Accounting information and stock returns in Vietnam securities market: Machine learning approach

Ngoc Hung Dang, Thi Thuy Van Vu, Thi Nhat Le Dao
Hanoi University of Industry, Vietnam
National Economics University, Vietnam
National Economics University, Vietnam

This paper studies the relationship between accounting information reflected in financial statements and stock return in Vietnam Stock Market. The authors propose a research model to define the relationship. Besides, machine learning algorithms are used in researching and forecasting, with data obtained from observed firms during the period from 2009 to 2020. Research results show that gradient boosting algorithm has the best self-reporting performance and financial ratios also have great impact on stock returns, including operating income growth, stock earnings volatility, dividend yield, earnings before tax-to-equity ratio, cash holding ratio, and accrual quality. Based on the research results, the authors make some recommendations for investors, firms, and policy makers.

Keywords: stock return, machine learning, decision tree, gradient boosting, elastic net, accounting information, accrual quality

Información contable y rendimiento de acciones en el mercado de valores de Vietnam: enfoque de aprendizaje automático

Este artículo estudia la relación entre la información contable reflejada en los estados financieros y el rendimiento de las acciones en el mercado de valores de Vietnam. Se propone un modelo de investigación para definir la relación. Además, los algoritmos de aprendizaje automático se utilizan en la investigación y la previsión, con datos obtenidos de empresas observadas durante

https://doi.org/10.18800/contabilidad.202201.004
Contabilidad y Negocios (17) 33, 2022, pp. 94-118 / e-ISSN 2221-724X

el período de 2009 a 2020. Los resultados de la investigación muestran que el algoritmo gradient boosting tiene el mejor rendimiento de autoinforme y los índices financieros también tienen un gran impacto en la rentabilidad de las acciones, al incluir el crecimiento de los ingresos operativos, la volatilidad de las ganancias de las acciones, el rendimiento de los dividendos, la relación entre las ganancias antes de impuestos y el capital, la relación de tenencia de efectivo, y la calidad de la acumulación. Sobre la base de los resultados de la investigación, se hacen algunas recomendaciones para inversionistas, empresas y formuladores de políticas.

Palabras clave: rentabilidad de acciones, aprendizaje automático decision tree, gradient boosting, elastic net, información de cuenta, calidad devengada

Informações contábeis e retornos de ações no mercado de ações do Vietnã: abordagem de aprendizado de máquina

Este artigo estuda a relação entre as informações contábeis refletidas nas demonstrações financeiras e o retorno das ações no mercado de ações do Vietnã. Os autores propõem um modelo de pesquisa para definir a relação. Além disso, algoritmos de aprendizado de máquina são usados em pesquisas e previsões, com dados obtidos de empresas observadas durante o período de 2009 a 2020. Os resultados da pesquisa mostram que o algoritmo gradient boosting tem o melhor desempenho de autorrelato e os índices financeiros também têm grande impacto nos retornos das ações, incluindo crescimento da receita operacional, volatilidade do lucro das ações, rendimento de dividendos, lucro antes do índice de impostos sobre o patrimônio líquido, índice de retenção de caixa e qualidade de acumulação. Com base nos resultados da pesquisa, os autores fazem algumas recomendações para investidores, empresas e formuladores de políticas.

Palavras-chave: retorno de ações, aprendizado de máquina, decision tree, gradient boosting, elastic net, informação contábil, qualidade da provisão

1. Introduction

Stock returns have been a topic of scientific research interest for many years (Ball & Brown, 1968; Basu, 1983; Freeman, 1987; Collins & Kothari, 1989; Easton & Harris, 1991) have built yield models from income and return variables, by testing the relationship between current accounting earning divided by stock price at the beginning of the period and stock return. Investors aim to maximize the value of their assets by investing in firm shares as they offer the most significant potential for long-term

returns based on a trade-off of return and risk. Anwaar (2016) believes that a firm's financial and accounting information can be used effectively to evaluate investment opportunities for investors. A firm's information is divided into internal information and external information, based on the source of the information (Emamgholipour et al., 2013). As the name implies, external information is taken from the stock market, while internal information appears on financial statements. Therefore, investors can rely on accounting information on financial statements to evaluate stock returns over time when making investment decisions.

Previous financial research literature contains debates about the effectiveness of accounting information in forecasting stock returns. Previous studies have been conducted over several decades, and there is still no consensus on which of the indicators of accounting information seems to be most closely associated with stock returns. For example, Barnes (1987) argues that the indicators commonly used to predict stock returns are relevant indicators to a firm's financial and operating performance and can help predict future performance of the firm. According to Lewellen (2004), stock returns are predictable. Similarly, a research by Delen et al. (2013) shows that the most important financial ratio used to predict firm performance is net profit margin. The authors use decision tree algorithms to evaluate the relationship between financial ratios and firm performance. In addition, the results from a research by Musallam (2018) show that earnings per share, dividend yield and earnings yield are strongly related to stock returns.

Recently, there is a promising research field which focuses more on artificial intelligence techniques and can handle non-linearity and stochasticity. Improvements in technology and computer science have spread implementations to a wider audience, spurring the growth of research in machine learning. Financial market prediction research is done by applying several machine learning algorithms, including gradient boosting, support vector machine, and random forest to predict price, profitability, direction and volatility of securities, stocks and commodities indicators (Henrique et al., 2019).

Research on evaluating stock returns of Vietnamese firms using machine learning algorithms, together with accounting information, has not been studied yet. Besides, the literature is not consistent in previous studies in the world about whether financial ratios are closely related to stock returns. Forecasting a firm's stock return is certainly a difficult task, and even more so when just simply applying a series of financial indicators take from accounting information. However, it is a demanding task for decision makers, including stocks analysts, creditors, portfolio managers, and investors,

to determine which indicators are suitable for making predictions. Therefore, there is a need to study the impact of accounting information on stock returns on Vietnam Stock Market.

The study aims to give an overview of the relationship between accounting information, through relevant financial ratios, and stock returns prediction by answering the research question and the following questions: "How is the accounting information on financial statements reflected in financial ratios associated with the stock returns on Vietnam Stock Market?". Furthermore, the following detailed questions are formulated to answer whether certain financial ratios have a stronger association with stock returns than others. To determine the association, we use and compare machine learning models to each other to answer the questions: Which financial ratios have the strongest association with stock returns? Which applied machine learning model has the highest predicting accuracy?

Literature review

Using accounting information is perhaps the most common way to evaluate firm performance and provides essential inside information about firm value and market capitalization. Accounting information is widely used for many years all over the world, especially to predict the future performance of firms. Additionally, it can be used as a tool to predict stock returns in the stock market (Lewellen, 2004). It is also used to determine a firm's creditworthiness.

Financial theory suggests that financial ratios can provide an insight into a firm's operating and financial performance by rearranging accounting information on financial statements such as returns and firm value, its ability to allocate cash flows to investments, its liquidity and its debt size. Financial ratios are also used for future performance prediction. For example, they are used as explanatory variables in predictive modeling to predict financial distress, failure, and bankruptcy (Sun et al., 2011; Zięba et al., 2016; Kim & Upneja, 2014). Similarly, financial ratios are used in modeling the relationship between them and stock returns, by using multiple or sets of individual financial variables or using statistical methods or machine learning techniques (Emamgholipour et al., 2013; Lewellen, 2004). Although most of the previous studies are successful in modeling the relationship between financial ratios and firm returns, they lack determination about the importance features of these ratios used to evaluate firm performance.

In Vietnam, there have been a number of studies by Hai et al. (2015), Ha et al. (2018), Hung et al. (2018), and Hung and Van (2020) showing the impact of accounting infor-

mation on stock returns. However, studies which are based on machine learning algorithms and the importance of financial indicators on stock returns have not been yet considered. Results from previous studies show evidence of relationship between accounting information and returns. Yet, those studies are based on various theories and models. Each model has its own advantages and disadvantages, and the data collection methods are different. This study will determine which algorithm is most suitable in determining the impact of accounting information on stock returns, and which financial ratios have the strongest impact.

3. Research methodology

3.1. Research model

From the studies presented in the literature review such as Lewellen (2004), Delen et al. (2013), Emamgholipour et al. (2013), Musallam (2018), Hai et al. (2015), Ha et al. (2018), Hung et al. (2018), Hung and Van (2020), and Metsomäki (2020), the authors establish the model as follows:

$$\textbf{STOCK\_RETURNs}_{it} = \alpha + \beta\textbf{Information Accouting}_{it} + \varepsilon_{it}$$

Stock return is a ratio that measures the percentage of a stock's net income per dollar of invested capital. The formula for the stock returns is the appreciation in the price plus any dividends paid, divided by price at the beginning of the period. This calculation method is frequently mentioned in studies on stock returns and in the original model of Easton and Harris (1991), which is:

$$\text{STOCK\_RETURNs}_{it} = \frac{(P_{it} - P_{it-1}) + D_{it}}{P_{it-1}}$$

P_{it-1} , P_{it}: price of share at the end of year t and t-1, respectively
D_{it}: dividend per share for the year t.

This study examines 11 financial indicators of accounting information on financial statements, including *Debt-to-equity ratio, Earnings before tax-to-equity ratio, Total assets turnover, Fixed assets turnover, Operating income growth, Stock earnings volatility, Cash holding ratio, Dividend yield, Earnings management, Accrual quality, Earnings per share*. The measurement of these financial indicators are described briefly in appendix 1. To conduct this study, we used Python 3.8 to analyze, forecast and evaluate the models.

3.2. Research algorithms

3.2.1. Linear models

Linear regression: As the first linear model, linear regression with the ordinary least square method is usually implemented. its aim is to minimize the sum of squares between the true and estimated values by fitting the linear model with coefficients (Pedregosa et al., 2011).

Ridge regression and ***Lasso regression*** are two regression models that apply regularization techniques to avoid overfitting. Overfitting is a phenomenon where the model only fits well on training dataset but does not predict well on test data. This is often the case when training machine learning models. This phenomenon has a negative effect and leads to inapplicable models because the predictions turn wrong when applied in practice. There are many causes of overfitting. One of the common causes is that the training dataset and the test dataset have different distribution, making the rules learned in the training dataset no longer valid in the test dataset, or it is because when a model has too many parameters, its data is not representative. Regularization is a technique to avoid overfiting by adding a regularization term to the loss function. Usually, this term is in the form of L1 norm or L2 norm of the coefficients. They are called ridge regression and Lasso regression respectively.

For these regressions, we need to fit coefficient α to find a coefficient that is best for each dataset. If data is severely overfitting, it is necessary to reduce overfitting by increasing the effect of the regularization term by increasing coefficient α. If the model is not overfitting, α can be close to 0. If $\alpha = 0$, the regression equation is equivalent to multivariable linear regression.

Elastic net: To add more explanatory power to the family of linear models, elastic net (Zou & Hastie, 2005) was applied. Firstly, it is a continuation of linear regression models trained with Lasso's L1 penalty and Ridge's L2 penalty. Combining the penalties of both methods in one model produces a regular, competitive model in which the weight of parameters is non-zero (Pedregosa et al., 2011).

3.2.2. Decision tree

Decision tree: Decision tree is a classification model introduced by Belson (1959), widely used in various fields. After the introduction of the machine learning method system, decision tree was further developed with C4.5 algorithm by J. Ross Quinlan (1996) and ID3 algorithm by J. Ross Quinlan (1986). Decision tree is a structured classification tree that classifies objects based on sequences of rules. Independent variables

and attributes can be of different data types such as binary, nominal, ordinal, quantitative data. To determine which variable is classified first, which variable is classified later, the weight of evidence (Entropy) for each variable is calculated, the higher the information value, the more classification information the variable carries.

Random forest: Random forest is an attribute classification method developed by Leo Breiman at the University of California, Berkeley. Breiman is also the co-author of CART (classification and regression trees) method, which is one of ten data mining methods. In Random Forest, a significant improvement in classification accuracy is the result from the growth of a set of trees, each of which will then "vote" for the most popular class. Normally, to develop these sets of trees, random vectors are generated, which will govern the growth of each tree in the aforementioned sets. For the k^{th} tree, a random vector Vk is generated, which is independent from previously generated vectors V1, V2, ..., Vk-1, although the distribution of these vectors is similar. A tree is grown based on the training set and vector Vk, resulting in a subclass h(x, Vk) where x is the input vector. After a large number of trees are created, these trees will "vote" for the most popular class.

AdaBoost: Boosting is a technique that uses a combination of machine learning algorithms on a sample space sequentially, then combines separate classification results to get an effective classifier. An efficient algorithm in boosting is AdaBoost (adaptive boosting), which uses classification error weights assigned to each sample. Firstly, the algorithm classifies equivalent weights on each training sample. In each iteration, the algorithm: (i) trains the sample by a weak classifier; (ii) checks whether the classification results on the training sample are correct; (iii) recalculates classification error weights on the sample by: increasing error weights on misclassified samples and decreasing error weights on properly classified samples. After finishing the loop, the algorithm will combine the sub-classifiers to construct a composite classifier.

Gradient boosting: Gradient boosting is also an algorithm that uses a combination of boosting methods to develop an advanced prediction tool. In many ways, gradient boosting is similar to AdaBoost, but with a few key differences: unlike AdaBoost which builds decision trees, gradient boosting builds trees that typically have 8–32 leaves. Gradient boosting views boosting problem as an optimization problem, where it uses a loss function and tries to minimize errors. This is why it is called gradient boost, as it is inspired by gradient descent. Finally, trees are used to predict the residuals of samples (prediction minus reality). Gradient boosting starts by building a tree to try to fit the data, and subsequent trees are built for the purpose of reducing residuals (errors). It does this to areas where existing learners are underperforming, which is similar to AdaBoost.

3.2.3. SVM and KNN models

The other two applied models are grouped together, although they are not as similar as previous models, they use the same method to evaluate the relationship between financial ratios and stock returns.

Support vector machine (SVM): SVM is a binary classification algorithm, SVM takes input data and classifies them into two different classes. Given a set of training examples of two given categories, SVM algorithm builds a SVM model to classify other examples into those two categories. SVM builds a hyperplane to classify a dataset into two separate classes. To do this, first SVM will construct a hyperplane or a set of hyperplanes in a multi-dimensional or infinite-dimensional space, which can also be used for classification, regression, or other tasks. For the best classification, it is necessary that optimal hyperplane is located as far away from the data points of all classes (margins) as possible, because the larger the margins, the smaller the generalization errors of the classification algorithm.

K-nearest neighbors (K-NN): K-nearest neighbors algorithm (K-NN) is commonly used in data mining. K-NN is a method to classify objects based on the closest distance between query points and all objects in training data. An object is classified based on its K neighbors. K is a positive integer determined before the execution of the algorithm. Euclidean distance is often used to calculate the distance between objects. This simple algorithm is capable of solving the regression problems proposed by Altman (1992). It assumes that similar observations exist in close proximity. K-NN predicts output by using number k of training data.

3.3. Feature importance

The technique of assigning scores to input features in a predictive model is called feature importance. Feature importance scores are an essential part of a predictive model because they can be used to enhance a model's performance and provide insight into the model and dataset. Provided relative importance scores can be used to determine which features are most relevant for a study. There are many types of feature importance scores in these techniques, those that are simple to calculate are statistical correlation coefficients such as Pearson's correlation and Spearman's rank, for linear and nonlinear correlation respectively.

There are three types of importance scores for more advanced features that are also implemented from model coefficients as part of the linear model, from decision model and permutation importance, which are described in detail by Pedregosa et al. (2011). We describe three important features as follows:

Coefficients as feature importance: After fitting a linear machine learning model on dataset, the coefficient of each input variable can be retrieved and stated as a feature importance score. Comparison is possible because the dataset is normalized and the variables have the same scale. This approach is applied to linear regression and elastic net model for importance scores retrieval.

Decision tree feature importance: Decision tree algorithms, like the CART algorithm used in this study, are provided in scikit-learning implementations with feature importance scores, based on criterion reduction used to select split points. This approach is applied to decision tree model and all the tree-based combination methods such as random forest, gradient boosting, and AdaBoost.

Permutation feature importance: This technique computes relative importance scores independently to the model used. After fitting a model on a dataset, a prediction is made, then this process is repeated five times for each feature in dataset, resulting in an average importance score for each input feature. This technique is suitable for models that do not provide original feature importance scores, such as k-Nearest Neighbors and SVM in this study. Important characteristics are identified as follows:

$$F_{n(fused)} = w_1 F_{1n} + w_2 F_{2n} + \ldots + w_m F_{mn}$$

$$W_i = \frac{\frac{1}{RMSE_i}}{\sum_j^n \frac{1}{RMSE_i}}, \text{ for } i = 1,\ldots n$$

F: absolute value of the relative importance score of the feature generated by the model

W: normalized weighted value based on the model's predictability

m: number of models

n: number of financial ratios (n = 11 in this study)

3.4. Model evaluation

Evaluating a regression model is perhaps the most important part of building a supervised machine learning model. This determines whether a model is ready for deployment. Indicators only show numbers, but if examined carefully, they can tell us about feature selection and feature engineering. In this study, we will look at the regression metrics and discover why we use these following regression metrics: mean absolute error, mean squared error and root mean squared error.

***Mean absolute error* (MAE)** is a measure of the mean absolute error between predicted and actual values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|yi - \hat{y}i\right|$$

Basically, MAE is a L1 norm. The smaller MAE, the smaller the distance between predicted and actual values, and the better the model. However, MAE value does not cover unit difference.

***The mean squared error* (MSE)** of an estimator measures the average squares of errors – that is, the average squared difference between predicted values and actual values. MSE is a risk function, corresponding to the expected value of squared error loss or quadratic loss. MSE is the second moment (about the origin) of errors.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left(yi - \hat{y}i\right)^2$$

***Root mean squared error* (RMSE)** is the standard deviation of residuals (prediction errors). Residuals are a measure of how far from the regression line data points are. RMSE is a measure of how spread-out residuals are. In other words, it tells you how concentrated data is around the line of best fit. RMSE is commonly used in forecasting and regression analysis to verify experimental results and is also a measure of how effective models are, by measuring the difference between predicted and actual values. The smaller RMSE, the smaller the errors, the higher the level of reliability estimation of models.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(yi - \hat{y}i\right)^2}$$

3.5. Research data

This study uses data collected from Vietnam Stock Exchange over the period of 2009-2020. Data is collected from audited financial statements of listed firms after excluding firms in banking, securities and insurance sectors. After determining the indicators, there are 3397 observations to be analyzed and forecasted, presented in table 1 by year and by industry.

Based on figure 1, the stock return of firms witnessed huge fluctuations over the period of 2009-2010, especially in 2010, stock return decreased to -9%, while during 2011-2013, it increased to 81% in 2013. Then from 2014 to 2017, the return was stable at around 20%. In 2019, the stock return dropped to -15%, but by 2020, it increased dramatically to 66%.

Table 1. Descriptive statistics among the variables in the model

Table A: Data by year			Table B: Data by industry		
Year	**Number of observations**	**Percentage**	**Industry**	**Number of observations**	**Percentage**
2009	139	4.1%	Real estate & construction	1,194	35.1%
2010	213	6.3%	Technology	102	3.0%
2011	301	8.9%	Industrial	400	11.8%
2012	314	9.2%	Service	371	10.9%
2013	334	9.8%	Consumer goods	309	9.1%
2014	323	9.5%	Energy	221	6.5%
2015	341	10.0%	Agriculture	293	8.6%
2016	349	10.3%	Materials	378	11.1%
2017	350	10.3%	Medical	129	3.8%
2018	330	9.7%			
2019	274	8.1%			
2020	129	3.8%			
Total	**3397**	**100.0%**	**Total**	**3397**	**100.0%**

Figure 1. Stock returns volatility by year

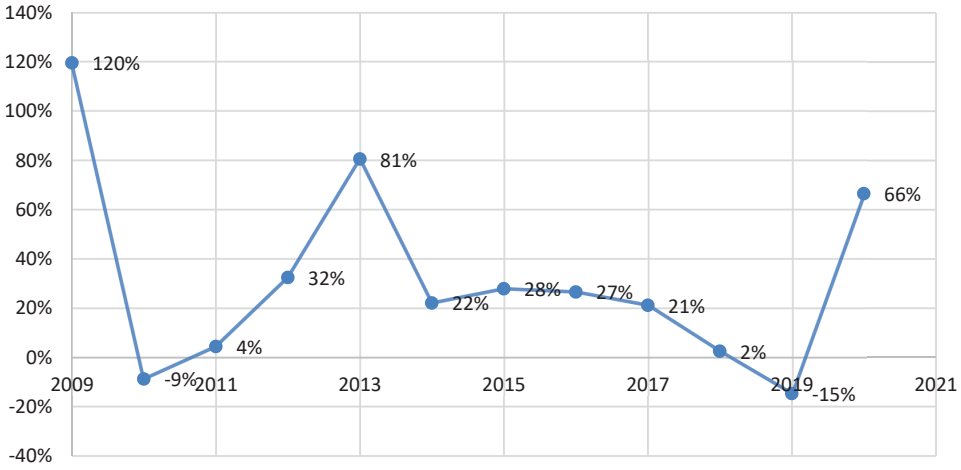
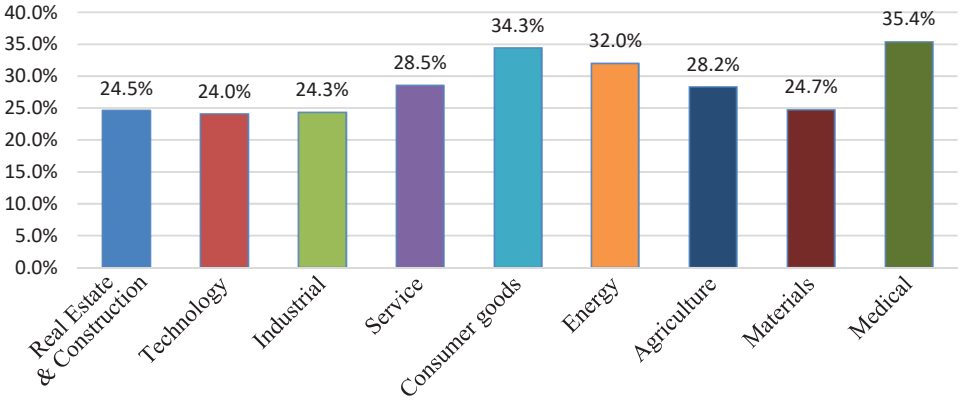


Figure 2 shows observed stock returns by industry in the period of 2009-2020, in which the industry with the highest stock return is the health industry, followed by the consumer goods industry. The industries with low stock return are technology and industrial.

Figure 2. Stock returns by industry



Statistical data in table 2 shows that the average stock return is 27.0%, the standard deviation is 54.4%, the lowest return is -59.9% and the highest is 244.9%. The accounting information is reflected through 11 financial ratios in terms of mean, standard deviation, minimum value and maximum value.

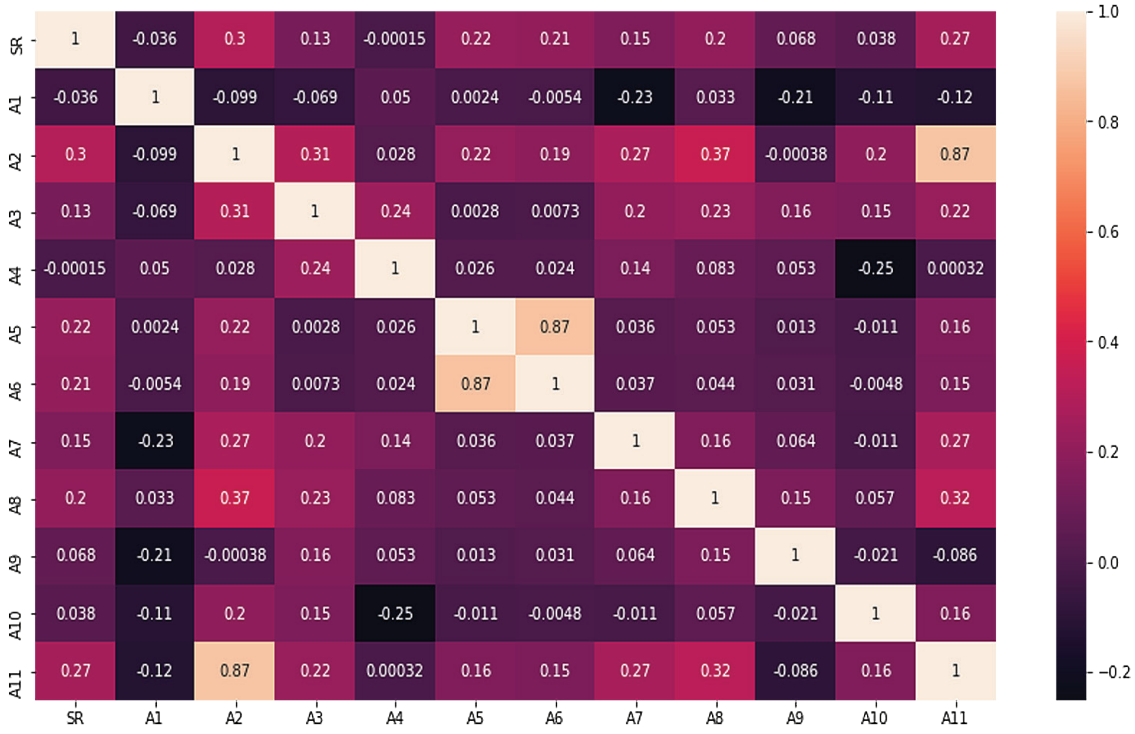
Table 2. Descriptive statistics among the variables in the model

Variable	Obs.	Mean	Std. Dev.	Min	Max
SR	3,397	0.270	0.544	-0.599	2.449
A1-Debt-to-equity ratio	3,397	1.485	1.287	0.074	7.022
A2-Earnings before tax-to-equity ratio	3,397	0.123	0.084	-0.108	0.398
A3-Total assets turnover	3,397	1.195	0.774	0.087	4.617
A4-Fixed assets turnover	3,397	17.222	29.542	0.303	273.554
A5-Operating income growth	3,397	1.120	0.908	-2.520	8.892
A6-Stock earnings volatility	3,397	0.029	0.893	-3.316	7.920
A7-Cash holding ratio	3,397	0.094	0.090	0.001	0.467
A8-Dividend yield	3,397	0.095	0.107	0.000	0.589
A9-Earnings management	3,397	0.577	0.815	0.004	5.843
A10-Accrual quality	3,397	-0.005	0.126	-0.236	0.657
A11-Earnings per share	3,397	2.275	1.941	-1.408	9.850

4. Result and discussion

Figure 3 shows the result of the correlation coefficient between variables. The purpose of testing the correlation between independent and dependent variables is to eliminate factors that can lead to multicollinearity, before running the regression model. There are no two variables that have correlation coefficient greater than 0.6, except for that of A2 - Earnings before tax-to-equity ratio and A11 - Earnings per share.

Figure 3. Autocorrelation matrix between variables in research model



To evaluate the effectiveness of algorithms in predicting the impact of accounting information on financial statements on stock returns, we use three following metrics: mean absolute error (MAE), mean squared error (MSE) and root mean squared error (RMSE), presented in table 3. The algorithm with the smallest value has the highest measurement performance.

Table 3. Algorithm's evaluation result

	Mean absolute error	Mean squared error	Root mean squared error
Linear regression	0.372	0.236	0.485
Lasso	0.375	0.238	0.488
Ridge	0.372	0.236	0.485
Elastic net	0.373	0.237	0.487
Random forest	0.370	0.230	0.480
Decision tree	0.487	0.425	0.652
AdaBoost	0.498	0.336	0.580
Gradient boosting	0.350	0.208	0.456
K-neighbor	0.368	0.239	0.489
SVM	0.339	0.219	0.468

With MAE measurement, AdaBoost algorithm has the highest value of 0.498, which proves that this algorithm has the lowest measurement performance, while SVM algorithm has a value of 0.339, which proves that it has the best measuremance performance among these ten algorithms in applied dataset. Meanwhile, with MSE and RMSE measurement, the algorithm with the lowest value, 0.208 and 0.456 respectively, is gradient boosting algorithm. Thus, gradient boosting algorithm has the best prediction performance among the algorithms used in predicting the impact of accounting information from financial indicators on stock returns.

When applying regression algorithms, we choose to use linear regression, Lasso, ridge, elastic net algorithms to determine importance coefficient of each financial ratio, presented in table 4. Based on RMSE, we determine F (Fused) for the importance regression coefficient for each ratio, shown in table 4 and figure 4.

Among 11 financial indicators of accounting information on financial statements, earnings before tax-to-equity ratio has the highest importance score of 0.093, the next one is operating income growth with a score of 0.063 and the third one which scores 0.053 is dividend yield. in contrast, the three indicators with the lowest importance scores are debt-to-equity ratio (0.003), accrual quality (0.010) and total assets turnover (0.011).

Table 4. Coefficients as feature importance

	Linear regression	Lasso	Ridge	Elastic net	F (Fused)
Debt-to-equity ratio	0.0063	0.0000	0.0063	0.0000	0.003
Earnings before tax-to-equity ratio	0.0921	0.0949	0.0921	0.0921	0.093
Total assets turnover	0.0154	0.0037	0.0154	0.0097	0.011
Fixed assets turnover	0.0222	0.0027	0.0222	0.0121	0.015
Operating income growth	0.0651	0.0599	0.0651	0.0617	0.063
Stock earnings volatility	0.0235	0.0188	0.0235	0.0218	0.022
Cash holding ratio	0.0432	0.0349	0.0432	0.0382	0.040
Dividend yield	0.0554	0.0499	0.0554	0.0529	0.053
Earnings management	0.0194	0.0099	0.0194	0.0139	0.016
Accrual quality	0.0162	0.0000	0.0162	0.0082	0.010
Earnings per share	0.0150	0.0061	0.0151	0.0115	0.012
Root mean squared error	0.485	0.488	0.485	0.487	1.945
Weight	0.250	0.249	0.250	0.250	

Figure 4. Coefficients as feature importance

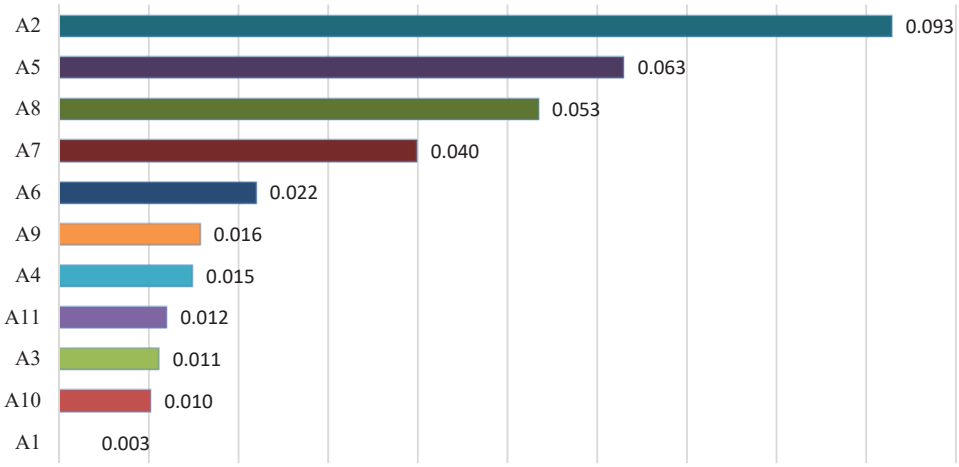
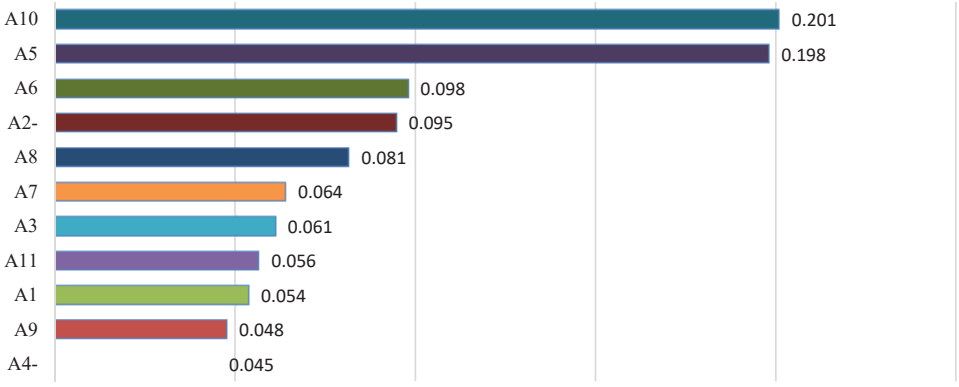


Table 5. Feature importance

	Random forest	Decision tree	AdaBoost	Gradient boosting	F (Fused)
Debt-to-equity ratio	0.075	0.068	0.042	0.033	0.054
Earnings before tax-to-equity ratio	0.100	0.093	0.066	0.113	0.095
Total assets turnover	0.075	0.067	0.065	0.040	0.061
Fixed assets turnover	0.064	0.064	0.042	0.015	0.045
Operating income growth	0.160	0.171	0.209	0.244	0.198
Stock earnings volatility	0.102	0.083	0.108	0.096	0.098
Cash holding ratio	0.074	0.077	0.053	0.054	0.064
Dividend yield	0.094	0.084	0.035	0.105	0.081
Earnings management	0.076	0.070	0.022	0.025	0.048
Accrual quality	0.113	0.139	0.305	0.246	0.201
Earnings per share	0.066	0.084	0.054	0.031	0.056
Root mean squared error	0.480	0.652	0.580	0.456	2.168
Weight	0.277	0.204	0.229	0.291	

Next, we used decision tree algorithm group including random forest, decision tree, AdaBoost, gradient boosting. Feature importance values of financial ratios are shown in table 5 and their permutation feature importance values are presented in figure 5.

Figure 5. Feature importance



As shown in figure 5, the three financial indicators that have the highest feature importance score are accrual quality, operating income growth and stock earnings volatility, with a value of 0.201, 0.198 and 0.098, respectively. The three financial ratios with the lowest importance score are fixed assets turnover (0.045), earnings management (0.0048) and debt-to-equity ratio (0.054).

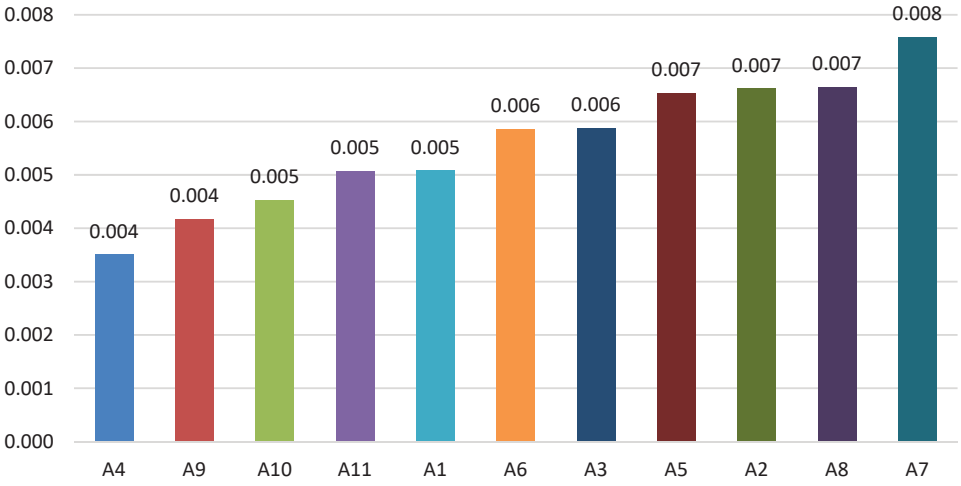
The feature importance of support vector machine model and K-nearest neighbors model are calculated as permutation feature importance. The initial values provided by this technique are shown in table 6. The result is the average weighted importance score for each financial indicator as determined in column F (Fused). The value of permutation feature importance is much smaller than feature importance and coefficients as feature importance presented in the previous sections, therefore they cannot be directly compared with each other. However, importance feature rankings from highest to lowest can still be presented as the average weighted importance score for each financial indicator.

Table 6. Permutation feature importance

	K-Neighbor	SVM	F (Fused)
A1-Debt-to-equity ratio	0.0063	0.0039	0.0051
A2-Earnings before tax-to-equity ratio	0.0061	0.0072	0.0066
A3-Total assets turnover	0.0065	0.0053	0.0059
A4-Fixed assets turnover	0.0035	0.0035	0.0035
A5-Operating income growth	0.0051	0.0080	0.0065
A6-Stock earnings volatility	0.0056	0.0061	0.0059
A7-Cash holding ratio	0.0089	0.0063	0.0076
A8-Dividend yield	0.0057	0.0075	0.0067
A9-Earnings management	0.0040	0.0043	0.0042
A10-Accrual quality	0.0043	0.0048	0.0045
A11-Earnings per share	0.0046	0.0056	0.0051
Root mean squared error	0.4889	0.4682	0.9571
Weight	0.4892	0.5108	

According to figure 6, the three most important financial indicators of accounting information on the financial statements according to the permutation feature importance, are cash holding (0.008), dividend yield (0.007) and earnings before tax-to-equity ratio (0.007). In contrast, the financial indicators of the lowest importance are fixed asset turnover (0.004), earnings management (0.004) and accrual quality (0.005).

Figure 6. Permutation feature importance

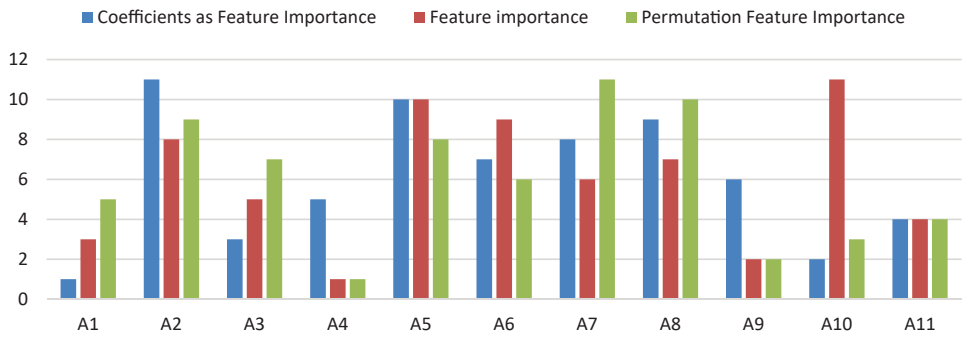


To evaluate the importance of financial indicators of accounting information on financial statements in a comprehensive way, we rank the results from three approaches, from 1 to 11, to examine the overall importance of all financial ratios. The results are based on the importance of financial ratios in the analysis above and are visualized as a bar graph below. The highest importance score takes the value 11 and the lowest one takes the value 1, which means the higher the bar, the higher the importance of the financial ratios. Each financial ratio is grouped with all three approaches: "Coefficients as feature importance", "Feature importance" and "Permutation feature importance" for a better approach-based comparison. Figure 7 shows five financial indicators of accounting information on financial statements that can be considered important: Operating income growth, dividend yield, cash holding ratio, earnings before tax-to-equity ratio, and stock earnings volatility. One indicator that can be listed in this group is accrual quality.

Earnings before tax-to-equity ratio is an important indicator of stock returns. Besides, earnings before tax-to-equity ratio coefficient is positive in all four linear models, which is consistent with the study of (Anwaar, 2016). Previous research has found out that net profit margin is an important indicator for both firm performance and stock return (Delen et al., 2013; Anwaar, 2016). The importance of operating income growth is consistent with previous research in both linear and decision tree models. Based on positive coefficient values, the results show that stock returns increase as Earnings before tax-to-equity ratio increases, which is consistent with the research results of Öztürk & Karabulut (2018). They found that net profit margin has a positive

effect on stock returns, and higher net profit margin generates higher profits for the next period. Previous research has also found that dividend yield can predict stock returns (Fama & French, 2021; Lewellen, 2004). Similarly, the research results of Musallam (2018) show that dividend yield is a highly important financial ratio, with positive coefficient values, they show a positive association with stock returns.

Figure 7. Feature importance ranking



Research results of Musallam (2018), and Emamgholipour et al. (2013) indicate that earnings per share has a positive and significant association with stock returns. This result is consistent with these following research: Easton and Harris (1991), Ball and Brown (1968), Basu (1983), and Hung et al. (2018). Stock earnings volatility also has an important impact on stock returns, this result is consistent with that of Freeman (1987), Collins and Kothari (1989), and Easton and Harris (1991). Besides, consistent with a research by Hung and Van (2020), accrual quality also has a significant impact on stock returns.

5. Conclusion and recommendation

To get experimental research results for the evaluation and measurement of the importance of accounting information on financial statements as financial indicators, on stock returns, we use panel data with 3397 observations from listed firms on Vietnam Stock Market from 2009-2020 and machine learning algorithms. The research results show that operating income growth, stock earnings volatility, dividend yield, earnings before tax-to-equity ratio ratio, cash holding ratio and accrual quality have an important and positive impact on stock returns. Based on the results, we propose some recommendations as follows:

- When deciding to invest in stocks, investors need to pay attention to accounting information because this type of information affects stock prices, such as operating income growth and stock earnings volatility on the income statement.

However, there are other factors that can affect stock returns and announced accounting information has not yet strongly affected stock return volatility on Vietnam Stock Market. Before making investment decisions, investors can consider earnings before tax-to-equity ratio, dividend yield because these indicators have an important and positive impact on stock returns. Investors should also refer to other additional information to make the best decisions. Empirical results of other indicators like cash holding ratio, on the relationship between earnings and stock returns, have suggested the difference between firms with the same earnings rate.

- Firms need to disclose complete and timely financial statements. According to the recommendations for investors above, investors can rely on accounting information to make investment decisions, so a firm itself needs to ensure the quality of disclosed financial statements to create and maintain investors' confidence in the firm. This not only creates a premise for investors' confidence in accounting and financial information that firms announce, but also helps firms gain a position and improve their firm value to attract potential investors. complete and timely disclosure of financial statements, audit reports, and board of directors' reports will create belief in investors about the transparency in information disclosure, which is a good signal to attract more investors.

Contribución de autores

Hung DN: Conceptualización, Metodología, Validación, Análisis formal, Investigación, Recursos, Redacción borrador original, Validación, Administración del proyecto. **Van VT:** Conceptualización, Metodología, Software, Validación, Análisis formal, Investigación, Curación de datos, Redacción borrador original, Supervisión, Validación. **Le DT:** Conceptualización, Metodología, Software, Análisis formal, Investigación, Curación de datos.

Declaración de conflicto de intereses

El (los) autor(es) declara(n) que, durante el proceso de investigación, no ha existido ningún tipo de interés personal, profesional o económico que haya podido influenciar el juicio y/o accionar de los investigadores al momento de elaborar y publicar el presente artículo.

References

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician, 46*(3), 175-185. https://doi.org/10.1080/00031305.1992.10475879

Anwaar, M. (2016). Impact of firms performance on stock returns (Evidence from listed companies of ftse-100 index London, UK). *Global Journal of Management and Business Research, 16*(1), 1-10.

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

Barnes, P. (1987). The analysis and use of financial ratios. *Journal of Business Finance dan Accounting, 14*(4), 449-461. https://doi.org/10.1111/j.1468-5957.1987.tb00106.x

Basu, S. (1983). The relationship between earnings' yield, market value and return for NYSE common stocks: Further evidence. *Journal of financial economics, 12*(1), 129-156. https://doi.org/10.1016/0304-405X(83)90031-4

Belson, W. A. (1959). Matching and prediction on the principle of biological classification. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 8*(2), 65-75. https://doi.org/10.2307/2985543

Collins, D. W., & Kothari, S. (1989). An analysis of intertemporal and cross-sectional determinants of earnings response coefficients. *Journal of Accounting and Economics, 11*(2-3), 143-181. https://doi.org/10.1016/0165-4101(89)90004-9

Dechow, P. M., & Dichev, I. D. (2002). The quality of accruals and earnings: The role of accrual estimation errors. *The Accounting Review*, 77, 35-59. https://doi.org/10.2308/accr.2002.77.s-1.35

Delen, D., Kuzey, C., & Uyar, A. (2013). Measuring firm performance using financial ratios: A decision tree approach. *Expert Systems with Applications, 40*(10), 3970-3983. https://doi.org/10.1016/j.eswa.2013.01.012

Easton, P. D., & Harris, T. S. (1991). Earnings as an explanatory variable for returns. *Journal of Accounting Research, 29*(1), 19-36. https://doi.org/10.2307/2491026

Emamgholipour, M., Pouraghajan, A., Tabari, N. A. Y., Haghparast, M., & Shirsavar, A. A. A. (2013). The effects of performance evaluation market ratios on the stock return: Evidence from the Tehran stock exchange. *International Research Journal of Applied and Basic Sciences, 4*(3), 696-703. https://doi.org/10.5539/ijef.v4n7p41

Fama, E. F., & French, K. R. (2021). Dividend yields and expected stock returns. University of Chicago Press.

Freeman, R. N. (1987). The association between accounting earnings and security returns for large and small firms. *Journal of Accounting and Economics, 9*(2), 195-228. https://doi.org/10.1016/0165-4101(87)90005-X

Ha, H. T. V., Hung, D. N., & Dung, T. M. (2018). Impact of accounting data on stock prices: the case of Vietnam. *International Journal of Accounting and Financial Reporting, 8*(1), 140-154. https://doi.org/10.5296/ijafr.v8i1.12671

Hai, T. T. T., Diem, N. N., & Binh, H. Q. (2015). The relationship between accounting information reported in financial statements and stock returns-empirical evidence from Vietnam. *International Journal of Accounting and Financial Reporting, 5*(1), 229-238. https://doi.org/10.5296/ijafr.v5i1.7473

Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications*, *124*, 226-251. https://doi.org/10.1016/j.eswa.2019.01.012

Hung, D. N., Ha, H. T. V., & Binh, Đ. T. (2018). Impact of accounting information on financial statements to the stock price of the energy enterprises listed on Vietnam's Stock Market. *International Journal of Energy Economics and Policy, 8*(2), 1-6.

Hung, D. N., & Van, V. T. T. (2020). Studying the impacts of earnings quality on stock return: Experiments in Vietnam. *International Journal of Advanced and Applied Sciences, 7*, 45-53. https://doi.org/10.21833/ijaas.2020.04.007

Kim, S. Y., & Upneja, A. (2014). Predicting restaurant financial distress using decision tree and AdaBoosted decision tree models. *Economic Modelling*, *36*, 354-362. https://doi.org/10.1016/j.econmod.2013.10.005

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

Metsomäki, J. (2020). *Relation between financial ratios and stock returns: Machine learning approach* [Master thesis, Lappeenranta-Lahti University of Technology], https://lutpub.lut.fi/handle/10024/161123.

Musallam, S. R. (2018). Exploring the relationship between financial ratios and market stock returns. *Eurasian Journal of Business and Economics, 11*(21), 101-116. https://doi.org/10.17015/ejbe.2018.021.06

Öztürk, H., & Karabulut, T. A. (2018). The relationship between earnings-to-price, current ratio, profit margin and return: an empirical analysis on Istanbul stock exchange. *Accounting and Finance Research, 7*(1), 109-115. https://doi.org/10.5430/afr.v7n1p109

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825-2830.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81-106. https://doi.org/10.1007/BF00116251

Quinlan, J. R. (1996). Bagging, boosting, and C4. 5. En *American Association for Artificial Intelligence* (Ed.), AAAI'96: Proceedings of the thirteenth national conference on artificial intelligence. Vol 1 (pp. 725-730). AAAI Press.

Sun, J., Jia, M.-y., & Li, H. (2011). AdaBoost ensemble for financial distress prediction: An empirical comparison with data from Chinese listed companies. *Expert Systems with Applications, 38*(8), 9305-9312. https://doi.org/10.1016/j.eswa.2011.01.042

Zięba, M., Tomczak, S. K., & Tomczak, J. M. (2016). Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. *Expert Systems with Applications*, *58*, 93-101. https://doi.org/10.1016/j.eswa.2016.04.001

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Appendix 1. Attributes in research model

Attribute	Notation	Measurement
Debt-to-equity ratio	A1	Total debts / Equity
Earnings before tax-to-equity ratio	A2	Earnings before tax / Equity
Total assets turnover	A3	Net income / Total assets
Fixed assets turnover	A4	Net income / Total fixed assets
Operating income growth	A5	Operating income$_t$ / Operating income$_{t-1}$
Stock earnings volatility	A6	Earnings before tax$_t$ - Earnings before tax$_{t-1}$ / Average number of outstanding shares
Cash holding ratio	A7	Cash / Total assets
Dividend yield	A8	Annual dividends per share / Price per share
Earnings management	A9	There are many ways to measure/ manage earnings, the authors use the model of Jones (1991). Variable earnings management (EM) is measured through a proxy, which is the residual of equation **(1).** Earnings quality (EQ) is inversely proportional to the residual of the following equation: $ACC_{it} = \beta_0 + \beta_1 (REV_{it} - AR_{it}) + \beta_2 PPE_{it} + \varepsilon_{it}$ **(1)** **In which:** ΔREV_{it} is the difference between revenue of firm i in year t and year t-1 PPE_{it} is the cost of fixed assets of firm i in year t AR_{it-1} is total assets of year t-1 α_1, α_2, α_3, are the parameters of each firm Earnings management will be taken as a residual of ε_{it} because it is an earnings-adjusting behavior, so whether the adjustment is increased or decreased (equivalent to the adjusted accruals variable with negative values or positive) are all behaviors. Thus, ε_{it} is the measurement of EM variable (Earnings quality), the higher the deviation εit, the lower earnings quality. Earnings quality is measured by Earnings management: EQ_EM=EM*(-1)

Attribute	Notation	Measurement
Accrual quality	A10	To estimate the quality of accruals, the authors used a model developed by Dechow and Dichev (2002), in which Accrual quality is presented as current Working capital accruals and it is calculated by Operating cash flows of the previous period, this period, and the following period, all divided by Total assets at the beginning of the period. To measure EQ, which is an AQ (Accrual quality) variable and is considered as the negative standard deviation of residual εi,t of equation **(2)** after regression. $CACC_{it} = \beta_0 + \beta_1 CFO_{it-1} + \beta_2 CFO_{it} + \beta_3 CFO_{it+1} + \varepsilon_{it}$ **(2)** **In which:** CFO_{it-1}, CFO_{it}, CFO_{it+1} are respectively the cash flows of year t-1, year t and year t+1, all divided by total assets at the beginning of the period (A_{it-1}). A higher value of AQ indicates poorer Accrual quality because less variation in current accruals is explained by cash flows performance. Since earnings are the sum of accruals and cash flows, and the cash flow component is generally considered to be objective and unmanipulated, the quality of earnings depends on the quality of accruals.
Earnings per share	A11	Net income / Average number of outstanding shares

Fecha de recepción: 21/09/2021
Fecha de aceptación: 21/02/2022
Correspondencia: dangngochung@haui.edu.vn