	1.16	0.546	0.531
10	5	1.08	0.604	0.481
10	9	1.10	0.584	0.512
20	0	1.14	0.562	0.521
20	5	1.05	0.617	0.470
20	9	1.08	0.598	0.503

The results for all 20 cubist models are too long; therefore, the results of the model 20 as an example are given in the Appendix. For example, Rule 20/2 uses 14 cases with an average 0.219, minimum –0.321, maximum 0.428, and estimated error is 0.203. If *logTOTA* > 4.2, *BOKV* <= 0.49 and *AGEB* > 38 the prediction of EPS comes from the equation

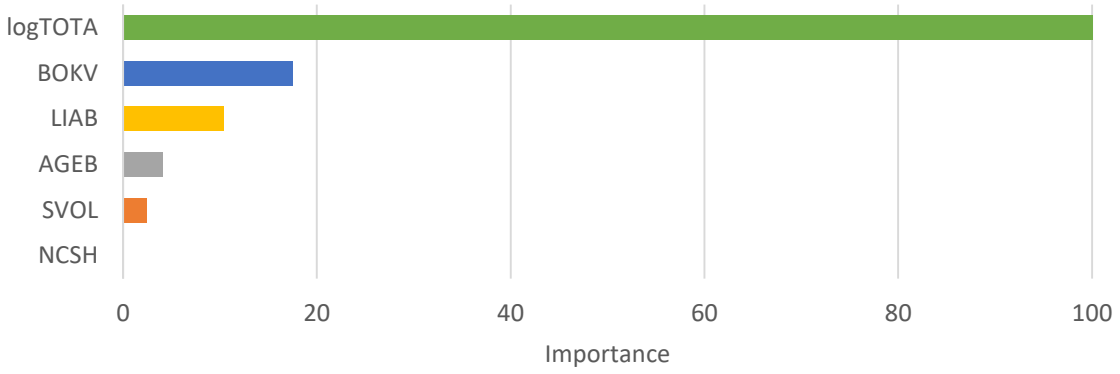


Figure 8. Conditional inference tree variable importance scores for EPS model

$$outcome = -5.016 + 1.15 \log TOTA - -0.95\, BOKV + 0.23\, SVOL.$$

Rule 20/2: [14 cases, mean 0.219289, range –0.32118 to 0.42896, est. err. 0.203703],

if

$logTOTA > 4.217362$

$BOKV <= 0.4932171$

$AGEB > 38$

then

$$outcome = -5.016187 + +1.15\, logTOTA - 0.95\, BOKV + +0.23\, SVOL$$

Table 8. Cubist regression approach variable importance scores and performance metrics for EPS model

	Predictor variables					
	logTOTA	**BOKV**	**LIAB**	**AGEB**	**SVOL**	**NCSH**
Score	100	47.4	39.8	34.8	25.7	0
Model performance						
	RMSE		**R-squared**		**MAE**	
Training data	0.375		0.958		0.175	
Testing data	1.917		0.543		0.882	

The results of the cubist regression performance metric are given in Table 8. The RMSE is 0.375 for training data, while it goes up to 1.917 for testing data. R-squared is about 95.8% for training data and goes down to about 54.3 % for testing data. MAE is about 0.175 for training data and goes up to 0.882 for testing data.

Figure 9 and Table 8 show the cubist regression variable importance for EPS model. *logTOTA*, *BOKV*, and *LIAB* transpire to the top of important variables. The important scores start receding with *AGEB* and *SVOL*. Note that the variable *NCSH* has no importance score. Consequently, *logTOTA*, *BOKV*, and *LIAB* have the biggest impact on EPS.

4. DISCUSSION

Table 9 gives comparisons among the approaches used in this study. In terms of root mean square error and training data, the best method is cubist regression 0.375, followed by regression tree 0.978, prune regression tree 1.058, conditional inference tree 1.131, and linear regression 1.409. With respect to testing data, the best method is cubist regression 1.917, followed by linear regression 2.183, conditional inference tree 2.190, prune regression tree 2.321, and regression tree 2.263. Similarly, it can rank models in terms of R-squared and MAE.

Table 9. Performance metrics for the study methods

Method	**Data**	**Performance metrics**		
		RMSE	**R-squared**	**MAE**
Linear regression	Training	1.409	0.294	0.837
	Testing	2.183	0.315	1.216
Regression tree	Training	0.978	0.659	0.463
	Testing	2.263	0.253	1.176
Prune regression tree	Training	1.058	0.602	0.535
	Testing	2.321	0.217	1.283
Conditional inference tree	Training	1.131	0.545	0.526
	Testing	2.190	0.280	1.110
Cubist regression	Training	0.375	0.958	0.175
	Testing	1.917	0.543	0.882

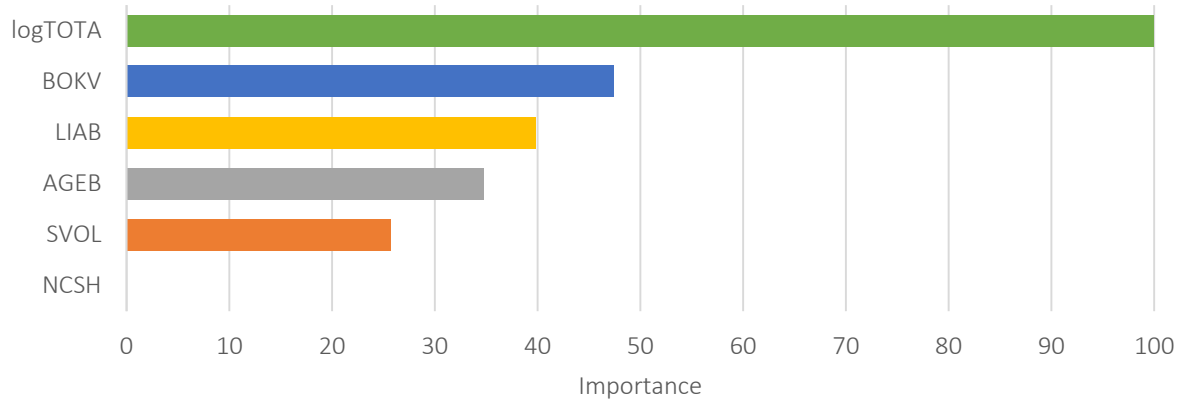


Figure 9. Cubist regression variable importance scores for EPS model

http://dx.doi.org/10.21511/imfi.17(2).2020.05

For testing data, one can also note that the best model is cubist regression 54%, followed by linear regression 31%, conditional inference tree 28%, regression tree 25%, and pruned regression tree 22%. Similarly, it can rank models in terms of MAE.

From Figure 9, the important scores show the impact of independent variables on EPS. For cubist regression, *logTOTA*, *BOKV*, and *LIAB* have the biggest impact on EPS, followed by *AGEB* and *SVOL*, with no important score for *NCSH*. From the above results, one can conclude that each method has a strong point on one side and a weak point on another side. Therefore, multi-dimensional data need different approaches to be modeled. Where the results of this study are limited to regression tree approaches, namely, LR, CART, CIT and CRT, it may be in the future these results are compared with the results of different approaches such as neural network.

CONCLUSION

Motivated by the importance of making investment decisions, the predictive models are built based on machine learning approaches, namely, linear regression, regression tree, pruned regression tree, conditional inference tree, and cubist regression to help in making such a decision. These models were carried out on the data from eight countries in MENA region, where sixty-three banks are selected, with a total of 630 observations. The sampled data are divided into training data (from 2008 to 2017) to build the predictive model and testing data (2018) to validate the model.

Root mean square error, R-squared, and mean absolute error are used to assess the performance of the models. The results show that the cubist regression is outperforming other methods in terms of three measures, namely, RMSE, R-squared, and MAE. R-squared in training data for cubist regression is about 96%, while for the second-best basic regression tree method, it is about 66%. This has given at least 30% (96%-66%) improvement over other methods. Root mean square in cubist regression for training data is 0.375, while it is 0.978 in a basic regression tree. This has given the improvement in root mean square error by at least 0.603 (0.978-0.375) over other methods. R-squared in testing data for cubist regression is about 54%, while for the second-best linear regression approach, it is about 31%. This has given at least 23% improvement over other methods. Important scores are used to know which variables have the biggest impact in predicting earnings per share.

In terms of the best results, the cubist regression has shown that the total assets, bank book value, and total liability have the biggest impact on predicting earnings per share. Because each approach has its strengths and weaknesses, multi-dimensional data should be approached by different techniques. Because this study identified a set of important variables, this may help bank's manger in increasing the stability of financial market. For example, the LIAB variable could give a good conception about how a bank is financially solid. This study can be extended when there are financial and non-financial data such as tone. Also, it can be extended to include data from other types of business, such as service companies.

AUTHOR CONTRIBUTIONS

Data curation: Elsayed A. H. Elamir.
Formal analysis: Elsayed A. H. Elamir.
Methodology: Elsayed A. H. Elamir.
Software: Elsayed A. H. Elamir.
Writing – original draft: Elsayed A. H. Elamir.

REFERENCES

1. Affes, Z., & Hentati-Kaffel, R. (2019). Forecast bankruptcy using a blend of clustering and MSRS model: case of US banks. *Annals of Operations Research, 281,* 27-64. https://doi.org/10.1007/s10479-018-2845-8
2. Alkhatib, B., & Al-Horani, A. (2012). Predicting financial distress of public companies listed in Amman Stock Exchange. *European Scientific Journal, 8*(15), 788-789. Retrieved from https://eujournal.org/index.php/esj/article/view/226/251
3. Altman, E. I., Iwanicz-Drozdowska, M., Laitinen, E. K., & Suvas, A. (2017). Financial Distress Prediction in an International Context: A Review and Empirical Analysis of Altman's Z-Score Model. *Journal of International Financial Management & Accounting, 28*(2), 131-171. https://doi.org/10.1111/jifm.12053
4. Altman, E. I., Sabato, G., & Wilson, N. (2010). The value of non-financial information in small and medium-sized enterprise risk management. *Journal of Credit Risk, 6*(2), 95-127. https://doi.org/10.21314/JCR.2010.110
5. Appiah, K. O., & Abor, J. (2009) Predicting corporate failure: some empirical evidence from the UK. *Benchmarking: An International Journal, 16*(3), 432-444. https://doi.org/10.1108/14635770910961425
6. Appiah, K. O., Chizema, A., & Arthur, J. (2015). Predicting corporate failure: a systematic literature review of methodological issues. *International Journal of Law and Management, 57*(5), 461-485. https://doi.org/10.1108/IJLMA-04-2014-0032
7. Balakrishnan R., Qiu, X. Y., & Srinivasan, P. (2010). On the predictive model ability of narrative disclosures in annual reports. *European Journal of Operation Research, 202*(3), 789-802. https://doi.org/10.1016/j.ejor.2009.06.023
8. Bellotti, A., Brigo, D., Gambetti, P., & Vrins, F. D. (2019). Forecasting Recovery Rates on Non-Performing Loans with Machine Learning. *Credit Scoring and Credit Control Conference XVI.* http://dx.doi.org/10.2139/ssrn.3434412
9. Bose, I., & Mahapatra, R. (2001). Business data mining – machine learning perspective. *Information & Management, 39*(3), 211-225. https://doi.org/10.1016/S0378-7206(01)00091-X
10. Breiman, L, Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees.* Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.
11. Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123-140. https://doi.org/10.1007/BF00058655
12. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324
13. Bulgurcu, B. (2012). Application of TOPSIS Technique for Financial Performance Evaluation of Technology Firms in Istanbul Stock Exchange Market. *Procedia - Social and Behavioral Sciences, 62,* 1033-1040. https://doi.org/10.1016/j.sbspro.2012.09.176
14. Canhoto, A., & Clear, F. (2020). Artificial intelligence and machine learning as business tools: A framework for diagnosing value destruction potential. *Business Horizons, 63*(2), 183-193. https://doi.org/10.1016/j.bushor.2019.11.003
15. Carmona, P., Climent, F., & Momparler, A. (2019). Predicting failure in the U.S. banking sector: an extreme gradient boosting approach. *International Review of Economic & Finance, 61*(2), 304-323. https://doi.org/10.1016/j.iref.2018.03.008
16. Chu, J., He, Y., Hui, K. W., & Lehavy, R. (2020). New Product Announcements, Innovation Disclosure, and Future Firm Performance. *SSRN.* https://doi.org/10.2139/ssrn.3543632
17. Dias, W., & Matias-Fonseca, R. (2010). The Language of Annual Reports as an Indicator of the Organizations' Financial Situation. *International Review of Business Research Papers, 6*(5), 206-215. Retrieved from https://www.academia.edu/728641/The_Language_of_Annual_Reports_as_an_Indicator_of_the_Organizations_Financial_Situation
18. Döpke, J., Fritsche, U., & Pierdzioch, C. (2017). Predicting recession with boosted regression trees. *International Journal of Forecasting, 33*(4), 745-759. https://doi.org/10.1016/j.ijforecast.2017.02.003
19. Friedman, J. (2002). Stochastic Gradient Boosting. *Computational Statistics and Data Analysis, 38(4),* 367-378. https://doi.org/10.1016/S0167-9473(01)00065-2
20. Gepp, A., Kumar, K., & Bhattacharya, S. (2010). Business failure prediction using decision trees. *Journal of Forecasting, 29*(6), 212-224. https://doi.org/10.1002/for.1153
21. Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning, 63*(1), 3-42. https://doi.org/10.1007/s10994-006-6226-1
22. Goel, S., & Gangolly, J. (2012). Beyond the numbers: Mining the annual reports for hidden cues indicative of financial statement fraud. *Intelligent Systems in Accounting, Finance and Management, 19*(2), 75-89. https://doi.org/10.1002/isaf.1326
23. Han, H., & Kwak, M. (2019). An alternative method in estimating propensity scores with conditional inference tree in multilevel data: A case study. *Journal of the Korean Data, 30*(4), 951-96. https://doi.org/10.7465/jkdi.2019.30.4.951
24. Hastie, T., & Pregibon, D. (1990). *Shrinking trees.* Retrieved from http://ww.web.stanford.edu/~hastie/Papers/shrink_tree.pdf

http://dx.doi.org/10.21511/imfi.17(2).2020.05

25. Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The Elements of Statistical Learning* (2nd ed.). New York: Springer. Retrieved from https://www.springer.com/gp/book/9780387848570

26. Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics, 15*(3), 651-674. https://doi.org/10.1198/106186006X133933

27. Kloptchenko, A., Eklund, T., Back, B., Karlsson, J., Vanharanta, H., & Visa, A. (2004). Combining data and text mining techniques for analyzing financial reports. *Intelligent Systems in Accounting, Finance and Management, 12*(1), 215-227. https://doi.org/10.1002/isaf.239

28. Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software, 28*(5), 1-26. Retrieved from https://www.jstatsoft.org/article/view/v028i05

29. Kuhn, M., & Johnson, K. (2013). *Applied predictive modelling.* New York: Springer. Retrieved from https://esl.hohoweiya.xyz/references/AppliedPredictiveModeling.2013.pdf

30. Lewellen, J. (2004). Predicting Returns with Financial Ratios. *Journal of Financial Economics, 74*(2), 209-235. https://doi.org/10.1016/j.jfineco.2002.11.002

31. Lin Yu-Cheng, Yu-Hsin Lu, Fang-Chi Lin, & Yi-Chen Lu. (2017). Net Losses and the Relationship between Auditor Independence and Client Importance: Evidence from a Cubist Regression-Tree Model. *Journal of Emerging Technologies in Accounting, 14*(1), 13-25. https://doi.org/10.2308/jeta-51673

32. McNichols, M. F. (2000). Research design issues in earnings management studies. *Journal of Accounting and Public Policy, 19*(4-5), 313-345. https://doi.org/10.1016/S0278-4254(00)00018-1

33. Olson, D. L., & Wu, D. (2020). Regression tree models. In *Predictive data mining models* (pp. 45-54). Singapore: Springer. Retrieved from https://www.springer.com/gp/book/9789811096457

34. Onder, E., & Altintas, A. T. (2017). Financial performance evaluation of Turkish construction companies in Istanbul Stock Exchange (BIST). *International Journal of Academic Research in Accounting, Finance and Management Sciences, 7*(3), 108-113. https://doi.org/10.6007/IJARAFMS/v7-i3/3237

35. Ou, J. A., & Penman, S. H. (1989). Financial statement analysis and the prediction of stock returns. *Journal of Accounting and Economics, 11*(4), 295-329. https://doi.org/10.1016/0165-4101(89)90017-7

36. Petropoulos, A., Siakoulis, V., Stavroulakis, E., & Vlachogiannakis, N. (2020). Predicting bank insolvencies using machine learning techniques. *International Journal of Forecasting.* https://doi.org/10.1016/j.ijforecast.2019.11.005

37. Qiu, X. Y., Srinivasan, P., & Hu, Y. (2014) Supervised Learning models to predict firm performance with annual reports: An empirical study. *Journal of the American Society for Information Science and Technology, 65*(2), 400-413. https://doi.org/10.1002/asi.22983

38. Quinlan, J. R. (1992). Learning with continuous classes. In *Proceedings of the 5th Australian Joint Conference on Artificial Intelligence* (pp. 343-348). Hobart, Tasmania.

39. Quinlan, J. R. (1993). Combining instance-based and model-based learning. In *Proceedings of the Tenth International Conference on Machine Learning* (pp. 236-243). Amherst, MA, USA.

40. Core Team R (2017). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/

41. Ridgeway, G. (2007). *Generalized Boosted Models: A guide to the gbm package*. Retrieved from http://cran.r-project.org/web/packages/gbm/vignettes/gbm.pdf

42. San Ong, T., Yichen, Y. N., & Teh, B. H. (2010). Can high price earnings ratio act as an indicator of the coming bear market in the Malaysia? *International Journal of Business and Social Science, 1*(1), 194-213. Retrieved from http://ijbssnet.com/journals/19.pdf

43. Strobl, C., Malley, J., & Tutz, G. (2009). An Introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods, 14*(4), 323-348. https://doi.org/10.1037/a0016973

44. Tsai, C., & Wang, S. (2009). Stock price forecasting by hybrid machine learning techniques. *Proceedings of the International Multi-Conference of Engineers and Computer Scientists, Vol I IMECS 2009, Hong Kong*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.149.970&rep=rep1&type=pdf

45. Westfall, P., & Young, S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*. Wiley. Retrieved from https://www.amazon.com/Resampling-Based-Multiple-Testing-Examples-Adjustment/dp/0471557617

46. Zekic-Susac, M., Sarlija, N., & Bensic, M. (2004). Small business credit scoring: a comparison of logistic regression, neural network, and decision tree models. *26th International Conference on Information Technology Interfaces* (pp. 265-270). Cavtat, Croatia. Retrieved from https://ieeexplore.ieee.org/document/1372413

47. Zhang, W., Cao, Q., & Schniederjans, M. (2004). Neural network earnings per share forecasting models: A comparative analysis of alternative methods. *Decision Sciences, 35*(2), 205-237. https://doi.org/10.1111/j.00117315.2004.02674.x

APPENDIX

Cubist regression run – Model 20
Call:
cubist.default(x = x, y = y, committees = param$committees)
Cubist [Release 2.07 GPL Edition]

Target attribute 'outcome'
Read 566 cases (7 attributes)

Model 20:
Rule 20/1: [343 cases, mean 0.146556, range −1.37092 to 3.23486, est. err. 0.216391]
if
logTOTA <= 4.217362
then
outcome = −0.564935 + 0.17 *logTOTA*

Rule 20/2: [14 cases, mean 0.219289, range −0.32118 to 0.42896, est. err. 0.203703]
if
logTOTA > 4.217362
BOKV <= 0.4932171
AGEB > 38
then
outcome = −5.016187 + 1.15 *logTOTA* − 0.95 *BOKV* + 0.23 *SVOL*

Rule 20/3: [11 cases, mean 0.635500, range 0.04462 to 1.3659, est. err. 0.520912]
if
logTOTA > 4.217362
logTOTA <= 4.835725
BOKV > 0.4932171
BOKV <= 0.7001268
then
outcome = −0.78711 − 6.59 *BOKV* + 1.69 *logTOTA* − 0.496 *LIAB* + 0.48 *SVOL*

Rule 20/4: [19 cases, mean 1.214434, range 0.03667 to 2.48613, est. err. 1.103157]
if
logTOTA > 4.217362
logTOTA <= 5.091389
BOKV > 0.980139
then
outcome = −45.665448 + 17.76 *BOKV* + 6.06 *logTOTA*

Rule 20/5: [24 cases, mean 1.507473, range 0.02826 to 5.63544, est. err. 2.141038]
if
logTOTA > 4.217362
logTOTA <= 5.091389
LIAB > 8.335691
BOKV > 0.7508169
BOKV <= 0.980139
SVOL > 0.913998
then
outcome = −34.242879 + 29.05 *BOKV* + 1.86 *logTOTA* + 0.33 *SVOL*

Rule 20/6: [46 cases, mean 1.521246, range 0.04462 to 6.05082, est. err. 1.406400]
if
logTOTA > 4.217362 б
logTOTA <= 4.825879
LIAB <= 8.335691
BOKV > 0.4932171
then
outcome = −23.272601 + 5.88 *logTOTA* − 0.765 *LIAB* + 2.26 *SVOL* − 0.57 *BOKV*

http://dx.doi.org/10.21511/imfi.17(2).2020.05

Rule 20/7: [20 cases, mean 1.684005, range 0.02826 to 6.05082, est. err. 1.349789]
if
logTOTA > 4.217362
logTOTA <= 5.091389
BOKV > 0.7508169
BOKV <= 0.8331338
SVOL > 0.913998
then
outcome = −6.820059 + 4.64 *SVOL* + 0.64 *logTOTA* − 0.84 *BOKV*

Rule 20/8: [20 cases, mean 1.749578, range 0.78797 to 3.49926, est. err. 0.828504]
if
logTOTA > 5.091389
AGEB > 43
then
outcome = −21.591626 + 5.03 *logTOTA* + 0.854 *LIAB* − 4.02 *BOKV*
− 0.071 *AGEB* − 1.32 *SVOL*

Rule 20/9: [43 cases, mean 1.902359, range −0.10022 to 5.021, est. err. 0.716480]
if
logTOTA > 5.091389
AGEB <= 43
then
outcome = 12.143369 − 3.24 *logTOTA* + 0.168 *AGEB* + 0.57 *SVOL*
− 2.9e-005 *NCSH*

Rule 20/10: [12 cases, mean 2.102828, range 0.27522 to 3.07909, est. err. 1.452597]
if
logTOTA > 4.217362
BOKV <= 0.3715467
AGEB <= 38
then
outcome = −2.542574 + 0.001681 *NCSH* + 1.71 *logTOTA* − 0.196 *LIAB*
− 0.06 *BOKV*

Rule 20/11: [125 cases, mean 2.263881, range 0.02826 to 13.53065, est. err. 1.708024]
if
logTOTA > 4.217362
logTOTA <= 5.091389
BOKV > 0.4932171
then
outcome = −1.787032 − 12.16 *BOKV* + 3.09 *logTOTA*

Rule 20/12: [9 cases, mean 2.705642, range 1.16072 to 5.65934, est. err. 1.951760]
if
logTOTA > 4.217362
BOKV > 0.3715467
BOKV <= 0.4932171
AGEB <= 38
then
outcome = 10.439658 − 32.02 *BOKV* + 0.432 *LIAB* + 0.17 *logTOTA*

Rule 20/13: [21 cases, mean 3.672726, range 0.93146 to 13.53065, est. err. 1.703290]
if
logTOTA > 4.825879
logTOTA <= 5.091389
LIAB <= 8.335691
BOKV > 0.7001268
then
outcome = 8.096202 − 1.13 *LIAB* + 0.46 *logTOTA* + 0.19 *SVOL* − 0.28 *BOKV*

Rule 20/14: [10 cases, mean 4.042699, range 1.18529 to 5.52219, est. err. 4.533568]

if
logTOTA > 4.217362
BOKV > 0.7001268
BOKV <= 0.7508169
SVOL > 0.913998
then
outcome = −22.225696 + 6.87 *logTOTA* − 5.24 *BOKV* + 0.93 *SVOL*

Rule 20/15: [12 cases, mean 4.381435, range 0.117 to 13.53065, est. err. 4.549168]
if
logTOTA > 4.217362
logTOTA <= 5.091389
BOKV > 0.4932171
SVOL > 0.6851702
SVOL <= 0.913998
then
outcome = −16.584116 + 27.34 *BOKV*

Rule 20/16: [6 cases, mean 7.318485, range 1.81329 to 13.52586, est. err. 3.257550]
if
logTOTA > 4.835725
logTOTA <= 5.091389
BOKV > 0.4932171
BOKV <= 0.7001268
SVOL > 0.913998
then
outcome = −525.245474 + 109.4 *logTOTA* − 11.24 *BOKV* − 0.444 *LIAB*
+ 0.81 *SVOL*

Evaluation on training data (566 cases):
Average |error| 0.252853
Relative |error| 0.23
Correlation coefficient 0.94

Attribute usage:
Conds model

91% 87% *logTOTA*
45% 30% *LIAB*
29% 37% *AGEB*
27% 61% *BOKV*
10% 41% *SVOL*
1% 6% *NCSH*

http://dx.doi.org/10.21511/imfi.17(2).2020.05