Check for updates

Original Research Article

SCIENCE PROGRESS

Science Progress
2024, Vol. 107(0) 1–20
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00368504231220766
journals.sagepub.com/home/sci

Sage

Machine learning approach for carbon disclosure in the Korean market: The role of environmental performance

Jeong Hwan Lee[1], Jin Hyung Cho[2], Bong Jun Kim[1] and Won Eung Lee[1]

[1]College of Economics and Finance, Hanyang University, Seoul, Korea
[2]Kakao, Seongnam-si, Gyeonggi-do, Korea

Abstract

Over the past few decades, scholars have employed a wide range of methodologies to determine the factors influencing firms' voluntary carbon disclosure. Most of these studies have been conducted in advanced markets. This article aims to examine the trend of voluntary carbon disclosure in the Korean financial market by utilizing machine learning models such as Random Forest and Gradient Boosted Decision Tree. Based on a set of hand-collected carbon disclosure data, we initially demonstrated significantly better performance of machine learning models compared to the traditional logistic model. Regarding the factors influencing disclosure, we consistently find the importance of environmental scores, emphasizing the role of the emerging mega-trend of ESG management practices in disclosure decisions. However, in contrast to recent studies, we do not find that the unique Korean governance structure, *chaebol*, has any significantly different implications in terms of prediction performance and variable importance in carbon disclosure decisions.

Keywords

Carbon emission, chaebol, CSR, ESG, machine learning, RF, GBDT

Introduction

To implement effective Environmental Social Governance (ESG) management practices, an increasing number of firms nowadays disclose their carbon emissions in an effort to comply with global standards. For example, international organizations such as the

Corresponding author:
Jeong Hwan Lee, College of Economics and Finance, Hanyang University, Seoul, Korea
Email: jeonglee@hanyang.ac.kr

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access page (https://us.sagepub.com/en-us/nam/open-access-at-sage).

Carbon Disclosure Project (CDP) and the Global Reporting Initiative have encouraged firms to disclose their environmental information, thereby enhancing transparency. These companies voluntarily disclose and report their carbon emission levels to national governments and international agencies, detailing their efforts to promote sustainability in business.[1–3]

In recent years, there has been a considerable amount of research on corporations' voluntary tendency to disclose carbon emissions at both international and regional levels.[4–7] However, while an increasing number of firms from emerging countries are engaging in carbon disclosure practices,[5,8] the majority of studies have focused on highly developed countries, relying exclusively on traditional logistic regression models to predict disclosure tendencies.

This article aims to address the aforementioned limitations in the existing literature. To achieve this, we first compare the performance of machine learning models with the traditional logit model in explaining the decision-making process behind voluntary carbon disclosure. Traditional linear models may have limited explanatory power for the disclosure decision because it is a voluntary choice, not a mandatory one. Such voluntary decisions inherently involve more complex managerial considerations. Next, we determine the factors influencing carbon disclosure decisions by analyzing the importance of each variable in machine learning models. Particularly, if machine learning models demonstrate superior performance, the conclusions drawn from previous research on voluntary disclosure, largely reliant on logit models, may need to be reconsidered.

To overcome regional limitations present in existing studies, we conduct these empirical analyses within the context of the Korean financial markets. The choice of this market is deliberate for several reasons. The Korean market is unique as it has transitioned from a developing market to an advanced market. Similar to other advanced markets, the Korean market offers an extensive dataset, including reliable ESG performance scores. Having access to a wide range of datasets is essential for applying machine learning algorithms. Furthermore, the Korean market is suitable for observing the impact of firm heterogeneity on determining ESG policies, especially concerning corporate governance structure. Recent studies on corporate environmental policies, including carbon disclosure,[5,9,10] underscore the significance of governance structure in shaping distinct ESG policies.

Our work yields several noteworthy results. First, we demonstrate that machine learning algorithms, including Random Forest (RF, hereafter) and Gradient Boosted Decision Tree (GBDT, hereafter) models, exhibit significantly better performance in identifying firms' carbon disclosure tendencies. Interestingly, the traditional logit model fails to accurately predict any instances of carbon disclosure in our test sample. Second, our results emphasize environmental performance as one of the most crucial factors in a firm's decision to disclose its carbon emissions. Variables related to firm size also prove their significance in determining voluntary carbon disclosure. Finally, we find no significant differences in the determinants of carbon disclosure decisions between *chaebol* and non-*chaebol* affiliates. Factors such as size, profitability, and environmental performance remain important regardless of corporate governance structures.

In particular, we deviate from previous research[5,6,11] by exclusively employing machine learning methodologies to predict firms' tendencies to disclose carbon emissions and to identify the factors influencing this disclosure. We incorporate unique firm

characteristics, such as affiliation with *chaebol*, into our analysis, which may influence corporate policies related to carbon disclosure. Consequently, our analysis not only demonstrates the superior performance of machine learning methodologies but also successfully considers firm heterogeneity—both *chaebol* and non-*chaebol*—which is distinctive in the Korean capital market.

This article contributes to the literature in several ways. Firstly, we directly showcase the superior performance of machine learning models in explaining carbon disclosure decisions. While a majority of existing studies rely on traditional logit models,[5,6] we strongly advocate for the potential role of machine learning models in analyzing disclosure determination. Additionally, we underscore the significance of environmental performance in the decision-making process of carbon disclosure. This implies that the recent trend of ESG management, including the management of environmental factors as key drivers of carbon disclosure decisions, contrasts with approaches focusing solely on economic efficiencies, such as carbon intensities.

Finally, our findings indicate a rather homogenous structure in carbon disclosure decisions across *chaebol* and non-*chaebol* affiliates, unlike recent studies on Korean ESG policies. For instance, Yoon et al. (2018) find a stronger valuation effect of ESG performance within the group of *chaebol* affiliates.[9] Yoon et al. (2021) also show that higher ESG performance limits tax avoidance in the group of *chaebol* affiliates.[10] However, our results emphasize the lack of significant differences in economic determinants influencing voluntary carbon disclosure between the two groups.

The remainder of our paper is structured as follows. The second section provides the literature background on carbon disclosure and machine learning algorithms. The third section introduces our data construction and the methodology of machine learning and traditional regression models. The fourth section presents the results of the data analysis. Finally, the fifth section concludes.

Literature review

Firm's tendency to carbon disclosure

Traditionally, a firm's inclination toward voluntary carbon disclosure is closely linked to stakeholder theory in management. It is believed that the resources of various stakeholders form the foundation of firm development, and firm management should meet the needs and expectations of stakeholders in business activities.[12] However, in the wake of climate change, the disclosure of carbon emission information has become a significant concern for energy conservation and environmental protection. It is believed that the quality of carbon information disclosure is positively related to institutional expectations.[13,14]

Recently, ESG practices themselves have been highlighted as important factors in the decision-making process for carbon emission disclosure. ESG management has become a major trend in managerial practice, especially after the emergence of ESG investments as effective tools for asset allocation.[9] Firms adopting ESG management practices may voluntarily disclose carbon emission information because such disclosure is considered the first step in addressing climate change, a key aspect of ESG practices.[5]

Empirical studies on carbon emissions often utilize data from the CDP. The CDP is a nonprofit institution that collects global data on firms' carbon disclosure. Specifically, third-party disclosure approaches like firms' CDP reporting minimize firm-specific biases and enhance the quality of disclosed information.[15] In fact, CDP reports are believed to provide more comprehensive and comparative information than individual firms' carbon disclosure reports typically do,[16–18] as the comparability of carbon emission information from individual firm reports is weak. For instance, after analyzing 19 Canadian firms between 2004 and 2015, Wegener et al. (2019) found that carbon data information revealed by different institutions was disconnected, resulting in a lack of comparability in the entire dataset.[19] Moreover, Depoers et al. (2016) examined 140 French-listed firms and found that greenhouse gas emission information disclosed in firm annual reports was lower than that in CDP reports.[20]

As such, previous studies utilize carbon disclosure data collected through CDP reports as the main variable to investigate the effect of firms' characteristics on their tendency to disclose carbon emissions.[5,21] In particular, existing studies highlight the positive relationship between a firm's disclosure of carbon emission information and its shareholder value at both international and regional levels. For instance, using CDP data from the United States, Brazil, Russia, India, and China, Yan et al. (2021) argue for the positive valuation effect of carbon disclosure.[21] They also demonstrate that this positive relationship is more significant in developing countries.

A growing number of studies emphasize the positive relationship between firms' environmental performance and their willingness to voluntarily disclose carbon information. Based on the analysis of CDP data from Global 500 firms between 2008 and 2011, Guenther et al. (2016) state that firms' carbon emission performance is positively related to the level of carbon emissions disclosure.[22] Dawkins et al. (2011) demonstrate a positive relationship between environmental performance and the disclosure of climate change by analyzing the U.S. financial market.[23]

Another line of empirical studies highlights the relationship between firm size, profitability, and voluntary carbon disclosure. For example, after analyzing annual reports of 37 listed firms in Indonesia, Faisal et al. (2018) conclude that large size, high profitability, and low leverage ratio of firms are positively associated with voluntary carbon emission disclosure.[24] Stanny and Ely (2008) analyzed 494 firms participating in CDP in 2007 and pointed out that firms with large size, new assets, and a history of carbon disclosure are more willing to disclose carbon emissions.[16]

Machine learning prediction

Existing studies commonly rely on traditional logistic regression models in the analysis of voluntary carbon disclosure.[5,7,15,21,25] This is because logistic models provide statistics for hypothesis testing purposes. In fact, these studies generally attempt to test existing theories related to voluntary carbon disclosure based on such statistics, often at the expense of adopting alternative models with superior explanatory power.

Recently, the use of machine learning techniques to analyze management trends has become widespread. For instance, various machine learning methodologies are applied to credit rating,[26,27] firm bankruptcy,[28,29] and optimal capital structure.[30] These studies

aim to enhance the predictive power of models and provide new insights into a wide range of research topics.

However, only a few studies on carbon disclosure have employed machine learning methodologies. Harker et al. (2022) examined 839 listed firms in Australia and used machine learning techniques to extract climate risks and risk management strategies identified by each firm.[31] Their results revealed that the most serious risk identified by Australian firms is physical climate risk, followed by market and regulatory risks. Quyen et al. (2021) employed various machine learning algorithms, such as multilayer perception neural networks, RF, and extreme gradient boosting, to explain the predictive power of carbon emissions patterns.[32] They found that prediction accuracy improved by incorporating features such as energy production and disclosures of firms in specific regions and sectors.

Korean market and firms' carbon disclosure

The Korean government has introduced various measures to counter the risk of carbon emissions, including the promotion of eco-friendly vehicles and the implementation of the Emissions Trading Scheme. In line with these efforts, numerous firms have taken various measures, including carbon disclosure, to address demands from external stakeholders. Particularly, an increasing number of Korean firms report their carbon emissions levels through CDP reports.

Several studies in the Korean market use CDP reports to determine whether firms disclose their carbon emissions. For example, using CDP data, Lee and Cho (2021) in Korea found that Korean firms with higher environmental performance are likely to voluntarily disclose their carbon emissions.[5] They also emphasize the importance of firm heterogeneity, such as corporate governance structure, in the decision-making process for voluntary disclosure. Choi and Noh (2016) in Korea observed a higher firm value and a better credit rating for firms voluntarily disclosing carbon information based on their hand-collected data on carbon disclosure.[11]

Data construction and methodologies

Data construction

CDP annually selects 250 Korean firms based on market capitalization and requests their annual carbon emission data (scopes 1, 2, and 3). From the report, we manually collect carbon emission data for each year to determine whether a firm discloses its carbon emissions voluntarily or not. The emission information in the report is left unmarked if a firm decides not to reveal the degree of carbon emission. We then exclude companies belonging to the financial or insurance industries from the sample analysis due to the distinctive nature of their business.

The sample period of our analysis spans from 2013 to 2020, reflecting the data availability of environmental score, one of our key variables of interest. We utilize the ESG score published by Sustinvest, the second-largest ESG rating company in the Korean financial market. Financial and macroeconomic variables are obtained from FnGuide, a

financial data-providing institution in Korea, and the Bank of Korea, respectively. To categorize firms into *chaebol* and non-*chaebol* affiliations, we employ the list of *chaebol* affiliates offered by the Korean Fair Trade Commission.

Figure 1 presents the results of voluntary disclosures over the sample period. After the exclusion process, the proportion of firms choosing to voluntarily disclose their carbon emissions between 2013 and 2020 is 109 out of 1355 (8.0%). The sample of voluntary disclosure is divided into 96 for *chaebols* and 13 for non-*chaebol* firms. Interestingly, *chaebol* affiliates are more likely to disclose their carbon emission information. In fact, the proportion of carbon emission disclosure is 30.6% (96 out of 314) for *chaebol* affiliates, which is significantly higher than that of non-*chaebol* affiliates at 1.2% (13 out of 1041).

To construct the features for our machine learning methodologies, we primarily follow the approach of Amini et al. (2021),[30] which suggests a wide range of firm-specific and macroeconomic variables to explain corporate policies. Specifically, we employ the following firm-specific and financial variables: Profitability (*Profit*), Firm Size (*Assets*), Market-to-Book ratio (*Mktbk*), Assets Growth (*ChgAsset*), Physical Investment (*Capex*), Assets Tangibility (*Tang*), Innovation Investment (*RD*), Nonproduction cost (*SGA*), Top Tax Rate (*Taxrate*), Depreciation (*Depr*), Stock Variance (*StockVar*), Bankruptcy Probability (*Zscore*), Logarithm of annual firm sales less capital expenditure (*Macroprof*), and Market Returns (*CrspRet*). Four classes of leverage variables such as TDM, TDA, LDM, and LDA are introduced based on the combinational use of total/long-term debt as the numerator and market/book value as the denominator. To control for industry effects, we include Industry Leverage (*Industlev*), Industry Growth (*Industrgr*), and Net Payout (*Netpay*). Lastly, to consider macroeconomic factors in our analysis, we add Term Spread, which is the difference between the returns of

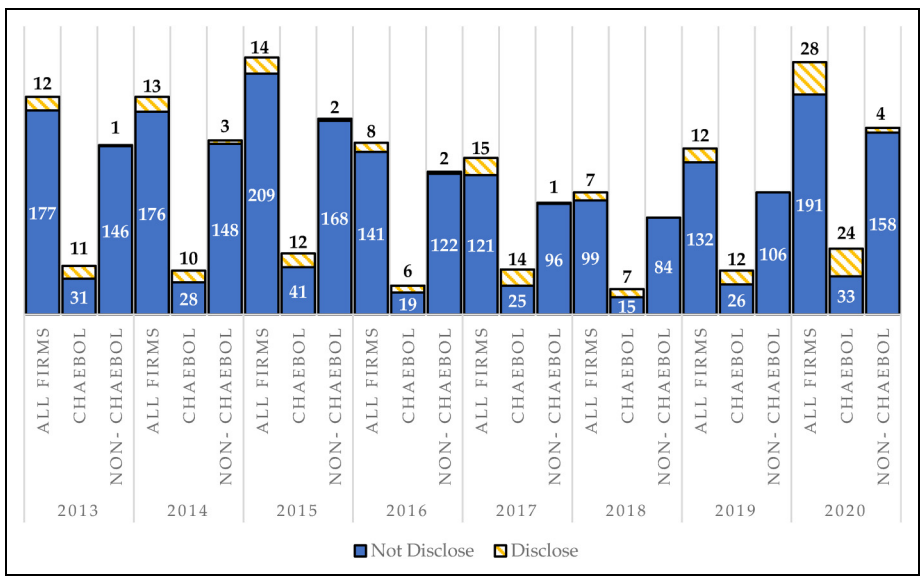

Figure 1. Tendency of carbon disclosure for all firms, chaebols, and non-chaebols.

Table 1. Variable definitions.

Variables	Description
Environment score (Env)	Environment score
Logarithm of Total Asset (LogTA)	Natural Logarithm of Total Asset
Return on Asset (ROA)	Firms' Earning/Total Asset
Total Leverage (Leverage)	Total Debt/Total Equity
Adveritsing Expense (AdvExp)	Advertising Expenses/Total Assets
Market Value of Equity (MVE)	The stock's close price in fiscal year times common shares outstanding
Market Value of Assets(MVA)	Debt in current liability, plus long-term debt, plus preferred stock liquidating value, minus deferred tax + MVE
Leverage (TDM)	(Debt in current liabilities plus long-term debt)/MVA
Leverage (TDA)	(Debt in current liabilities, plus long-term debt)/total asset
Leverage (LDM)	Long-term debt/MVA
Leverage (LDA)	Long-term debt/total assets
Profitability (Profit)	Operating income before depreciation/total asset
Firm Size (Assets)	The logarithm of total asset
Market-to-Book (Mktbk)	Market value of assets/total asset
Assets Growth (ChgAsset)	Change in the logarithm of total asset
Physical Investment (Capex)	Capital expenditure/total asset
Assets Tangibility (Tang)	Net property, plant and equipment/total assets
Innovation Investment (RD)	R&D expenses/total sales
Nonproduction cost (SGA)	Selling, general and administrative expenses/total sales
Cash Holdings (Cash)	Cash and short-term investments/total asset
Top Tax Rate (Taxrate)	The top corporate tax rate
Depreciation (Depr)	Depreciation and amortization/total assets
Stock Variance (StockVar)	The annual variance of daily stock returns
Bankruptcy Probability(Zscore)	Altman's Z-score
Annual stock return change(Stock)	Change in annual stock return
Market Returns (CrspRet)	Cumulative annual market returns using monthly raw returns
Industry Leverage (Industlev)	The median value of firm leverage (TDM)
Industry Growth (industgr)	The median value of assets growth (ChgAsset)
Term Spread (Termsprd)	The difference between 10-year bond returns and 1-year bond returns
Inflation	Expected one-year change in the Consumer Price Index
Logarithm of annual firm sales less capital expenditure (Macroprof)	Change in the logarithm of annual firm sales with inventory valuation and capital consumption adjustments
Growth in GDP(Macrogr)	Change in the logarithm of real GDP
Net Payout (Netpay)	(Cash dividends, plus purchase of common and preferred stock, minus sale of common and preferred stock)/total assets

For emerging countries including Korea, the calculation methodology for Z-score is calculated as follows (Meeampol et al., 2014).[33]

$Z = 3.25 + 6.56X_1 + 3.26X_2 + 6.72X_3 + 1.05X_4$

X_1 = (Current Asset – Current Debt) / Total Asset

X_2 = Current Profit/Total Assets

X_3 = EBIT/Total Assets

X_4 = Total Capital/Total Debt

1-year bonds and 10-year bonds (*Termsprd*), Inflation (*Inflation*), and GDP Growth (*Macrogr*) to both logistic and machine learning models. Table 1 presents detailed definitions for each financial, firm-specific, and macroeconomic variable used in our empirical analysis.

Methodologies

Logistic regression

The logistic regression model is commonly utilized in the literature on voluntary carbon disclosure. This model is developed as a statistical analysis tool that models the probability of an event taking place by making the log-odds for the event a linear combination of independent variables. A single dependent variable coded as 0 and 1 in the model represents "no disclosure" and "disclosure" of carbon emission information, respectively. The logistic function converts the log odds to probability using the estimated coefficients. To be specific, the log odds ratio is defined as follows:

$$\log \text{Odds} = \log\left(\frac{P(X)}{1 - P(X)}\right) \tag{1}$$

Then, logistic regression predicts the probability for the positive class $P(y_i = 1|X_i)$ by employing the logistic function:

$$p\,(X_i) = \text{expit}(X_i w + w_0) = \frac{1}{1 + \exp\,(-X_i w - w_0)} \tag{2}$$

The loss for each data sample is calculated and then averaged to evaluate the adequacy of the model. The cost function of the logistic model is:

$$\min_w C \sum_{i=1}^{n} (-y_i \log(\hat{p}(X_i)) - (1 - y_i)\log(1 - \hat{p}(X_i))) + r(w) \tag{3}$$

To minimize it, the regularization term *r(w)* is introduced in the following analysis.

Random forest

Next, we introduce the RF model as the first machine learning algorithm model. RF is an ensemble-based machine learning model which is capable of performing both classification and regression. The RF is a variant of the decision tree model, a supervised machine learning mechanism that builds upon a tree-splitting procedure in the classification. While the decision tree is simple and intuitive in conducting classification tasks, the model is widely known to derive overfitting problems in the training sample.

To enhance the performance of classification, RF adopts an ensemble learning scheme based on the original decision tree model. The algorithm of RF adopts a general technique of bootstrap aggregating, or bagging that conducts repeated random sampling with replacement. To be specific, the bagging procedure constructs B times of bootstrapped samples with size N from the training dataset. For each bootstrapped sample, RF builds a new decision tree, conducts classification task, and obtain forecasts.

Apart from the original bagging mechanism described above, RF employs another type of bagging mechanism so-called "feature bagging." Given the number of p features, typically, $\sqrt{p}$ features are randomly selected to build up for each decision tree. This feature bagging procedure prohibits the generation of correlated decision trees that are considered in the model.

Finally, the majority voting rule is applied to determine the final prediction of the RF model. Each tree generates a different classification result, and by taking the majority vote among this set of classifications, the RF model finalizes its predictions. In other words, the hard voting result from each decision tree's classification is chosen as the final classification result. In this way, RF feeds fresh data into a variety of decision trees.[34–36]

In constructing Bagging estimates,[37] the following equation (4) illustrates the bagging procedure, generating bootstrap samples by randomly drawing with replacements, $b = 1, \ldots, B$. In each of our bootstrap samples, the estimate $\hat{g}_b(x)$ through minimizing the loss function is given as:

$$\min_{\hat{g}_b(x)} \sum_{i=1}^{N} \left(y_i^b - \hat{g}_b(x_i^b)\right)^2 \tag{4}$$

In this way, we put all of the estimated forecast $\hat{g}_1(x), \ldots, \hat{g}_B(x)$ in order to make a final Bagging estimate, as in the following equation (5):

$$\hat{g}(x)^{\text{bag}} = \frac{1}{B}\sum_{b=1}^{B} \hat{g}_b(x) \tag{5}$$

Gradient boosted decision trees

We employ another machine learning model, GBDT, which generalizes boosting to arbitrary differentiable loss functions.[38] Similar to RF, GBDT is generally considered an ensemble of decision tree models that are subject to substantial overfitting problems.

The GBDT algorithm can be easily understood in the context of the least square method. In the least square method, the objective is to minimize the mean squared error (MSE) of predictions, that is, $\sum_{i}^{n} (y_i - \widehat{y_i})^2 / n$, where y_i is the true value and $\widehat{y_i}$ is the estimated value from a predictive model.

GBDT adopts M steps of the learning procedure. At each stage m, the algorithm incorporates a new estimator $h_m(x_i)$ to the estimate of mth stage $F_m(x_i)$ to improve the performance of predictors. GBDT fits h_m to the loss or the residual of prediction $y_i - F(x_i)$. It can be numerically shown that h_m is proportional to the negative gradients of the MSE. Accordingly, gradient boosting can be a specialized version of a gradient descent algorithm. This logic can be generalized to any type of loss function used in classification and ranking problems as well.

Specifically, when a training dataset $D = \{x_i,\ y_i\}_1^N$ is given, gradient boosting finds an approximation $\hat{g}^{gbm}(x)$ in the function $\hat{g}^{gbm*}(x)$, which maps instances x to the output values y, by minimizing the expected value of any given loss function, $L(y,\ \hat{g}^{gbm}(x))$. Thus, an estimate of $\hat{g}^{gbm}(X)$ is the weighted sum of individual estimates of each tree,

as in the equation (6):

$$\hat{g}_m^{gbm}(x) = \hat{g}_{m-1}^{gbm}(x) + \rho_m h_m(x) \tag{6}$$

where ρ_m is the weight of the mth function, which is $h_m(x)$. The approximation is then constructed iteratively. To be specific, a constant approximation of $\hat{g}^{gbm*}(x)$ is obtained as in the equation (7):

$$\hat{g}_0^{gbm}(x) = \arg\min_{\alpha} \sum_{i=1}^{n} L(y_i, \alpha) \tag{7}$$

In consideration of the aforementioned equations, the following model is expected to minimize:

$$(\rho_m, h_m(x)) = \arg\min_{\rho,h} \sum_{i=1}^{N} L(y_i, \hat{g}_{m-1}^{gbm}(x_i) + \rho h(x_i)) \tag{8}$$

Additionally, instead of solving the optimization issue, each h_m is a greedy step in a gradient descent optimization for $\hat{g}^{gbm*}$. Each model, h_m, is then trained on a new dataset, which is $D = \{x_i,\ r_{mi}\}_{i=1}^{N}$, where the pseudo-residuals, r_{mi}, are calculated by the following equation (9):

$$r_{mi} = \left[\frac{\partial L(y_i, \hat{g}^{gbm}(x))}{\partial \hat{g}^{gbm}(x)}\right]_{\hat{g}^{gbm}(x)=\hat{g}_{m-1}^{gbm}(x)} \tag{9}$$

The main advantage of gradient boosting is a unique approach of sequentially connecting a large number of relatively shallow decision trees. This results in significantly enhanced performance for each tree, as each decision tree accurately predicts specific data points.

Empirical results

We first compared the prediction performance of machine learning models with that of the traditional logistic regression (logit) model. Prediction performance was assessed using several metrics. Firstly, we considered accuracy, defined as the ratio of correct predictions to all predictions made by a specific algorithm. Accuracy indicates the quality of a machine learning model's binary classification predictions and can be expressed as follows:

$$\text{Accuracy} = \frac{\text{Correct predictions}}{\text{All predictions}} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{10}$$

Here, true positive (TP), true negative (TN), false positive (FP), and false negative (FN) represent the correctly predicted positive cases, correctly predicted negative cases, incorrectly predicted positive cases, and incorrectly predicted negative cases, respectively.

Precision, on the other hand, is defined as the ratio of true positives to the sum of true positives and false positives. Precision measures the accuracy of positive predictions and

can be calculated as follows (equation 11):

$$\text{Precision} = \frac{\text{True positive}}{\text{Positive predictions}} = \frac{\text{TP}}{\text{TP} + \text{TN}} \tag{11}$$

Recall is calculated as the ratio of true positives to the sum of true positives and false negatives. It represents the proportion of actual positives that were correctly identified and is commonly known as sensitivity or specificity.

$$\text{Recall} = \frac{\text{True positive}}{\text{Actual positives}} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{12}$$

Finally, the F1-score, as shown in the following equation (13), combines recall and precision into a single metric ranging from 0 to 1.

$$\text{F1} - \text{score} = 2\frac{\text{Precision} * \text{recall}}{\text{Precision} + \text{recall}} \tag{13}$$

Table 2 presents the prediction performance of logistic models and machine learning models. These performance measures were calculated using the testing sample, where the training sample and the testing sample constituted 75% and 25% of the entire sample, respectively. The logit model, RF, and GBDT models were applied to samples of all firms, *chaebols*, and non-*chaebols*.

The table clearly shows that the logistic model lacks reliable predictive power for firms' voluntary carbon disclosure. This finding holds true for both *chaebol* and non-*chaebol* affiliates. In fact, the logistic model records 0 for both precision and recall rate, indicating its failure to accurately predict a firm's carbon disclosure when it voluntarily discloses carbon emission information. Such low prediction performance raises significant doubts about previous studies relying on traditional logistic regression methods.

In contrast, machine learning models demonstrate superior performance in predicting voluntary carbon disclosure. For the RF model, the precision for voluntary disclosure (class = 1) is 0.95 for the entire firm's sample and 0.88 for the *chaebol* affiliates sample. GBDT also exhibits a similar pattern of prediction performances, with precision for voluntary carbon disclosure being 1.00 for all firms and 0.95 for *chaebol*.

The superior performance of machine learning models might be closely related to the aforementioned "voluntary" nature of carbon disclosure decisions. Unlike other disclosures mandated by regulations, carbon disclosure is a voluntary decision that can be significantly influenced by high-dimensional managerial considerations. The linear structure of the logistic model may be too restrictive to capture such a complex mechanism.

Notably, the prediction accuracy is reported as 0.99 for both the logistic model and the machine learning models in the case of non-*chaebol*. However, this similarity may not demonstrate that the prediction performances are identical across the logit and machine learning models. The dataset of non-*chaebol*, which mostly does not disclose their carbon emissions, results in only four samples remaining in the testing sample. This limitation hinders a comprehensive analysis of performance comparison.

Now we turn to examine the feature importance and permutation importance of the logistic and the machine learning models. A higher value for these importance measures

Table 2. Performance of predictions.

		All firms				Chaebol				Non-Chaebol			
		Precision	recall	F1-score	Support	Precision	Recall	F1-score	Support	Precision	Recall	F1-score	Support
Logistic	0	0.89	1.00	0.94	323	0.62	1.00	0.77	59	0.99	1.00	0.99	264
	1	0.00	0.00	0.00	40	0.00	0.00	0.00	36	0.00	0.00	0.00	4
	Macro average	0.44	0.50	0.47	363	0.31	0.50	0.38	95	0.49	0.50	0.50	268
	Weighted avgaverage	0.79	0.89	0.84	363	0.39	0.62	0.48	95	0.97	0.99	0.98	268
		Accuracy		0.89	363	Accuracy		0.62	95	Accuracy		0.99	268
RF	0	0.94	1.00	0.97	323	0.81	0.95	0.88	59	0.99	1.00	0.99	264
	1	0.95	0.53	0.68	40	0.88	0.64	0.74	36	0.00	0.00	0.00	4
	Macro avgaverage	0.95	0.76	0.82	363	0.85	0.79	0.81	95	0.49	0.50	0.50	268
	Weighted average	0.95	0.94	0.94	363	0.84	0.83	0.82	95	0.97	0.99	0.98	268
		Accuracy		0.94	363	Accuracy		0.83	95	Accuracy		0.99	268
GBDT	0	0.92	1.00	0.96	323	0.79	0.98	0.88	59	0.99	1.00	0.99	264
	1	1.00	0.30	0.46	40	0.95	0.58	0.72	36	0.00	0.00	0.00	4
	Macro average	0.96	0.65	0.71	363	0.87	0.78	0.80	95	0.49	0.50	0.50	268
	Weighted average	0.93	0.92	0.90	363	0.86	0.83	0.82	95	0.97	0.99	0.98	268
		Accuracy		0.92	363	Accuracy		0.83	95	Accuracy		0.99	268

GBDT: Gradient Boosted Decision Tree; RF: Random Forest.

generally indicates a more significant role in the decision of voluntary disclosure. The feature importance measure is usually provided by a class of decision tree models and is defined as the mean decrease in impurity. This impurity, in terms of Gini, Log Loss, or MSE, is calculated by the splitting criterion of the decision trees. The impurity-based feature importance is significantly biased toward high cardinality features such as continuous variables. To address such problems, the permutation importance is defined as the decrease in a model score if a single feature value is randomly changed. Unlike the feature importance, permutation importance can be computed using either the training set or testing set.

Figure 2 below presents the variable importance of all firms for the logistic regression model, RF model, and GBDT. The results of the logistic regression model are reported in panel A. The results for RF and GBDT are depicted in panels B and C, respectively.

Figure 2 highlights the importance of environmental performance in the decision of voluntary carbon disclosure. Most importantly, the permutation importance in the machine learning models indicates that a corporation's environmental performance (*ENV*) is the most crucial determinant in the decision of voluntary disclosure. Both RF and GBDT rank environmental performance as the most significant factor. Furthermore, the feature importance draws the same conclusion. Environmental performance ranks first in the logistic model and GBDT model, and it ranks second in the RF model.

This finding emphasizes the significance of ESG practices in determining voluntary disclosures. Before the 2010s, the major factor driving voluntary disclosure was considered to be the economic benefits and costs related to carbon intensity. However, in the 2010s, ESG management practices have emerged as key management objectives that corporations seek to achieve, especially for large firms. The disclosure of carbon information is generally seen as an initial step toward the adoption of ESG management practices. Firms with strong ESG practices tend to exhibit higher environmental performance with voluntary disclosure.

It is also noteworthy that other firm-specific factors related to firm size are shown to be important in the decision of voluntary carbon disclosures. The logarithm of annual firm sales less capital expenditure (*Macroprof*), the market value of assets (*MVA*), and the market value of equity (*MVE*) are significant in both feature and permutation importance across all examinations. These variables generally represent sales volume and the size of book/market equity values, all of which are closely associated with the firm's size. In other words, large firms have a strong tendency to disclose carbon information voluntarily.

The significance of firm size is generally in line with the predictions of stakeholder theory. Large firms tend to have a vast network of stakeholders and thus more strongly need to meet the expectations of these stakeholders in their business activities. Carbon information disclosure might be one of the important demands from stakeholders under the rapidly growing trend of the green economy. Accordingly, large firms may disclose their carbon emissions voluntarily to satisfy the demand of stakeholders for a green economy.

Figures 3 and 4 show the feature and permutation importance for *chaebol* and non-chaebol affiliates, respectively. The importance of these variables is reported for the logistic regression model, RF model, and GBDT. The results of the logistic regression model

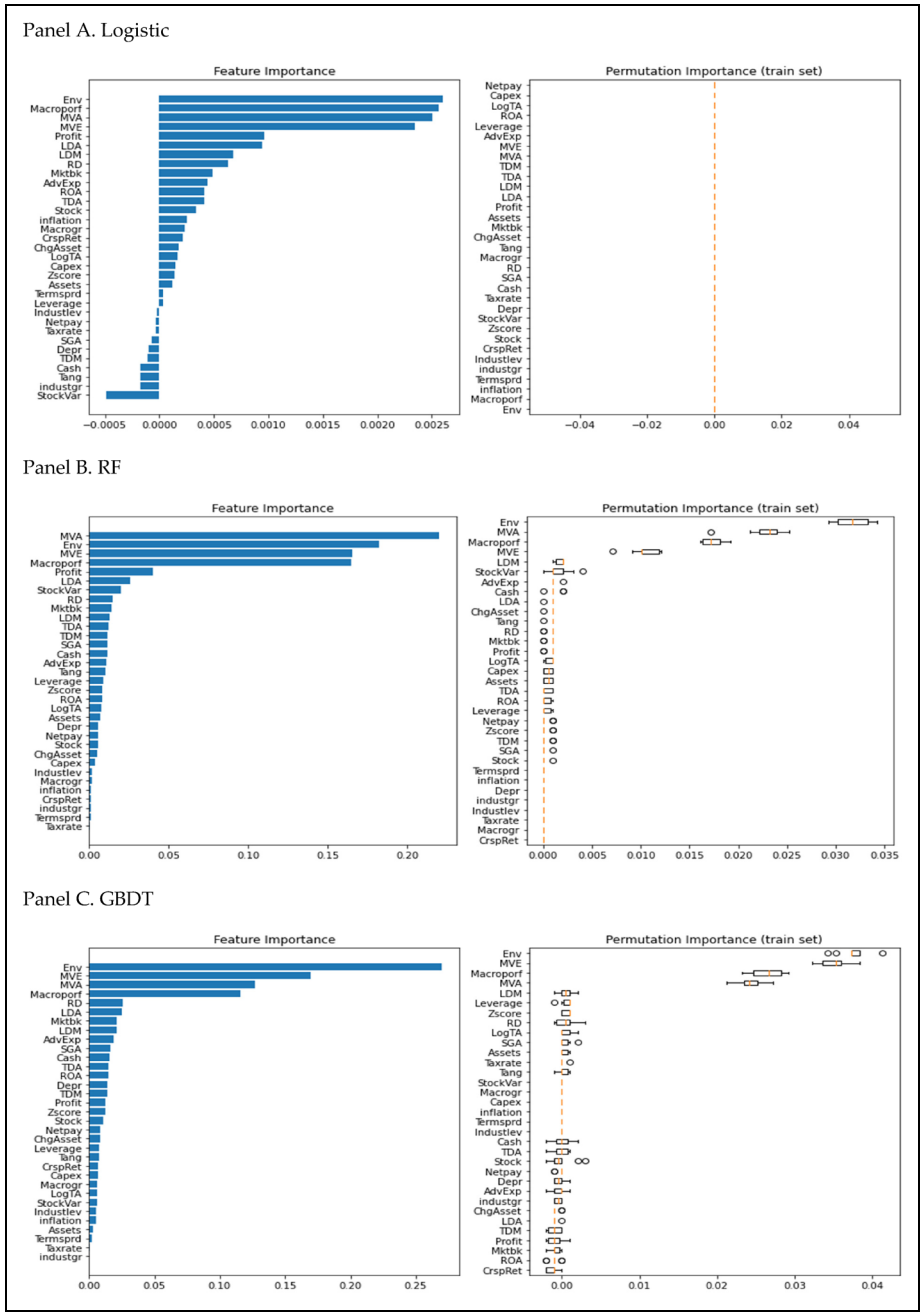

Figure 2. Variable importance: entire sample.

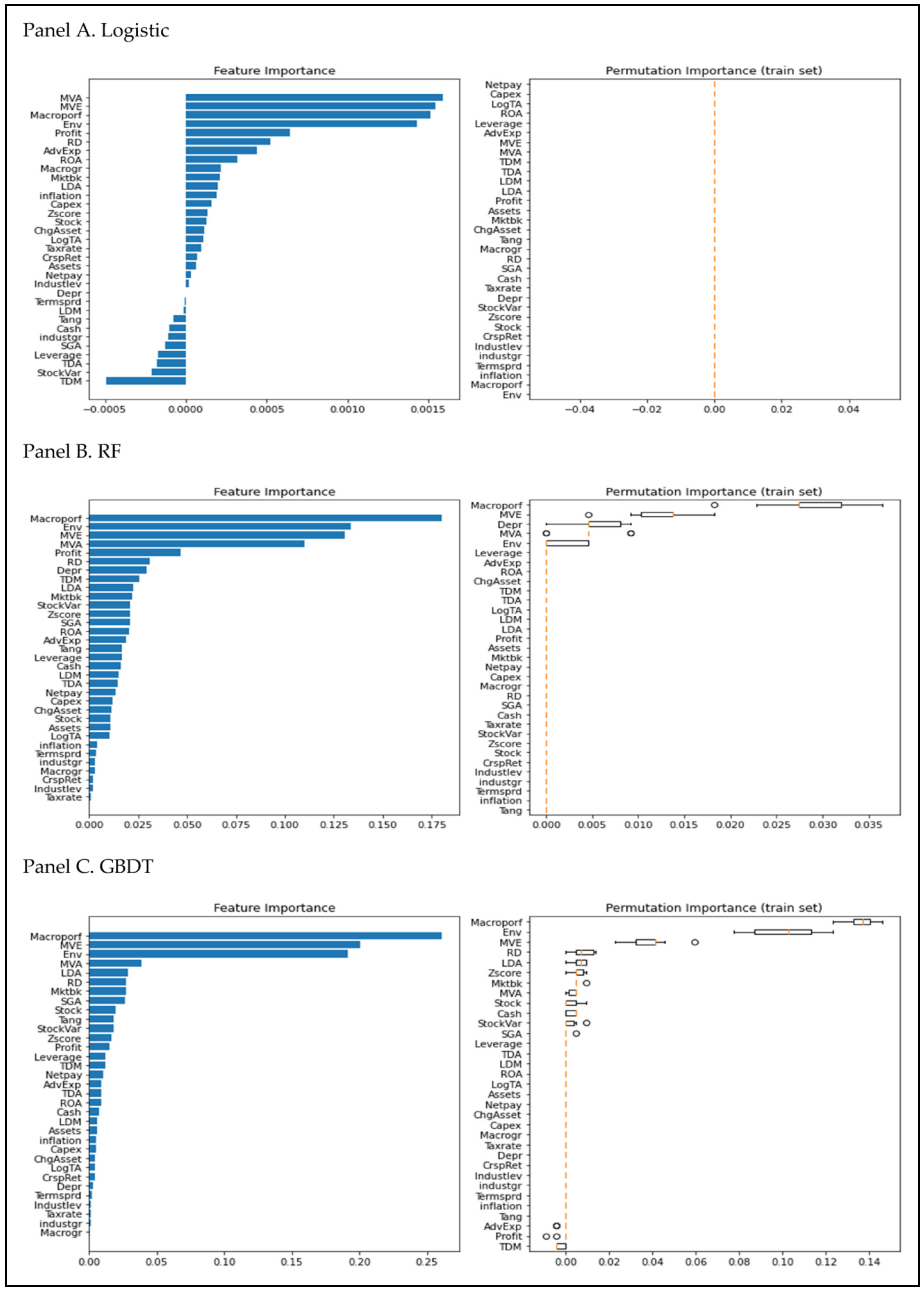

Figure 3. Variable importance: chaebol affiliates.

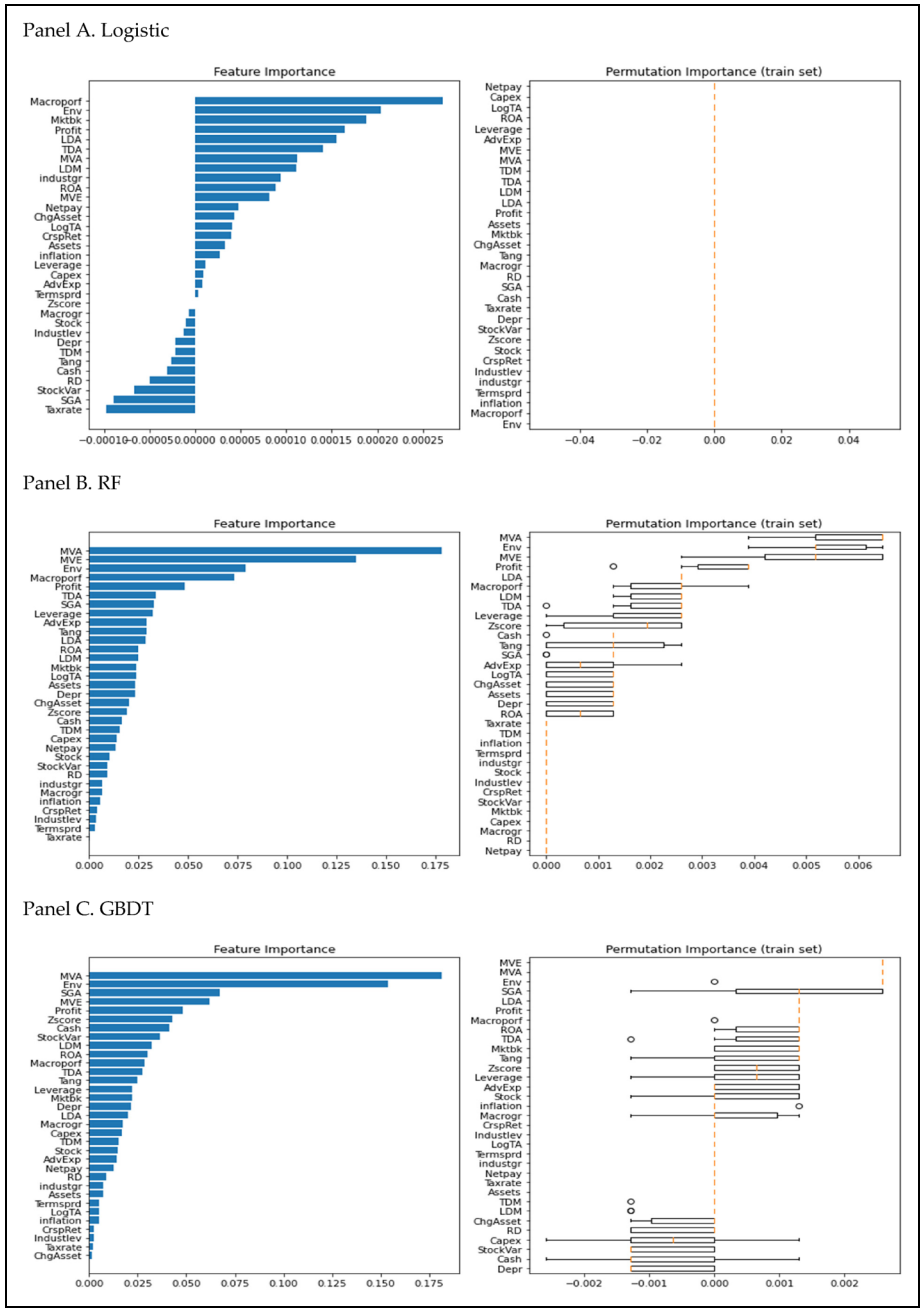

Figure 4. Variable importance: non-chaebol affiliates.

are reported in panel A, while the results for RF and GBDT are described in panels B and C, respectively.

The figures clearly show no significant differences in variable importance across the groups of *chaebol* and non-*chaebol* affiliates. Specifically, the logarithm of annual firm sales less capital expenditure (*Macroprof*), the *MVE*, and the MVA still show significance in the determination of voluntary carbon disclosure. The role of environmental performance is robust across these two groups of firms as well, although its significance decreases compared to the results of the entire sample analysis.

Such a finding is not well aligned with recent studies emphasizing the significant heterogeneity in shaping corporate ESG policies. In contrast to studies finding different valuation effects of ESG policies or carbon disclosure policies across *chaebol* and non-*chaebol* affiliates in the Korean market,[5,9] our results suggest no significant differences in variable importance across these groups in the case of voluntary carbon disclosure. This finding could be surprising because the tendency of voluntary disclosures is shown to be quite distinctive across *chaebol* and non-*chaebol* affiliates; *chaebol* affiliates are highly likely to disclose information on carbon emissions, as shown in Figure 1.

Discussion

Our research presents various approaches to predicting a firm's carbon disclosure performance and variable importance by adopting machine learning and traditional logistic models. Our results convey some important implications, which are built upon previous literature on carbon information disclosure and firm-specific characteristics.[13,14,39–41]

First, our results align well with previous literature[13,14] in that the transparent disclosure of carbon information is positively related to the expectations of institutions. As demonstrated in the prediction of a firm's carbon disclosure, the superior performance of the machine learning model, which considers managerial decisions in various dimensions, may result from firms' conformity to the needs of shareholders, creditors, and other related groups.[39]

Second, our findings on the prediction of variable importance in carbon disclosure suggest that machine learning models better capture the firm-specific factors that determine the voluntary tendency to disclose carbon information. Notably, our machine learning models not only capture environmental performance but also specific factors, such as the size of the firm. These findings align well with previous research[39,40] indicating that the disclosure of carbon information is positively related to firm size. In other words, firms with a large size tend to make more extensive disclosures compared to firms with a small size. This could further affect the credibility and extent of carbon information that firms may be more willing to disclose to stakeholders in the future.[41] In this sense, it is plausible to argue that our machine learning models effectively capture the high-dimensional characteristics of managerial discretion.

Our study has several limitations. First, although this study employs a variety of machine learning models, we do not explore other machine learning models or deep learning techniques. Second, our research does not consider the firm value aspects of carbon emission disclosure.[5,6] These aspects need to be addressed in future studies.

Conclusion

Our analysis mainly confirms that the machine learning models, RF and GBDT, show better performance over the traditional logit models in the disclosure decision of carbon emissions. The traditional logit model is not able to predict any voluntary disclosure correctly in the testing sample and all the performance measures including precision, accuracy, recall rate, and F1-score show a superior performance of the machine learning models. Furthermore, our results highlight the significance of environmental performance in the decision of carbon disclosure while the results also show the importance of firm size variables as well. However, we did not find any significantly different patterns across the groups of *chaebol* and non-*chaebol* affiliates in both terms of prediction performance and variable importance.

Our research is particularly important for the exclusive use of machine learning methodologies in analyzing a firms' tendency for carbon disclosure and firm factors in the determination of the carbon disclosure tendency.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iD


Jeong Hwan Lee https://orcid.org/0000-0002-9382-8952

Supplemental material

Supplemental material for this article is available online.

References

1. Bragdon J and Marlin J. Is pollution profitable? *Risk Manag* 1972; 19: 9–18.
2. Chava S. Environmental externalities and cost of capital. *Manag Sci* 2014; 60: 2223–2247.
3. Sengupta PP. Corporate disclosure quality and the cost of debt. *Account Rev: Q J Am Account Assoc* 1998; 73: 459–474. https://www.jstor.org/stable/pdfplus/248186.pdf.
4. Hardiyansah M, Agustini AT and Purnamawati I. The effect of carbon emission disclosure on firm value: environmental performance and industrial type. *J Asian Finance Econ Bus* 2021; 8: 123–133.
5. Lee JH and Cho JH. Firm-value effects of carbon emissions and carbon disclosures—evidence from Korea. *Int J Environ Res Public Health* 2021; 18: 12166.
6. Matsumura EM, Prakash R and Vera-Muñoz SC. Firm-value effects of carbon emissions and carbon disclosures. *Account Rev* 2013; 89: 695–724.
7. Saka C and Oshika T. Disclosure effects, carbon emissions and corporate value. *Sustain Account Manag Policy J* 2014; 5: 22–45.

8. Li Y, Eddie I and Liu J. Carbon emissions and the cost of capital: Australian evidence. *Rev Account Finance* 2014; 13: 400–420.
9. Yoon BH, Lee JH and Byun R. Does ESG performance enhance firm value? Evidence from Korea. *Sustainability* 2018; 10: 3635.
10. Yoon BH, Lee JH and Cho JH. The effect of ESG performance on tax avoidance—evidence from Korea. *Sustainability* 2021; 13: 6729.
11. Choi JS and Noh JH. Usefulness of voluntarily disclosed carbon emission information. *Korean Account Rev* 2016; 41: 105–157.
12. Deegan C and Blomquist C. Stakeholder influence on corporate reporting: an exploration of the interaction between WWF-Australia and the Australian minerals industry. *Account Organ Soc* 2006; 31: 343–372.
13. Cotter J and Najah MM. Institutional investor influence on global climate change disclosure practices. *Aust J Manag* 2012; 37: 169–187.
14. Liesen A, Hoepner AGF, Patten DM, et al. Does stakeholder pressure influence corporate GHG emissions reporting? Empirical evidence from Europe. *Account Audit Account*. 2015; 28: 1047–1074.
15. Giannarakis G, Zafeiriou E and Sariannidis N. The impact of carbon performance on climate change disclosure. *Bus Strategy Environ* 2017; 26: 1078–1094.
16. Stanny E and Ely K. Corporate environmental disclosures about the effects of climate change. *Corp Soc Responsib Environ Manag* 2008; 15: 338–348.
17. Reid EM and Toffel MW. Responding to public and private politics: corporate disclosure of climate change strategies. *Strateg Manag J* 2009; 30: 1157–1178.
18. Luo L, Lan YC and Tang Q. Corporate incentives to disclose carbon information: evidence from the CDP global 500 report. *J Int Financ Manag Account* 2012; 23: 93–120.
19. Wegener M, Labelle R and Jerman L. Unpacking carbon accounting numbers: a study of the commensurability and comparability of corporate greenhouse gas emission disclosures. *J Cleaner Prod* 2019; 211: 652–664.
20. Depoers F, Jeanjean T and Jérôme T. Voluntary disclosure of greenhouse gas emissions: contrasting the carbon disclosure project and corporate reports. *J Bus Ethics* 2014; 134: 445–461.
21. Yan J, Luo L, Xu J, et al. The value relevance of corporate voluntary carbon disclosure: evidence from the United States and BRIC countries. *J Contemp Account Econ* 2021; 17: 100279.
22. Guenther E, Guenther T, Schiemann F, et al. Stakeholder relevance for reporting: explanatory factors of carbon disclosure. *Bus Soc* 2016; 55: 361–397.
23. Dawkins C and Fraas JW. Coming clean: the impact of environmental performance and visibility on corporate climate change disclosure. *J Bus Ethics* 2011; 100: 303–322.
24. Faisal F, Andiningtyas ED, Achmad T, et al. The content and determinants of greenhouse gas emission disclosure: evidence from Indonesian companies. *Corp Soc Responsib Environ Manag* 2018; 25: 1397–1406.
25. Desai R. Determinants of corporate carbon disclosure: a step towards sustainability reporting. *Borsa Istanb Rev* 2022; 22: 886–896.
26. Li J, Mirza N, Rahat B, et al. Machine learning and credit ratings prediction in the age of fourth industrial revolution. *Technol Forecast Soc Change* 2020; 161: 120309.
27. Wallis M, Kumar K and Gepp A. Credit rating forecasting using machine learning techniques. In: *Advances in data mining and database management book series*. IGI Global, 2019, pp.180–198.
28. Kim H, Cho H and Ryu D. Corporate bankruptcy prediction using machine learning methodologies with a focus on sequential data. *Comput Econ* 2021; 59: 1231–1249.

29. Lombardo G, Pellegrino M, Adosoglou G, et al. Machine learning for bankruptcy prediction in the American stock market: dataset and benchmarks. *Future Internet* 2022; 14: 244.
30. Amini S, Elmore R, Öztekin Ö, et al. Can machines learn capital structure dynamics? *J Corp Finance* 2021; 70, 102073.
31. Harker C, Hassall M, Lant P, et al. What can machine learning teach us about Australian climate risk disclosures? *Sustainability* 2022; 14: 10000.
32. Nguyen Q, Diaz-Rainey I and Kuruppuarachchi D. Predicting corporate carbon footprints for climate finance risk analyses: a machine learning approach. *Energy Econ* 2021; 95:105129.
33. Meeampol S, Lerskullawat P, Wongsorntham A, et al. Applying Emerging Market Z-Score model to predict bankruptcy: A case study of listed companies in the Stock Exchange of Thailand (SET). *Human Capital Without Borders: Knowledge and Learning for Quality of Life; Proceedings of the Management, Knowledge and Learning International Conference* 2014: 1227–1237. https://ideas.repec.org/h/tkp/mklp14/1227-1237.html
34. Altman N and Krzywinski M. Ensemble methods: bagging and random forests. *Nat Methods* 2017; 14: 933–934.
35. Breiman L. Bagging predictors. *Mach Learn*. 1996; 24: 123–140.
36. Breiman L. Random forests. *Mach Learn*. 2001; 45: 5–32.
37. Lee T, Ullah A and Wang R. Bootstrap aggregating and random forest. In: *Advanced studies in theoretical and applied econometrics*. Springer Cham, 2019, pp.389–429.
38. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat* 2001; 29: 1189–1232.
39. Freedman M and Jaggi B. Global warming, commitment to the Kyoto protocol, and accounting disclosures by the largest global public firms from polluting industries. *Int J Account* 2005; 40: 215–232.
40. Prado-Lorenzo J, Rodríguez-Domínguez L, Gallego-Álvarez I and García-Sánchez I. Factors influencing the disclosure of greenhouse gas emissions in companies world-wide. *Manag Decis* 2009; 47: 1133–1157.
41. Rankin M, Windsor C and Wahyuni D. An investigation of voluntary corporate greenhouse gas emissions reporting in a market Governance System: Australian evidence. *Soc Sci Res Network* 24(8): 1037–1070. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1961820. Published online November 1, 2011.

Author biographies

Jeong Hwan Lee is an associate professor at the Department of Economics and Finance in Hanyang University. His research area covers ESG, green finance and financial intermediaries.

Jin Hyung Cho holds a PhD in Economics. His area of research is ESG, corporate finance and artificial intelligence (AI).

Bong Jun Kim is a MS student in data science. His research interests are ESG and machine learning.

Won Eung Lee is an undergraduate student majoring in Economics. His area of research is ETF, ESG and momentum strategy.