ORIGINAL RESEARCH ARTICLE

J.F. Mkt. Gov. (2024) Vol. 3, Issue-1

DOI:https://doi.org/10.54728/JFMG.202310.00072

Value Relevance of Key Accounting Information in Predicting Market Performance: A Machine Learning Approach

Imran Mahmud*
Bangladesh Institute of Capital Market
Email: imran.mahmud@bicm.ac.bd

Faima Akter
Bangladesh Institute of Capital Market
Email: faima.akter@bicm.ac.bd

Mostafiz Ahammed
Chandpur Science and Technology University
Email: mostafiz@cse.cstu.ac.bd

Keywords
Tobin's Q, Market price Price prediction, Machine learning, Random Forest Regression, Long short term memory

JEL Classification
G14, C23, C58, M41

Manuscript Number:
JFMG-202310-00072

Received: 02 October, 2023
Accepted: 10 June, 2024
Published Online: 09 September 2024
Published in Print: 20 September 2024

ISSN (Online): 2958-9290
ISSN (Print): 2958-9282

***Corresponding author**

Abstract

This study aims to examine the value relevance of key accounting information in predicting market performance and to explore which accounting information has the most value relevance in predicting market performance. The dataset is comprised of a total of 1401 observations from 117 companies over a period of 2010 to 2021. To conduct the study, total 12 indicators were taken from financial statements to represent the major accounting information (AI) and year-end market price and Tobin's Q were taken as a proxy for market performance. Under the principal component analysis (PCA), it was found that only 3 AIs namely earnings per share (EPS), book value per share (BVPS) and cash flow per share (CFPS) could contribute to the whole explained variance ratio of market performance. The study employs several machine learning approaches: random forest regression (RFR) was chosen as the base line model and the result was compared with decision tree regression (DTR), long short term memory (LSTM), neural network model and multivariate regression model. After splitting the total observation into a 70:30 training-testing dataset and controlling for noise reduction, it was found that over any other models, LSTM and RFR models can predict the market performance with higher R-square value of 62% and 65% respectively along with the lowest MSE of all other models. It was found that EPS has the highest value relevance (factor importance) in predicting market profitability whereas BVPS and CFPS were found to have less than 10% factor importance in predicting market profitability.

1.0 Introduction

Over the last decades, the capital market and world economy have integrated to a great extent and investors and shareholders are more aware of high-quality recognized financial reporting standards to use as a guidebook. Investors at home and abroad will look for reliable and relevant financial information to protect them from fraud or any misleading financial data while investing or disinvesting (Wagdy, 2001). The information on how the firm performs is communicated to the investors, shareholders, creditors, and stakeholders through financial reporting. The main objective of financial statements is to disseminate information about the company so that users of information particularly investors can make better decisions (Germon and Meek, 2001). The convergence of IFRS with GAAP ensures transparent, comparable, and consistent financial information to provide investors with the optimal investment decision-making capacity (Jacob et al, 2009).

Financial statements provide different types of information: Accounting Information (AI) and non-accounting information. Accounting Information is information that describes an account for a utility. It processes financial transactions to provide external reporting to outside parties such as stockholders, creditors, and investors. In contrast, non-accounting information cannot be measured in monetary terms provided in the management discussion in the notes to the statements (Perera and Thrikawala, 2010). Accounting Information (AI) i.e. earnings growth, the net profit, and the cash flows of the firm are the critical source of information that allows the investors to make decisions to buy or hold a share. In Bangladesh, after the share market collapse of 1995 and 2010 specifically, the Securities Commission along with the government has been promoting investment education nationwide to make people aware of the investment guidelines and the use of accounting information in decision-making. Two institutes i.e. Bangladesh Institute of Capital Market (BICM) and Bangladesh Academy of Securities Market (BASM) were created to provide and disseminate investment education to various stakeholders of the capital market. However, despite all the efforts to develop the capital and financial market, there are incidences where the stock market experiences hyped trades of stocks with fundamentally lower accounting values. So it raises a conspicuous inquiry regarding the relevance of critical accounting information in determining the market performance of stocks.

There is different literature available regarding the value relevance of accounting information in predicting market prices in emerging countries and other Asian countries(Ball and brown 1968, Vishnani S., Shah B.K., 2008, Dontoh A., Radhakrishnan S., Ronen J., 2000, Dung, 2010, Hadi, 2004, OyerindeD.T., 2009), to the best of our knowledge, there is no empirical studies conducted to dig out the relevance of different accounting information published in annual reports in predicting the market performance of the stocks.

Thus, this study aims to identify the key critical AIs that are likely to contribute much to market performance among all the published AIs and how important those AIs are in predicting market performance.

2.0 Literature review

Within the framework of creating and expressing a company's wealth, accounting plays a crucial role. The most significant source of externally verifiable information about corporations is still financial statements. There is some concern that accounting theory and practice have not kept up with the rapid economic developments and high technological advances despite their widespread usage and ongoing progress (Meyer et al, 2007). The relevance of accounting information is affected by this situation.

A value relevance investigation evaluates how accounting information and capital market values relate to one another. According to Beaver (2002), the theoretical underpinnings of value relevance studies using a measurement approach are a combination of valuation theory and contextual accounting and financial reporting arguments (accounting theory), which enables the researcher to forecast how accounting variables and other information relating to market value will act. According to Holthausen and Watts (2001), value relevance studies can conclude two separate theories of accounting and standard setting: (I) "direct valuation" theory and (ii) "inputs-to equity-valuation" theory. The theory of direct valuation suggests a connection between accounting earnings and stock market value. The latter approach follows that the valuation of stock is dependent on accounting information.

Over four decades ago, the first studies attempting to demonstrate a correlation between accounting figures and equity values were published. Miller and Modigliani (1966) published the first such article, which used data from the electricity sector to show that capitalized earnings on assets influence most to market value. The foundational works on the information value of accounting numbers are typically regarded as Ball and Brown (1968) and Landsman et al. (1968). Landsman noted both price and volume responses to earnings announcements in addition to the relationship between the information content of the earnings number and stock prices that Ball and Brown demonstrated. Accounting researchers have been documenting the diminishing value relevance of accounting data, such as net income and return on equity, in predicting stock prices for many years. Lev and Gu (2016) contend that poor intellectual property accounting standards, notably the expensing of R&D, are to blame for this fall in relevance.

On the one hand, previous research has revealed that accounting information's value and relevancy have deteriorated recently (Core et al, 2003; Marquardt & Wiedman., 2004). On the Vietnamese stock market, Dung (2010) examined the value-relevance of financial statement information. The study revealed that accounting's value relevance was statistically significant, albeit considerably less so than in other established and emerging markets. Accounting data, according to Amir and Lev (1996), is useless for businesses that rely heavily on intangible assets.

On the other hand, several studies also have been carried out in recent years that showed the value relevance of accounting information has increased. Cooke et al. (2009) investigated the time series relationships of five Japanese conglomerates during the years 1950–2004 to determine the extent of the long-run explanatory power of the book value of net assets for market value. Their findings demonstrated that there is evidence of a long-term association between market value and net book value of assets in four out of the five firms. In Sri Lanka, Perera and Thrikawala (2010) discovered relations between the market price per share and particular accounting data for 5 years. Their research revealed a connection between accounting information and market price per share.

For 38 companies, Glezakos et al. (2012) connected book value and EPS to share prices on the Athens Stock Exchange. hey demonstrated that book value and EPS became more valuable over time. According to Alali and Foote (2012), earnings have a positive correlation with cumulative returns, and earnings per share and book value per share have a positive correlation with share prices on the Abu Dhabi Stock Exchange. Collins et al. (1997), on the other hand, discovered that the value relevance of accounting information has increased over time and is now equal to or higher than that for industries with a significant concentration of intangible assets.

Moreover, there have been numerous studies that focus on predicting the market price of shares using artificial intelligence that is, using different machine learning approaches.

Agarwal et al. (2022) used deep learning models to predict stock prices based on technical analysis. Baheti et al. (2021) used social media and news to analyze the market performance using a machine learning algorithm. Cagliero et al (2023) and Hu et al. (2019) tried to provide recommendations on stock trading using candlestick pattern recognition. Khairi et al. (2019) showed that using technical analysis, fundamental analysis and news in machine learning can improve the prediction of stock market performance. However, few studies have compared all the machine learning models and identified which one outperform in identifying market price. Also, few studies have incorporated accounting information from financial statements to predict the market price.

The most recent study on value relevance using a machine learning approach also found a more complicated, but not diminishing, relationship between accounting information and share price (Barth et. all, 2023). The study examined additional accounting elements and discovered no change in overall value relevance from 1962 to 2018. The study also examined the progression of each item's value relevance and discovered rises, most notably for items connected to intangible assets, growth prospects, and alternative performance measurements, all of which are significant in the new economy. The quantity of relevant items also grows. They looked at the new economy, old economy profit, and old economy loss enterprises separately. The changes were more obvious for new economy enterprises, but they extended beyond them. Inferences were based on a non-parametric technique that did not require the valuation connection to be specified.

This study is unique in identifying the machine learning model that can best predict the market price of the shares. This study also differentiates itself from other stock price predicting studies in that it incorporates different accounting information as input to predict market price compared to historical time series trade-related information and sentiment-reflecting information (Lim and Tan, 2021). This study also incorporates a huge dataset comprising data from 117 companies of all manufacturing companies of Bangladesh making the study more representative of the market performance.

3.0 Research questions

The study's main goal is to investigate the impact of accounting information on forecasting market price of stocks. Based on the objective of the study, the main research questions are:

► Does critical accounting information have any value in predicting market performance?

► Which accounting information has more importance in predicting market performance?

4.0 Research methodology

4.1 Data source

For the study, we have employed several machine-learning approaches. The first step in implementing machine learning models is collecting relevant data. The data for this study is collected from the annual reports of all manufacturing companies existing from the period 2010-2021. In total 19 sectors have been covered in this study. Table 1 illustrates the list of sectors and the number of companies used in this study. Accounting information (AI) was collected from the annual reports. The price data have been collected from the Dhaka Stock Exchange. Companies that were listed after 2010 were not incorporated into the study due to the fractional nature of the datasets.

In total 1402 observations were taken for the study.

Table 1: Sectors and number of companies used in the study.

Name of Sectors	Number of Companies	Name of Sectors	Number of Companies
Paper & Printing	1	Textiles	22
IT Sector	5	Food & Allied	13
Travel & Leisure	1	Pharmaceuticals and Chemicals	18
Telecommunications	1	Cement	5
Engineering	17	Miscellaneous	8
Jute	2	Ceramics	5
Fuel & Power	11	Tannery	5
Services and Real Estate	3	**Total**	**117**

Source: Author's own

4.2 Variable selection

The raw data collected is not suitable for machine learning models. Missing values or outliers are needed to be fixed to improve the accuracy of the model. For the dataset, the closing market price was considered raw and it was transformed into a log version of market price. There are several accounting information that are needed to understand the financial position and financial performance of the company. Among them, we have taken 12 accounting information as independent variables for the study using the methodology of Barth et al. (2023). Some accounting information was dropped due to the unavailability of the data. In Table 2 the list of variables and their expected relationship with market performance is illustrated. All 12 variables including EPS, CFPS, and BVPS are all in ratio format. For the robustness of the study, Tobin's Q was also tested as the proxy of market performance. Due to the fragmented nature of the data between 1994 and 2009, the period was discarded from the study. Data were collected from 2010 to 2021 so that the dataset does not have any undue weightage on the scam of 2011 thereby 2010 was taken as the period of stability. For the robustness of the dataset, predictor variables were standardized using the RobustScaler function of Python.

Table 2: Operational definition of variables

Measures	Operational variables	Conceptual formula	Expected sign	Sources
Key Accounting Information (Predictor)	EPS	Net income available for common stock holders/no of shares outstanding	(+)	Miller and Modigliani (1966); Core at al. (2003)
	CFPS	Cash flow from operations/no of shares outstanding	(+)	Palepu and Healy (2008)
	BVPS	Book value of total equity/no of shares outstanding	(+)	Barth et al. (2023)
	R&DPS	Research & Development/ no of shares outstanding	(+)/ (-)	Core at al. (2003)
	SPS	Sales/ no of shares outstanding	(+)	Kothari and Shanken (2003)
Key Accounting Information (Predictor)	COGSPS	COGS/ no of shares outstanding	(-)	Barth et al. (2023).
	OEPS	Operating expense/ no of shares outstanding	(-)	Kothari and Shanken (2003)
	PPES	Property, plant and equipment (PPE)/ no of shares outstanding	(+)	Floyd, Li, and Skinner (2015)
	CAPS	Current Assets (CA)/ no of shares outstanding	(+)	Palepu and Healy (2008)
	Dividend%	Dividend declared in percentage of face value	(-)	Floyd, Li, and Skinner (2015)
	DPS	Total Debt/ no of shares outstanding	(-)/ (+)	Barth et al. (2023).
	APS	Total Assets/ no of shares outstanding	(+)	Floyd, Li, and Skinner (2015)

Measures	Operational variables	Conceptual formula	Expected sgn	Sources
Dependent Variables (predicted)	Tobin's Q	(Market Value of Equity+ Total debt)/ Total tangible Assets		Mahmud et al. (2021)
	Price	Ln price of a stock at the end of the accounting year		Barth et al. (2023)

Source: Author's own

4.3 Machine learning approach

In this study we have used a five step approach to machine learning using the methodology of Sonkavde et al. (2023) illustrated in figure 1. The steps described in the following section:

Step 1: Loading the dataset into python from a csv file and preprocessing the data by removing the null value, duplicates and scaling using robust scalar.

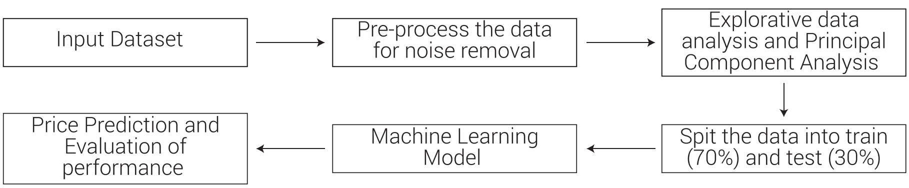
Step 2: Performing an explorative data analysis using mean, median, standard deviation to understand the basic pattern of the dataset. Running Principal Component Analysis (PCA) to select the variables that can explain the most variance of the predicted variables.

Step 3: Dividing the preprocessed data into training and testing data with 70% of the data used to train the model and the remaining 30% to test the model prediction power.

Step 4: Employing different basic machine learning algorithm and ensemble methods on the training dataset and creating cross validation model with K-fold validation score to calculate the accuracy.

Step 5: Evaluating the model performance using accuracy score, mean absolute error, mean square error and root mean square error.

Figure 1: Machine learning flow chart



4.3.1 Principal Component Analysis

Principal Component Analysis (PCA) is a data dimensionality reduction technique extensively employed in diverse domains, including statistical analysis of variables, pattern recognition, feature extraction, data compression, and the visualization of high-dimensional datasets (Hotelling, 1933; Jolliffe, 2011). Principal Component Analysis, or PCA, is a mathematical method for rearranging data into a different coordinate system. The first coordinate, or first principal component, in this new approach corresponds to the largest variation in the data. This trend holds true for the next components, with the second most significant variance matching the second coordinate. In essence, by converting the data into a more understandable and straightforward representation, PCA assists us in identifying and concentrating on the most significant variables in the dataset leaving the less significant variables outside the periphery. Moreover, PCA finds applications in the exploration of financial time series, the development of dynamic trading strategies, the computation of financial risk measures, and the practice of statistical arbitrage (Ince, 2007; Fung and Hsieh, 1997; Alexandar, 2009; Shukla and Trzcinka, 1990). Within the context of this research, we leverage PCA to forecast forthcoming stock price values by identifying the variables with highest variance-explanation power.

4.4 Machine learning models

After initial prepossessing the dataset, the whole 1402 observations were split into training and testing dataset using a split of 70:30 (Joseph, 2022). After the split, we have run different machine learning models using Python.

4.4.1 Linear regression

Linear regression is a simple type of supervised learning that aims to find a relationship between independent variables and price to provide continuous valued output. Machine learning involves teaching machines to identify patterns in data (Mustafa, 2019). The equation of linear regression for the model of the study is given as

$$lnprice = a + B1 * EPS + B2 * CFPS + B3 * BVPS$$

This regression is different and most robust compared to Ordinary least squares. Although they are closely related, linear regression and OLS are not the same thing. OLS is a technique used to estimate the parameters of a linear regression model, whereas linear regression is the more general idea of modelling the connection between a dependent variable and one or more independent variables. In machine learning, linear regression is used to create the independent variable parameters from the training dataset and then utilise those parameters for accurately predicting the price.

4.4.2 Decision tree regression

One of the popular data mining methodologies is the decision tree, which concurrently performs classification and prediction tasks. Using inductive reasoning and the given data, it creates a model of a tree-shaped structure. (Chang & Chen, 2009).The primary algorithm for creating decision trees, called ID3, was created by J. R. Quinlan and incorporates a top-down strategy, greedy search, and a space of potential branches with no backtracking. Repositioning information gain with Standard Deviation Reduction allows the ID3 algorithm to construct the tree in a decision tree for regression. (SDR). At the onset, Decision tree regressor algorithm predicts the target variable (lnPrice) using the AI given using standard deviation. Then, it splits the data into distinct attributes, and selecting the attribute with the highest standard deviation reduction as the decision node thus creating the first node. The data is then split based on the selected attribute, and the process is repeated. If the coefficient of deviation is less than the threshold, the subset of the dataset does not need further splitting, and the related leaf node has the average of the subset dataset. The process stops when the number of data points for all branches is equal or less than the number of branches, and the related leaf node is assigned the average of each branch (Polamuri et al, 2019).

4.4.3 Random forest regressor (RFR)

Random forest regressor is an advanced version of classification and regression tree (CART). It generates hundreds of decision tree predicted regression using bootstrap samples of the testing dataset given. The final output of the RFR is the average of predictions from the individual trees. Since individual trees produce multidimensional step functions, their average is again a multidimensional step function that can nevertheless predict smooth functions because it aggregates a large number of different trees (Breiman, 2017). Being a non-parametric model, it follows similar methodology like Decision tree regression. It provides a better prediction performance since it construct multiple de-correlated decision trees based on randomized subset of predictors and hence called “random” forest (Breiman, 2001). It controls over-fitting and improve the predictive accuracy. RFR algorithm is very user friendly that requires two parameters to run: the number of trees (n_{trres}) and the number of random variables or features for each split (m_{try}). In general the higher the number of trees in the forest, the excellent the prediction accuracy. It has been reported that that the default value of m_{try}, one-third of the number of all predictor variables, is often a good choice (Liaw et al, 2002). For the study n_{trres} = 500 and m_{try} = 1 has been used since the variables used in the study was 3. To summarize, the basic steps of RFR are given in the Table.

Step 1: Take ntree bootstrap samples from the original data
Step 2: randomly sample mtry of the predictors and choose the best split from among those variables
Step 3: Based on n trees, get predicted values of the price from all the decision trees.
Step4: Average the regression results of predicted values from all regression trees.

4.4.4 LSTM Model

Long short-term memory neural network is one of the advanced models to predict future values especial when data has longer term trend. This model is designed in such a way so that long term dependency problem due to recurrent neural networks can be overcome. LSTMs differ from more conventional feed forward neural networks in that they feature feedback connections. With the use of this property, LSTMs may process whole data sequences without considering each data point separately. Instead, they can process new data points by using the information from earlier data in the sequence to aid in their processing. Because of this, LSTMs excel at processing data sequences like text, audio, and general time-series.

4.4.5 Neural network

Neural networks learn (or are trained) by analyzing samples that have a known "input" and "result," creating probability-weighted connections between the two, which are stored within the net's data structure. Neural networks have found extensive application in the field of finance, revolutionizing various aspects of the industry. They excel in risk assessment and management by analyzing large datasets to detect patterns and anomalies, making them indispensable for fraud detection and prevention. Neural networks also play a crucial role in algorithmic trading, where they analyze market data in real-time to identify profitable trading opportunities. Furthermore, neural networks are employed in forecasting financial market trends, aiding investors in making informed decisions. LSTM and Neuron models have been used in this study just to check the robustness of other used models in the study.

When applying the models to the financial dataset, we encountered several challenges. The dataset consisted of 1401 observations and included accounting information from different sectors and companies. Obtaining this information and inputting it in a noise and error-free way proved to be difficult. Additionally, the financial dataset exhibited non-linearity and non-stationary behavior, making it challenging for linear regression models to capture the relationship. In contrast, LSTM and CNN models were better suited for capturing non-linear patterns. Lastly, financial market datasets are highly linked with regulatory and ethical concerns. However, the models cannot consider these factors when predicting market prices.

4.5 Model evaluation tools

To compare the model performance, actual price of the dataset must be compared with predicted price. Different evaluation metrics like R-squared, Mean squared error, Mean absolute error, root mean square error and k fold validation score is used in this study to evaluate the models.

Mean Absolute Error (MAE)

It is calculated by taking the average of the absolute differences between the predicted and actual price. The smaller the MAE value, the higher the prediction accuracy. The equation for calculating MAE is:

MAE = $(1/n) * \sum_{1}^{n}|y_i - \hat{y}_i|$; Where n is the number of observations, Yi is the actual value of the i-th observation, ŷi is the predicted value of the i-th observation.

Mean Square Error (MAE)

The MSE is the average of the squared differences between the actual and predicted numbers. It is determined by dividing the total number of data points by the sum of the squared differences. A lower MSE number signifies that the model is more effective at forecasting the values in the dataset.

$MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}$; Where n is the number of observations, Yi is the actual value of the i-th observation, ŷi is the predicted value of the i-th observation.

Root Mean Square Error (RMSE)

When comparing values predicted by a model with values actually observed, the Root Mean-Square Error (RMSE) is commonly used as a measurement. It is very similar to MAE, but it penalizes bigger absolute values by giving them more weight than the MAE. The variance in the individual errors increases as MAE and RMSE diverge more widely. RMSE is defined by the following equation:

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

Goodness of fit (R2)

R2 is a different metric for gauging how closely a model's projected outcomes match actual values. The MAE, ME, and RMSE are nearly zero when the predicted values are close to the measured values. On the other hand, an R2 value near 1 denotes good agreement between observed and predicted data.

$$R2 = 1 - \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{(y_i - \bar{y})^2};$$

Goodness of fit (R2)

R2 is a different metric for gauging how closely a model's projected outcomes match actual values. The MAE, ME, and RMSE are nearly zero when the predicted values are close to the measured values. On the other hand, an R2 value near 1 denotes good agreement between observed and predicted data.

Where n is the number of observations, yi is the actual price of the i-th observation, ŷi is the predicted price of the i-th observation and ȳ is the mean of actual price.

Adjusted R-Square

Unlike R-squared, which may increase even if new predictors do not enhance the model, adjusted R-squared takes into account the number of predictors and adjusts the R-squared value accordingly to compensate for the inclusion of extraneous predictors.

k-fold validation score

K-fold cross-validation is a technique for training and assessing a model K times after dividing a dataset into K identical "folds" of equal size. Each fold serves as the training set, and the remaining K-1 folds serve as the validation set. As a result, the model is tested on various data subsets and

and its performance is more thoroughly assessed. After K-fold cross-validation, a performance statistic known as the K-fold validation score is created. This score often shows how effectively the model generalizes to new data. The average of the results from all K iterations constitutes the final score. To cross validate the testing result, 5-folds validation dataset have been used in this study. While performing the K-fold cross validation test, the entire dataset was a split on time series basis taking a moving average of 8 years to prevent the dataset from look ahead bias.

5.0 Empirical results

5.1 Descriptive statistics

For a total 1401 observations of all manufacturing companies listed in Dhaka Stock Exchange, the study first shows the mean, standard deviation, lowest and maximum values for 117 companies in total. The optimum machine learning model for predicting the price is then determined by employing all the findings from training dataset into the testing dataset.

Table 3: Descriptive statistics before scaling

Variables	Measurement scale	Obs	mean	std	min	max
LnReturn	Logarithmic	1401	4.48	1.48	1.28	9.47
Tobins Q	Ratio	1401	7.53	87.90	0.00	2489.62
EPS	Ratio	1401	7.36	23.50	-166.90	269.80
BVPS	Ratio	1401	59.21	124.69	-1114.52	1045.11
CFPS	Ratio	1401	10.65	37.87	-211.07	371.82
R&DPS	Ratio	1401	0.09	1.03	0.00	23.95
Dividend	Percentage	1401	0.39	1.21	0.00	16.50
cogsps	Ratio	1401	130.74	455.23	0.00	6708.68
opexps	Ratio	1401	35.55	147.45	0.00	3515.09
ppeps	Ratio	1401	65.16	116.97	0.00	1933.24
caps	Ratio	1401	108.50	230.65	0.00	3780.12
sps	Ratio	1401	161.64	504.80	0.00	7164.42
assetps	Ratio	1401	181.18	316.56	0.00	4157.37
debtps	Ratio	1401	63.52	168.76	0.00	2227.52

Source: Author's calculation

Analyzing the data used in the study, the descriptive statistics namely mean, standard deviation, minimum and maximum value for all the accounting information across all year and across all companies is shown in table 3. It is evident that the trend of all the variables are significantly disperse with minimum value being 0 since many companies had suspended operations a number of years.

5.2 Principal Component Analysis (PCA)

Given that there was 12 accounting information with 1401 observations, the primary objective of our study was to identify the principal components that best captured the most relevant information of the dependent variable (lnreturn). To accomplish this, we performed PCA to determine the most crucial and representative variables from the selected accounting information that would be used to form the training dataset. The results of PCA were measured by the explained variance ratio, and we utilized the PCA transformed package of Python for this analysis. Table 4 depicts the results of PCA and indicates that EPS, BVPS, and CFPS alone cover almost 99% of the variance of the market price. Therefore, in the subsequent stages of the study, we focused on these three variables to test and forecast the market price.

Table 4: Result of PCA

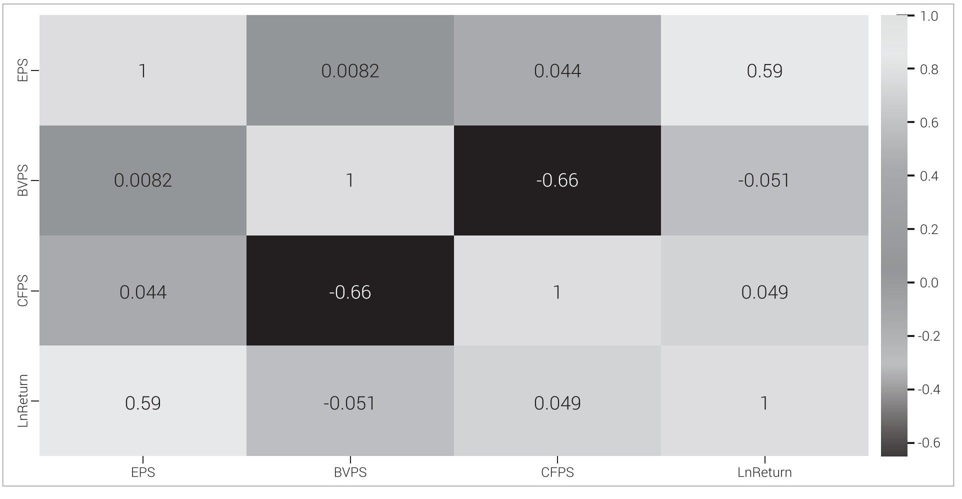
Variables	Explained variance ratio	Variables	Explained variance ratio
EPS	0.902	opexps	0.001
BVPS	0.075	ppeps	0.000
CFPS	0.010	caps	0.000
R&DPS	0.006	sps	0.000
Dividend	0.003	assetps	0.000
cogsps	0.002	debtps	0.000

Source: Author's calculation

5.3 Correlation heat map

The correlation heat map is created after performing the PCA test which recommended that EPS, CFPS and BVPS explains almost 98% of the variance of the dependent variable. The heat map in figure 2 shows a substantial positive relationship between EPS and market performance which is about 0.59. The market price is positively correlated with cash flow per share as well however a negative relationship exists between book value per share and price. It suggests that the investors somewhat don't concert with improving market performance in terms of all the crucial variables. According to the heat map, BVPS and CFPS exhibit a significant negative correlation of -0.66. To ensure that multicollinearity was not an issue when performing linear (OLS) regression, we evaluated the VIF score. The VIF test yielded a score of 1.65, which is less than 2, indicating that there was no multicollinearity when regressed against lnreturn.

Figure 2: Correlation Heatmap



Source: Author's calculation

5.4 Findings of machine learning models

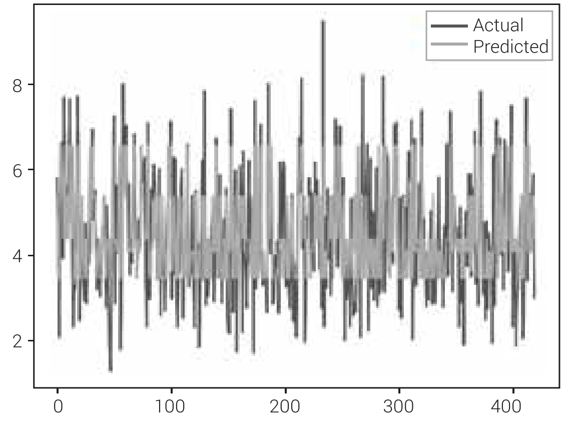
In this section, the findings from our study with regard to the ensemble and deep learning models used have been shown. First, the performance evaluation of each models have been discussed and then the variable importance [the most value relevant] accounting information in price forecasting is explained.

Table 5: Result of training and testing dataset

Performance Measurement	Training dataset	Testing dataset
R^2	0.63	0.66
Adjusted R^2	0.63	0.45
MSE	0.81	0.72
RMSE	0.9	0.84
MAE	0.67	0.62

Source: Author's calculation

Figure 3: Plotting of actual price and predicted price under Random Forest Model



Source: Author's calculation

Random forest regression (RFR)

Table 5 and Figure 3 present the results of price prediction using the RFR model. The table indicates that the training dataset was well-trained, as evidenced by the R-squared and adjusted R-squared values, which reflect the accuracy of the prediction. These values were 0.63 and 0.63, respectively, and align with the performance measurement indicators of the testing dataset.

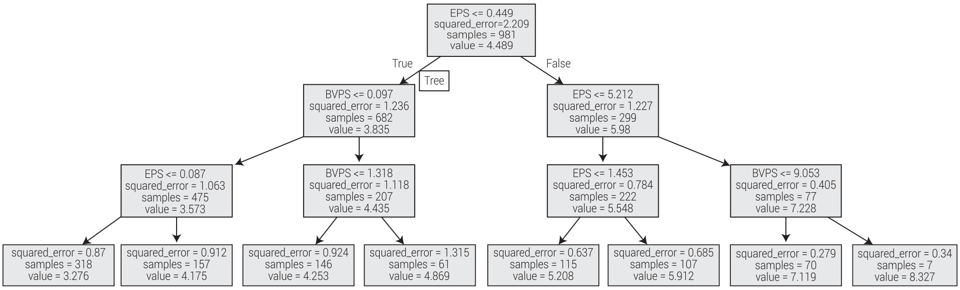
Decision tree regression (DTR)

Table 6: Result of training and testing dataset

Performance Measurement	Training dataset	Testing dataset
R^2	0.61	0.64
Adjusted R^2	0.60	0.42
MSE	0.85	0.76
RMSE	0.92	0.87
MAE	0.7	0.64

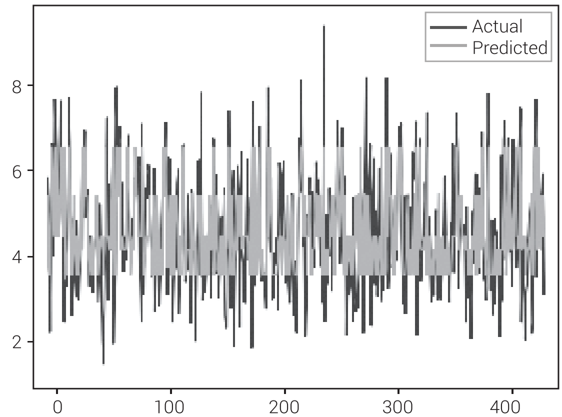
Source: Author's calculation

Figure 4: Diagram of Decision Tree



Source: Author's calculation

Figure 5: Plotting of actual price and predicted price under Random Forest Model



Source: Author's calculation

Table 6 present the results of price prediction using the DTR model. The table indicates that the training dataset was well-trained, as evidenced by the R-squared and adjusted R-squared values, which reflect the accuracy of the prediction. These values were 0.61 and 0.60, respectively, and align with the performance measurement indicators of the testing dataset.

Figure 4 displays the decision tree that was developed from the dataset. The root node of the tree begins with EPS, which is the prime feature and the initial split value that minimizes the squared error. Based on EPS as the root node, two child nodes use BVPS as subsequent features for the split. The leaf nodes in the decision tree consist of the mean value of the price for all the training observations in that node.

Figure 5 represents the graphical plotting of the actual price and the predicted price obtained from the testing dataset. The figure shows that there is a high scale of association between actual and predicted prices.

Figure 4 displays the decision tree that was developed from the dataset. The root node of the tree begins with EPS, which is the prime feature and the initial split value that minimizes the squared error. Based on EPS as the root node, two child nodes use BVPS as subsequent features for the split. The leaf nodes in the decision tree consist of the mean value of the price for all the training observations in that node.

Figure 5 represents the graphical plotting of the actual price and the predicted price obtained from the testing dataset. The figure shows that there is a high scale of association between actual and predicted prices.

Linear regression (LR)

Table 7: Result of training and testing dataset

Performance Measurement	Training dataset	Testing dataset
R^2	0.35	0.36
Adjusted R^2	0.35	0.63
MSE	1.41	1.34
RMSE	1.19	1.16
MAE	0.95	0.93

Source: Author's calculation

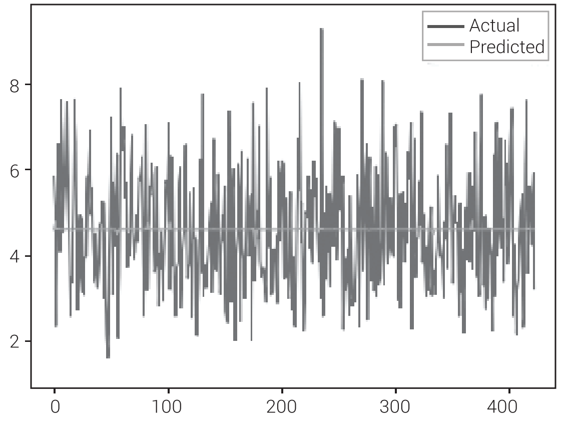
Table 8: Result of Linear Regression (OLS)

Variable	Coefficient(P-value)
EPS	0.22 (0.000)
BVPS	0.0066 (0.004)
CFPS	0.0006 (0.97)
R Square	0.28
Adjusted R Square	0.24
F-statistics	6.290 (0.000)

Source: Author's calculation

The LR model's results for price prediction are presented in Table 7 and Table 8. These tables indicate that the training dataset may have been poorly trained, as reflected by the low R-squared and adjusted R-squared values, which were both 0.35. The regression coefficient reveals that only EPS and BVPS are significant, with an overall R square of the regression of 0.28. However, the model was still significant, as evidenced by the p value of less than 5% for the F statistics.

Figure 6: Plotting of actual price and predicted price under Random Forest Model



Source: Author's calculation

Figure 6 represents the graphical plotting of the actual price and predicted price obtained from testing dataset. The figure shows that there is low level of accuracy of prediction between actual and predicted price.

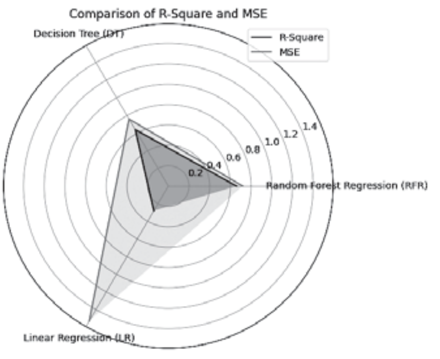
5.4.1 Performance evaluation

Table 9: Prediction performance evaluation of different ML models

	Testing dataset prediction evaluation		
Regression Model	**Random Forest Regression (RFR)**	**Decision Tree (DT)**	**Linear regression (LR)**
R^2	0.66	0.64	0.36
Adjusted R^2	0.45	0.42	0.63
MSE	0.72	0.76	1.34
RMSE	0.84	0.87	1.16
MAE	0.62	0.64	0.93
K folds validation score	0.62	0.59	0.34

Source: Author's calculation

Figure 7: Radar chart for comparison of performance under different ML models



Source: Author's calculation

Table 9 and Figure 7 show the performance evaluation of all the selected ML models. Random Forest Regression has the lowest MAE of 0.62 when compared to all of the specified machine learning models. It suggests that crucial accounting information's [EPS, BVPS, and CFPS] projections of the market price are on average 0.62 off. It also demonstrates that the model did well in predicting market performance. The mean squared error matrices are used to assess the performance of a machine learning model. The greater the MSE number, the worse the model's prediction power. Random Forest Regression produced the lowest MSE in this study when compared to other models. It suggests that key accounting information's market price estimates are on average 0.72 off. RMSE is more comparable to MAE, but it penalizes bigger absolute values by giving them more weight. The greater the disparity between MAE and RMSE, the greater the variation in individual errors. According to the study, RFR has the lowest RMSE of all other models, providing greater predictability of market performance through key accounting information. The higher the accuracy R2 score, the better the model predicts the actual market price using the testing dataset. RFR has the highest R2 value of 0.66, indicating that key accounting information can be utilized to significantly forecast market price. To cross-validate the testing result, a 5-fold validation dataset has been used in this study and a k-fold cross-validation score, which is the average of the 5 predictions from each validation testing dataset, is reported. According to the k-fold validation score, RFR has the highest prediction accuracy.

5.4.2 Value relevance of accounting information

Table 10: Value relevance of key accounting information under different ML models

Factor Importance [Value relevance]	Random Forest Regression	Decision Tree	Ordinary Least Squares (coefficient)
EPS	0.90	0.92112	0.22
BVS	0.09	0.05729	0.0066
CFPS	0.01	0.02157	0.0006

Source: Author's calculation

In the study, value relevance refers to the factor importance of selected accounting information in predicting market price. The study reveals, in Table 10, regardless of all three models, market price can be predicted accurately with only EPS for 90% of the time followed by BV per share and CF per share with least impact on the forecasting.

5.4.3 Robustness check

5.4.3.1 Alternate models

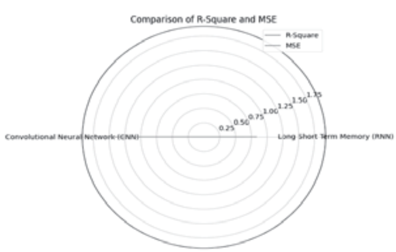
To test whether the RFR model used in our study provides the best performance matrix compared to other models, this study also employs deep learning models like recurrent neural network (RNN) and convolutional neural network (CNN). Table 11 and Figure 8 show the result of the testing prediction score for these two models. It is visible from the result that RNN can be a better tool compared to CNN in terms of predicting market price with 62% prediction accuracy score.

Table 11: Prediction performance evaluation of alternative deep learning ML models

Regression Model	Long Short Term Memory (RNN)	Convolutional Neural Network (CNN)
R^2	0.62	0.16
MSE	0.83	1.84
RMSE	0.91	1.35
MAE	0.93	0.93

Source: Author's calculation

Figure 8: Radar Chart for comparison of performance under alternative deep learning ML models



Source: Author's calculation

5.4.3.2 Alternate variable

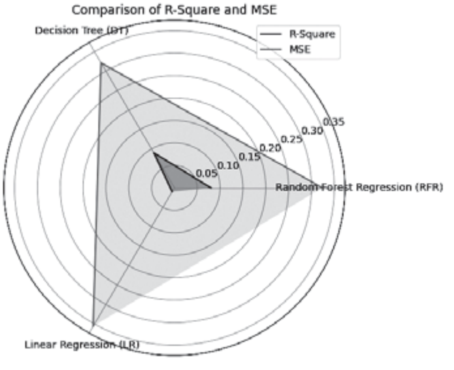
To test further, another popular measure of market performance, Tobin's Q, was used in the study. Tobin's Q assesses the relationship between a firm's market value and the replacement cost of its assets. It is a key indicator that helps investors and analysts evaluate whether a company's stock is undervalued or overvalued (Ali, et al, 2016). Table 12 shows the performance of prediction of Tobin's q as the predicted value along with EPS, BVPS and CFPS as the predictors.

Table 12: Prediction performance evaluation of alternate variable (Tobin's Q)

Alternate Variable (Tobin's Q)			
Regression Model	**Random Forest Regression (RFR)**	**Decision Tree (DT)**	**Linear regression (LR)**
MSE	324.03	321.31	356.05
RMSE	18.0	17.93	18.87
MAE	7.45	7.47	9.35
R^2	0.08	0.09	0.01

Source: Author's calculaion

Figure 9: Radar chart for comparison of performance for alternate variable (Tobin's Q)



Source: Author's calculation

It is evident from the figure that in terms of all parameters of performance evaluation, these models can't predict the value of Tobin's q with only 10% of the accuracy score at maximum.

Table 13: Value relevance of key accounting information under different ML models

Regression Model	Random Forest Regression	Decision Tree (DT)	Linear regression (Co-efficient)
EPS	0.56	0.85	0.72
BVS	0.40	0.15	-0.02
CFPS	0.04	0.0	-0.02

Source: Author's calculation

From Table 13 it can be deducted that Tobin's q can be predicted accurately with only EPS for 85% of the time followed by BV per share and CF per share with least impact on the forecasting only when DT model is used.

5.5 Summary of findings

In our study, Principal Component Analysis (PCA) revealed that only three variables – EPS, BVPS, and CFPS – could collectively account for 100% of the variance among the 12 independent accounting indicators (AIs). This suggests a strong explanatory power of these three variables in the context of our analysis.

Moving on to machine learning models, we found that Random Forest Regression (RFR), Decision Tree (DT), and Long Short-Term Memory (LSTM) models can predict market performance and share prices with over 60% accuracy, using the aforementioned three key AIs. However, Linear Regression (LR) showed significantly lower predictive capability with an accuracy below 30%.

Among the three key AIs, EPS emerged as the most influential factor in predicting price, followed by BVPS and CFPS, emphasizing its strong value relevance. This finding suggests that market price is predominantly dependent on EPS, implying that market participants attribute unequal importance to the various AIs presented in financial statements.

To assess the robustness of our models, we also tested LSTM and Neural Network. Only LSTM displayed a higher accuracy score of 62%. Notably, Random Forest Regression (RFR) stood out as the most accurate model for predicting lnPrice in our study.

Furthermore, we evaluated the market replacement value (Tobin's Q) and market price under each model. We observed that accuracy in the testing dataset was significantly lower compared to the training dataset for all three models: RFR, DTR, and LR when using Tobin's Q and market price as predictive variables. This highlights potential limitations in the predictive power of these models when applied to these specific indicators.

6.0 Conclusions and limitations

Value relevance of accounting information in predicting market price is a crucial issue that needs to be understood for investors to take the proper investment decision. This study aims to deduct the crucial accounting information that can play a forceful role in predicting market price by utilizing different machine learning methods. Moreover, the study aims to categorize the important accounting factor that holds the most predicting power. Having total 117 companies' accounting information for 2010-2021, we deploy different machine learning methods and found that random forest regression (RFR) model can predict the market price way better than any other ensemble models. Using PCA, the study also found only EPS, CFPs and BVPS are the most crucial accounting information among all the accounting item used in the study. Additionally, in terms of value relevance in prediction, EPS stands out to be the foremost important variable followed by BVPS and CFPS.

This investigation holds unique significance as it establishes a nexus between machine learning methodologies applied in the realms of finance and accounting research, aiming to discern their ramifications on the prediction of market prices. Notwithstanding, it is imperative to note that the dataset employed in this study is confined to a subset of 117 companies, thereby engendering potential challenges in extrapolating the outcomes to the broader context of share price prediction within the

Dhaka Stock Exchange (DSE). This limitation underscores the necessity for future research endeavors to expand the dataset, incorporating a more extensive array of companies, and employing customized models to further refine the predictive efficacy of market prices on the DSE.

7.0 References

Agrawal, Manish, Piyush Kumar Shukla, Rajit Nair, Anand Nayyar, and Mehedi Masud. 2022. "Stock Prediction Based on Technical Indicators Using Deep Learning Model." Computers, Materials & Continua 70: 287–304.

Ali, M. R., Lima, R. P., & Mahmud, M. S. 2016. "Analyzing Tobin's Q ratio of banking industry of Bangladesh: A comprehensive guideline for investors." Asian Business Review, 6(2), 85-90.

Alali, F. & Foote, P. (2012). The Value Relevance of International Financial Reporting Standards: Empirical Evidence in an Emerging Market. The International Journal of Accounting. 47. 10.1016/j.intacc.2011.12.005.

Alexander C. 2009. "Market risk analysis, value at risk models." Vol. 4. John Wiley & Sons.

Amir, E. & Lev, B. (1996). Value-Relevance of Nonfinancial Information: The Wirless Communications Industry.. Journal of Accounting and Economics. 22. 3-30. 10.1016/S0165-4101(96)00430-2.

Baheti, Radhika, Gauri Shirkande, Sneha Bodake, Janhavi Deokar, and Archana K. 2021. "Stock Market Analysis from Social Media and News using Machine Learning Techniques." International Journal on Data Science and Machine Learning with Applications 1: 59–67.

Barth, M. E., Li, K., & McClure, C. G. 2023. "Evolution in value relevance of accounting information." The Accounting Review, 98(1), 1-28.

Ball, R., & Brown, P. 1968. An Empirical Evaluation of Accounting Income Numbers. Journal of Accounting Research, 6(2), 159–178. https://doi.org/10.2307/2490232

Beaver, W.H. (2002) Perspectives on Recent Capital Market Research. The Accounting Review, 77, 453-474. https://doi.org/10.2308/accr.2002.77.2.453

Breiman, L. 2001. "Random forests." Machine learning, 45, 5-32.

Breiman, L. 2017. "Classification and regression trees." Routledge.

Cagliero, Luca, Jacopo Fior, and Paolo Garza. 2023. "Shortlisting machine learning-based stock trading recommendations using candlestick pattern recognition." Expert Systems with Applications 216: 119493.

Cooke, Terence & Omura, Teruyo & Willett, Roger. (2009). Consistency, Value Relevance and Sufficiency of Book for Market Values in Five Japanese Conglomerates Over the Period 1950–2004. Abacus-a Journal of Accounting Finance and Business Studies - ABACUS. 45. 88-123. 10.1111/j.1467-6281.2009.00279.x.

Chen, M. Y., & Du, Y. K. 2009. "Using neural networks and data mining techniques for the financial distress prediction model." Expert Systems with Applications, 36(2), 4075-4086.

Core, J.E., W.R. Guay, and A. Van Buskirk. 2003. "Market valuations in the New Economy: An investigation of what has changed." Journal of Accounting and Economics 34: 43-67.

Collins, D.W., Edward L. Maydew, Ira S. Weiss. 1997. Changes in the value-relevance of earnings and book values over the past forty years, Journal of Accounting and Economics, Volume 24(1), 39-67, ISSN 0165-4101, https://doi.org/10.1016/S0165-4101(97)00015-3.

Dung, Nguyen. (2010). Value-Relevance of Financial Statement Information: A Flexible Application of Modern Theories to the Vietnamese Stock Market. Development and Policies Research Center (DEPOCEN), Vietnam, Working Papers.

Dontoh, Alex & Radhakrishnan, Suresh & Ronen, Joshua. (2000). The Declining Value Relevance of Accounting Information and Non-Information-Based Trading: An Empirical Analysis. Contemporary Accounting Research. 21. 10.2139/ssrn.230826.

Floyd, E., N. Li, and D.J. Skinner. 2015. "Payout policy through the financial crisis: The growth of repurchases and the resilience of dividends." Journal of Financial Economics 118(2): 299-316.

Germon, H., Meek, G..,(2001), Accounting: An international perspective. McGraw Hill.

Glezakos, Michalis & Mylonakis, John & Kafouros, Charalampos. (2012). The Impact of Accounting Information on Stock Prices: Evidence from the Athens Stock Exchange. International Journal of Economics and Finance. 4. 10.5539/ijef.v4n2p56.

Hadi, M. M. (2004). The Importance of Accounting Information to the Investors in Banking sector: Kuwaiti Evidence By Kuwait University.

Hotelling H. 1933. "Analysis of a complex of statistical variables into principal components." Journal of Educational Psychology, 24(6): 417.

Hu, Weilong, Yain-Whar Si, Simon Fong, and Raymond Yiu Keung Lau. 2019. "A formal approach to candlestick pattern classification in financial time series." Applied Soft Computing 84: 105700.

Ince H, Trafalis TB. 2007. "Kernel principal component analysis and support vector machines for stock price prediction." IIE Transactions, 39(6): 629–637.

Jacob, Rudy & Madu, Christian. (2009). International financial reporting standards: An indicator of high quality?. International Journal of Quality & Reliability Management. 26. 712-722. 10.1108/02656710910975778.

Jolliffe I. "Principal component analysis." In: International Encyclopedia of Statistical Science. Springer; 2011. 1094–1096.

Khairi, Teaba W. A., Rana M. Zaki, and Wisam A. Mahmood. 2019. "Stock Price Prediction using Technical, Fundamental and News based Approach." Paper presented at 2019 2nd Scientific Conference of Computer Sciences (SCCS), Baghdad, Iraq, March 27–28.

Kothari, S.P., and J. Shanken. 2003. "Time-series coefficient variation in value-relevance regressions: A discussion of Core, Guay, and Van Buskirk and new evidence." Journal of Accounting and Economics 34: 69-87.

Lev, B. & Gu, F. (2016). Accounting: Facts or Fiction?. 10.1002/9781119270041.ch9.

Landsman, Wayne R. and Maydew, Edward L., Beaver (1968) Revisited: Has the Information Content of Annual Earnings Announcements Declined in the Past Three Decades? (May 2001). Available at SSRN: https://ssrn.com/abstract=204068 or http://dx.doi.org/10.2139/ssrn.204068

Liaw, A., & Wiener, M. 2002. "Classification and regression by randomForest." R news, 2(3), 18-22.

Marquardt, C.A. & Wiedman, C.I. 2004. The Effect of Earnings Management on the Value Relevance of Accounting Information. Journal of Business Finance & Accounting. 31, 3-4, p.297-332. https://doi.org/10.1111/j.0306-686X.2004.00541.x

Meyer, Christopher & Schwager, Andre. (2007). Understanding Customer Experience. Harvard business review. 85. 116-26, 157.

Mahmud, I., Fahad, H. A., & Rahman, A. N. 2021. "The Nexus between Firm Specific Factors, Macroeconomic Factors And Firm Performance of Textile Sector of Bangladesh." Asian Finance & Banking Review, 5(1), 54-74. https://doi.org/10.46281/asfbr.v5i1.1415

Mustafa Goçken et al. 2019. "Stock price prediction using hybrid soft computing models incorporating parameter tuning and input variable selection." Volume: 31, Issue: 2, Date 2019.

Oyerinde, D.T. 2009. Value relevance of accounting information in emerging stock market: The case of Nigeria. In Repositioning African Business and Development for the 21st Century, Simon Sigue (Ed.).

Perera, A. & Thrikawala, S. (2010). An Empirical Study of the Relevance of Accounting Information on Investor's Decisions. Accessed through: https://www.researchgate.net/publication/265438598_An_Empirical_Study_of_the_Relevance_of_Accounting_Information_on_Investor's_Decisions

Palepu, K.G., and P.M. Healy. 2008. "Business analysis and valuation: Using financial statements, text and cases." 4th ed. Cengage learning.

Shukla R, Trzcinka C. 1990. "Sequential tests of the arbitrage pricing theory: a comparison of principal components and maximum likelihood factors." The Journal of Finance, 45(5): 1541–1564.

Sonkavde G, Dharrao DS, Bongale AM, Deokate ST, Doreswamy D, Bhat SK. 2023. "Forecasting Stock Market Prices Using Machine Learning and Deep Learning Models: A Systematic Review, Performance Analysis and Discussion of Implications." International Journal of Financial Studies; 11(3):94. https://doi.org/10.3390/ijfs11030094

Subba Rao Polamuri, K. Srinivas, A. Krishna Mohan. 2019. "Stock Prices Prediction using random forest and extra tree regression." International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-3. Pages 1224-1228.

Vishnani, S. and Shah, K. (2008) Value Relevance of Published Financial Statements with Special Emphasis on Impact of Cash Flow Reporting. International Research Journal of Finance and Economics, 17, 84-90.

V. Roshan Joseph. 2022. "Optimal ratio for data splitting." Wiley Online Library. Volume15, Issue4. Pages 531-538. DOI: https://doi.org/10.1002/sam.11583.

Wagdy M, Abdallah (2001). Managing multinationals in the Middle East : accounting and tax issues: Quontm Books, 88 Post Road West, Westport, An imprint of Greenwood hblishing Group, Inc. Printed in the United States of America.

8.0 Appendix

Table A1: Descriptive statistics after using Robust Scaler

Variables	Measurement scale	count	mean	std	min	max
LnReturn	Logarithmic	1401	0.00	1.00	-2.17	3.38
TobinssQ	Ratio	1401	0.00	1.00	-0.09	28.25
EPS	Ratio	1401	0.00	1.00	-7.42	11.17
BVPS	Ratio	1401	0.00	1.00	-9.42	7.91
CFPS	Ratio	1401	0.00	1.00	-5.86	9.54
R&DPS	Ratio	1401	0.00	1.00	-0.08	23.14
Dividend	Percentage	1401	0.00	1.00	-0.32	13.31
cogsps	Ratio	1401	0.00	1.00	-0.29	14.45
opexps	Ratio	1401	0.00	1.00	-0.24	23.61
ppeps	Ratio	1401	0.00	1.00	-0.56	15.98
caps	Ratio	1401	0.00	1.00	-0.47	15.92
sps	Ratio	1401	0.00	1.00	-0.32	13.88
assetps	Ratio	1401	0.00	1.00	-0.57	12.56
debtps	Ratio	1401	0.00	1.00	-0.38	12.83

Source: Author's calculation