Received: 12 July 2021 | Revised: 3 January 2022 | Accepted: 20 February 2022

DOI: 10.1002/isaf.1505

RESEARCH ARTICLE

WILEY

Corporate governance performance ratings with machine learning

Jan Svanberg[1] | Tohid Ardeshiri[2] | Isak Samsten[3] | Peter Öhman[4] | Presha E. Neidermeyer[5] | Tarek Rana[6] | Natalia Semenova[7] | Mats Danielson[3,8]

[1]Centre for Research on Economic Relations, and The Royal Melbourne Institute of Technology, University of Gävle, Gävle, Sweden

[2]University of Gävle, Gävle, Sweden

[3]Department of Computer and Systems Sciences, Stockholm University, Stockholm, Sweden

[4]Department of Economics, Geography, Law and Tourism, Centre for Research on Economic Relations, Mid Sweden University, Sundsvall, Sweden

[5]West Virginia University, Morgantown, West Virginia, USA

[6]The Royal Melbourne Institute of Technology, School of Accounting, Information Systems & Supply Chain, RMIT University, Melbourne, VIC, Australia

[7]Department of Accounting and Logistics, School of Business and Economics, Linnaeus University, Växjö, Sweden

[8]International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria

Correspondence
Jan Svanberg, University of Gävle, Centre for Research on Economic Relations, SE-801 76 Gävle, Sweden, and The Royal Melbourne Institute of Technology, School of Accounting, Information Systems & Supply Chain, Melbourne, VIC 3001, Australia.
Email: jan.svanberg@hig.se

Funding information
Stiftelsen Länsförsäkringsbolagens Forskningsfond, Grant/Award Number: P 18/08

Summary

We use machine learning with a cross-sectional research design to predict governance controversies and to develop a measure of the governance component of the environmental, social, governance (ESG) metrics. Based on comprehensive governance data from 2,517 companies over a period of 10 years and investigating nine machine-learning algorithms, we find that governance controversies can be predicted with high predictive performance. Our proposed governance rating methodology has two unique advantages compared with traditional ESG ratings: it rates companies' compliance with governance responsibilities and it has predictive validity. Our study demonstrates a solution to what is likely the greatest challenge for the finance industry today: how to assess a company's sustainability with validity and accuracy. Prior to this study, the ESG rating industry and the literature have not provided evidence that widely adopted governance ratings are valid. This study describes the only methodology for developing governance performance ratings based on companies' compliance with governance responsibilities and for which there is evidence of predictive validity.

KEYWORDS
artificial intelligence, ESG, governance controversies, machine learning, performance of ESG ratings, prediction, socially responsible investment

1 | INTRODUCTION

Corporate governance as an accountability mechanism is a core research area in the accounting and finance disciplines (Zengul et al., 2019) and an essential component of the environmental, social, and governance (ESG) principles (Christensen et al., 2021). Although researchers have commented on the need for corporate governance performance (CGP) ratings, the potential of using machine learning (ML) to leverage the gains of predictive modeling has not been investigated in the ESG rating literature (Chatterji et al., 2016). We, therefore, develop a comprehensive measure of CGP as part of the overall ESG performance metric.

The CGP rating issue is important because institutional investors and other stakeholders need to assess this performance of companies and avoid the costs of non-compliance with fiduciary duties of the board, management, and/or auditors (Arnold & de Lange, 2004; Asthana et al., 2010; Canada et al., 2009; Nofsinger et al., 2019). Previous research on socially responsible investment (SRI) has suggested that ESG ratings assist investors in predicting future ESG risks[1] of portfolio companies (Oikonomou et al., 2018); consequently, research and such ratings have been adopted to determine the stock market returns of SRI portfolios. However, no study provides any comforting support for the claim that ESG ratings are valid measures of corporate social performance. On the contrary, a body of findings consistently indicates a lack of validity (Chatterji et al., 2016; Chen & Delmas, 2011; Christensen et al., 2021; Semenova & Hassel, 2015; Trumpp et al., 2013).

There are several possible reasons why ESG ratings do not represent the extent that companies are sustainable. The use of linear models is a key disadvantage of traditional rating methods (Abhayawansa & Tyagi, 2021; Chen & Mussalli, 2020; Chen & Delmas, 2011), because a complex and multifaceted concept such as ESG unlikely has linear relationships with its constituting feature indicators. Berg et al. (2019) found that six major ESG ratings are produced with linear models, which disables them from representing nonlinear feature contributions to ESG. Moreover, the ratings are ad hoc weighted averages, which means that they presume that the relative importance of all aspects of ESG are appropriately represented by the model weights set by the rater (Berg et al., 2019; Chen & Delmas, 2011; Delmas et al., 2013). Traditional ESG rating methodology also fails to consider and weigh qualitative and nonfinancial information (Kotsantonis & Serafeim, 2019), which is problematic because a great deal of sustainability reporting consists of such information. Another possible reason is that the underlying conceptual model used by traditional ratings makes such ratings difficult or impossible to interpret and, therefore, unverifiable for users.

The motivation for SRI is high among institutional investors (Krueger et al., 2020), but the screens they use may be so poor that they achieve the same financial results and the same level of responsible investment as if they were oblivious of SRI (Hartzmark & Sussman, 2019). In the same vein, Chatterji et al. (2016) concluded their examination of convergent validity of ESG ratings from several commercial suppliers of such ratings by arguing that the results of academic studies using these metrics should be reassessed. Another issue for practice is that top managers in large businesses respond to the pressure from stakeholders to address ESG issues based on questionable rating methodologies (Crilly et al., 2012) or strive toward a lower cost of capital. To overcome this lack of validity, we propose a new type of CGP rating that addresses the weaknesses of current ratings by adopting an ML approach that enables less subjectivity in the weighting of indicator importance and predictive validity valuable to institutional investors.

The core problem with ESG rating validity lies not in lack of data or poor data quality but in the type of models used for computing the ratings. While credit ratings improve with the amount of credit data on which they are based, the subjective methods with which indicators are aggregated to form holistic ESG ratings actually cause a reduction of convergent validity when the amount of ESG data increases (Christensen et al., 2021). Raters are attempting to estimate a complex and multifaceted concept for which they have no general proxy, and therefore no way of validating their ratings other than comparing with each other. Relying on subjective indicator weights means that the more indicators that raters aggregate the greater is the destructive impact of the subjectivity of indicator weighting schemes on the validity of their ratings and the more that ratings diverge. However, the estimation difficulties are less daunting for predictive modeling than for explanatory modeling. While explanatory modeling is limited by the theory-driven design and the use of statistical methods, predictive modeling is data driven (e.g., indicator weights do not need to be known a priori) with potentially unlimited capacity for representing complexity, multidimensionality, and nonlinearity, and the meaning of the contribution of various indicators to the aggregated CGP construct is developed ex post or in the process of data analysis (Shmueli, 2010). Using a general proxy for CGP and employing ML in a supervised learning setting enables estimation of an optimal set of indicator weights that is, if not an objective weighting scheme, then a far less subjective weighting scheme than the manual guesswork of traditional ratings.

Our study discusses shortcomings with current ESG ratings and extends previous findings by illustrating that the governance component of ESG ratings is an inappropriate classifier of companies' likelihood of incurring a governance controversy. We introduce ML to the context of CGP ratings and demonstrate how the use of ML solves central problems with traditional ratings. ML has been successfully used to solve similar problems in accounting and finance; for example, predicting credit risk and failure (Butaru et al., 2016; Khandani et al., 2010; Sigrist & Hirnschall, 2019), internal controls evaluation (Changchit & Holsapple, 2004), predicting management fraud (Fanning & Cogger, 1998; Goel & Gangolly, 2012), predicting dividend policy (Longinidis & Symeonidis, 2013), and obviously corporate failure (O'Leary, 1998; Telmoudi et al., 2011). The new CGP rating methodology that rates companies based on predicted governance controversy likelihood contributes to the literature by finding that governance controversies can be predicted across multiple measures and that a CGP rating methodology, therefore, can be developed from the information contained in governance indicators. Whereas

traditional ESG ratings have all been questioned for completely lacking validity (Berg et al., 2019; Chatterji et al., 2016) due to subjective weighting schemes (Christensen et al., 2021), our rating methodology has predictive validity and suggests a path to an objective ESG rating methodology that ensures institutional investors' valid SRI screening of investment portfolios and rids the investment industry of the traditional no-validity ESG ratings.

The next section presents the literature used and the motivation of the study. The method is then outlined and the results are presented. A concluding discussion ends the paper.

2 | RELATED LITERATURE AND MOTIVATION OF THE STUDY

2.1 | Background: Problems with current ESG ratings

Despite the popularity of ESG ratings, they have been critically assessed with discouraging evidence of what these aggregated ratings actually represent (Chatterji & Levine, 2006; Chen & Delmas, 2011; Delmas et al., 2013; Delmas & Blass, 2010; Trumpp et al., 2013). As indicated, we propose three main reasons why CGP assessment with holistic ESG ratings is a challenge.

First, a linear combination of indicators is unlikely to be a valid representation of CGP because such estimation methods cannot represent a multidimensional, complex, and issue-contingent construct, which is how CGP is described in the leading theoretical accounts (Wood, 2010). For example, Semenova and Hassel (2015) and Delmas et al. (2013) argued that the multidimensionality of such ratings most likely obscures their content. Similarly, Mattingly and Berman (2006) claimed that aggregating nonconvergent metrics in the construction of a combined measure masks an underlying association between the metrics and other variables that then confuses the interpretation of observed relationships. Assessments of such diverse issues as philanthropic activity, the natural environment, support of local society, corporate governance, and human rights (Griffin & Mahon, 1997) is difficult to capture in a few metrics owing to the heterogeneity of the issues involved (Carroll, 1999; Delmas & Blass, 2010; Graves & Waddock, 1994). The challenge with measuring such a diverse collection of topics is more similar to image recognition than it is to most accounting research problems that are examined with statistical models. Applied to the CGP rating problem, this means that an estimator capable of capturing multidimensionality and nonlinearity, such as interaction, and the most diverse data should be adopted.

Second, the use of linear models assigning weights to CGP indicators means that the raters determine the importance of each indicator for their aggregated ratings by choosing coefficients to their rating formula. There is no objective justification for the indicator coefficients, so raters essentially tend to define CGP ratings according to their own preferences (Chatterji & Levine, 2006; Hillman & Keim, 2001). The raters have no reference point for assessments of the indicator importance weights and can therefore not know whether it is best to use equal weighting for simplicity (e.g., Refinitiv Eikon) or unequal weights (e.g., MSCI, Sustainalytics). Empirical evidence supports that the use of subjective indicator weights is the main cause of ratings' lack of convergent validity (Christensen et al., 2021).

Third, the linear models used in traditional CGP ratings do not assess the CGP relative to a performance standard when assessing how feature indicators are associated with having good or bad CGP. This lack of performance standard makes the indicator weighting scheme arbitrary, and raters have to define ratings by comparing companies within a group (e.g., an industry). Within-group comparisons, however, are no solution to the subjectivity problem because ratings become dependent on the choice of groups and samples of firms (Kotsantonis & Serafeim, 2019). This rating practice is inconsistent with how performance is assessed in other areas of performance assessment. Unfortunately, the lack of standard for overall CGP makes the estimation of the contribution of individual features to the holistic rating an indeterminable problem. With such a standard (i.e., a proxy for CGP), the contributions of individual indicators can be estimated with an ML model and the equivalent of indicator weights uniquely determined.

2.2 | Labeling CGP with governance controversies

Traditional ESG ratings are not only biased as a consequence of subjective weighting schemes but also inconsistent with institutional investor information needs. Recent research has found that institutional investors have selective preferences for CGP. They are indifferent to the extent that companies perform on CGP features that are not governance responsibilities but underweight stocks in companies that do not comply with such obligatory governance responsibilities (Nofsinger et al., 2019). The reason for this asymmetrical preference is that noncompliance with legal or moral social responsibilities leads to adverse effects, such as litigation, government punishment, customer boycotts, and disrupted production (Benabou & Tirole, 2010), whereas the extent that companies perform on issues that are not obligatory does not incur noncompliance costs. Doing good on discretionary issues may have benefits to the company (e.g., lower cost of capital and lower risk; Dyck et al., 2019), but the benefits are completely offset by higher costs of accomplishing the deeds in most conditions (Nofsinger et al., 2019), making the net effect financially irrelevant for institutional investors. Some illustration to this information preference is that markets react strongly to bad ESG news but very little to good news (Capelle-Blancard & Petit, 2019; Krüger, 2015). Institutional investors would therefore benefit from CGP ratings that indicate the extent that companies comply with compulsory governance responsibilities. The information preference found by Nofsinger et al. (2019) is consistent with the Krueger et al. (2020) finding that institutional investors' strongest motivation for considering ESG risks in their investment process is reputation protection. This means that these investors predominantly avoid investing in companies that are controversy prone. They focus on avoiding this investment risk rather than

on strengthening their investments' competitive position on ethical motives (Amir & Serafeim, 2018).

Another motivation to consider ESG in institutional investment is the emergent fiduciary obligation to invest in a responsible manner (Krueger et al., 2020). The traditional view has been that fund management regulation would prevent them from adopting SRI (Sandberg, 2011), but the strong growth of ESG considerations in all fund management is about to make this view obsolete, and expectations on institutional investment are rapidly tilting toward an obligation to consider ESG. This is indicated by the number of signatories, above 3,000 in 2020, of the United Nations Global Compact Principles for Responsible Investment. The principles state that "businesses should support and respect the protection of internationally proclaimed human rights." As a fiduciary obligation, it should not suffice for institutional investors to use a subjective or idiosyncratic view of ESG, or CGP for that matter, because the duty would then amount to very little. The duty to consider ESG is a strong case for assessing ESG consistent with a societal conception of it that would have to reflect companies' compliance with social responsibilities, because they are the ESG concerns that society has institutionalized.

As a response to these information needs of institutional investors, we suggest a compliance-based CGP rating methodology that is conceptually developed from Wood's (1991, 2010) leading conceptualization, and which is closely related to the work of Carroll (1979). According to Wood (2010), the definition of corporate social performance, of which we view CGP as a component, is a set of descriptive categorizations of business activity, focusing on the impacts and outcomes for society, stakeholders, and the firm. Types of outcomes are determined by the linkages, both general and specific, defined by the structural principles of corporate social responsibility. The processes by which these outcomes are produced, monitored, evaluated, compensated, and rectified (or not) are defined by the processes of corporate social responsiveness. With this definition, ESG, and therefore CGP, includes company features (behaviors, structures) relevant for how it performs on the structural principles of social responsibility, which are legitimacy, public responsibility, and managerial discretion. The three concepts that make up the principles of social responsibility collectively match the main thrust of Carroll's (1979) pyramid of responsibilities that companies have toward stakeholders and society. This overlap underlines the importance of legal and moral responsibilities as the conceptual foundation of CGP. Following Wood (2010), we use a norms-based definition of CGP and concur with the critique that "doing good" according to a subjective set of preferences is an inappropriate conceptual foundation of social performance because subjective preferences may be inconsistent with social responsibilities for the same reasons as subjective weighting schemes are the main reason ESG ratings lack validity (Christensen et al., 2021). We therefore restrict our definition of CGP to companies' compliance with legal and moral governance responsibilities.

We view governance controversies as holistic labels for companies' noncompliance with governance responsibilities. A governance controversy is an event or situation that involves the employment of financial resources, a questionable ethical behavior (judged against legal or moral responsibilities), and is covered by media. In contrast to the CGP indicators that describe governance features (e.g., board member composition, chief executive officer compensation), the controversies are reactions to features (e.g., inappropriate board member composition or unfair chief executive officer compensation). Using the court of law as a metaphor, the CGP indicators are facts in the case, whereas controversies are assessments of the facts with consequences for legitimacy and legitimacy loss (Deegan, 2019, 2002). In predictive modeling, the controversies, therefore, give meaning to the indicators as governance performance relative to standards. This is a proposition not previously discussed in the ESG rating literature, although some raters use controversies as a reason to deduct some amount from a company's rating if it has had controversies (e.g., Refinitiv Eikon has its Eros STX Global Corporation rating). We use controversies more radically because we use ML to learn the meaning of CGP indicators from their association with breaches of governance responsibilities, collectively conveying information about systematic weaknesses in companies. When companies are pressured by high financial expectations, they tend to cut corners and not comprehensibly comply with governance responsibilities (Fiaschi et al., 2017; Surroca et al., 2013). The individual characters of managers and ethical climate are obviously not reported in annual reports, but a controversy prediction model can use a broad set of governance structure indicators.

3 | DATA AND EXPERIMENTS

3.1 | Research design and ML algorithms

In this study we explore the possibility to predict governance controversy in the classification sense from the information that companies voluntarily disclose in annual reports. The main question is whether this can be done, because if it can then a refined version of this methodology can be developed to obtain CGP ratings that mirror companies' compliance with moral and legal governance responsibilities. As our study is explorative, we investigate the predictive ability of nine ML algorithms on the task of predicting the risk of governance controversies by learning to identify CGP indicator patterns typical of controversy companies. The experiments adopt a cross-sectional research design, and we predict the likelihood of incurring a governance controversy in a 10-year window. The predictive ability is evaluated according to five distinct performance measures: precision, recall, *F*-measure, area under the receiver operating characteristic (ROC) curve, and area under the precision–recall curve (PRC). Our design and execution of computational experiments follow established methodological practices in computer science (Alpaydin, 2010).

Because we are exploring a rating methodology, we prefer predictive modeling to explanatory modeling in line with Shmueli (2010). Moreover, ESG data are a too large and rich dataset to be analyzed with explanatory modeling, and we are investigating the development of CGP ratings for which explanatory understanding is secondary to predictive accuracy. Predictive modeling with ML is an analysis of data that finds patterns and relationships in the data that are less

TABLE 1 The companies included in the dataset described by the industry sector, total assets, and number of controversies

GICS sector	No. companies	No. controversies	No. companies with controversies	No. of controversies if >0						Assets ($10^9$$)	Asset per company ($10^9$$)					
				Min.	Max.	Mean	*SD*	Skewness	Kurtosis		Min.	Max.	Mean	*SD*	Skewness	Kurtosis
Financials	398	310	101	1	20	3.07	3.83	2.65	7.48	79,534	0.24	2,509	199	393	3	11
Health care	141	110	45	1	11	2.44	2.45	2.17	4.27	2,257	0.05	144	16	24	2	7
Consumer discretionary	297	103	58	1	13	1.78	1.83	4.66	26.05	4,153	0.11	331	13	35	6	46
Information technology	201	103	43	1	8	2.4	1.8	1.5	1.66	2,387	0.16	182	11	24	4	24
Industrials	420	99	59	1	9	1.68	1.46	3.4	13.22	5,770	0.25	457	13	35	10	122
Communication services	161	74	41	1	7	1.8	1.29	2.15	5.83	3,415	0.06	284	21	40	3	14
Energy	183	73	32	1	9	2.28	2.16	1.87	2.69	4,581	0.07	291	25	51	3	12
Materials	277	54	30	1	7	1.8	1.27	2.67	9.16	2,703	0.04	96	9	15	3	14
Consumer staples	156	48	29	1	10	1.66	1.72	4.4	21.39	2,377	0.29	160	15	23	3	15
Real estate	155	16	10	1	3	1.6	0.7	0.78	−0.15	1,667	0.42	78	10	12	2	7
Utilities	124	14	11	1	3	1.27	0.65	2.42	5.51	3,494	0.13	259	28	36	3	15

The table shows an overview of the distribution of equities across Global Industry Classification Standard (GICS) sector and data for companies with more than one controversy disclosed separately. For example, there are 310 controversies in Financials, and of these companies the mean number of controversies is 3.07. Assets refers to total assets.

obtainable with traditional statistical methods (Collopy et al., 1994; Gurbaxani & Mendelson, 1990, 1994). Though predictive modeling is less popular for developing new theory, it is a methodology suitable for developing new measures that can interplay with theory (Van Maanen et al., 2007) owing to its complexity and nonlinearity handling capacity. No other method known to us can aggregate even most of the 114 governance behavior indicators we investigate and detect patterns in those that are associated with the likelihood of having a governance controversy. For example, regression in explanatory modeling would prove difficult due to multicollinearity, but this problem does not restrict the predictive power of algorithmic prediction (Vaughan & Berry, 2005), which allows the ML models to include a broad set of indicators. Predictive modeling also serves the purpose of quantifying the level of predictability of, in our case, governance controversies (Ehrenberg & Bound, 1993). This level can later be compared with what can be achieved with explanatory models. When any such model is achieving even a fraction of what the best predictive models do in terms of predictive power, this is evidence of theoretical refinement (Shmueli, 2010).

We obtained data from the Refinitiv Eikon database, which is often used in ESG research owing to its wide indicator coverage and transparency (Semenova & Hassel, 2015), but we constrained our sample to companies that have ESG data for a period of 10 years by requiring, as our only inclusion criterion, that companies have an ESG rating for all 10 of the years to be included in our sample. The complete list of corporate governance indicators adopted in this study is provided in Appendix A. Refinitiv Eikon contains approximately 400 indicators of ESG and 23 types of controversies, of which governance controversies is one. We collected 114 indicators classified as governance data. The dataset contains $n = 2{,}517$ firms, out of which approximately 20% have been involved in at least one governance controversy over the period between 2009 and 2018. Table 1 provides an overview of the firms included from various sectors.

TABLE 2 Hyper-parameters

Classifier	Description	Notes
NN	Nearest neighbors	Three nearest neighbors
Linear SVM	Linear support vector machine	Linear kernel with $C = 0.025$
RBF SVM	RBM support vector machine	RBF kernel with $C = 0.025$
RF	Random forest	100 trees
LR	Logistic regression	Ridge regularization with $C = 1$
ANN	Artificial neural network	Four hidden layers of size 100 using the RELU activation function
GB	Gradient boosting	Learning rate of 0.1
NB	Naive Bayes	No hyper-parameters
QDA	Quadratic discriminant analysis	No hyper-parameters

RBF: radial basis function; RBM: restricted Boltzmann machine; RLU: rectified linear unit.

Cross-sectional experiments were performed using nine ML algorithms selected based on their wide applicability and varying ability at discovering linear or nonlinear attribute interactions. We categorized the firms into two (distinct) categories based on their lack of or involvement in a controversy during any of the 10 years and denote the firms with a controversy as positive cases and those that have not been involved as negative cases. To capture the longitudinal aspect of the dataset, we employ a simple strategy in which the indicators were averaged if numerical or encoded using dummy variables if binary (i.e., one per year). Notably, our goal was not to model a particular firm's risk of a controversy given past indicators. Instead, the goal was to capture the attribute interactions that describe a firm that has a high risk of being involved in a controversy. The nine ML algorithms enumerated in Table 2 are briefly described in Appendix B.

3.2 | Experiments

3.2.1 | Hyper-parameters

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

We evaluate the predictive performance of our classifier using a previously unseen set of test instances (i.e., the governance indicators for a firm). The common approach to evaluate the predictive performance of a classifier is to partition the dataset into a training set (used for training the ML algorithm) and a test set (used for evaluating its performance on independent test data). However, another method may be employed if data are scarce or if a more reliable estimate of the generalization performance of the ML algorithm is required. One approach to accomplish this is to employ k-fold cross-validation, which is employed in the current research. This modification is a cross-validation approach to reliably estimate the performance of a learning algorithm. It partitions the dataset into k disjoint folds and trains the learning model iteratively on $k - 1$ folds, leaving one of the folds for testing. Though there are many approaches for selecting the number of folds k, we employ the most often used ML research technique of 10-fold cross-validation (Azadeh et al., 2011; Laha et al., 2015; Lu et al., 2019; Safa & Samarasinghe, 2011; Xu et al., 2009).

As indicated, we focused on five of the most frequently used and trusted measures of performance in research: precision, recall, F-measure, area under the ROC curve, and area under the PRC (Alpaydin, 2010). Evaluation and comparison of algorithm performance was done with the precision and recall for the estimators, as defined in Equations (1) and (2), using true positive (TP), false positive (FP), true negative (TN), and false negative (FN). Precision is the fraction of TPs in relation to the total number of positive case predictions. Precision, Equation (1), measures the sensitivity of the classifier (i.e., its accuracy in predicting the controversy and

noncontroversy classes expressed as the fraction of correct positives divided by the total number of positive predictions), and recall, Equation (2), is the fraction of TP predictions in relation to all positive cases in the data:

$$\text{Precision}_{\text{positive}} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (1)$$

$$\text{Recall}_{\text{positive}} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2)$$

Precision and recall are conflicting measures, but the *F*-measure, Equation (3), captures the trade-off between the two measures:

$$F\text{-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

The area under the ROC curve measure is the area under the curve defined as a plot of TP, Equation (4), versus the FP, Equation (5):

$$\text{TP}_{\text{rate}} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4)$$

$$\text{FP}_{\text{rate}} = \frac{\text{FP}}{\text{FP} + \text{TN}} \quad (5)$$

This area displays the probability of a classifier ranking a TP instance before an FP instance. Similarly, the PRC is a curve showing the precision for multiple thresholds of recall and is, similar to the *F*-measure, used to measure the trade-off between precision and recall. The area under PRC is defined as the area under the plot of precision versus the recall. The main benefit of both these areas is that they are insensitive to the class distribution of the training and testing data.

4 | RESULTS

4.1 | Predictive performance of algorithms

An overview of the nine algorithms' predictive performance is shown in Table 3. On most of the performance measures, three methods outperform the others. Random forest (0.75), gradient boosting, (0.65), and artificial neural network (ANN; 0.62) have a markedly higher precision than the other algorithms, which is important in investment because false predictions have high costs. Random forest and several other algorithms predict governance controversies with high precision, which is evidence that the models have predictive validity. The findings in Table 3 are promising because they indicate that a CGP rating methodology can be developed using a comprehensive set of CGP indicators as in Appendix A. The findings indicate that CGP indicators collectively contain substantial information of the weaknesses signaling noncompliance with governance responsibilities.

Ten-fold cross-validation is used for computing the results in Table 3. Cross-validation economizes with data, ensures that the same firm is not included both in training and testing data, and avoids overtraining. Throughout the experiments, identical training and testing partitions ensure comparability of results for all algorithms. Each training and testing fold included identical firms for all algorithms.

Next, we investigate how Refinitiv Eikon's governance component ESG rating would function as a classifier for controversy/noncontroversy. The histogram in Figure 1 discloses the distributions of companies and company-years with controversies (blue) and without controversies (yellow). The histogram suggests that the rating does not distinguish between controversy companies and noncontroversy companies. Figure 2 illustrates the same information for company-years. Both histograms suggest that these governance ratings are not able to separate the distributions in a meaningful way, which is an indication, although not conclusive evidence, that these governance ratings do not represent information about the extent that companies are likely to have governance controversies.

TABLE 3 Results for predicting controversies using the nine ML algorithms

	Precision	Recall	*F*-measure	Area under ROC curve	Area under PRC
Nearest neighbors	0.3620	0.1958	0.2507	0.5995	0.2564
Linear SVM	0.3023	**0.6885**	**0.4195**	0.7374	0.4226
RBF SVM	0.2526	**0.7820**	0.3817	0.7075	0.3737
Random forest	**0.7506**	0.1700	0.2756	**0.7787**	**0.5110**
Logistic regression	0.3387	0.6513	**0.4444**	0.7531	**0.4699**
ANN	**0.6186**	0.1851	0.2787	**0.7537**	0.4649
Gradient boosting	**0.6509**	0.2700	0.3777	**0.7846**	**0.5206**
Naive Bayes	0.2579	**0.7886**	0.3885	0.7155	0.3865
QDA	0.3017	0.5839	**0.3971**	0.7016	0.4028

Note: The three best performance numbers for each ML algorithm are in bold.
ANN: artificial neural network; QDA: quadratic discriminant analysis; RBF: radial basis function; SVM: support vector machine.

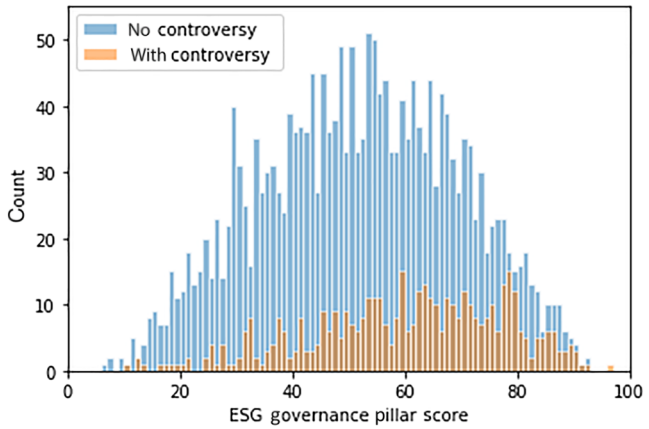

FIGURE 1 The distribution of governance controversies over companies and corporate governance performance ratings. Companies with at least one governance controversy in the 10-year window are classified as "with controversy" on the *y*-axis and companies with no controversies are "no controversy." The *x*-axis is the environmental, social, governance (ESG) rating. The total number of companies is 2,517

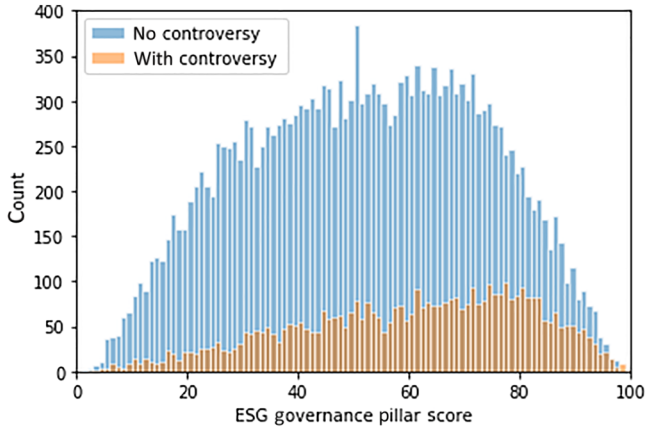

FIGURE 2 The distribution of governance controversies over company-years and corporate governance performance (CGP) ratings. Companies with at least one governance controversy in any year are classified as "with controversy," which means that a single company can be classified up to 10 times as "with controversy," possibly with different CGP ratings. Companies with no controversy in a particular year are classified as "no controversy," which means that any single company can be classified as "no controversy" up to 10 times, possibly with different CGP ratings. The *x*-axis is the environmental, social, governance (ESG) rating. The total number of company-years is 25,170

As the next step of the analysis, we examine how the models manage to learn the particular CGP indicator patterns typical of governance controversies. We compute the ROC curve and the PRC (Figures C1 and C2, respectively, in Appendix C). The purpose of these graphs is to evaluate the ranking and predictive performance of the algorithms. The ROC for each algorithm and class is shown in Figure C1. The interpretation of Figure C1 is that a curve closer to the top-left corner is preferable to a curve close to the diagonal. Comparisons of the ROC curves in Figure C1 with the results in Table 3 confirm that our interpretation of the results in Table 3 is consistent; that is, random forest, gradient boosting, and ANN have stronger performance than more simple models like logistic regression and linear support vector machine (SVM). The relative advantage of the complex algorithms emerges most clearly when considering the importance of precision to investment. Furthermore, the ROC curves provide information on predictive performance for both controversy and noncontroversy predictions. This is important, because SRI screens may be designed targeting either class. We find that the performance advantage of random forest, gradient boosting, and ANN extends to noncontroversy class prediction. Interestingly, relatively simple models like logistic regression and linear SVM perform well on the ROC curve but fall short if precision is required.

A PRC, as presented in Figure C2, describes the trade-off between precision and recall. Naturally, this curve should be as close as possible to the top-right corner of the graph and as far away as possible from the bottom-left corner, because it is beneficial to the model's usefulness to provide as high a precision as possible without sacrificing recall. The PRC also illustrates that the trade-off that can be attained if the model's own threshold probability is not used for determining the precision–recall combination, thus making the various algorithms more comparable. The graphs testify that random forest, gradient boosting, and ANN perform well, but also that logistic regression has a good overall trade-off. In particular, random forest and gradient boosting, which consider nonlinear attribute dependencies, perform well for the controversy cases (the yellow line), whereas radial basis function (RBF) SVM, nearest neighbors, naive Bayes, and quadratic discriminant analysis (QDA) perform significantly worse.

4.2 | Governance controversy prediction as CGP rating

To provide some perspective on how the controversy prediction model functions as a CGP rating, we study the correlation between the rating and the number of controversies a company has during the 10-year window. This correlation is important because our definition of CGP as compliance with governance legal and moral responsibilities and because governance controversies are company-specific instances of noncompliance. In line with the expectations, we find that the correlation is negative for all the prediction models, described in Figure 3.[2] On the *x*-axis, we have the number of controversies and on the *y*-axis the CGP rating for each company. The nine models all testify that companies with several controversies tend to be in the low end of the CGP rating. This suggests that the CGP indicator patterns of the controversy-prone companies are associated with the likelihood of having a controversy and that the assessment of this likelihood distinguishes between more or less controversy-prone companies.

Some algorithms may be ruled out due to their rating distributions. Nearest neighbors has a discrete distribution with few levels, which is obviously not appropriate for an ESG rating. Naive Bayes and

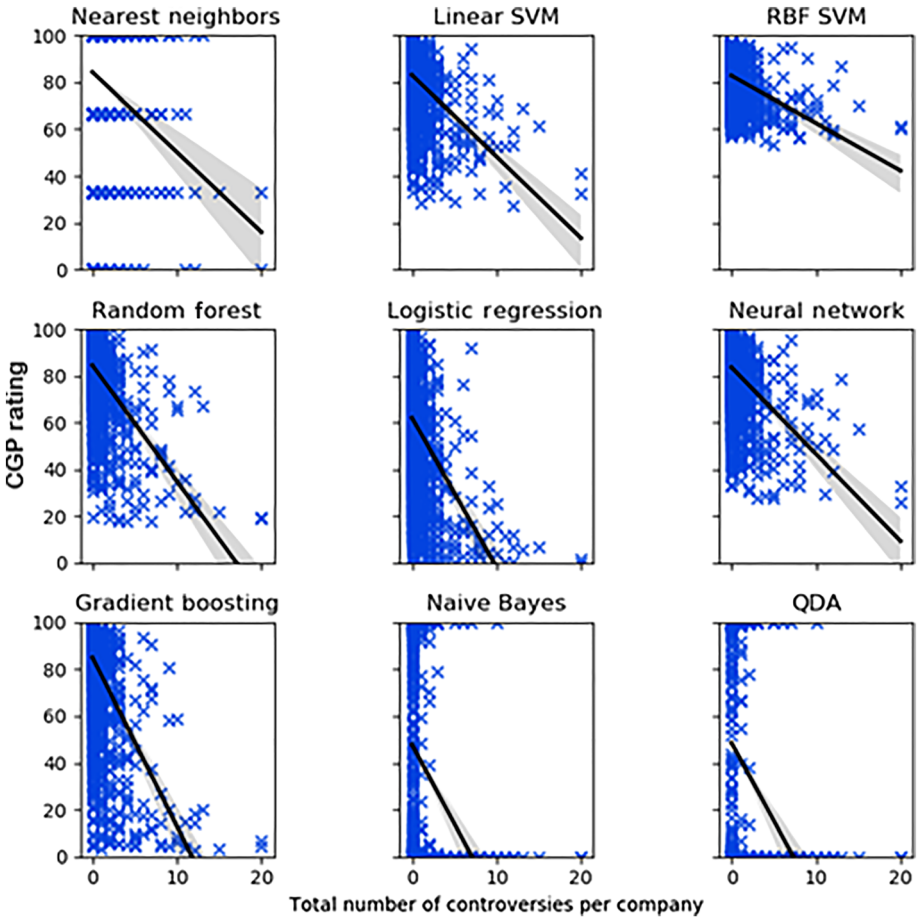

FIGURE 3 Correlations between the prediction-based corporate governance performance (CGP) rating and controversies. The graphs present correlations between CGP ratings produced by our mchine-learning algorithms and the number of years with controversies for each company. The figures show that our environmental, social, governance rating penalizes firms with more controversies by assigning lower ratings to them; that is, there is a negative correlation with all algorithms. QDA: quadratic discriminant analysis; RBF: radial basis function; SVM: support vector machine

QDA likewise produce discontinuous distributions. The smoothly distributed random forest generates a cautious rating with comparably few companies with a rating under 40 and with the number of companies increasing more sharply at about 50. The difference between the ML-based CGP ratings and a wrongdoing index (which describes the amount of fault committed by controversy companies) is illustrated by the random forest's rating of several companies in the range 60–100 despite some of these companies having between 5 and 15 controversies and by its rating of several companies in the range 20–40 despite them having had no controversies. These companies are rated low because they have indicator patterns similar to companies with controversies, not because they have had any controversies themselves.

4.3 | Model interpretation

The interest in explaining or interpreting ML models has been intensified due to the advancement of these models (Diakopoulos, 2016). An explanation is commonly defined as justification for an action or belief, with inferential schemas (Keil, 2006), but there is no conceptual clarity as to what explanation, interpretation, or transparency means (Lipton, 2018). We follow Lipton (2018) and distinguish between transparency as the technical understanding of a method's inner logic, whereas explanation or interpretation refers to the domain-specific conceptual, post hoc, interpretation of the model outputs, without much knowledge of computations. A common view is that linear models, such as logistic regression, have the advantage that they can be easily interpreted, whereas nonlinear models tend to be opaque (Johansson et al., 2011). Random forest, gradient boosting, and ANN would not be as interpretable as linear models because they are more complex and because their nonlinearity means that they represent variable interactions that make model weights virtually impossible to interpret. However, random forest and gradient boosting make use of trees for which post hoc interpretations can be developed. When data are so complex that linearity cannot be assumed, tree-based models are actually more interpretable than linear regressions (Lundberg et al., 2019). We discuss interpretation limitations of traditional ESG ratings and how our ML-based CGP rating can be made more interpretable.

We need to address the challenge with interpreting traditional ESG ratings, which is indicated by users feeling uncomfortable when interpreting ratings (Wong & Petroy, 2020). As experts attempt to understand ESG ratings, they attempt to reconcile model weights with expert knowledge, resulting in disbelief if the model does not appear consistent with their knowledge. Let us focus on two causes of inconsistency. One cause is computational. According to Lundberg et al. (2019), a linear model applied to nonlinear data will attribute importance to irrelevant features, thus not only generating inaccurate (invalid, false) predictions but also false explanation. The second is conceptual. Traditional ESG ratings do not provide any explicit

definition of the object they claim to measure, although they indicate relevant features (Berg et al., 2019), and appeal to the user's common-sense notion of sustainable performance as "doing good," as if it would be self-evident what it would be. This is problematic, because different ratings will result from different preference for good company features, as evidenced in recent studies (Berg et al., 2019; Chatterji et al., 2016; Semenova & Hassel, 2015). Because the weighting schemes used in traditional ESG ratings are selected by each rater based on their perceptions of feature materiality or importance, no logic can justify why one weighting scheme is superior to another (Callan & Thomas, 2009). The consistent divergence between ratings reported in previous research indicates that expert users may find interpretation of traditional ratings implausible.

The solution to the interpretation problem has three related components: Our use of (1) an explicit definition of CGP that corresponds to the most widespread conceptual development of ESG and CGP, and (2) our use of state-of-the-art, nonlinear, ML models, ensuring accuracy in predicting performance on the definition of CGP, for which (3) variable importance diagrams can be generated. Feature importance diagrams can be computed with several methods that resemble each other but we use the SHAP (Shapley additive explanations) values developed by Lundberg and Lee (2017), which is a unified framework for feature importance measures. We exemplify global variable importance for the random forest model in Figure 4.

A number of observations from the feature importance ranking for the global model (all companies) can be made. Several indicators of company size appear as important indicators in Figure 4, suggesting that the controversy prediction provides a size-biased rating, thus calling for size adjustment. Size adjustment cannot be done but can be accomplished in later versions of this methodology by, for example, using random forest in regression mode, with the dependent variable being a wrongdoing index (Fiaschi et al., 2020). Furthermore, in the rating application, the feature importance diagrams for single companies would be of interest.

The plausibility of the model, as well as of individual company ratings, can be assessed by an ESG expert by assessing whether the most important features for the rating appear to be reconcilable with the literature. In the global model, there are four key feature groups: top management, the board of directors, the audit function, and the company's stress level in terms of litigation and financial performance. The first group is an indicator of managers' financial incentives: "Total senior executive compensation," "Highest remuneration package,"

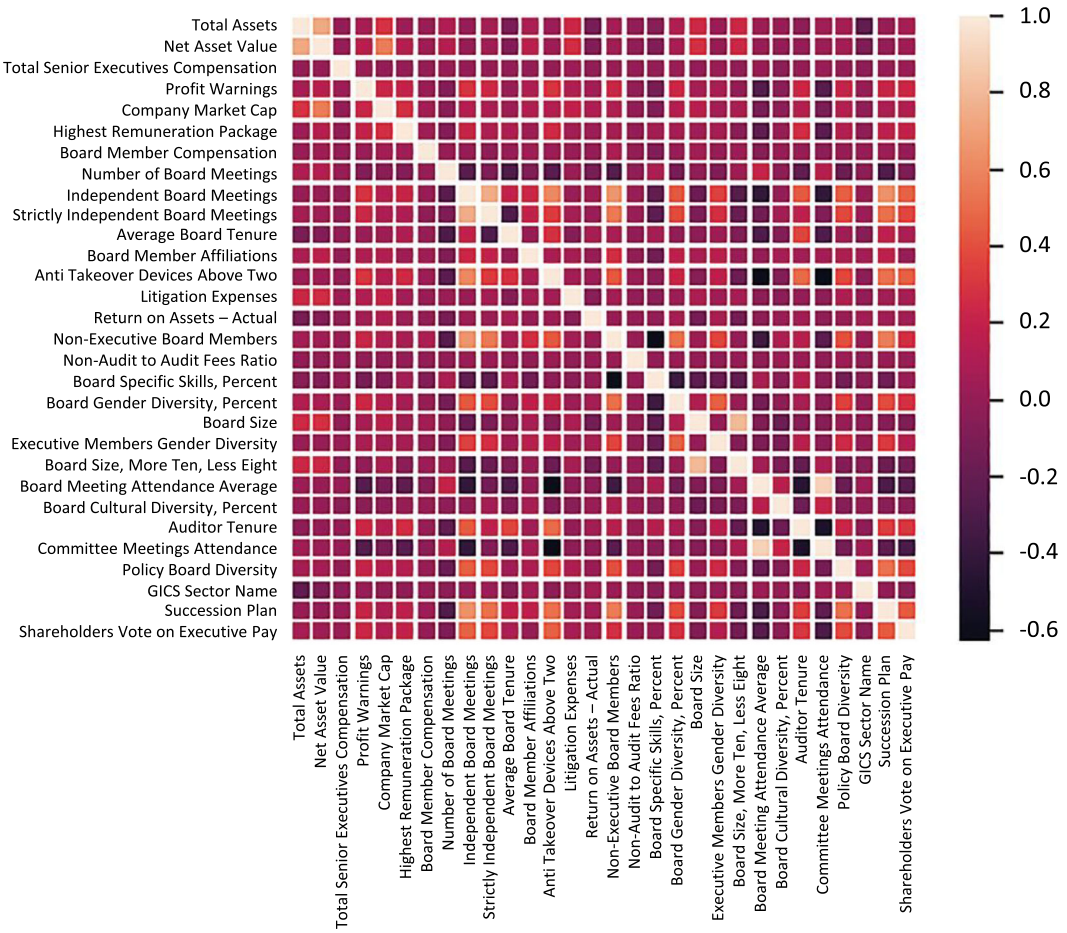

FIGURE 4 Indicator importance for the controversies prediction. The figure shows a ranking of the importance of individual indicators for the prediction of the likelihood that a company has a governance controversy. The higher on the list, the stronger the impact is on the classification. The colored graph also displays correlation between indicators from weak to strong

"Board member compensation," and a different kind of indicator "Profit warnings," which can be seen as a symptom of financial or operational instability. The compensation indicators measure monetary self-interest of top management. "Anti-takeover devices above two" is a factor, not directly related to compensation. It represents the extent that top management maintains control versus the equity market and can be seen as a sign of management's attempt to stay in power. Common such devices are poison pill, stock repurchase, and staggered board, all of which can be signs of inappropriate governance structures (Jory et al., 2015).

The focus on the board resonates well with the controversy literature: "Board member composition," "Number of board meetings," "Independent board members," "Strictly independent board members," "Average board tenure," "Board member affiliations," and so on. These show how the board operates and is composed in terms of integrity, frequency, and skills. The board of directors, which is a central governance mechanism, dominates the top 30 list with 14 entries. Another theme that appears high on the list is the audit function with indicators of well-known governance issues (e.g. "Auditor tenure" and "Non-audit to audit fees-ratio"). The last of the fourth group of indicators appears to represent the stress level of the company. It is consistent with the governance controversies literature because difficulties meeting expectations is a key driver of governance controversies. Thus, "Litigation expenses" and "Return on assets" represent the level of financial challenge the corporate governance structures need to deal with. For example, companies struggling with defending immaterial rights or companies themselves the subject of heavy lawsuits are, per definition, in a state of emergency that is either caused by or may cause inappropriate governance. The same applies to companies who, for example, have a low or decreasing return on assets. Finally, the model considers "GICS sector name" (Global Industry Classification Standard), suggesting that the likelihood of having a controversy varies substantially between sectors.

In a rating application the use of SHAP diagrams would assist the post hoc interpretation of the ratings because the ESG expert would controversy risk. This is an advantage compared with a subjective rating, because feature importance measures would not represent the subjective assessments of a rater but their relative importance determined by the ML model based on collective level scrutiny of companies' compliance with governance responsibilities. Our solution replaces the endless debate about the meaning of "doing good" with the challenge of providing a broad enough set of CGP indicators and as comprehensive coverage of governance controversies as possible. Lipton's (2018) conceptual distinction between transparency and interpretation accentuates that even perfect transparency is of little use if the model is inappropriate or the underlying construct is not defined.

5 | DISCUSSION AND CONCLUSIONS

Traditional CGP ratings are constructed as ad-hoc-weighted arithmetic averages of indicators, and these aggregated ratings lack convergent validity (Berg et al., 2019; Chatterji et al., 2016; Christensen et al., 2021; Semenova & Hassel, 2015; Trumpp et al., 2013). Despite this concerning evidence, there is little discussion in the literature of what ratings measure or should measure. Summarized succinctly by Wood (2010), ESG tends to be defined in empirical measurement as "doing good," with the effect that ESG is defined by rating companies' subjective assessments of the relative wrongdoing of the many ESG features used in the ratings. This arbitrary and idiosyncratic element to ratings destabilizes or proliferates the empirical representation of the ESG concept and persistently fuels the inconclusive evidence of the financial effects of ESG. Accordingly, there is a need to design an ESG rating methodology with better promise of validity and with less rater dependency (Berg et al., 2019; Chatterji et al., 2016; Chen & Delmas, 2011; Christensen et al., 2021; Kotsantonis & Serafeim, 2019).

Another limitation with traditional ratings is that ad hoc indicator weighting does not produce ratings that match the primary information demand of institutional investors. There are many variants of investment strategy, but, according to recent findings, institutional investors prefer to invest in companies that comply with ESG responsibilities (Nofsinger et al., 2019) because such investments protect investor reputation (Krueger et al., 2020; Nofsinger et al., 2019; Zavyalova et al., 2012). Avoiding controversial, noncomplying portfolio companies also contributes to fulfilling institutional investors' fiduciary duties toward capital owners and toward society and is ultimately driven by expected favorable financial results.

Our findings suggest a solution to the arbitrary weighting scheme problem described in the ESG rating literature (Berg et al., 2019; Callan & Thomas, 2009; Chen & Delmas, 2011; Christensen et al., 2021; Kotsantonis & Serafeim, 2019; Mitchell et al., 1997) and to the information needs of institutional investors (Nofsinger et al., 2019). The most important component of our rating methodology is to use a holistic label of CGP that enables an assessment relative to a standard, and therefore the estimation of a large set of CGP indicators' contributions to a company's likelihood of performing according to the standard.

Having defined CGP as the comprehensive compliance with governance responsibilities, we investigate the possibility of using ML to estimate a model of companies' compliance with governance responsibilities. Such a model would have to predict whether a company has a controversy by looking at governance indicator patterns alone. The results from our ML experiments suggest that governance indicator patterns contain information on governance behaviors associated with governance controversies.

In addition to the advantage of being able to predict governance controversies, our rating methodology offers an important contribution compared with traditional ratings. Whereas traditional ESG-type ratings use manually selected subjective feature indicator weights with little or no validity, we use data-driven ML to assess feature importance, enabling the equivalent of a low-bias (Lundberg et al., 2019), nonsubjective indicator importance weighting scheme. Our rating methodology is therefore far less rater dependent than is traditional ESG rating methodology, which means that a widespread

use of our methodology would eliminate much of the critique of current ESG ratings. The difference in meaning between, for example, "Board member compensation" and "Inappropriate board member compensation" is not a matter of subjective assessment, but of fact. The same applies to the difference between compliance with CGP responsibilities and noncompliance.

The results from our ML experiments suggest a methodology to construct CGP ratings with predictive validity, because model weights are the outputs of the ML algorithm's learning from the labeled instances. Our methodology also makes ratings easy to interpret conceptually, because feature importance is the contribution of a feature value to the estimation of controversy risk. The plausibility of individual ratings can be assessed relative to the substantial body of accounting research on antecedents of governance failures. As a contrast, the accounting literature has little to offer an ESG expert attempting to interpret a traditional ESG rating with references to financially material ESG risks (Christensen et al., 2021), which are conceptually unexplained and for which there are no valid and reliable estimates.

A practical finding is that three ML algorithms with exceptional complexity handling capacity (i.e., random forest, gradient boosting, and ANN) are superior to the other estimators in learning to estimate CGP. The best prediction of governance controversies is achieved with methods capable of representing nonlinearity, indicator interaction, and multidimensionality. This suggests that the underlying construct poses a challenge that may be difficult to address with explanatory modeling and traditional statistical methods.

We recognize the limitations that several steps of our modeling cause to the interpretation of our results. We make a number of simplifying assumptions described in Section 3, such as treating the 10-year time window as a single event and defining indicator values using their averages. One simplifying assumption is that we do not take into account that the large companies in our sample may suffer more media scrutiny than the smaller ones do. Future development of this method into an actual rating can adjust for this condition by constructing a wrongdoing index that normalizes corporate wrongdoing with a method that measures the relative media exposure of companies (Fiaschi et al., 2020). Another limitation that relates to a wrongdoing index is that we treat all governance controversies as though they would be of equal nature and equal importance. Future research may explore how this limitation could be avoided. Research on wrongdoing indices, as well as recent research on the short-term financial effects of various kinds of controversies (Cui & Docherty, 2020), could provide some insights into this problem. Further research on our method may also explore the use of different kinds of databases, such as media databases, through which a rich coverage of governance controversies may be obtained.

ACKNOWLEDGMENTS

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

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from Refinitiv Eikon. Restrictions apply to the availability of these data, which were used under license for this study.

ORCID

Jan Svanberg https://orcid.org/0000-0002-4436-5920
Tarek Rana https://orcid.org/0000-0003-2050-7004

ENDNOTES

[1] ESG risks refer to a company's financial material risks caused by noncompliance with standards for ESG performance. Such standards include binding ESG norms, such as national laws on pollution, labor law, and company law, but also include international standards such as human rights tractates, some of the European Union Agenda 2030 goals, and international accounting standards. We refer to obligatory ESG norms as the norms that, if not complied with, may cause an ESG controversy. Our definition of the compulsory ESG norms is therefore related to what society through its media views as most important and most obligatory for companies to comply with. Discretionary sustainability norms (see Wood, 2010), which refer to acts that companies are praised for conducting but are not harshly criticized for not conducting, fall outside our concept.

[2] In contrast to the plausible negative correlation, the correlation between Refinitiv Eikon governance ratings and the number of controversies is even slightly positive.

REFERENCES

Abhayawansa, S., & Tyagi, S. (2021). Sustainable investing: The black box of environmental, social and governance (ESG) ratings. *Journal of Wealth Management*, *24*(1), 49–54. https://doi.org/10.3905/jwm.2021.1.130

Alpaydin, E. (2010). *Introduction to machine learning* (2nd ed.). The MIT Press.

Amir, A. Z., & Serafeim, G. (2018). Why and how investors use ESG information: Evidence from a global survey. *Financial Analysts Journal*, *74*(3), 87–103. https://doi.org/10.2469/faj.v74.n3.2

Anagnostopoulos, F., Liolios, E., Persefonis, G., Slater, J., Kafetsios, K., & Niakas, D. (2012). Physician burnout and patient satisfaction with consultation in primary health care settings: evidence of relationships from a one-with-many design. *Journal of Clinical Psychology in Medical Settings*, *19*, 401–410.

Arnold, B., & de Lange, P. (2004). Enron: An examination of agency problems. *Critical Perspectives on Accounting*, *15*(6–7), 751–765. https://doi.org/10.1016/j.cpa.2003.08.005

Asthana, S. C., Balsam, S., & Krishnan, J. (2010). Corporate governance, audit firm reputation, auditor switches, and client stock price reactions: The Andersen experience. *International Journal of Auditing*, *14*(3), 274–293. https://doi.org/10.1111/j.1099-1123.2010.00417.x

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 Systems with Applications*, *38*(3), 1364–1373. https://doi.org/10.1016/j.eswa.2010.07.033

Benabou, R., & Tirole, J. (2010). Individual and corporate social responsibility. *Economica*, *77*(305), 1–19. https://doi.org/10.1111/j.1468-0335.2009.00843.x

Berg, F., Kölbel, J., & Rigobon, R. (2019). Aggregate confusion: The divergence of ESG ratings, available from SSRN. https://doi.org/10.2139/ssrn.3438533

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W., & Siddique, A. (2016). Risk and risk management in the credit card industry. *Journal of Banking and Finance*, *72*, 218–239. https://doi.org/10.1016/j.jbankfin.2016.07.015

Callan, S. J., & Thomas, J. M. (2009). Corporate financial performance and corporate social performance: An update and reinvestigation. *Corporate Social Responsibility and Environmental Management*, *16*(2), 61–78. https://doi.org/10.1002/csr.182

Canada, J., Sutton, S. G., & Randel Kuhn, J. (2009). The pervasive nature of IT controls: An examination of material weaknesses in IT controls and audit fees. *International Journal of Accounting & Information Management*, *17*(1), 106–119. https://doi.org/10.1108/18347640910967753

Capelle-Blancard, G., & Petit, A. (2019). Every little helps? ESG news and stock market reaction. *Journal of Business Ethics*, *157*(2), 543–565. https://doi.org/10.1007/s10551-017-3667-3

Carroll, A. B. (1979). A three-dimensional conceptual model of corporate performance. *Academy of Management Review*, *4*(4), 497–505. https://doi.org/10.5465/amr.1979.4498296

Carroll, A. B. (1999). Corporate social responsibility: Evolution of a definitional construct. *Business and Society*, *38*(3), 268–295. https://doi.org/10.1177/000765039903800303

Changchit, C., & Holsapple, C. W. (2004). The development of an expert system for managerial evaluation of internal controls. *Intelligent Systems in Accounting, Finance & Management*, *12*(2), 103–120. https://doi.org/10.1002/isaf.246

Chatterji, A., & Levine, D. (2006). Breaking down the wall of codes: Evaluating non-financial performance measurement. *California Management Review*, *48*(2), 29–51. https://doi.org/10.2307/41166337

Chatterji, A. K., Durand, R., Levine, D. I., & Touboul, S. (2016). Do ratings of firms converge? Implications for managers, investors and strategy researchers. *Strategic Management Journal*, *37*(8), 1597–1614. https://doi.org/10.1002/smj.2407

Chen, C. M., & Delmas, M. (2011). Measuring corporate social performance: An efficiency perspective. *Production and Operations Management*, *20*(6), 789–804. https://doi.org/10.1111/j.1937-5956.2010.01202.x

Chen, M., & Mussalli, G. (2020). An integrated approach to quantitative ESG investing. *The Journal of Portfolio Management Ethical Investing*, *46*, 65–74.

Christensen, D. M., Serafeim, G., & Sikochi, A. (2021). Why is corporate virtue in the eye of the beholder? The case of ESG ratings. *The Accounting Review*, *97*, 147–175. https://doi.org/10.2308/tar-2019-0506

Collopy, F., Adya, M., & Armstrong, J. S. (1994). Principles for examining predictive validity: The case of information systems spending forecasts. *Information Systems Research*, *5*(2), 170–179. https://doi.org/10.1287/isre.5.2.170

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297.

Crilly, D., Zollo, M., & Hansen, M. T. (2012). Faking it or muddling through? Understanding decoupling in response to stakeholder pressures. *Academy of Management Journal*, *55*(6), 1429–1448. https://doi.org/10.5465/amj.2010.0697

Cui, B., & Docherty, P. (2020). Stock price overreaction to ESG controversies. Monash Business School Working Paper. https://www.monash.edu/business/mcfs/our-research/stock-price-overreaction-to-esg-controversies

Deegan, C. (2002). Introduction: The legitimising effect of social and environmental disclosures—A theoretical foundation. *Accounting, Auditing & Accountability Journal*, *15*(3), 282–311. https://doi.org/10.1108/09513570210435852

Deegan, C. (2019). Legitimacy theory: Despite its enduring popularity and contribution, time is right for a necessary makeover. *Accounting, Auditing and Accountability Journal*, *32*(8), 2307–2329. https://doi.org/10.1108/AAAJ-08-2018-3638

Delmas, M., & Blass, V. D. (2010). Measuring corporate environmental performance: The trade-offs of sustainability ratings. *Business Strategy and the Environment*, *19*(4), 245–260. https://doi.org/10.1002/bse.676

Delmas, M. A., Etzion, D., & Nairn-Birch, N. (2013). Triangulating environmental performance: What do corporate social responsibility ratings really capture? *Academy of Management Perspectives*, *27*(3), 255–267. https://doi.org/10.5465/amp.2012.0123

DeTienne, K. B., DeTienne, D. H., & Joshi, S. A. (2003). Neural networks as statistical tools for business researchers. *Organizational Research Methods*, *6*, 236–265.

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM*, *59*(2), 56–62. https://doi.org/10.1145/2844110

Du, K. L., & Swamy, M. N. S. (2006). *Neural networks in a softcomputing framework*. Springer-Verlag.

Dyck, A., Lins, K. V., Roth, L., & Wagner, H. F. (2019). Do institutional investors drive corporate social responsibility? International evidence. *Journal of Financial Economics*, *131*(3), 693–714. https://doi.org/10.1016/j.jfineco.2018.08.013

Ehrenberg, A. S. C., & Bound, J. A. (1993). Predictability and prediction. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *156*(2), 167–206. https://doi.org/10.2307/2982727

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

Fanning, K. M., & Cogger, K. O. (1998). Neural network detection of management fraud using published financial data. *Intelligent Systems in Accounting, Finance & Management*, *7*, 21–41. https://doi.org/10.1002/(SICI)1099-1174(199803)7:1

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. *Journal of World Business*, *52*(4), 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. *Business Horizons*, *63*(3), 287–299. https://doi.org/10.1016/j.bushor.2019.09.004

Ghritlahre, H. K., & Prasad, R. K. (2018). Investigation of thermal performance of unidirectional flow porous bed solar air heater using MLP, GRNN, and RBF models of ANN technique. *Thermal Science and Engineering Progress*, *6*, 226–235.

Goel, S., & Gangolly, J. (2012). Beyond the numbers: Mining the annual reports for hidden cues indicative of financial statement fraud. *Intelligent Systems in Accounting, Finance and Management*, *19*(2), 75–89. https://doi.org/10.1002/isaf.1326

Graves, S. B., & Waddock, S. A. (1994). Institutional owners and corporate social performance. *Academy of Management Journal*, *37*(4), 1034–1046. https://doi.org/10.5465/256611

Griffin, J. J., & Mahon, J. F. (1997). The corporate social performance and corporate financial performance debate: Twenty-five years of incomparable research. *Business and Society*, *36*(1), 5–31. https://doi.org/10.1177/000765039703600102

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. In C. Schulz & D. Liew (Eds.), *2015 Imperial College computing student workshop (ICCSW 2015)* (pp. 21–28). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Gurbaxani, V., & Mendelson, H. (1990). An integrative model of information systems spending growth. *Information Systems Research*, *1*(1), 23–46. https://doi.org/10.1287/isre.1.1.23

Gurbaxani, V., & Mendelson, H. (1994). Modeling vs. forecasting: The case of information systems spending. *Information Systems Research*, *5*(2), 180–190. https://doi.org/10.1287/isre.5.2.180

Hartzmark, S. M., & Sussman, A. B. (2019). Do investors value sustainability? A natural experiment examining ranking and fund flows. *Journal of Finance*, *74*(6), 2789–2837. https://doi.org/10.1111/jofi.12841

Hillman, A. J., & Keim, G. D. (2001). Shareholder value, stakeholder management, and social issues: What's the bottom line? *Strategic Management Journal*, *22*(2), 125–139. https://doi.org/10.1002/1097-0266(200101)22:2<125::AID-SMJ150>3.0.CO;2-H

Hu, L.-Y., Huang, M.-W., Ke, S.-W., & Tsai, S.-W. (2016). The distance function effect on k-nearest neighbor classification for medical datasets. *SpringerPlus*, *5*(1304), 2–9.

Johansson, U., Sönströd, C., Norinder, U., & Boström, H. (2011). Trade-off between accuracy and interpretability for predictive in silico modeling. *Future Medicinal Chemistry*, *3*(6), 647–663. https://doi.org/10.4155/fmc.11.23

Jory, S. R., Ngo, T. N., Wang, D., & Saha, A. (2015). The market response to corporate scandals involving CEOs. *Applied Economics*, *47*(17), 1723–1738. https://doi.org/10.1080/00036846.2014.995361

Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*, *57*, 227–254. https://doi.org/10.1146/annurev.psych.57.102904.190100

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking and Finance*, *34*(11), 2767–2787. https://doi.org/10.1016/j.jbankfin.2010.06.001

Kotsantonis, S., & Serafeim, G. (2019). Four things no one will tell you about ESG data. *Journal of Applied Corporate Finance*, *31*(2), 50–58. https://doi.org/10.1111/JACF.12346

Krueger, P., Sautner, Z., & Starks, L. (2020). Importance of climate risks for institutional investors. *The Review of Financial Studies*, *33*(3), 1067–1111. https://academic.oup.com/rfs/article-abstract/33/3/1067/5735302

Krüger, P. (2015). Corporate goodness and shareholder wealth. *Journal of Financial Economics*, *115*(2), 304–329. https://doi.org/10.1016/j.jfineco.2014.09.008

Laha, D., Ren, Y., & Suganthan, P. N. (2015). Modeling of steelmaking process with effective machine learning techniques. *Expert Systems with Applications*, *42*(10), 4687–4696. https://doi.org/10.1016/j.eswa.2015.01.030

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31–57. https://doi.org/10.1145/3236386.3241340

Longinidis, P., & Symeonidis, P. (2013). Corporate dividend policy determinants: Intelligent versus a traditional approach. *Intelligent Systems in Accounting, Finance and Management*, *20*(2), 111–139. https://doi.org/10.1002/isaf.1338

Lu, H. J., Zou, N., Jacobs, R., Afflerbach, B., Lu, X. G., & Morgan, D. (2019). Error assessment and optimal cross-validation approaches in machine learning applied to impurity diffusion. *Computational Materials Science*, *169*, 109075. https://doi.org/10.1016/j.commatsci.2019.06.010

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2019). Explainable AI for trees: From local explanations to global understanding. *ArXiv*. http://arxiv.org/abs/1905.04610

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, & R. Fergus (Eds.), *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 4766–4775). Curran Associates.

Maione, C., Batista, B. L., Campiglia, A. D., Barbosa, F., & Barbosa, R. M. (2016). Classification of geographic origin of rice by data mining and inductively coupled plasma mass spectrometry. *Computers and Electronics in Agriculture*, *121*, 101–107.

Mattingly, J. E., & Berman, S. L. (2006). Measurement of corporate social action: Discovering taxonomy in the Kinder Lydenburg Domini ratings data. *Business and Society*, *45*(1), 20–46. https://doi.org/10.1177/0007650305281939

Mitchell, R. K., Agle, B. R., & Wood, D. J. (1997). Toward a theory of stakeholder identification and salience: Defining the principle of who and what really counts. *The Academy of Management Review*, *22*(4), 853–886. https://doi.org/10.2307/259247

Mustapha, I. B., & Saeed, F. (2016). Bioactive molecule prediction using extreme gradient boosting. *Molecules*, *21*(983), 1–11.

Nofsinger, J. R., Sulaeman, J., & Varma, A. (2019). Institutional investors and corporate social responsibility. *Journal of Corporate 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? *British Accounting Review*, *50*(4), 379–401. https://doi.org/10.1016/j.bar.2017.10.003

O'leary, D. E. (1998). Using neural networks to predict corporate failure. *Intelligent Systems in Accounting, Finance & Management*, *7*, 187–197. https://doi.org/10.1002/(SICI)1099-1174(199809)7:3

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Rumelhart, D. E., Widrow, B., & Lehr, M. A. (1994). The basic ideas in neural networks. *Communications of the ACM*, *37*, 87–92.

Safa, M., & Samarasinghe, S. (2011). Determination and modelling of energy consumption in wheat production using neural networks: A case study in Canterbury province, New Zealand. *Energy*, *36*(8), 5140–5147. https://doi.org/10.1016/j.energy.2011.06.016

Sandberg, J. (2011). Socially responsible investment and fiduciary duty: Putting the Freshfields report into perspective. *Journal of Business Ethics*, *101*(1), 143–162. https://doi.org/10.1007/s10551-010-0714-8

Semenova, N., & Hassel, L. G. (2015). On the validity of environmental performance metrics. *Journal of Business Ethics*, *132*(2), 249–258. https://doi.org/10.1007/s10551-014-2323-4

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330

Shrivastava, P., & Addas, A. (2014). The impact of corporate governance on sustainability performance. *Journal of Sustainable Finance & Investment*, *4*, 21–37.

Sigrist, F., & Hirnschall, C. (2019). Grabit: Gradient tree-boosted tobit models for default prediction. *Journal of Banking and Finance*, *102*, 177–192. https://doi.org/10.1016/j.jbankfin.2019.03.004

Sözen, A. (2009). Future projection of the energy dependency of Turkey using artificial neural network. *Energy Policy*, *37*, 4827–4833.

Surroca, J., Tribó, J. A., & Zahra, S. A. (2013). Stakeholder pressure on MNEs and the transfer of socially irresponsible practices to subsidiaries. *Academy of Management Journal*, *56*(2), 549–572. https://doi.org/10.5465/amj.2010.0962

Telmoudi, F., El Ghourabi, M., & Limam, M. (2011). RST-GCBR-clustering-based RGA-SVM model for corporate failure prediction. *Intelligent Systems in Accounting, Finance and Management*, *18*(2–3), 105–120. https://doi.org/10.1002/isaf.323

Trumpp, C., Endrikat, J., Zopf, C., & Guenther, E. (2013). Definition, conceptualization, and measurement of corporate environmental performance: A critical examination of a multidimensional construct. *Journal of Business Ethics*, *126*(2), 185–204. https://doi.org/10.1007/s10551-013-1931-8

Van Maanen, J., Sørensen, J. B., & Mitchell, T. R. (2007). The interplay between theory and method. *Academy of Management Review*, *32*(4), 1145–1154. https://doi.org/10.5465/AMR.2007.26586080

Vaughan, T. S., & Berry, K. E. (2005). Using Monte Carlo techniques to demonstrate the meaning and implications of multicollinearity. *Journal of Statistics Education*, *13*(1), 1–9. https://doi.org/10.1080/10691898.2005.11910640

Wong, C., & Petroy, E. (2020). *Rate the raters 2020: Investor survey and interview results*. https://sustainability.com/wp-content/uploads/2020/03/sustainability-ratetheraters2020-report.pdf

Wood, D. J. (1991). Corporate social performance revisited. *The Academy of Management Review*, *16*(4), 691–718. https://doi.org/10.2307/258977

Wood, D. J. (2010). Measuring corporate social performance: A review. *International Journal of Management Reviews*, *12*(1), 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 Systems with Applications*, *36*(2 Part 2), 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. *Information Sciences*, *254*, 98–110.

Zavyalova, A., Pfarrer, M. D., Reger, R. K., & Shapiro, D. L. (2012). Managing the message: The effects of firm actions and industry spillovers on media coverage following wrongdoing. *Academy of Management Journal*, *55*(5), 1079–1101. https://doi.org/10.5465/amj.2010.0608

Zengul, F. D., Byrd, J. D., Oner, N., Edmonds, M., & Savage, A. (2019). Exploring corporate governance research in accounting journals through latent semantic and topic analyses. *Intelligent Systems in Accounting, Finance and Management*, *26*(4), 175–192. https://doi.org/10.1002/isaf.1461

How to cite this article: Svanberg, J., Ardeshiri, T., Samsten, I., Öhman, P., Neidermeyer, P. E., Rana, T., Semenova, N., & Danielson, M. (2022). Corporate governance performance ratings with machine learning. *Intelligent Systems in Accounting, Finance and Management*, 1–19. https://doi.org/10.1002/isaf.1505

APPENDIX A.

THE COMPLETE LIST OF VARIABLES ACQUIRED

Board Functions Policy.
Corporate Governance Board Committee.
Nomination Board Committee.
Audit Board Committee.
Compensation Board Committee.
Board Structure Policy.
Policy Board Size.
Policy Board Independence.
Policy Board Diversity.
Policy Board Experience.
Policy Executive Compensation Performance.
Policy Executive Compensation ESG Performance.
Policy Executive Retention.
Compensation Improvement Tools.
Internal Audit Department Reporting.
Succession Plan.
External Consultants.
Audit Committee Independence.
Audit Committee Mgt Independence.
Audit Committee Expertise.
Audit Committee Non-Executive Members.
Compensation Committee Independence.
Compensation Committee Mgt Independence.
Compensation Committee Non-Executive Members.
Nomination Committee Independence.
Nomination Committee Mgt Independence.
Nomination Committee Involvement.
Nomination Committee Non-Executive Members.
Board Attendance.
Number of Board Meetings.
Board Meeting Attendance Average.
Committee Meetings Attendance Average.
Board Structure Type.
Board Size More Ten Less Eight.
Board Size.
Board Background and Skills.
Board Gender Diversity, Percent.
Board Specific Skills, Percent.
Average Board Tenure.
Non-Executive Board Members.
Independent Board Members.
Strictly Independent Board Members.
CEO-Chairman Separation.
CEO Board Member.
Chairman is ex-CEO.
Board Member Affiliations.
Board Individual Re-election.
Board Member Membership Limits.
Board Member Term Duration.
Executive Compensation Policy.
Executive Individual Compensation.
Total Senior Executives Compensation.
Highest Remuneration Package.
CEO Compensation Link to TSR.
Executive Compensation LT Objectives.
Sustainability Compensation Incentives.
Shareholder Approval Stock Compensation Plan.
Board Member Compensation.
Board Member LT Compensation Incentives.
Board Cultural Diversity, Percent.
Executive Members Gender Diversity, Percent.
Chief Diversity Officer.
Executives Cultural Diversity.
Shareholder Rights Policy.

Policy Equal Voting Right.
Policy Shareholder Engagement.
Different Voting Right Share.
Equal Shareholder Rights.
Voting Cap.
Voting Cap Percentage.
Minimum Number of Shares to Vote.
Director Election Majority Requirement.
Shareholders Vote on Executive Pay.
Public Availability Corporate Statutes.
Veto Power or Golden share.
State Owned Enterprise SOE.
Anti Takeover Devices Above Two.
Poison Pill.
Poison Pill Adoption Date.
Poison Pill Expiration Date.
Unlimited Authorized Capital or Blank Check.
Classified Board Structure.
Staggered Board Structure.
Supermajority Vote Requirement.
Golden Parachute.
Limited Shareholder Rights to Call Meetings.
Elimination of Cumulative Voting Rights.
Pre-emptive Rights.
Company Cross Shareholding.
Confidential Voting Policy.
Limitation of Director Liability.
Shareholder Approval Significant Transactions.
Fair Price Provision.
Limitations on Removal of Directors.
Advance Notice for Shareholder Proposals.
Advance Notice Period Days.
Written Consent Requirements.
Expanded Constituency Provision.
Earnings Restatement.
Profit Warnings.
Litigation Expenses.
Non-audit to Audit Fees Ratio.
Auditor Tenure.
CSR Sustainability Committee.
Integrated Strategy in MD&A.
Global Compact Signatory.
Stakeholder Engagement.
CSR Sustainability Reporting.
GRI Report Guidelines.
CSR Sustainability Report Global Activities.
CSR Sustainability External Audit.
CSR Sustainability External Auditor Name.
ESG Reporting Scope.
UNPRI Signatory.

APPENDIX B.

ML ALGORITHMS

Nearest neighbors

Nearest neighbors is an instance or distance-based classifier that relies on a distance or similarity measure to predict controversies. It differs from other ML algorithms in that the nearest neighbor algorithm does not have a training phase. It records all firms and their CGP indicators and controversies and queries the database and finding the *k* closest firms, in terms of CGP indicators, and calculates the probability of a controversy as a fraction of the *k* closest firms' controversy status. In this study, we employ the simple Euclidean distance as our measure of distance (Hu et al.,

Linear and RBF SVMs

This study uses CGP estimators separating low from high CGP using the specified broad set of indicators. A linear SVM separates data using an $(n - 1)$-dimensional plane. Introduced by Cortes and Vapnik (1995), it has been effective for solving many pattern-recognition and prediction problems (e.g., stock and bankruptcy prediction). This method, however, requires that the data are linearly separable (Xu et al., 2009). If the data are not linearly separable then an RBF SVM can be employed instead. The RBF is a kernel function to SVM that is used to enable nonlinear classification, and a form of kernels that provide windows for mapping the nonlinearity in the *n*-dimensional original space onto a higher order space in which the classifier is linear.

Random forest

The random forest technique was first developed by Breiman (2001) and has since then become the state-of-the-art ML algorithm in many applications. It constructs an ensemble model from decision trees that are trained using random samples of training data. To increase the predictive performance, each tree is also constructed by sampling a limited number of indicators at each node. The controversy prediction is the trees' majority vote. The use of majority vote among the trees avoids data overfitting and provides precise forecasts (Breiman, 2001).

Governance indicators contain a lot of noise due to greenwashing and lack of accounting standards, which makes the prediction of controversies more difficult. Random forest, however, works well with outliers and noise in the training set (Yeh et al., 2014). A further benefit compared with other methods is that it computes the importance of each indicator for the classification results, which can be displayed in graphs for rating interpretation (see Maione et al., 2016).

Logistic regression

The logistic regression, also used in statistics, classifies data after estimating the coefficients of a regression equation. Logistic regression relates controversies to the CGP indicators. Its goal is to find the best fit set of regression coefficients. As a classifier, each feature is multiplied by the estimated regression coefficient and then added together. The result is passed through a sigmoid function, which produces the binary output.

Artificial neural network

- ANNs are composed of multiple stacked layers with an output layer consisting of a logistic, softmax, or linear regression model. ANNs can approximate any linear or nonlinear function. In our case, the ANN relates CGP indicators to controversies using multiple layers with a logistic regression output layer. ANNs consist of a large number of artificial neurons, which are densely interconnected, simple computational elements that operate in parallel. ANNs cope with noisy data, generalize well, learn nonlinear relationships, and training data may have any distribution. The neurons are connected by corresponding links between layers with numeric weights, which represent long-term memory in the ANN (Ekonomou, 2010). The advantages of ANNs over traditional statistical methods are well documented: better forecasting (Rumelhart et al., 1994), better operation on fuzzy and complex data (Ghritlahre and Prasad, 2018), and better coping with unknown interactions (Azadeh et al., 2011; DeTienne et al., 2003; Du and Swamy, 2006; Safa & Samarasinghe, 2011; Sözen, 2009).

Gradient boosting

Gradient boosting combines multiple weak models into a strong ensemble model by reweighting the training data of CGP indicators to focus the learning on those cases that the algorithm cannot predict correctly. Gradient boosting defines a loss function and uses the gradient in the loss function to reweight the cases in order to focus on the misclassifications using the logistic loss. Gradient boosting is a high-performing classifier for a wide range of tasks (compare Mustapha & Saeed, 2016; Sigrist & Hirnschall, 2019).

Naïve Bayes

Naive Bayes classifiers consist of simple probabilistic classifiers developed from Bayes' theorem. Naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. Naive Bayes classifiers are effective in a supervised learning setting despite their constraining assumptions. The naive Bayes classifier computes the conditional probability of a controversy and a noncontroversy given a set of CGP indicators. Under the (naive) assumption that, in our case, the CGP indicators are independent, the naive Bayes classifier can be expressed as the conditional probability of a controversy multiplied by the product of the conditional probability of each CGP indicator given a controversy (see Gulo, Rúbio, Tabassum, & Prado, 2015).

Quadratic discriminant analysis

QDA creates a model based on the conditional densities of the data and generates a quadratic decision boundary. As a contrast to linear discriminant analysis, QDA does not assume equal class covariance. QDA captures nonlinear dependencies between indicators and labels (Anagnostopoulos et al., 2012).

APPENDIX C.

Figures C1 and C2

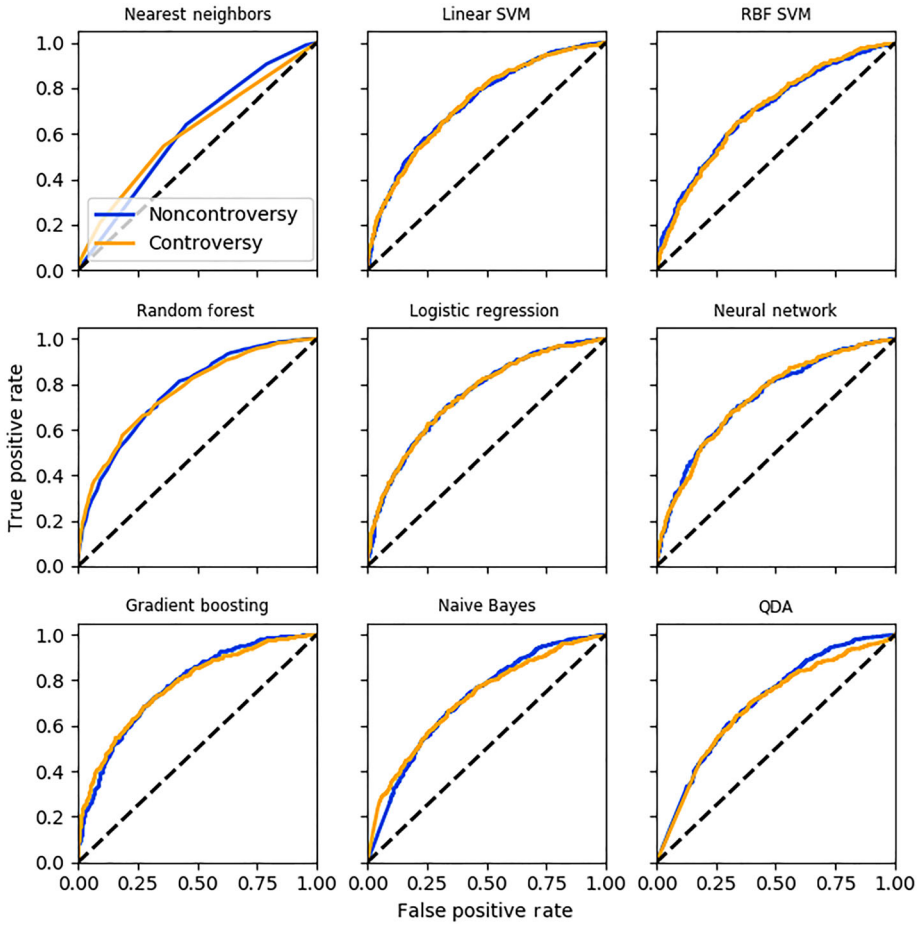

FIGURE C1 Area under receiver operating characteristic (ROC) curve for the nine learning algorithms. The blue line represents the ROC for predicting noncontroversy and the yellow line represents the ROC for predicting controversy. Note that classifiers that produce ROC curves which lie above the horizontal dashed line provide predictions that are better than random guessing. QDA: quadratic discriminant analysis; RBF: radial basis function; SVM: support vector machine

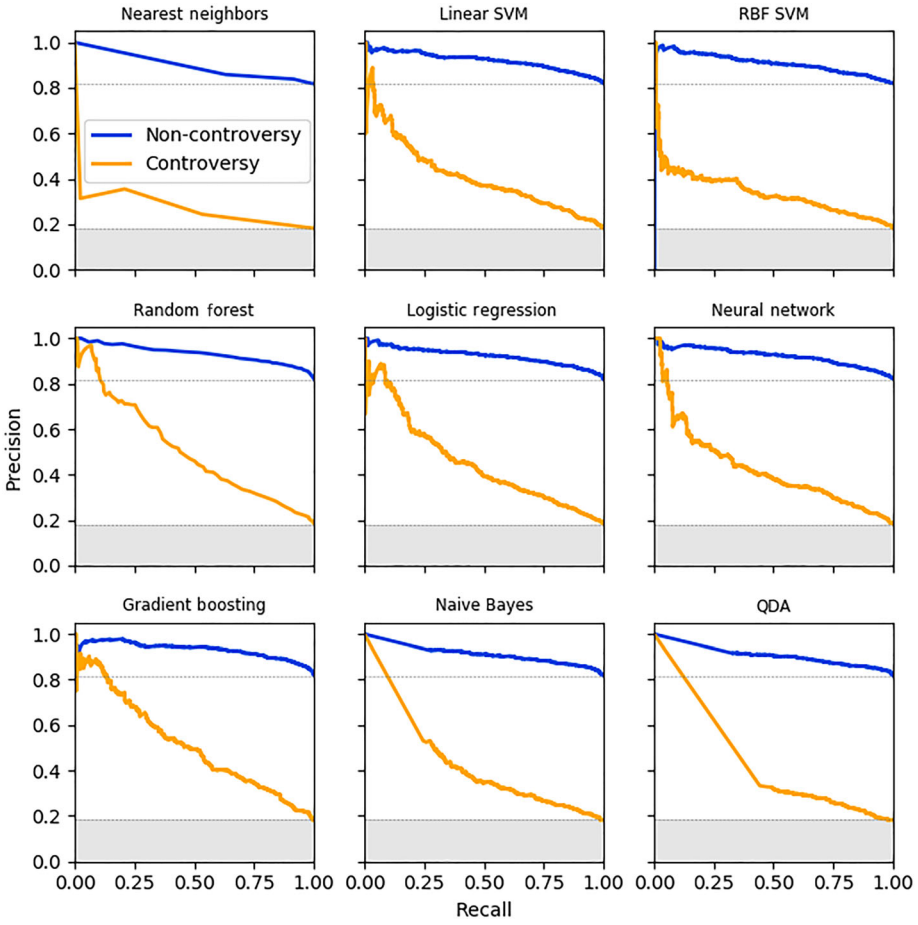

FIGURE C2 Precision–recall (PRC) curves. The blue lines represent the PRC for predicting noncontroversy, and the yellow line represents the PRC for predicting controversy. The PRC shows the precision of a classifier as the recall increases. The top region (defined by the dashed gray line) shows the region for which a classifier performs better than random guessing for the noncontroversy cases, and the region between the bottom and top region shows the region where a classifier performs better than random guessing for the controversy cases. QDA: quadratic discriminant analysis; RBF: radial basis function; SVM: support vector machine