Journal of Cleaner Production 344 (2022) 130979

Contents lists available at ScienceDirect

Journal of Cleaner Production

journal homepage: www.elsevier.com/locate/jclepro

Prediction of environmental controversies and development of a corporate environmental performance rating methodology

Jan Svanberg [a,*], Tohid Ardeshiri [a], Isak Samsten [b], Peter Öhman [c], Tarek Rana [d], Mats Danielson [b]

[a] University of Gävle Mid Sweden University Centre for Research on Economic Relations, SE-801 76, Gävle, Sweden
[b] Stockholm University, Department of Computer and Systems Sciences, SE-164 07, Kista, Sweden
[c] Mid Sweden University, Department of Economics, Geography Law and Tourism, Centre for Research on Economic Relations, SE-851 70, Sundsvall, Sweden
[d] The Royal Melbourne Institute of Technology, School of Accounting, Information Systems & Supply Chain, RMIT University, GPO Box 2476, Melbourne, VIC, 3001, Australia

ARTICLE INFO

Handling editor: Tomas B. Ramos

Keywords:
Corporate environmental performance
Environmental controversies
ESG
Machine learning
Prediction
Socially responsible investing

ABSTRACT

Institutional investors seek to make environmentally sustainable investments using environment, social, governance (ESG) ratings. Current ESG ratings have limited validity because they are based on idiosyncratic scores derived using subjective, discretionary methodologies. We discuss a new direction for developing corporate environmental performance (CEP) ratings and propose a solution to the limited validity problem by anchoring such ratings in environmental controversies. The study uses a novel machine learning approach to make the ratings more comprehensive and transparent, based on a set of algorithmic approaches that handle nonlinearity when aggregating ESG indicators. This approach minimizes the rater subjectivity and preferences inherent in traditional ESG indicators. The findings indicate that controversies as proxies for non-compliance with environmental responsibilities can be predicted well. We conclude that environmental performance ratings developed using our machine learning framework offer predictive validity consistent with institutional investors' demand for socially responsible investment screening.

1. Introduction

This paper addresses the problems associated with the validity of environment, social, governance (ESG) ratings, focusing on the environmental component referred to as corporate environmental performance (CEP). The rationale for the study is that socially responsible investment (SRI) principles can be used to transform society through institutional investors' influence on portfolio companies. The paper presents a conceptual problematization of the validity of current ESG ratings and three methodological problematizations of the main rating systems.

Environmental responsibilities have a substantial impact on companies' conduct (Heath and Gifford, 2002; Schultz et al., 2014). In parallel, the idea of CEP is attracting attention, as compliance with environmental responsibilities (Wood, 2010) is recognized as important for stakeholders. Institutional investors (e.g., banks, insurance companies, and pension funds) measure CEP with ESG ratings and exert substantial influence on companies through their investment/divestment campaigns, which affect companies' access to and cost of capital. In 2018, USD 12 trillion (about 25% of capital in the USA) and EUR 22 trillion (about 50% of capital in Europe) were considered SRI (Christensen and Clawson, 2018). Divestment campaigns based on environmental issues could therefore trigger a decrease in stock value of 8–10%. Simultaneously, SRI practices could cause a 5% reduction in greenhouse gas (GHG) emissions (Choi et al., 2020) through their influence on high-emission companies. However, institutional investors have difficulties selecting SRI portfolios because ESG ratings, with their limited validity, are inadequate measures of CEP. Regarding convergent validity, leading ratings (i.e., Refinitiv, MSCI, and Sustainalytics) differ by up to 30% (Chatterji et al., 2016). In addition, current traditional ESG rating methodologies presuppose idiosyncratic preferences or beliefs (Callan and Thomas, 2009), producing discretionary and disparate ratings. Considerable unexplainable differences between measurements of any scientific construct are inconsistent with the scientific requirement

* Corresponding author.
E-mail addresses: Jan.Svanberg@hig.se (J. Svanberg), tohid.ardeshiri@hig.se (T. Ardeshiri), isak.samsten@dsv.su.se (I. Samsten), Peter.Ohman@miun.se (P. Öhman), tarek.rana@rmit.edu.au (T. Rana), mats.danielson@su.se (M. Danielson).

https://doi.org/10.1016/j.jclepro.2022.130979
Received 1 June 2021; Received in revised form 11 February 2022; Accepted 13 February 2022
Available online 15 February 2022

that measurements be reproducible.

Chatterji et al. (2016) called for research to develop ESG ratings with non-subjective weighting schemes and predictive validity. The present study addresses this call by demonstrating the basic principles by which machine learning can be used in developing CEP ratings with predictive validity. We discuss a new direction for developing CEP ratings and propose a solution to the ESG rating validity problem by developing a machine learning framework for CEP anchored in environmental controversies. We contribute to critical investigations of ESG ratings by discussing how subjective indicator weights lead to skewed ESG ratings and how linear modelling further constrains traditional rating methodologies. We demonstrate how machine learning can advance the development of comprehensive and transparent CEP ratings. For that matter, our study introduces a set of algorithmic approaches to handling nonlinearity when aggregating CEP indicators. One contribution is to show that controversies can be predicted using machine learning. The evidence regarding the predictive performance of our methodology suggests that CEP ratings can be computed as the likelihood of comprehensive compliance with environmental responsibilities. Although rating issues are not new, to our knowledge, no previous study has investigated the predictive validity of CEP ratings.

The paper continues as follows: the next section presents the frame of reference, followed by a section on the data and research design; thereafter, the results section is followed by the discussion and conclusion section.

2. Frame of reference

2.1. ESG-type CEP ratings

There are theoretical arguments in favour of screening investments for CEP. SRI portfolio managers may benefit from choosing from a pool of equities superior to the overall market and therefore more likely to produce favourable financial performance over time (Oikonomou et al., 2018). Stakeholder theory supports this claim by stating that the better a company manages its relationships with stakeholders, the better its financial performance will be over time (Donaldson and Preston, 1995; Freeman, 2015).

Empirically, a large body of research indicates only a slightly positive effect of CEP screening on financial performance. However, the evidence for this is inconsistent and it is also possible to reconcile these results with the view that investing according to SRI is associated with lower returns (Cornell, 2020). The lack of evidence regarding the financial performance of SRI portfolios versus other portfolios suggests that the ESG ratings with which such portfolios are selected are questionable.

The validity issues caused by raters' prolific use of discretionary weighting schemes have been empirically confirmed (Chatterji et al., 2009, 2016; Christensen et al., 2022). The inaccuracy revealed by discretionary disclosures of raw ESG data (Orlitzky, 2013) and inconsistent measurements of ESG features by raters (Berg et al., 2019) constrain the validity of ratings. However, the main issue with rating validity is not data availability. Christensen et al. (2022) found that as data availability increases from the least ESG-disclosing quartile to the most ESG-disclosing quartile of companies, differences between ratings for the same company increase by 30%. Their findings indicate that the main problem with ESG ratings is the weights with which raters aggregate ESG indicators (e.g., CO_2 emissions) to form holistic metrics. There is no method for determining the relative importance of the many disparate features of an ESG rating, and leading raters cannot even agree on the three most important categories (Berg et al., 2019). This means that a rater must arbitrarily decide, for example, how important hazardous waste should be compared with direct CO_2 emissions for CEP ratings. The lack of answers to such questions makes CEP rating validity questionable. Popular rating methodologies create arithmetic averages of CEP indicators from discretionarily weighted indicators (Chen and Delmas, 2011; Delmas and Blass, 2010). Semenova and Hassel (2015) examined three CEP ratings and found that, while the components of these ratings are correlated, the aggregated CEP ratings are not.

We have identified several limitations of ESG rating methodologies. First, the arithmetic averaging of indicators precludes the capturing of heterogeneous features, for which a multidimensional construct is necessary (Mattingly and Berman, 2006). As Delmas et al. (2013) and Semenova and Hassel (2015) observed, aggregated CEP ratings' neglect of factor structure distorts the information contained in indicators. Second, ad hoc weighting assumes that the importance of features can be assessed by the rater, which is not true (Callan and Thomas, 2009). For example, CEP ratings may overestimate the meaning of programmes and policies at the expense of substantial outcomes (Delmas and Burbano, 2011). Ad hoc weighting may dilute and distort the impact of the most substantial outcome indicators, such as GHG emissions, water discharged, total energy use, fleet fuel consumption, total waste, and pollution, because it does not rationalize the actual weight differences. Third, an arithmetic average of indicators does not capture nonlinearity (Ding et al., 2020). There may be nonlinearity in the relationships between individual CEP features and total CEP, if not because of the behaviour of individual features, then due to interaction between them. For example, the importance of a GHG emission reduction policy is more significant the more such emissions a company produces. Furthermore, if CEP is defined as compliance with the responsibilities defined by society, CEP will display a threshold form at the indicator level because, for example, emissions are allowed up to a certain level beyond which they are illegal or immoral, making CEP drop sharply for a company that crosses this line. Traditional CEP ratings account for none of this nonlinearity.

If CEP ratings have questionable validity, they should be irrelevant to institutional investors, and there are indications that investors are increasingly dissatisfied with such ratings the more they "look under the hood" (Wong and Petroy, 2020). Institutional investors have selective preferences regarding CEP for information not covered by ESG-type ratings (Nofsinger et al., 2019). They are indifferent to whether companies have CEP features not required by environmental responsibilities, but underweight stocks in companies with CEP features that indicate non-compliance with these responsibilities. Traditional CEP ratings do not distinguish between, for example, emissions within acceptable limits and illegal/immoral emission levels because such ratings are linear functions of indicators (Berg et al., 2019). The lack of discrimination between compulsory and voluntary aspects of CEP in traditional rating methodologies makes such ratings problematic.

Nofsinger et al. (2019) argued that this asymmetry is accentuated for longer-term investors, and that the potential benefits of performing beyond what is required or of features not required are offset by the costs of accomplishing this. The long-term financial damages caused by companies' non-compliance with environmental responsibilities, however, are so high that they cannot be offset by any cost savings. According to Nofsinger et al. (2019), such costs stem from serious controversies when, for example, customers boycott the company, employees strike, or there is strong public opposition to the company's environmental impact, forcing the company to withdraw from investments. Institutional investors therefore pay close attention to such severe non-compliance with responsibilities that may cause environmental controversies, but very little, if any, attention to CEP features that do not relate to these financially material CEP risks.

A similar consideration is that traditional CEP ratings do not relate to the primary motivations for integrating environmental risk assessments into the investment process. In a survey of institutional investors, Krueger et al. (2020) found that the strongest motivation to use a measure of CEP is to protect the investor's reputation, which is most effectively accomplished by avoiding environmental controversy-prone portfolio companies. Institutional investors prefer CEP ratings that rate companies according to their ability to comply with environmental standards, because the only way a company can avoid controversies with certainty is to avoid violations of environmental responsibilities.

Investors' interest in environmental responsibility compliance and the avoidance of controversy-prone companies is, according to such theoretical arguments (Amel Zadeh and Serafeim, 2018), driven by the assessment that such a CEP construct has financial materiality.

2.2. Environmental controversies as indicators of CEP

Considering the methodological weaknesses of CEP ratings, institutional investors need a rating that distinguishes one company's overall CEP from another's. Furthermore, institutional investors would not benefit from an idiosyncratic CEP rating that captures the preferences of a particular rater rather than how companies perform according to society's preferences. There are two main reasons for this. First, if the owners of assets in the care of an institutional SRI investor found that investments were made in low-CEP companies, these owners would be dissatisfied with how the investor fulfilled the role of a responsible investor and would withdraw their assets from the fund (Grappi et al., 2013). Such events would damage the investor's reputation, suggesting considerable incentives to use CEP screening that coincides with the expectations of the investor's beneficiaries. This practice is confirmed by evidence of the information preferences of institutional investors (Nofsinger et al., 2019). Second, portfolio companies that do not comply with environmental responsibilities are more likely to be the subject of environmental controversies, to which financial markets would react by punishing non-compliant companies. Institutional investors appreciate neither the blame nor the financial repercussions of owning such portfolio companies (Krueger et al., 2020). Thus, institutional investors benefit from a holistic CEP rating that assesses the extent to which companies comply with environmental responsibilities.

Based on these observations reported in previous literature, we want to develop a rating methodology that scores companies based on a measure of CEP that captures their comprehensive compliance with environmental responsibilities. The most suitable construct to use as a holistic proxy for this is environmental controversies. The instances of such controversies are indicators of non-compliance with environmental responsibilities because controversies result from companies being identified as breaching such requirements (Faulkner, 2011; Nieri and Giuliani, 2018). We suggest that there is a distinction between the description of environmental features (i.e., behaviour and structures), which we call CEP indicators and which are the focus of traditional CEP ratings, and the assessment of these features relative to compliance or non-compliance with responsibilities. This is the focus of our novel rating methodology and for which environmental controversies provide data labels. Consistent with institutional investors' information preferences, we emphasize that CEP should be defined as performance relative to environmental responsibilities because it is companies' behaviour relative to their responsibilities that confers investor-relevant meaning on the CEP construct. We thus view controversies as a marker signalling that such performance is considered environmentally illegitimate from an ethical, regulatory, or legal point of view.

From this perspective, controversies provide information different from that provided by environmental behaviour indicators (e.g., waste recycling and hazardous waste production). Controversies are adverse reactions to environmental behaviours designating them as inappropriate from the perspective of societal responsibilities (e.g., inappropriate or illegal waste management). Controversies thus convey information about a particular instance of inappropriate CEP. They also indicate that there may be problems with the transgressor's structures and processes or its lack thereof (Nieri and Giuliani, 2018), because controversies seldom result from unintended accidents but more often from systematic breaches of moral and legal responsibilities in the pursuit of profit (Fiaschi et al., 2017, 2020).

3. Data and research design

3.1. ESG-type CEP ratings as predictors of environmental controversies

Before describing the research design, we describe data and discuss how CEP ratings, exemplified by the environmental pillar of the Refinitiv ESG rating, perform as evidence for predicting environmental controversies. The Refinitiv ratings were one of the three rating systems investigated in a study of ESG rating validity by Christensen et al. (2022). Refinitiv is the second largest provider of finance terminals to institutional investors such as large banks and insurance companies, just behind the market leader Bloomberg. Refinitiv (2019) is therefore a good reference point with which we can compare our proposed rating methodology.

We collected 112 indicators classified by us as environmental data and consistent with Refinitiv's classification (the complete list of indicators is provided in the Appendix). In addition to the environmental data, we included descriptors of the features total assets, net assets, market capital, return on assets, Global Industry Classification Standard (GICS) sector, and country of headquarters. The dataset covers 2517 companies, of which approximately 13% had at least one environmental controversy over the 10-year window. Table 1 provides an overview of the included companies.

We examine the extent to which the Refinitiv CEP ratings could serve as a predictor of the risk of incurring an environmental controversy, because predicting such controversies would indicate predictive validity. We first examine the histogram in Fig. 1a, which shows the distribution of companies with and without a controversy across CEP ratings. No clear-cut division of the two distributions is indicated. That means they cannot be partitioned by any CEP rating number, so that a substantial fraction of the companies without controversies is singled out from those with at least one controversy. The same message is revealed in Fig. 1b, which shows company-years. No CEP rating can be found that separates the blue from the yellow distributions.

Although the histograms do not consider that there may be a time lag between ratings and controversies, Fig. 1a and b indicate that the plotted CEP ratings are an inappropriate predictor of environmental controversies. There is even a slight skewness in the distributions, suggesting that companies with high CEP ratings are more likely than other companies to have a controversy, contrary to the ratings.

3.2. Research design and machine learning algorithms

Our anticipation of CEP's heterogeneity, nonlinearity, and multidimensionality suggests that an estimation method with a high capacity for capturing complexity and nonlinearity should be used. For these reasons, we find machine learning and predictive modelling more favourable for assessing the extent to which companies comply with environmental responsibilities than the conventional statistical methods used in explanatory modelling (Shmueli, 2010). Machine learning allows the prediction of CEP when distributions are unknown and when the estimated relationships are severely nonlinear (Duda et al., 2001). This methodology has been used in similar areas, for example, in finance for credit scoring (Cleofas-Sánchez et al., 2016) and bankruptcy prediction (Gerlein et al., 2016), which are applications having features in common with our problem.

In nine experiments with separate machine learning algorithms, the algorithms' abilities to predict environmental controversies were evaluated according to five different performance measures, i.e., precision, recall, *f*-measure, area under the receiver operating characteristic (ROC) curve, and the precision-recall curve (PRC). We adhere to established methodological practices in computer science for the design and execution of the experiments, which are described step by step in section 3.3.

We sort the companies into two (disjoint) categories based on whether they have been involved in a controversy during the past ten

Table 1
The companies included in the dataset described by the sector, percentage of companies in each sector included, average total assets per company, and number of controversies (the table reveals the sector distribution of companies and describes the distribution of controversies for companies that have had at least one).

	Companies,	Controversies,	Companies with	Controversies, when $n > 0$					Assets in USD	Assets in USD billions per company				
GICS Sector	*n*	*n*	controversies, *n*	Min	Max	Mean	Std. dev.	Skewness	Kurtosis	billions	Min	Max	Mean	Std. dev.	Skewness	Kurtosis
Energy	183	130	59	1	7	2.2	1.69	1.47	1.19	4581	0.07	291	25	51	3	12
Materials	277	96	48	1	7	2	1.43	1.83	3.2	2703	0.04	96	9	15	3	14
Utilities	124	59	38	1	4	1.55	0.89	1.53	1.37	3494	0.13	259	28	36	3	15
Consumer discretionary	297	36	21	1	4	1.71	0.9	1.09	0.4	4153	0.11	331	13	35	6	46
Industrials	420	19	18	1	2	1.06	0.24	4.24	18	5770	0.25	457	13	35	10	122
Consumer staples	156	16	12	1	2	1.33	0.49	0.81	−1.65	2377	0.29	160	15	23	3	15
Health care	141	7	6	1	2	1.17	0.41	2.45	6	2257	0.05	144	16	24	2	7
Financials	398	5	5	1	1	1	0	0	0	79534	0.24	2509	199	393	3	11
Information technology	201	2	2	1	1	1	0	0	0	2387	0.16	182	11	24	4	24
Communication services	161	1	1	1	1	1	0	0	0	3415	0.06	284	21	40	3	14
Real estate	155	1	1	1	1	1	0	0	0	1667	0.42	78	10	12	2	7

years and denote the companies involved in a controversy as positive cases and those not involved as negative cases. To capture the longitudinal aspect of the dataset, we employ a simple strategy in which the indicators are averaged if numerical (e.g., CO_2 emission) or encoded using dummy variables if binary (i.e., one per year). Our goal is not to model a particular company's risk of a controversy given past indicators; instead, our goal is to capture the attribute interactions that describe a company at high risk of being involved in a controversy.

The nine machine learning algorithms described below capture different aspects of the learning problem, such as linearity and nonlinearity. Further techniques for estimating CEP that might have been considered include deep learning or recurrent neural networks. However, for the purpose of examining the possibilities of using machine learning for predicting the likelihood of incurring an environmental controversy, we consider our selection of methods adequate to provide insights into the extent to which interactions and nonlinearity are necessary elements of the estimation tools.

Nearest neighbour is an instance or distance-based classifier that relies on a distance or similarity measure to predict controversies. In contrast to other machine learning algorithms, the nearest neighbour algorithm does not have a training phase. Instead, all companies and their CEP indicators and controversies are recorded. Then, a prediction is performed by querying the database and finding the k closest companies, in terms of CEP indicators, with the output being the probability of controversy as a fraction of the k closest companies' controversy status. Although any distance or similarity measure can be used, we employ the simple Euclidean distance (Hu et al., 2016).

This study uses CEP estimators separating low from high CEP using specified indicators. In addition, it employs the *linear* support *vector machine (SVM)* for analysis, separating data using an $(n - 1)$-dimensional plane. This linear SVM, as employed by Vapnik (1995), has been effective in identifying pattern recognition and prediction problems, for example, stock and bankruptcy prediction, in which the data are linearly separable (Xu et al., 2009).

When required, the *radial basis function (RBF) SVM* can be employed if the classification is not linear. The RBF is a kernel function of SVM used to enable nonlinear classification. It is a form of kernel serving as a window for mapping the nonlinearity in the n-dimensional original space onto a higher-order space in which the classifier can be linear.

Random forest, created by Breiman (2001), is one of the most widely applied machine learning algorithms and is often considered the current state of the art in many domains. The algorithm constructs an ensemble model by creating decision trees trained using (random) samples of training data (i.e., companies). We construct each tree by sampling a limited number of CEP indicators at each node to increase the variability and predictive performance. The final controversial prediction of the model is a majority vote among the trees. Random forest works well with outliers and noise in the training set (Yeh et al., 2014), which are good features of a prediction model for environmental controversies, which can be expected to be noisy due to greenwashing and the lack of internationally binding reporting standards. Another benefit of random forest is that it calculates the importance of each indicator for the classification results (Maione et al., 2016). Using the majority of votes among the trees avoids data overfitting and provides precise forecasts (Breiman, 2001).

Logistic regression, which is the same method used in statistics, classifies data by estimating the coefficients of a regression equation. Logistic regression relates environmental controversies to the CEP indicators, and its goal is to find the best-fit regression coefficients (Ding et al., 2020). In this classifier, each feature is multiplied by a weight and then all features are added together; the result is passed on to a sigmoid function, which produces the binary output. Logistic regression generates the coefficients with which to predict a logit transformation of the probability. It is named after the function used at the method's core, i.e., the logistic function.

Artificial neural networks comprise multiple stacked layers, with an

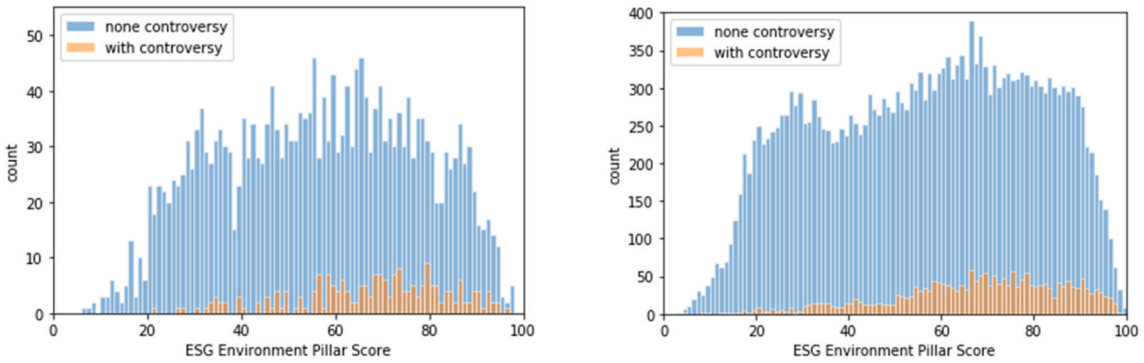

Fig. 1. (a) The distribution of environmental controversies over companies and Refinitiv CEP ratings. Companies with at least one controversy in the ten-year window are classified as "with controversy" on the *y*-axis and companies with no controversies are classified as "no controversy". The *x*-axis shows the CEP ratings. (b) The distribution of environmental controversies over company-years and Refinitiv CEP ratings. Company-years with at least one controversy in the ten-year window are classified as "with controversy" on the *y*-axis and company-years with no controversies are classified as "no controversy". The *x*-axis shows the Refinitiv CEP ratings.

output layer consisting of a logistic, softmax, or linear regression model. By organizing the model in layers with an activation function between each one, it is possible to show that the artificial neural network can approximate any linear or nonlinear function. Thus, an artificial neural network can relate CEP indicators to controversies using multiple layers with a logistic regression output layer. Artificial neural networks consist of many simple computational elements called artificial neurons, which are densely interconnected and operate in parallel. They are robust in that they can cope with noisy data, have good generalization capabilities, learn highly nonlinear relationships, and make no distribution assumptions about the training data. The neurons are connected by corresponding links between layers with numeric weights – the fundamental means of long-term memory in the artificial neural network (Ekonomou, 2010). The advantages of artificial neural networks over traditional statistical methods such as regression have been well substantiated in the literature. Artificial neural networks are known to produce better forecasting results than those obtained using statistical regression models (Rumelhart et al., 1994) and are specifically suited to finding solutions for complicated problems with fuzzy data (Ghritlahre and Prasad, 2018). Furthermore, artificial neural networks are naturally nonlinear nonparametric models that also cope with unknown interactions (Azadeh et al., 2011; Wong et al., 2010).

Gradient boosting is an ensemble approach that combines multiple weak models into a robust ensemble model by reweighting the training data on CEP indicators to focus the learning on those companies that the algorithm cannot predict correctly. It defines a loss function and uses the gradient of the loss function to reweight the companies to focus on the misclassifications using the logistic loss. Gradient boosting is one of the strongest classifiers for various tasks (Sigrist and Hirnschall, 2019).

Naïve Bayes classifiers are a family of simple probabilistic classifiers developed from Bayes' theorem. The method refers to a family of algorithms based on a shared principle: all naïve Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. Despite this simplifying assumption, Naïve Bayes classifiers can be trained effectively in a supervised learning setting. The Naïve Bayes classifier computes the conditional probability of a controversy and a non-controversy given a set of CEP indicators. Under the (naïve) assumption that the CEP indicators are independent, the classifier can be expressed as the conditional probability of a controversy multiplied by the product of the conditional probability of each CEP indicator, given a controversy (Gulo et al., 2015).

Quadratic discriminant analysis is a machine learning algorithm that generates a model based on the conditional densities of the data and constructs a quadratic decision boundary. As opposed to linear discriminant analysis, quadratic discriminant analysis does not assume that the covariance of each class is identical. As such, the algorithm can capture nonlinear dependencies between the governance controversies and the CEP indicators (cf. Yuan et al., 2017).

3.3. Experiments

To ensure the reproducibility of our study, we provide the hyper-parameters of each algorithm in Table 2. The default values for SciKit-learn version 0.22 (Pedregosa et al., 2011) have been used for the remainder of the hyper-parameters.

A previously unseen set of test instances (i.e., the CEP indicators for a company) should be employed to evaluate the predictive performance (Alpaydin, 2010). A simple and common approach to evaluate the predictive performance is to partition the dataset into two parts, using the first part (i.e., the training set) for training the machine learning algorithm and the second part (i.e., the test set) for evaluating its performance on independent test data. However, if data are scarce or a more reliable estimate of the generalization performance of the algorithm is sought, a different method is appropriate. One way to deal with this is to employ *k*-fold cross-validation. This cross-validation is an approach to reliably estimate the performance of a learning algorithm that partitions the dataset into *k* disjoint partitions (called folds) and trains the learning model iteratively on $k - 1$ folds leaving one fold for testing. The procedure results in *k* performance measures, and their mean represents a reliable estimate of the generalization performance. Although there are many approaches for selecting the number of folds, *k*, the most common value employed in this study is 10, resulting in 10-fold cross-validation (Guyon, 1997). In this study, we employ stratified *k*-fold

Table 2
Hyper-parameters of the machine learning algorithms.

Classifier	Notes
Nearest neighbour	Three nearest neighbours
Linear support vector machine (SVM)	Linear kernel with C = 0.025
Radial basis function (RBF) support vector machine (SVM)	RBS kernel with C = 0.025
Random forest	100 trees
Logistic regression	Ridge regularization with C = 1
Artificial neural network	Four hidden layers of size 100 using the RELU activation function
Gradient boosting	Learning rate of 0.1
Naïve Bayes	No hyper-parameters
Quadratic discriminant analysis	No hyper-parameters

cross-validation to ensure an equal number of positive and negative cases in each test set.

The performance of machine learning models can be estimated using various measures, each measuring a different aspect of the learning ability (Alpaydin, 2010). As mentioned, we use five commonly employed performance measures, i.e., precision, recall, *f*-measure, the area under the ROC curve, and PRC. To evaluate and compare the performances of the machine learning algorithms, we first calculate the precision and recall for the estimators. Precision and recall are defined using true positive (TP), false positive (FP), true negative (TN), and false negative (FN) as shown in equations 1–5 (see Table 3), with precision being the fraction of true positives relative to the total number of positive case predictions and recall being the fraction of true-positive predictions relative to all positive cases in the data. Precision then represents the number of times a dart player's arrow hits the target relative to the number of attempts, and recall represents the ability of the predictor to identify as large a fraction as possible of all the controversy companies in our data.

Precision measures the sensitivity of the classifier, i.e., its accuracy in predicting the controversy and non-controversy classes expressed as the correct positive fraction divided by the total number of positive predictions (Eq. 1). Since precision and recall are conflicting measures – for example, a classifier that predicts every company as having a controversy has a recall positive of 100% – the *f*-measure (Eq. 3) captures the trade-off between the two measures.

The measure area under the ROC curve is the area under the curve defined as a plot of true positives (Eq. 4) versus false positives (Eq. 5). Area under the ROC curve estimates the probability of a classifier ranking a true-positive instance ahead of a false-positive instance, and is thus a measure of its ranking performance. Similarly, the PRC estimates the mean precision for multiple thresholds of recall and is, like the *f*-measure, used to measure the trade-off between precision and recall. The area under the PRC is defined as the area under the plot of precision (Eq. 1) versus the recall (Eq. 2). The main benefit of the area under the ROC curve and the PRC is that they are insensitive to the class distribution of the training and testing data, as opposed to the accuracy.

4. Results

4.1. Predictive performance of different learning approaches

The experiments were conducted to investigate the potential to predict controversies from environmental behaviour indicators using machine learning algorithms and to examine the possibility of developing CEP ratings based on the prediction models. Here, we present the results in terms of the static measures precision, recall, and *f*-measure and the dynamic measures area under the ROC curve and area under PRC for the nine algorithms. Furthermore, CEP ratings were generated by the algorithms in an effort to establish the correlation between our machine-learning–based CEP ratings and the Refinitiv CEP ratings.

The predictive performance of the learning algorithms is presented in Table 4. The performance indicators demonstrate that, for most of the measures, the random forest, artificial neural network, and gradient boosting algorithms outperform or perform in the high end compared with the other models in predicting controversies. The models have significantly higher precision than do most of the other approaches, and random forest leads with a value of 0.70. Random forest typically produces models with uncalibrated probability estimates, meaning that the model has high confidence in a particular prediction. Notably, the three models with the highest precision also have the lowest recall. Thus, there is a trade-off between these measures and between the information types indicated by the precision and recall columns.

Table 3
Basic performance measures.

Measures of performance	Equation
$Precision = \frac{TP}{TP + FP}$	(Eq. 1)
$Recall = \frac{TP}{TP + FN}$	(Eq. 2)
$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$	(Eq. 3)
$True_{positive} = \frac{TP}{TP + FN}$	(Eq. 4)
$False_{positive} = \frac{FP}{FP + TN}$	(Eq. 5)

Table 4
Prediction results for environmental controversies using the nine machine learning algorithms (the highest value in each column is shown in bold).

	Precision	Recall	*F*-measure	ROC curve	PRC
Nearest neighbour	0.5438	0.2942	0.3769	0.7132	0.2929
Linear SVM	0.3026	**0.7630**	**0.4326**	0.8773	0.4594
RBF SVM	0.2526	0.7866	0.3820	0.8609	0.4380
Random forest	**0.6994**	0.1942	0.2990	0.8849	**0.5090**
Logistic regression	0.3045	0.7253	0.4282	0.8727	0.4814
Artificial neural network	0.6275	0.2039	0.3010	0.8761	0.4866
Gradient boosting	0.5731	0.3130	0.3993	**0.8889**	0.4839
Naïve Bayes	0.2759	0.6394	0.3847	0.7926	0.3160
Quadratic discriminant analysis	0.1734	0.3184	0.2133	0.5717	0.1770

In Table 4, the results are computed using 10-fold cross-validation to ensure that the same company is not used for both training and testing purposes. To ensure comparability of results for all algorithms, identical training and testing partitions are used, meaning that identical companies were included in each training and testing fold for all algorithms.

To evaluate the ranking and predictive performance of the algorithms, we computed the curves mentioned earlier (see Figs. 2 and 3, respectively). Fig. 2 shows the ROC for each algorithm and class. An ROC curve closer to the top-left corner is better, and closer to the diagonal is worse. The experiments reveal that the results in Table 4 and Fig. 2 are consistent, with random forest, artificial neural network, and gradient boosting having strong performance compared with the other models.

The difference between the machine learning algorithms in terms of precision is significant for an investment application because investing is, per definition, the high-precision selection of individual stocks, and it is not necessary to have a strong opinion about all stocks. The three above algorithms, having a high capacity to represent complexity, perform well for both non-controversies and controversies. Linear SVM and logistic regression perform well in terms of the ROC curve but have lower precision than do the best models.

Fig. 3 presents the PRC for all algorithms and classes. The curve should be close to the top-right corner of the graph, and as far away as possible from the bottom-left corner. The graphs reveal that random forest, gradient boosting, artificial neural network, and logistic regression are the best-performing models. In particular, random forest and gradient boosting consider nonlinear attribute dependencies and perform well for the controversial cases (the yellow line); nearest neighbour, naïve Bayes, and quadratic discriminant analysis perform significantly worse.

The graphs in Figs. 2 and 3 confirm that logistic regression is a simple model that performs well as a predictor, although it does not provide the high precision that would be desirable.

4.2. Controversy prediction as CEP rating

To investigate the use of the controversy prediction to produce CEP ratings, we first examine the correlation between the number of controversies a company has during the 10-year window and its CEP rating. CEP ratings should, on average, have negative correlations with

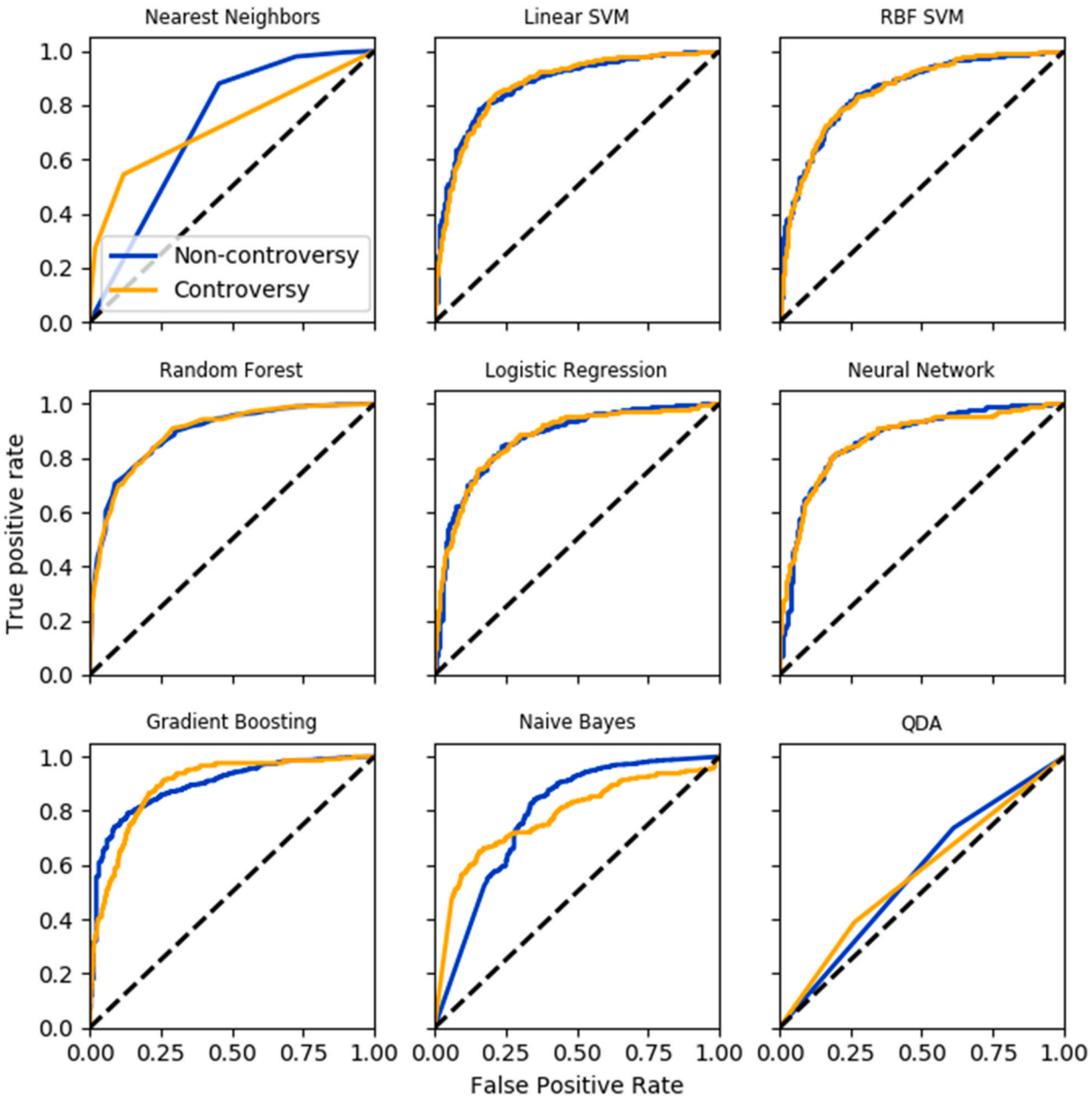

Fig. 2. Area under the ROC curve for the nine learning algorithms. The blue line represents the ROC for predicting non-controversy and the yellow line represents the ROC for predicting controversy. Note that classifiers that produce ROC curves that lie above the dashed line provide predictions that are better than random guessing. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

controversies because, as a whole, the controversies signal societal reactions to companies' non-compliance with environmental responsibilities (cf. Faulkner, 2011). Fig. 4 presents negative correlations between the machine-learning–based CEP ratings and the number of controversies for each company. The *x*-axis shows the number of controversies per company, and the *y*-axis the machine-learning–based CEP ratings. As expected, the slopes in the graphs are all downwards and there are some differences between the distributions of the ratings, as shown in the graphs. In contrast to the Refinitiv ratings, which did not assign controversy-prone companies lower ratings than those of non-controversy companies (Fig. 1), the nine models all punish companies by assigning lower ratings the more controversies they have. Since controversies likely identify systematic flaws in the companies' ability to comply with environmental responsibilities, validity is confirmed if the ratings of "repeat offenders" are lower than those of companies with fewer or no controversies.

Fig. 4 shows that our CEP ratings penalize companies with more controversies by assigning them lower scores. There is a negative correlation, meaning that lower CEP ratings indicate more controversies. The nearest neighbour distribution is inappropriate for a CEP rating methodology, because it should not be discrete with only a few levels. We adopted the machine learning algorithms for classification, meaning that the output is a class membership or a probability estimate of class membership. For nearest neighbour, a company is classified by a plurality vote of its neighbours. The model assigns a company to the class most common among its nearest neighbours. The peculiar structure of the output is a product of grouping companies into classes based on the model's definition of neighbourhoods. Companies with similar structures of their CEP indicators and similar numbers of controversies tend to be grouped together. Although the model produces a distribution with lower CEP ratings the more controversies companies have, the graph shows that the model is inappropriate for this prediction task. Naïve

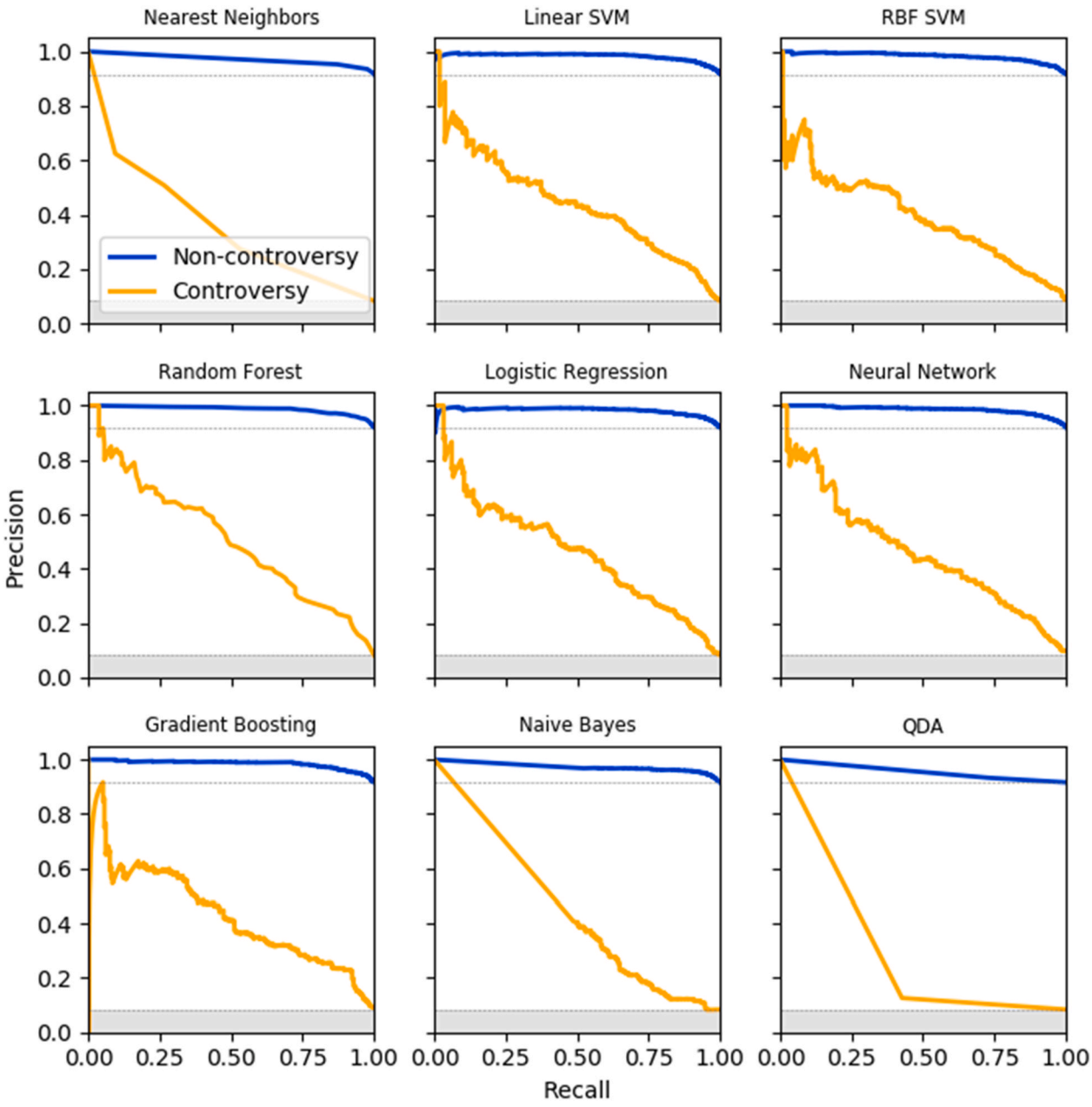

Fig. 3. PRCs for the nine learning algorithms. The blue line represents the PRC for predicting non-controversy and the yellow line represents the PRC for predicting controversy. The curves show the precision of a classifier as the recall increases. The top region (defined by the dashed grey line) shows the region in which a classifier performs better than random guessing for the non-controversy cases, and the region between the bottom and top regions shows where a classifier performs better than random guessing for the controversy cases. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Bayes and quadratic discriminant analysis likewise produce somewhat counterintuitive distributions attributable to their ways of functioning. Thus, although all algorithms should theoretically be able to use the differences between CEP indicator patterns to find differences in the likelihoods of controversies in corporations, the algorithms mentioned above do not appear to produce a continuous distribution of CEP ratings across the number of controversies.

The algorithms with the highest precision (i.e., random forest, artificial neural network, and gradient boosting) all produce a smooth distribution of companies over the two-dimensional space. For example, random forest generates a cautious rating with comparatively few companies under 40, but has a sharp increase at 50. This cautious rating of companies as low-performing indicates that this method has high precision in identifying the worst CEP companies.

The difference between a machine-learning–based CEP rating methodology, Refinitiv-type ratings, and a controversy index is illustrated by the random forest rating of several companies in the 60–100 region, although some of them have controversies in the 5–15 range, and by its rating of several companies in the 20–40 range. Although these companies have had no controversies, they are rated low because they have indicator patterns similar to those typically associated with controversies. The ratings are thus based on the knowledge extracted from the association between CEP indicator patterns and the frequency of environmental controversies. However, discretionary assessment of CEP indicators with no regard for the likelihood of controversies produces traditional ratings like those of Refinitiv, while actually occurring controversies are related to a wrongdoing index (cf. Fiaschi et al., 2020; Nieri and Giuliani, 2018).

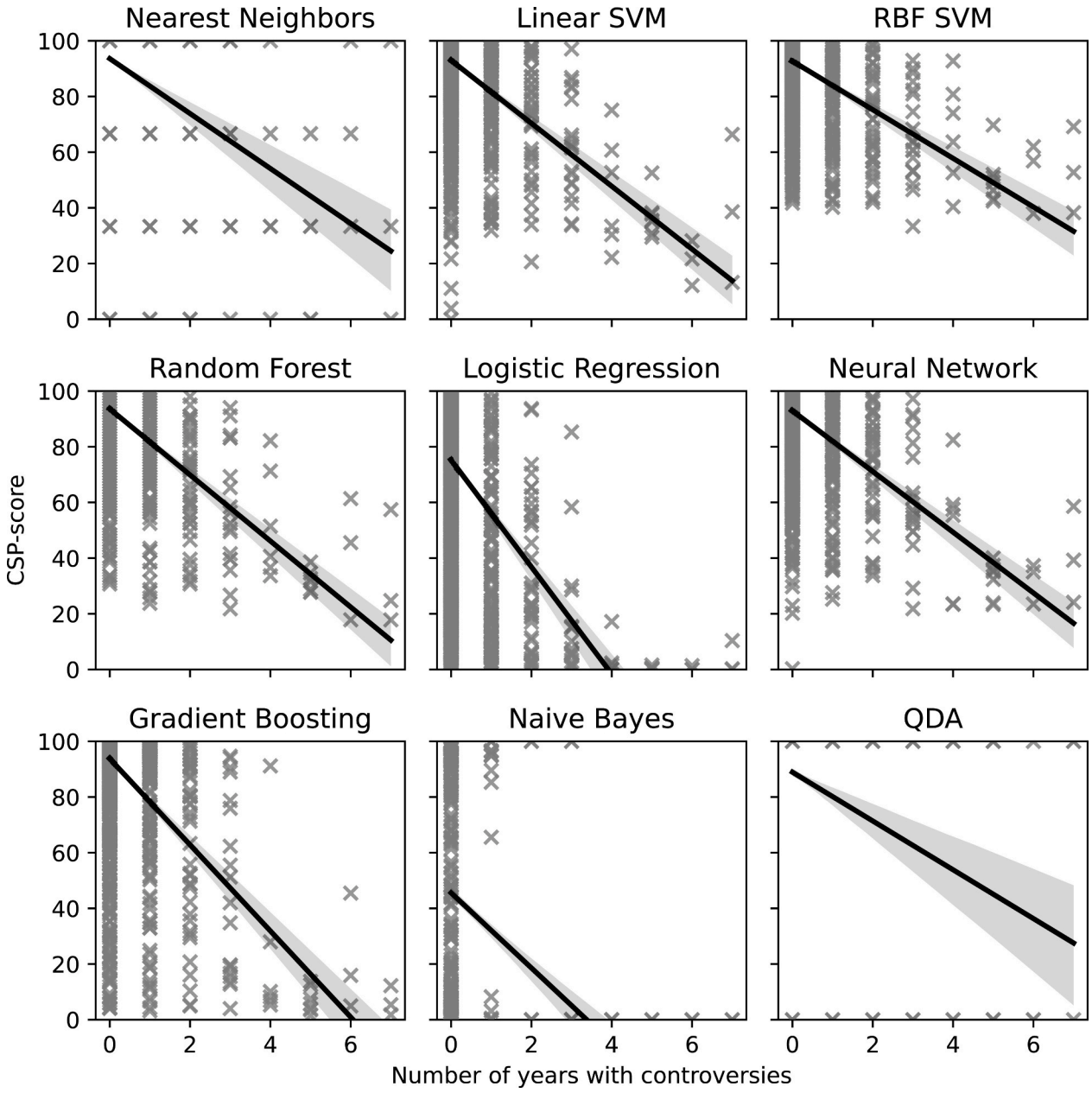

Fig. 4. Correlation between the CEP ratings produced by the nine machine learning algorithms and the number of years with controversies.

The differences between distributions show that further development of this machine learning technique to create ESG ratings would benefit from the comparison of several algorithms. One difference between the more- and less-complex methods is that the more complex ones (i.e., random forest, artificial neural network, and gradient boosting) give higher ratings to some companies with several controversies than does logistic regression. This difference suggests that the three complex methods tend to emphasize other aspects of indicator patterns than does logistic regression. A departure from giving all companies with many controversies very low ratings is apparently necessary in order to achieve high precision.

5. Discussion and conclusion

This study problematizes the inadequacy of most of the traditional ESG ratings. We highlight theoretical and methodological limitations caused mainly by the ineffective aggregation of indicators in the current ratings. Although institutional investors exert considerable influence on company activity in relation to environmental issues, a prerequisite for sustainable capital allocation is the availability of holistic ESG ratings based on comprehensive measures of CEP. For example, assessing the environmental harm of any company in comparison with other companies requires trade-offs between up to 100 parameters, and there is currently no valid method of determining the relative harm or benefit of a company's environmental features (Berg et al., 2019).

Our proposed CEP rating methodology is consistent with institutional investors' information needs, as described by Krueger et al. (2020). A central finding of their study was that reputation protection is achieved by excluding controversy-prone companies from portfolios, suggesting that a rating methodology with high validity in predicting environmental controversies would be preferred to traditional rating methodologies. In line with this, we suggest that a CEP rating methodology should capture the extent to which companies comply with environmental responsibilities and that environmental controversies indicate non-compliance with responsibilities, as described in the wrongdoing index literature (Fiaschi et al., 2020; Giuliani et al., 2015).

We demonstrate the possibility of developing such a rating by investigating whether controversies can be predicted from the company-specific patterns of environmental indicators. Although predicting controversies is only the first step towards developing such ratings, our results confirm that the suggested rating methodology is possible.

This study finds that environmental controversies can be predicted with high precision, and that the nonlinear models predict controversy slightly better than do the linear models. This indicates that institutional investors can systematically avoid controversy-prone companies. The findings are also consistent with an exclusive focus on companies' environmental requirement compliance due to its substantial financial materiality, as found by Nofsinger et al. (2019). Our proposed machine learning methodology is considerably less discretionary and subjective than are traditional ESG ratings because it assesses company environmental behaviour relative to a standard of environmental responsibilities established by society. The advantages of data-driven methodologies such as the proposed machine learning approach have not previously been discussed in the ESG literature, except for studies with a narrow focus on a particular issue, such as a prediction study of the carbon footprint (Nguyen et al., 2021) and a call for tests of ESG ratings' predictive performance (Chatterji et al., 2016).

Our study is limited because it does not design the actual CEP ratings but rather outlines blueprints for future method development, and we propose that this development path would enable future ratings to be validity tested against environmental controversies as holistic proxies for CEP. Such tests are essential for the credibility of commercially used ratings.

Our findings have exposed some challenges that future development of a machine-learning–based CEP rating methodology could address. A first limitation is that we did not investigate the information conveyed by each controversy. The nature of individual controversies may differ, so that one controversy may be of greater or lesser importance than another, and controversies might signal situation- and company-specific information that we did not account for. Further development and analyses may shed light on how environmental controversies differ and how this can be accounted for in a machine-learning–based model. A second limitation is that we did not adjust the model for the unequal media attention paid to companies, which likely leads to different companies having different likelihoods of having a controversy, not corresponding to differences in their non-compliance with environmental responsibilities. It is also possible that the media reporting differs between countries depending on how concerning environmental issues are perceived to be: what is perceived as alarming in one part of the world may be more or less ignored in another.

Future refinement of this novel approach to CEP rating should investigate how to replace controversies with a corporate-level wrongdoing index, i.e., a scaled and filtered metric of the wrongdoing signalled by controversies (cf. Fiaschi et al., 2020; Nieri and Giuliani, 2018). Such an index could be used with machine learning algorithms in regression models to accomplish a more sensitive and finer-tuned estimation of the extent to which companies comply with environmental responsibilities.

CRediT authorship contribution statement

Jan Svanberg: Conceptualization, Writing – original draft, Writing – review & editing. **Tohid Ardeshiri:** Data curation, Methodology, Writing – original draft. **Isak Samsten:** Data curation, Methodology, Writing – original draft. **Peter Öhman:** Writing – original draft, Writing – review & editing. **Tarek Rana:** Writing – original draft, Writing – review & editing. **Mats Danielson:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We acknowledge the financial support from Länsförsäkringars Forskningsfond and the Asset Management Department at Länsförsäkringar. We are grateful to Lars Höglund, Kristofer Dreiman, Alexander Elving, Peter Griepenkerl Lööf, and Mari Sparr.

APPENDIX. The ESG indicators acquired

No.	Variable name	No.	Variable name
1	Resource reduction policy	**57**	VOC emissions
2	Policy water efficiency	**58**	Total waste to revenues
3	Policy energy efficiency	**59**	Waste recycled to total waste
4	Policy sustainable packaging	**60**	Total hazardous waste to revenues
5	Policy environmental supply chain	**61**	Waste total
6	Resource reduction targets	**62**	Non-hazardous waste
7	Targets water efficiency	**63**	Waste recycled total
8	Targets energy efficiency	**64**	Waste recycling ratio
9	Environment management team	**65**	Hazardous waste
10	Environment management training	**66**	Waste reduction initiatives
11	Environmental materials sourcing	**67**	e-Waste reduction
12	Toxic chemicals reduction	**68**	Water pollutant emissions to revenues
13	Total energy use to revenues	**69**	Water discharged
14	Renewable energy use ratio	**70**	Water pollutant emissions
15	Renewable energy supply	**71**	ISO 14000 or EMS
16	Energy use total	**72**	EMS certified percent
17	Energy purchased direct	**73**	Environmental restoration initiatives
18	Energy produced direct	**74**	Staff transportation impact reduction
19	Indirect energy use	**75**	Accidental spills
20	Electricity purchased	**76**	Environmental expenditures investments
21	Electricity produced	**77**	Self-reported environmental fines
22	Cement energy use	**78**	Environmental partnerships
23	Renewable energy purchased	**79**	Internal carbon pricing
24	Renewable energy produced	**80**	Internal carbon price per tonne
25	Renewable energy used	**81**	Emission reduction target percentage

(*continued on next page*)

(*continued*)

No.	Variable name	No.	Variable name
26	Green buildings	**82**	Emission reduction target year
27	Water use to revenues	**83**	Environmental products
28	Water withdrawal total	**84**	Eco-design products
29	Fresh water withdrawal	**85**	Environmental R&D expenditures to revenues
30	Water recycled	**86**	Environmental R&D expenditures
31	Environmental supply chain management	**87**	Noise reduction
32	Environmental supply chain monitoring	**88**	Fleet fuel consumption
33	Environmental supply chain partnership termination	**89**	Hybrid vehicles
34	Land environmental impact reduction	**90**	Fleet CO_2 emissions
35	Total renewable energy	**91**	Environmental assets under management
36	Policy emissions	**92**	Equator principles
37	Targets emissions	**93**	Equator principles or environmental project financing
38	Biodiversity impact reduction	**94**	Environmental project financing
39	Total CO_2 equivalent emissions to revenues	**95**	Nuclear
40	CO_2 equivalent emissions total	**96**	Nuclear production
41	CO_2 equivalent emissions direct	**97**	Labeled wood production
42	CO_2 equivalent emissions indirect	**98**	Labeled wood
43	Carbon offsets/credits	**99**	Organic products initiatives
44	Estimated CO_2 emission equivalents total	**100**	Product impact minimization
45	CO_2 estimation method	**101**	Take-back and recycling initiatives
46	Emissions trading	**102**	Product environmental responsible use
47	Cements CO_2 equivalents emission	**103**	GMO products
48	Climate change commercial risks opportunities	**104**	Agrochemical products
49	Flaring gases	**105**	Agrochemical 5% revenue
50	Ozone-depleting substances	**106**	Animal testing
51	NOx and SOx emissions reduction	**107**	Animal testing cosmetics
52	NOx emissions	**108**	Animal testing reduction
53	SOx emissions	**109**	Renewable/clean energy products
54	VOC or particulate matter emissions reduction	**110**	Water technologies
55	VOC emissions reduction	**111**	Sustainable building products
56	Particulate matter emissions reduction	**112**	Real estate sustainability certifications

References

Alpaydin, E., 2010. Introduction to Machine Learning, second ed. The MIT Press, London.

Amel Zadeh, A., Serafeim, G., 2018. Why and how investors use ESG information: evidence from a global survey. Financ. Anal. J. 74, 87–103.

Azadeh, A., Saberi, M., Moghaddam, R.T., Javanmardi, L., 2011. An integrated data envelopment analysis-artificial neural network-rough set algorithm for assessment of personnel efficiency. Expert Syst. Appl. 38, 1364–1373.

Berg, F., Kölbel, J., Rigobon, R., 2019. Aggregate confusion: the divergence of ESG ratings. SSRN Electron. J. https://doi.org/10.2139/ssrn.3438533.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Callan, S.J., Thomas, J.M., 2009. Corporate financial performance and corporate social performance: an update and reinvestigation. Corp. Soc. Responsib. Environ. Manag. 16, 61–78. https://doi.org/10.1002/csr.182.

Chatterji, A.K., Durand, R., Levine, D.I., Touboul, S., 2016. Do ratings of firms converge? Implications for managers, investors and strategy researchers. Strat. Manag. J. 37, 1597–1614. https://doi.org/10.1002/smj.2407.

Chatterji, A.K., Levine, D.I., Toffel, M.W., 2009. How well do social ratings actually measure corporate social responsibility? J. Econ. Manag. Strat. 18, 125–169. https://doi.org/10.2307/41166337.

Chen, C.M., Delmas, M., 2011. Measuring corporate social performance: an efficiency perspective. Prod. Oper. Manag. 20, 789–804. https://doi.org/10.1111/j.1937-5956.2010.01202.x.

Choi, D., Gao, Z., Jiang, W., 2020. Global carbon divestment and firms' actions. SSRN Electron. J. https://doi.org/10.2139/ssrn.3589952.

Christensen, D., Serafeim, G., Sikochi, A., 2022. Why is corporate virtue in the eye of the beholder? The case of ESG ratings. Account. Rev. 97, 147–175. https://doi.org/10.2308/tar-2019-0506.

Christensen, M., Clawson, S., 2018. European SRI Study EUROSIF, Revised edition.

Cleofas-Sánchez, L., García, V., Marqués, A.I., Sánchez, J.S., 2016. Financial distress prediction using the hybrid associative memory with translation. Appl. Soft Comput. J. 44, 144–152.

Cornell, B., 2020. ESG Preferences, Risk and Return. European Financial Management eufm, p. 12295.

Delmas, M., Blass, V.D., 2010. Measuring corporate environmental performance: the trade-offs of sustainability ratings. Bus. Strat. Environ. 19, 245–260. https://doi.org/10.1002/bse.676.

Delmas, M.A., Burbano, V.C., 2011. The drivers of greenwashing. Calif. Manag. Rev. 54, 64–87.

Delmas, M.A., Etzion, D., Nairn-Birch, N., 2013. Triangulating environmental performance: what do corporate social responsibility ratings really capture? Acad. Manag. Perspect. 27, 255–267. https://doi.org/10.5465/amp.2012.0123.

Ding, K., Lev, B., Peng, X., Sun, T., Vasarhelyi, M.A., 2020. Machine learning improves accounting estimates: evidence from insurance payments. Rev. Account. Stud. 25, 1098–1134. https://doi.org/10.1007/s11142-020-09546-9.

Donaldson, T., Preston, L.E., 1995. The stakeholder theory of the corporation: concepts, evidence, and implications. Acad. Manag. Rev. 20, 65–91.

Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classification. John Wiley, New York.

Ekonomou, L., 2010. Greek long-term energy consumption prediction using artificial neural networks. Energy 35, 512–517.

Faulkner, R.R., 2011. Corporate Wrongdoing and the Art of the Accusation. Anthem Publishers. https://doi.org/10.7135/UPO9780857284204.

Fiaschi, D., Giuliani, E., Nieri, F., 2017. Overcoming the liability of origin by doing no-harm: emerging country firms' social irresponsibility as they go global. J. World Bus. 52, 546–563. https://doi.org/10.1016/j.jwb.2016.09.001.

Fiaschi, D., Giuliani, E., Nieri, F., Salvati, N., 2020. How bad is your company? Measuring corporate wrongdoing beyond the magic of ESG metrics. Bus. Horiz. 63, 287–299. https://doi.org/10.1016/j.bushor.2019.09.004.

Freeman, R.E., 2015. Strategic Management: A Stakeholder Approach. Cambridge University Press (CUP), Cambridge.

Gerlein, E.A., McGinnity, M., Belatreche, A., Coleman, S., 2016. Evaluating machine learning classification for financial trading: an empirical approach. Expert Syst. Appl. 54, 193–207.

Ghritlahre, H.K., Prasad, R.K., 2018. Application of ANN technique to predict the performance of solar collector systems - a review. Renew. Sustain. Energy Rev. 84, 75–88. https://doi.org/10.1016/j.rser.2018.01.001.

Giuliani, E., Nieri, F., Fiaschi, D., 2015. BRIC companies seeking legitimacy through Corporate Social Responsibility. Transnatl. Corp. 22, 5–42. https://doi.org/10.18356/e13d5a2e-en.

Grappi, S., Romani, S., Bagozzi, R.P., 2013. Consumer response to corporate irresponsible behavior: moral emotions and virtues. J. Bus. Res. 66, 1814–1821. https://doi.org/10.1016/j.jbusres.2013.02.002.

Gulo, C.A.S.J., Rúbio, T.R.P.M., Tabassum, S., Prado, S.G.D., 2015. Mining scientific articles powered by machine learning techniques. OpenAccess Series in Informatics 49, 21–28.

Guyon, I., 1997. A Scaling Law for the Validation-Set Training-Set Size Ratio. https://doi.org/10.4230/OASIcs.ICCSW.2015.21. AT&T Bell Laboratories 1–11.

Heath, Y., Gifford, R., 2002. Extending the theory of planned behavior: predicting the use of public transportation. J. Appl. Soc. Psychol. 32, 2154–2189.

Hu, L.Y., Huang, M.W., Ke, S.W., Tsai, C.F., 2016. The distance function effect on k-nearest neighbor classification for medical datasets. SpringerPlus 5.

Krueger, P., Sautner, Z., Starks, L., 2020. Importance of climate risks for institutional investors. Rev. Financ. Stud. 33, 1067–1111. https://doi.org/10.1093/rfs/hhz137.

Maione, C., De Paula, E.S., Gallimberti, M., Batista, B.L., Campiglia, A.D., Barbosa, F., Barbosa, R.M., 2016. Comparative study of data mining techniques for the authentication of organic grape juice based on ICP-MS analysis. Expert Syst. Appl. 49, 60–73.

Mattingly, J.E., Berman, S.L., 2006. Measurement of corporate social action: discovering taxonomy in the Kinder Lydenburg Domini ratings data. Bus. Soc. 45, 20–46. https://doi.org/10.1177/0007650305281939.
Nguyen, Q., Diaz-Rainey, I., Kuruppuarachchi, D., 2021. Predicting corporate carbon footprints for climate finance risk analyses: a machine learning approach. Energy Econ. 95, 105129. https://doi.org/10.1016/J.ENECO.2021.105129.
Nieri, F., Giuliani, E., 2018. International business and corporate wrongdoing: a review and research agenda. In: Castellan, D., Narula, R., Nguyen, Q., Surdu, I., Walker, J. (Eds.), Contemporary Issues in International Business. Springer International Publishing, pp. 35–53. https://doi.org/10.1007/978-3-319-70220-9_3.
Nofsinger, J.R., Sulaeman, J., Varma, A., 2019. Institutional investors and corporate social responsibility. J. Corp. Finance 58, 700–725. https://doi.org/10.1016/j.jcorpfin.2019.07.012.
Oikonomou, I., Platanakis, E., Sutcliffe, C., 2018. Socially responsible investment portfolios: does the optimization process matter? Br. Account. Rev. 50, 379–401. https://doi.org/10.1016/j.bar.2017.10.003.
Orlitzky, M., 2013. Corporate social responsibility, noise, and stock market volatility. Acad. Manag. Perspect. 27, 238–254. https://doi.org/10.5465/amp.2012.0097.
Pedregosa, F., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.
Refinitiv, 2019. Environmental, Social and Governance (ESG) Scores from Refinitiv. ESG Scores Methodology.
Rumelhart, D.E., Widrow, B., Lehr, M.A., 1994. The basic ideas in neural networks. Commun. ACM 37, 87–92.
Schultz, P.W., Messina, A., Tronu, G., Limas, E.F., Gupta, R., Estrada, M., 2014. Personalized normative feedback and the moderating role of personal norms: a field experiment to reduce residential water consumption. Environ. Behav. 48, 686–710.
Semenova, N., Hassel, L.G., 2015. On the validity of environmental performance metrics. J. Bus. Ethics 132, 249–258. https://doi.org/10.1007/s10551-014-2323-4.
Shmueli, G., 2010. To explain or to predict? Stat. Sci. 25, 289–310. https://doi.org/10.1214/10-STS330.
Sigrist, F., Hirnschall, C., 2019. Grabit: gradient tree-boosted tobit models for default prediction. J. Bank. Finance 102, 177–192. https://doi.org/10.1016/j.jbankfin.2019.03.004.
Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer, Berlin. https://doi.org/10.1007/978-1-4757-2440-0.
Wong, C., Petroy, E., 2020. Rate the Raters 2020: Investor Survey and Interview Results.
Wong, S.L., Wan, K.K.W., Lam, T.N.T., 2010. Artificial neural networks for energy analysis of office buildings with daylighting. Appl. Energy 87, 551–557.
Wood, D.J., 2010. Measuring corporate social performance: a review. Int. J. Manag. Rev. 12, 50–84. https://doi.org/10.1111/j.1468-2370.2009.00274.x.
Xu, X., Zhou, C., Wang, Z., 2009. Credit scoring algorithm based on link analysis ranking with support vector machine. Expert Syst. Appl. 36, 2625–2632. https://doi.org/10.1016/j.eswa.2008.01.024.
Yeh, C.C., Chi, D.J., Lin, Y.R., 2014. Going-concern prediction using hybrid random forests and rough set approach. Inf. Sci. 254, 98–110. https://doi.org/10.1016/j.ins.2013.07.011.
Yuan, L., Yong, F., Wei, Z., Shan, K., 2017. Using quadratic discriminant analysis to predict protein secondary structure based on chemical shifts. Curr. Bioinf. 12, 52–56. https://doi.org/10.2174/1574893611666160628074537*.

Jan Svanberg is Associated Professor (Ph.D) of Business Administration at University of Gävle and the Centre for Research on Economic Relations (CER). His research interests are behavioral issues, primary in accounting and auditing.

Tohid Ardeshiri has a degree of Doctor of Philosophy in Statistical Signal Processing from Linköping University. He conducts machine learning research on financial applications at University of Gävle and the Centre for Research on Economic Relations (CER).

Isak Samsten has a degree of Doctor of Philosophy in Computer and Systems Sciences from Stockholm University. He conducts machine learning research on financial applications at Stockholm University.

Peter Öhman is Professor (Ph.D) of Business Administration at Centre for Research on Economic Relations (CER) at Mid Sweden University. His research interests are primary in accounting and auditing.

Tarek Rana has a degree of Doctor of Philosophy in Accounting and is Senior Lecturer at The Royal Melbourne Institute of Technology. He conducts research on ESG ratings and management accounting.

Mats Danielson is Professor of Computer and Systems Sciences at Stockholm University. He conducts research on computer-based decision-making.