Chapter 4
Prediction of Controversies and Estimation of ESG Performance: An Experimental Investigation Using Machine Learning

Jan Svanberg, Tohid Ardeshiri, Isak Samsten, Peter Öhman, and Presha Neidermeyer

Abstract We develop a new methodology for computing environmental, social, and governance (ESG) ratings using a mode of artificial intelligence (AI) called machine learning (ML) to make ESG more transparent. The ML algorithms anchor our rating methodology in controversies related to non-compliance with corporate social responsibility (CSR). This methodology is consistent with the information needs of institutional investors and is the first ESG methodology with predictive validity. Our best model predicts what companies are likely to experience controversies. It has a precision of 70–84 per cent and high predictive performance on several measures. It also provides evidence of what indicators contribute the most to the predicted likelihood of experiencing an ESG controversy. Furthermore, while the common approach of rating companies is to aggregate indicators using the arithmetic average, which is a simple explanatory model designed to describe an average company, the proposed rating methodology uses state-of-the-art AI technology to aggregate ESG indicators into holistic ratings for the predictive modelling of individual company performance.

J. Svanberg · T. Ardeshiri
University of Gävle and Centre for research on Economic Relations, Sundsvall, Sweden
e-mail: jan.svanberg@hig.se

T. Ardeshiri
e-mail: tohid.ardeshiri@ri.se

I. Samsten
Stockholm University, Stockholm, Sweden
e-mail: samsten@dsv.su.se

P. Öhman (✉)
Centre for Research on Economic Relations, Mid Sweden University, Sundsvall, Sweden
e-mail: peter.ohman@miun.se

P. Neidermeyer
West Virginia University, Morgantown, WA, USA
e-mail: presha.neidermeyer@mail.wvu.edu

T. Rana et al. (eds.), *Handbook of Big Data and Analytics in Accounting and Auditing*,
https://doi.org/10.1007/978-981-19-4460-4_4

Predictive modelling using ML enables our models to aggregate the information contained in ESG indicators with far less information loss than with the predominant aggregation method.

Keywords Artificial Intelligence · Controversies · Corporate Social Performance · ESG · Machine Learning · Socially Responsible Investment

1 Introduction

Global capital is by far the strongest force in the world, more powerful than any government or political order. Nothing can be done without capital being allocated, at least not in the commercial world, indicating that capital allocation can help stop unsustainable activities. For a long time, however, there was no clear evidence of whether institutional investors have any impact on their portfolio companies' compliance with corporate social responsibility (CSR) demands. Recent findings indicate that divestment campaigns decrease stock value by 8–10 per cent and may achieve five per cent greenhouse gas (GHG) emission reductions in high-emission companies (Choi et al., 2020).

The portion of institutional investors considering the divestment of CSR-noncompliant companies is growing rapidly. Early estimates held that about a quarter of market capital would have to be allocated according to such preferences to have any real effect on capital costs (Ambec & Lanoie, 2008). We are above that level, with the total amount of socially responsible investment (SRI) in 2018 being one in every four USD (USD 12 trillion) of investment in the USA and one in every two euros (EUR 22 trillion) in Europe (EUROSIF, 2018; USSIF, 2018). Such an important task as directing global capital towards sustainable projects requires a sophisticated navigation system. Unfortunately, there has been no good method to measure or rate the level of sustainable performance of individual companies. The most advanced measures of corporate social performance (CSP) are the environmental, social, and governance (ESG) ratings, which are considered the standard measures of CSP, i.e. the extent to which companies comply with compulsory CSR policies (cf. Drempetic et al., 2020; Oikonomou et al., 2018). A Google Scholar search on 4 October 2020 using the terms' corporate social performance' and KLD (representing Kinder, Lydenberg, Domini) found 10,200 hits for one such rating.

As popular as ESG ratings are, their aggregation methodology has several shortcomings. First, research has found that aggregated ESG ratings are invalid (Chatterji et al., 2016; Delmas et al., 2013; Semenova & Hassel, 2015; Trumpp et al., 2015). These studies unequivocally concluded that the methodology with which information-rich ESG indicators are aggregated to form composite ratings is inappropriate for the broad, heterogenous, and nonlinear construct of CSP. Notably, the critique articulated in the cited studies does not claim that the ESG indicators do not contain a treasure trove of information about CSP, but that the methods with which indicators are aggregated and holistic ratings computed are far from accurate.

Second, the ESG aggregation methodology does not generate ratings that meet the information needs of institutional investors. Nofsinger et al. (2019) found that these investors were indifferent to whether companies had features that were not compulsory environmental, social, or governance requirements, but were underweight stocks in companies that did not meet compulsory requirements. This information preference for compulsory CSR compliance is inconsistent with the lack of discrimination between obligatory and discretionary aspects of CSR produced with conventional ESG aggregation methodology. The aggregated ratings are an arithmetic average of both aspects (Chen & Delmas, 2011), making these ratings more or less irrelevant for informing investment decisions, even though the information underlying ESG indicators likely contains information that would be useful to institutional investors.

This chapter summarizes the findings of a research project using artificial intelligence (AI) to develop a new type of ESG rating methodology with predictive validity and relevance to institutional investors (see e.g. Svanberg et al., 2022). Because of the broad, heterogenous, and nonlinear CSP concept, it is preferable to develop an ESG rating methodology using predictive rather than explanatory modelling. In this regard, we quote Wood (2010, p. 57): "CSP is a complex system and its measurement requires complex tools". Furthermore, a rating describes the state of a single observation (i.e. the performance of an individual company), which is what predictive modelling provides, whereas explanatory modelling generates relationships between variables generalizable to all observations, but no precise prediction of the state of a particular case. Deriving a rating is therefore an inappropriate task for explanatory modelling but an ideal application of predictive modelling with machine learning (ML).

The purpose of this chapter is to describe a rating methodology with which high-precision ESG ratings can be generated within a project including various studies. Because predictive validity and the choice of institutional-investor–relevant proxies for CSP are the keys to a rating that is useful to institutional investors, we discuss the proxies and their predictive performance on several performance measures.

2 The ESG Indicator Aggregation Problem

2.1 Limitations of ESG Indicator Aggregation

Institutional investors need to know if one company is more sustainable than another to compose portfolios that consider ESG risks. The more ESG criteria an investment assessment covers, the more difficult the process becomes for the investment manager, because the investment decision ultimately merges all criteria into one, and ESG contains too many disparate features to be considered separately (Gond & Crane, 2010). For example, how should an investment manager determine the relative financial materiality of two such different ESG features as 'employee turnover' and 'hazardous waste'? If considering disaggregated raw data, such weighing of one feature against another would have to be done with hundreds of ESG features. A central problem with ESG ratings is therefore how to aggregate all features without losing too much information.

In the early ESG literature, ESG was measured using questionnaires (Aupperle, 1984), regulatory compliance (Wokutch & Spencer, 1987), and content analysis (Wolfe, 1991). Another method was indexing, as in the Fortune corporate reputation index (Sharfman, 1996), but the problem of aggregating measures to form composite rating indexes has not been convincingly solved. For example, some studies have used stakeholder questionnaires to assess the relative importance of ESG features, but this method is considered arbitrary and is unpopular (Chatterji & Levine, 2006; Hillman & Keim, 2001). Another method is to use the 'data envelope' technique from operations management to estimate an aggregated ESG measure (Chen & Delmas, 2011). The fact that this method generates identical ratings for most companies makes it useless in practice. In reality, there is no alternative to using the large ESG indicator databases that raters supply. No other information source comes close to compiling hundreds of ESG indicators, most of them company self-reported. The question, then, is how to aggregate these data.

ESG rating methodology is as simple as it is popular. ESG ratings are in effect the arithmetic average of ESG indicators. However, due to scale differences (i.e. binary versus continuous), ranking formulas need to be used, and company performance is computed via within-industry comparison between companies. Previous reviews of ESG methodology confirm our assessment: "Most empirical studies on CSP use simple linear aggregations, weighted or non-weighted, to derive a composite CSP score from a selection of CSP metrics" (Chen & Delmas, 2011, p. 789).

It is surprising that accounting and finance articles in high-ranked journals such as The Accounting Review refer to this aggregation methodology as providing a fairly unproblematic measure of ESG. There are few critical studies of the validity of ESG aggregation methodology (Berg et al., 2019; Chatterji et al., 2016; Drempetic et al., 2020; Semenova & Hassel, 2015; Trumpp et al., 2015), but one indication that something is wrong with the way ESG indicators are aggregated in research is the amount of conflicting evidence regarding the returns of sustainable performance (Margolis et al., 2012; Orlitzky et al., 2003). For example, some find that the financial performances of portfolios with and without SRI screening are indistinguishable (Statman, 2006; Renneboog et al., 2008).

The few studies of the validity of ESG ratings tell the story of the harm that inappropriate aggregation does to a composite rating. Semenova and Hassel (2015) investigated the aggregation of indicators of the environmental part of three ESG ratings and found that the ratings were uncorrelated. Similarly, Trumpp et al. (2015) found that environmental performance is a multidimensional and aggregate construct as opposed to a subordinate construct and claimed that the reduction of this collection of uncorrelated features to one dimension causes information loss. Sharfman (1996) examined the correlation between ESG ratings based on KLD data and the Fortune reputation index, finding that they were not correlated. Among the most thorough studies of aggregated ESG rating validity is that of Chatterji et al. (2016), who examined the convergent validity of six different ratings and found that none of them was correlated with any other rating. Their results indicate that different ratings judge the same company fairly differently, leading to the conclusion that most or all of the ratings must be wrong and that much empirical research relying on ESG ratings must be reassessed.

Two methodological problems with ESG ratings may explain why it is difficult to find evidence that aggregated ESG ratings have validity. The first problem with this rating methodology is its aggregation of indicators using arithmetic averaging. Arithmetic averaging is a linear estimator, but the underlying CSP construct is likely nonlinear due to feature interaction (Chen & Delmas, 2011; Oikonomou et al., 2018). This incapacity is related to arithmetic averaging's low capacity to represent complexity, because averaging treats ESG indicator patterns as if they were expressions of a uniform underlying construct (Semenova & Hassel, 2015) and as if there were no interactions between indicators (Sigrist & Hirnschall, 2019). As a reference point from which to understand the complexity of the rating task, our rating assesses more than 100 indicators per ESG component, among which there may be patterns that exhibit their own distinct behaviors. The arithmetic average is an aggregation method that accounts for none of this.

The second problem is the equal weighting of indicators. The assumption that all aspects of ESG are equally important for overall sustainability performance is clearly wrong (Callan & Thomas, 2009; Chen & Delmas, 2011; Sharfman, 1996). The equal weighting scheme has survived because there has been no viable alternative. From a legitimacy perspective on ESG, a weighting scheme would have to mirror the relative importance of areas of CSP as they appear to society.

These two constraints are due to the manner in which ESG indicators are aggregated and could apply to any complex hypothetical construct. A different constraint arising from the aggregation of ESG indicators is that it does not produce ratings that disclose information relevant to institutional investors. Notably, this problem is not because the ESG indicators lack this information, but because aggregation by means of averaging cannot represent it. As evidenced in several studies, institutional investors view ESG as financially relevant, but they are focused exclusively on ESG weaknesses. The weaknesses refer to features related to the risk of ESG controversies. Nofsinger et al. (2019) claimed that the reason for institutional investors' asymmetric information preference is that any positive effects of delivering ESG performance beyond minimum compliance with CSR demands are so small that they are offset by the costs of achieving those benefits. They argued, in contrast, that non-compliance with binding, compulsory CSR, which we for clarity refer to as corporate social *responsibilities* because we emphasize the obligatory, non-discretionary aspect of this part of CSR, causes irreparable financial damage.

The adverse economic effects associated with CSR controversies include lawsuits, disrupted production, and consumer boycotts (Luo & Balvers, 2017) as well as strikes and government investigations. These effects are relevant only to the extent to which company behaviors are compliant with compulsory CSR, but not at all to the discretionary 'doing good' part of ESG. The cost asymmetry resonates with how institutional investors use ESG information, which is predominantly for managing investment risk rather than out of moral concern with doing good (Amir & Serafeim, 2018). Similar findings were reported in a recent survey study (Krueger et al., 2020): investor reputation protection is the primary motivation to use ESG data when making investments, and a reputation is best protected by avoiding investment in controversy-prone companies. Financial incentives and a desire to protect one's reputation may explain why institutional investors underweight stocks in companies prone to ESG controversies and why their information preferences are one-sided (Nofsinger et al., 2019). Consistent with these observations is the evidence that stock markets show small or no reactions to good news but large negative reactions to bad news (Capelle-Blancard & Petit, 2019; Krüger, 2015).

These findings have consequences for how ESG indicators should be aggregated. An ESG rating needs to be a high-precision estimate of the extent to which individual companies comply with those parts of CSR that could trigger controversy if violated by a company. As an arithmetic average of indicators is highly unlikely to be such an estimate, this aggregation method is inappropriate for ESG ratings.

2.2 Controversies as Proxies for Non-Compliance with Compulsory CSR

In contrast to the current ad hoc explanatory approach to aggregating ESG indicators, we claim that ratings should be constructed using predictive modelling because such modelling generates ratings that are far more accurate than those computed using explanatory modelling (cf. Ding et al., 2020; Fiaschi et al., 2020; Shmueli, 2010). Predictive modelling is data driven and therefore not evaluated relative to how consistent it is with theory. The validity of predictive modelling is evaluated using performance measures that indicate the extent to which the model predicts a variable of interest. Our variable of interest is determined by the information preferences of institutional investors. As they are interested in predicting the extent to which potential portfolio companies comply with the compulsory parts of CSR, and because non-compliance with such norms is associated with the risk of controversies, our variable of interest is CSR controversies.

As noted, institutional investors' strongest motivation to integrate ESG assessment in investment is to protect investor reputation. Negative legitimacy caused by social controversies in portfolio companies spills over to the owners (Zavyalova et al., 2012), so avoidance of controversy-prone companies is a priority. An ESG rating predicting the likelihood that a company will become embroiled in a social controversy would serve this purpose. Krueger et al. (2020) found that the second strongest motivation for integrating ESG risk assessment into the investment process is institutional investors' perceived fiduciary obligation towards beneficiaries to invest responsibly. When beneficiaries are dissatisfied with how a fund invests their money, there is a risk that they will withdraw their money from the fund (Grappi et al., 2013). Withdrawing money as an expression of dissatisfaction with the sustainability of investments can be expected when beneficiaries feel that too many portfolio companies are associated with CSR controversies. An ESG rating predicting CSR controversies would therefore not only protect the institutional investor's reputation but also help them perform in line with beneficiaries' expectations.

Another reason for institutional investors to value an ESG rating that predicts CSR controversies is that stocks associated with recent controversies tend to experience high volatility and price declines (Cui & Docherty, 2020; Muller & Kräussl, 2011). This would be viewed as negative by the investors, because investment managers are often evaluated based on their short-term performance (Harmes, 2011). Nofsinger et al. (2019) argued, however, that the avoidance of companies that have ESG weaknesses is more pronounced among institutional investors with a long-term than a short-term investment focus. Taken together, these pieces of evidence and theoretical arguments indicate that CSR controversies should be the proxies for ESG performance when developing an ESG rating using predictive modelling.

The use of CSR controversies as the variable of interest and ESG indicators as input variables produces an ML model conceptually different from conventional ESG ratings. The conventional aggregation of ESG indicators does not assess ESG relative to a performance standard. Describing behaviors is not the same as assessing the extent to which behaviors constitute performance, because a prerequisite for performance is that behaviors should exceed or comply with a standard. We therefore refer to ESG indicators as company behavior indicators and to our estimation of the risk of controversy as our assessment of these behaviors relative to a performance standard of corporate social requirements. As previously indicated, these requirements refer to obligatory CSR norms. Some of those are legally binding, for example, as set forth in labor, environmental, and corporate law; some are 'soft law', such as international standards, for example, for accounting; and some are moral norms, such as the emerging norm to recognize the threat of climate change and reduce GHG emissions. The binding or obligatory nature of these requirements suggests that noncompliance with them is associated with a risk of controversy, which may arise, for example, through litigation initiated by government or international organizations or through boycotts and strikes initiated by stakeholders.

Our definition of ESG performance as formed by behaviors assessed relative to obligatory CSR (referred to below as corporate social *responsibilities*, CS*R*), means that while our ML ratings estimate companies' compliance with norms, conventional aggregation methodology simply summarizes behavioral indicators. For example, having a 'salary gap' and having an 'inappropriate salary gap' are not the same thing, although conventional ratings treat them as if they were. There has been no criticism of the methodological aspects of this conceptual distinction in the ESG rating literature. In contrast, worker performance is assessed by observing achievement relative to a performance standard (Groover, 2007). Judges in all courts of law judge behaviors relative to a standard called law without which their judgments would be arbitrary, questionable, or meaningless. Auditors assess company accounting relative to accounting standards and recommendations (Öhman & Wallerstedt, 2012), and so on.

When aggregating ESG indicators to form a composite rating, it is common practice to define performance as the sum of behaviors, with no reference to any performance standard (cf. Wood, 2010). The lack of performance standard in ESG rating methodology deprives the assessment of such ratings the guidance offered by norms, leaving institutional investors and empirical researchers with an unanswered question: What does the unanchored aggregated ESG indicator actually measure?

Our use of CS*R* controversies as proxies for ESG performance anchors our ML ratings to a system of binding standards and makes the ratings reflect the risk of legitimacy loss described by legitimacy theory (Deegan, 2019). A company non-compliant with CS*R* is one that per definition risks its legitimate right to pursue its business and makes itself a target of controversy. A rating anchored in ESG controversies regarding non-compliance with CS*R* would therefore be a rating that sets out to measure the most fundamental antecedents of legitimacy at the same time as it is consistent with institutional investors' information needs.

Controversies are assessments and reactions to CS*R*-relevant company behaviors (e.g., inappropriate or illegal waste management). As such, they provide different information from that provided by ESG indicators. CS*R* controversies also tell something about whether the transgressor has the structures and processes necessary to ensure CS*R* compliance (Nieri & Giuliani, 2018) regarding, for example, management quality and supply-chain management structures (Chiu & Sharfman, 2018). Identifying the lacking structures is a complex task because they can be related to errors in many areas, each complex in itself, for example, the environment, anti-competitive behavior, patents and intellectual property, lack of respect for human rights, poor labor relations, tax fraud, child labor, and inappropriate management compensation. Predictive modelling accomplishes this through associating indicator patterns with meanings conveyed by controversies. These associated meanings of controversies can be generalized because controversies often arise from intended, systematic breaches of CS*R*. Companies embroiled in controversies often violate moral and legal norms when pursuing financial goals (Fiaschi et al., 2017; Surroca et al., 2013) because performance pressures cause corners to be cut and reward inappropriate behavior, eventually normalizing it (Earle et al., 2010). Predictive modelling extrapolates the likelihood of controversies from known controversies to other companies exhibiting ESG indicators similar to those of the controversial companies.

3 Research Design and Measures

We adopt cross-sectional designs in predictive modelling. The predictive design sacrifices theoretical explanation for substantially higher empirical precision in describing individual companies' performances (Collopy et al., 1994; Gurbaxani & Mendelson, 1990). While explanatory modelling investigates relationships between features in a population, predictive modelling examines the state of specific companies. A rating unable to distinguish between individual companies is as useless to an investor as a poor diagnostic instrument is to a medical practitioner. Differences between explanatory and predictive modelling are described by Shmueli (2010) and Bzdok et al. (2018). Predictive modelling is superior to explanatory modelling for developing diagnostic methods such as image assessment in rheumatology (Hügle et al., 2020), for credit assessment (Kruppa et al., 2013), and for predicting bankruptcy (Heo & Yang, 2014). In such diagnostic applications, it would be inappropriate to use explanatory modelling. Predictive modelling is therefore the most effective ESG rating methodology.

The three studies in our research project use nine ML algorithms and the task is to predict CS*R* controversies in companies by examining ESG indicators. The achievement of this task is evaluated using five measures of predictive performance: precision, recall, Fmeasure, area under receiver operating characteristic (ROC) curve (hereafter AUC), and precision recall curve (PRC).

We obtain data from Refinitiv Eikon and use all available indicators—nearly 400 in the selected ten-year window. An enumeration of the indicators used in each of the three studies is provided in the appendices to the three working papers that this chapter summarizes. In addition, we include the additional indicators ‘market capital’, ‘return on assets’, ‘industry’, ‘country of headquarters’, ‘total assets’, and ‘net assets’. The three studies are based on a sample of 2517 companies for the 2009–2018 period.

In a first step, the companies are divided into two groups: one of companies having experienced controversies (i.e. environment and governance controversies) or having experienced more than an average number of social controversies, and another of companies having experienced no controversies (i.e. environment and governance controversies) or having experienced fewer than average social controversies. While environmental and governance controversies are rare, social controversies are more common, with an average number of 5.17 during the ten-year window. We therefore use a different approach to define the disjoint classes. For environmental and governance controversies, we use the absence of controversy as indicating high ESG. For social controversies, the definition of high ESG is that a company has experienced fewer than average controversies. The high-ESG class includes negative cases and companies in the low-ESG class are considered positive cases.

The idea behind the cross-sectional research design is not to model the risk of future controversies based on a company’s past ESG indicators, but to model the likelihood that an indicator pattern is associated with a company’s risk of experiencing a controversy. The longitudinal aspect of the data is reduced to a cross-sectional format, with indicators averaged over the ten years if numerical or encoded with dummy variables if binary.

The models are developed in ML experiments in which one algorithm at a time is extracting information from ESG indicators by associating patterns of indicators with CS*R* controversies. The nine ML algorithms, representing a full range of algorithms for supervised learning, have different functionalities and therefore capture different aspects of this learning task. The algorithms used are nearest neighbors, linear support vector machine (linear SVM), radial basis function support vector machine (RBF

SVM), random forest, logistic regression, artificial neural network, gradient boosting, naïve Bayes, and quadratic discriminant analysis (QDA).

4 Experiments

An overview of the settings with which the nine ML algorithms are executed is presented in Table 1. The hyper-parameter settings ensure that our studies can be reproduced and provide additional background to support their interpretation. For simplicity, we have used the default settings of scikit-learn (Pedregosa et al., 2011), version 0.22.

Predictive modelling evaluates the trained algorithms using an unseen set of test instances. A method to do this that economizes on scarce data is k-fold cross-validation. It partitions the data in k disjoint folds and conducts training iteratively on k – 1 folds, with one fold for testing. This repeated training/testing uses the entire dataset for training and testing, with the advantage that predictive performance is evaluated from k different angles. In total, k performance measures are obtained, the mean of which is a better estimate of the generalization performance than if the predictive performance were calculated using only one partition. This study employs stratified tenfold cross-validation. Stratification ensures an equal number of positive and negative cases in each test set.

As mentioned, five measures of performance—precision, recall, F-measure, AUC, and PRC—are used to evaluate predictive performance. The definition of these measures must be understood from initial standard distinctions described in a similar way in all ML literature (see Table 2). Precision (Eq. 1) is the sensitivity of the predictor, i.e. the fraction of positive cases in relation to all predicted positive cases. Recall (Eq. 2) measures the specificity of the classifier, which is its ability to identify as large a fraction as possible of the positive cases. The true positive is the number of times the predictor can correctly label a controversy. The false positive is the number

Table 1 Hyper parameters

Algorithm	Notes
Nearest neighbor	Three nearest neighbors
Linear SVM	Linear kernel with $C = 0.025$
RBF 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
QDA	No hyper-parameters

Table 2 Basic measures

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

of times the predictor falsely believes a company has experienced a controversy. The true negative is the number of correct observations of a company as not having experienced a controversy. The false negative is the number of contrary-to-fact predictions that a company has not experienced a controversy.

There is a performance trade-off between precision and recall. For example, a predictor that predicts every company as a positive case and having a recall of 100 per cent would often have a very low precision, in our case zero. The F-measure (Eq. 3) captures the trade-off.

Another measure of the relationship between two other measures is the AUC. This is the area under the true positive rate versus false positive rate curve as defined in Eqs. 4 and 5. This measure estimates the probability of a predictor rating a true positive instance ahead of a false positive instance and is therefore a measure of its aggregated rating performance. The PRC shows the mean precision for multiple thresholds of recall and is therefore an average of the trade-off between precision and recall. The PRC is obtained by plotting precision versus recall.

5 Results

5.1 Predictive Performance of Different ML Algorithms

Before examining the prediction results, we investigate whether an ESG rating would provide appropriate data for classifying companies as involved or not involved in ESG controversies. The following is a simple way to illustrate whether a conventional ESG rating contains any information of the type demanded by institutional investors. Figure 1 displays two distributions of companies across overall ESG ratings. The yellow bars show company-years with more than an average number of social controversies on the *y*-axis and the total ESG rating from Refinitiv on the *x*-axis. Company-years with fewer than average social controversies are shown by the blue bars. The distributions are similar with no cut-off point defining the two classes. The ESG rating is particularly insensitive to the likelihood of controversies in the ESG range where most companies are situated. We find that the same types

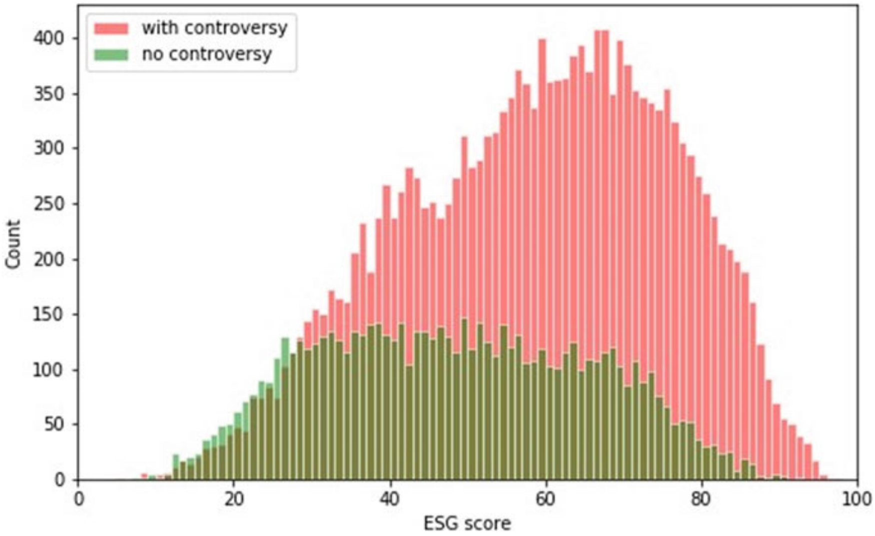

Fig. 1 The distribution of social controversies over company-years and ESG ratings. Company-years with fewer than average social controversies in a ten-year window are classified as 'no controversy' on the *y*-axis and company-years with more than an average number of controversies are classified as 'with controversy'

of overlapping and non-distinguishable distributions are repeated for environmental, social, and governance component ESG ratings and for environmental, social, and governance controversies.

5.2 Prediction Results

We first address the static measures precision, recall, and F-measure, which show the performance of each ML algorithm at a given probability threshold for predicting a positive case. Table 3 shows the performance of two of the ML algorithms for the three types of controversies: environmental, social, and governance. We are contrasting the most capable model, random forest, with one of the less capable but still well-performing ones, logistic regression. We pay special attention to the precision column, because precision should be the priority for institutional investors who are about to allocate large sums of money based on a decision to include a company in their portfolios. Random forest is the most precise of the two models for all three types of controversies (70–84 per cent of the controversies are correctly predicted), but the two models perform more similarly for the dynamic measures and logistic regression has higher recall. Random forest's high precision can be linked to its learning capabilities, because it produces models that have uncalibrated

Table 3 Predictive and learning performance

	Precision	Recall	F-measure	AUC	PRC
Environmental					
Random forest	0.6994	0.1942	0.2990	0.8849	0.5090
Logistic regression	0.3045	0.7253	0.4282	0.8727	0.4814
Social					
Random forest	0.8417	0.5756	0.6823	0.9165	0.8104
Logistic regression	0.6132	0.7788	0.6852	0.8982	0.7898
Governance					
Random forest	0.7506	0.1700	0.2756	0.7787	0.5110
Logistic regression	0.3387	0.6513	0.4444	0.7531	0.4699

probability estimates, requiring that the model have high confidence in predictions. Identical training and testing folds were used for all nine algorithms under study.

In the separate studies we find that random forest, artificial neural networks, and gradient boosting are the best for precision, with the bottom group of naïve Bayes, RBF SVM, and QDA far behind with small variations among them. We also find that the differences between the high- and lower-precision predictors over the five performance measures are similar to those in Table 3. All the lower-precision models (i.e. naïve Bayes, RBF SVM, and QDA) have the same pattern of low precision and relatively higher recall. From the investment manager's point of view, this difference is important because it indicates that the lower-precision algorithms are prepared to guess wrong many more times to get one guess right than the more cautious high-precision models random forest, artificial neural networks, and gradient boosting. This wrong screening of the lower-precision algorithms would limit diversification for institutional investors and encourage investment on false premises. As a reference point, an insurance company may own 2000–5000 equities, which can be compared with the approximately 8000 companies making up 80 per cent of global capital for which Refinitiv provides ESG indicators. A negative screening tool that reduces the investment universe too much due to low precision would not satisfy investment decision support requirements.

Furthermore, we evaluate the ranking and predictive performance of the algorithms using the AUC and PRC in Figs. 2 and 3. Discrimination, equivalent to the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance, is measured by the AUC, i.e. it is equivalent to the two-sample Wilcoxon rank-sum statistic.

The AUC should be as close to the top-left corner as possible. If the curve is below the dashed line, it represents a prediction worse than a random guess. Overall, the plots in Fig. 2, the yellow curve for predicting positive cases and the blue curve for negative cases, confirm our interpretation of Table 3. The differences between the models are more difficult to discern graphically, but we find that random forest has a steeper rise in the true positive rate than does logistic regression, corresponding to

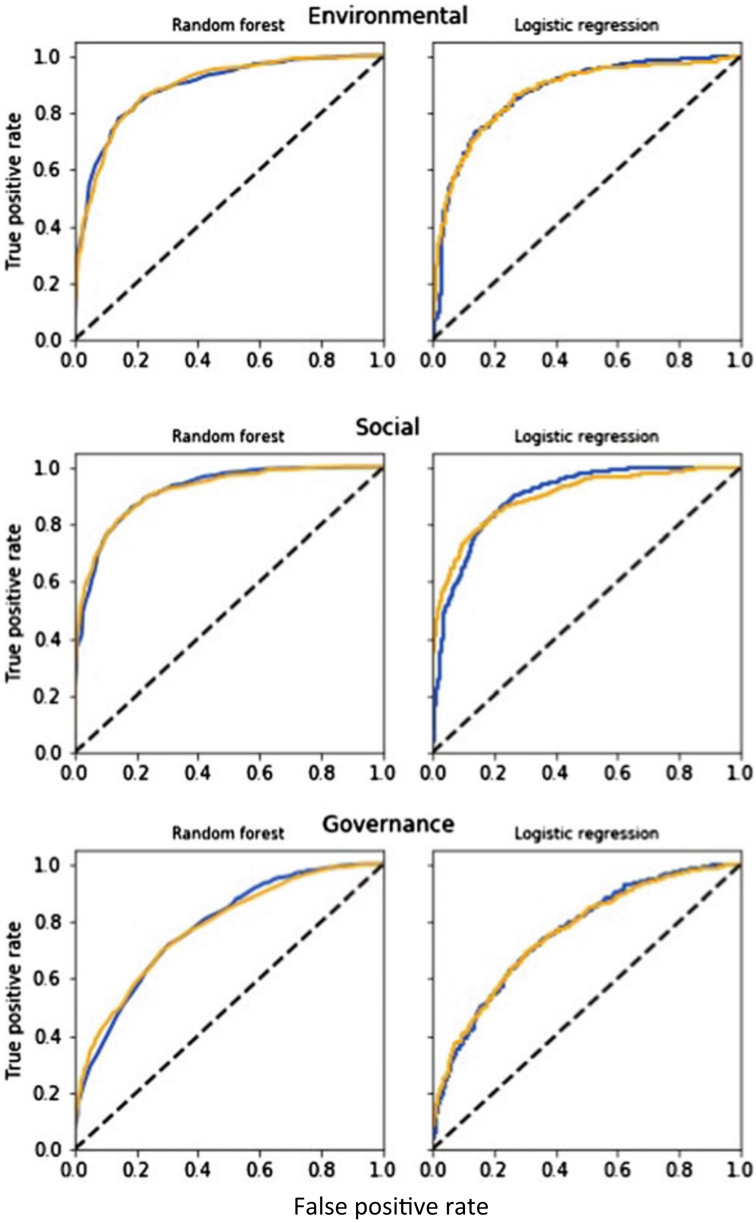

Fig. 2 Area under the ROC curve (AUC) for three different ratings (environmental, social, and governance) and for two algorithms (random forest and logistic regression). The blue line represents the AUC for predicting non-controversy proneness and the yellow line represents the AUC for predicting controversy proneness

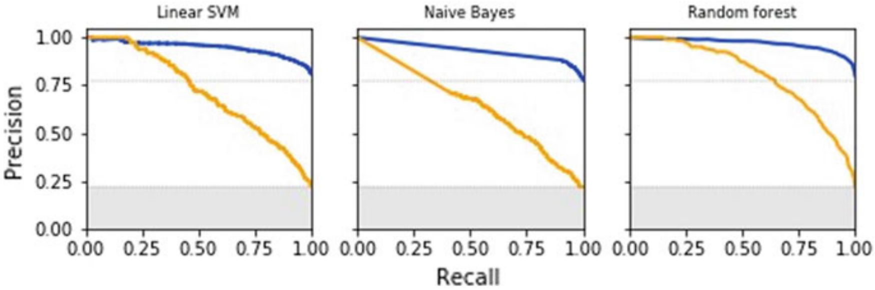

Fig. 3 Precision recall curves (PRC). The blue lines represent the PRC for predicting non-controversyproneness and the yellow lines represent the PRC for predicting controversy-proneness for social controversies. The PRC shows the precision of a classifier as the recall increases. The top region (above the grey area) is where a predictor performs better than random guessing for the non-controversy cases, and the region between the bottom and top region show the region where a predictor performs better than random guessing for the controversy cases

the difference between the high-precision and lower-precision algorithms found in our separate studies, i.e. logistic regression is a linear model with limited complexity representation capacity. Any increase in the false positive rate is costly considering the small fraction of companies defined as controversy prone. It is therefore important that the model have a steep rise in the AUC.

Figure 3 contains PRC and contrasts linear SVM, naïve Bayes, and random forest for predicting positive and negative cases of social controversies. The PRC curve is a dynamic measure of the predictive performance of the algorithms, showing the trade-off between precision and recall. For the investment screening application, it is ideal to have both high precision and high recall, but as Table 3 demonstrates, this is not attainable because none of the models can quite produce the ideal output.

In terms of the graphic presentation in Fig. 3, the ideal is a PRC curve in the topright corner and as far as possible from the bottom-left corner. As long as the algorithms need to predict controversy-prone companies with absolute certainty, they can maintain high precision, but as they are also required to find a large proportion of the total number of such companies, the algorithms sacrifice the certainty of predictions to pursue more positive cases. The PRC curves confirm what we see in Table 3. Both the non-controversy and controversy prediction curves are in the top-right corner for random forest, and our separate studies demonstrate that this applies to the whole group of competent algorithms, i.e. random forest, artificial neural network, and gradient boosting, as well as to logistic regression. They accomplish a more efficient trade-off between precision and recall than do the other algorithms. As seen in Fig. 3, the trade-off drops off significantly more steeply with linear SVM and naïve Bayes than with random forest. The grey area at the bottom of each graph represents the class distribution. Several algorithms offer precision of around 0.8 at a recall well above 0.5 for social controversies, which are high numbers.

5.3 Controversy Prediction as ESG Rating

Having seen that the predictive modelling of environmental, social, and governance controversies is possible and that the results indicate that controversy-prone companies can be detected with our high-precision screening, we now focus on how the corresponding MLbased methodology performs at rating. We expect that any ESG rating and our rating in particular should be negatively correlated with the number of controversies a company experiences. While many company CSR activities can be symbolic and inauthentic, controversies are instances in which companies are accused of not following required practices in the CSR area and ratings measuring this non-compliance are what institutional investors need.

Our best model extrapolates from what it has learnt by studying the association between indicator patterns and controversies. It expects certain patterns to be associated with controversies and should typically give lower ratings to companies experiencing more controversies. However, there may be many deviations from the normal. For example, companies experiencing more controversies than average may have ESG indicator patterns typical of very compliant companies. They 'look good' from the rating model's point of view and are therefore awarded high ratings. Other companies may have experienced no or fewer than average controversies despite exhibiting a pattern of features typical of companies that are controversy prone. They are awarded low ML ratings. On average, however, a negative correlation between an ESG rating and the number of controversies is expected.

The rating distributions are described in Fig. 4. The *x*-axis in the graph shows the number of controversies per company and the *y*-axis the ratings. We illustrate the results of our studies by showing the performance of four algorithms for predicting social controversies. (The graphs for environmental and governance controversies are similar.) How closely the rating models are dictated by actual controversies may depend on the complexity of the model. A simplistic rating model would typically have a less tight alignment between the ESG ratings and controversies because it misses relevant interactions between features and controversies, which is called under-fitting to the training data with possible bias as a result. With under-fitting we would anticipate finding some models giving high ratings to companies experiencing many controversies, and this should be expected for linear SVM and logistic regression, but we observe none of that. A complex nonlinear model may instead lead to over-fitting to the data with lower generalizability as a result. There seems to be a mixed pattern in Fig. 4, however, with logistic regression and gradient boosting being unwilling to give any company experiencing many controversies a high rating, whereas random forest (and artificial neural networks) of the complex models and linear SVM (and RBF SVM) of the less complex models appear to use more discretion and sometimes give companies with many controversies a rating around 40–50.

Figure 4 reveals, as expected, a negative correlation between the four ML-based ESG ratings and the number of social controversies experienced by companies. The models with the highest predictive performance have smooth distributions. There is a difference between logistic regression and gradient boosting, on one hand, and

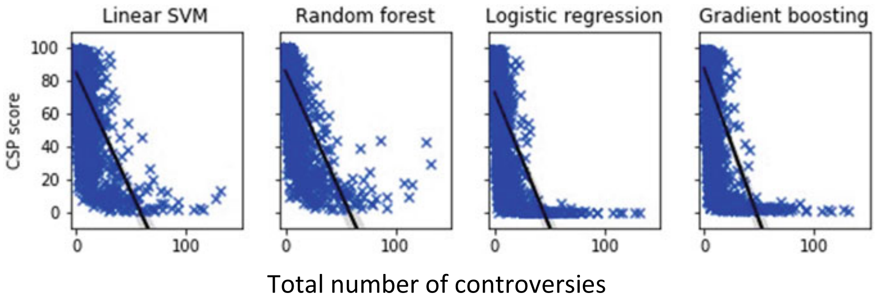

Fig. 4 Correlation between the social component ESG rating produced by four of the machine learning algorithms and the number of years with social controversies. (Total number of controversies)

random forest and linear SVM, on the other, in assigning low ratings to all companies experiencing many controversies. There is a considerably wider distribution in the ratings for companies experiencing many controversies for random forest and linear SVM than for logistic regression and gradient boosting. Because random forest and gradient boosting represent two of the more complex models, whereas logistic regression and linear SVM are linear and therefore less complex, strict adherence to a many controversies–low rating rule can be accomplished using linear and nonlinear models. All the models have a steep downward slope of the correlation, which is evidence that all models operate according to the logic we require of an ESG rating. Except for relatively few companies in which there are good reasons to believe that controversies are not associated with structural weaknesses in governance, many controversies should result in a low ESG rating because CS*R* controversies are the best available proxies for noncompliance with CSR.

6 Conclusions

SRI is endorsed by a large fraction of institutional investors in most parts of the world. Those investors believe that they have the power to make companies more sustainable through the way they exercise ownership power, and recent finance research (Choi et al., 2020) gives them right. However, it is difficult for institutional investors in good faith to try to invest responsibly and influence portfolio companies to become more sustainable if they do not know what a sustainable company looks like.

The problem is that institutional investors need to assess sustainable performance manually by scrutinizing the huge amounts of ESG information companies now tend to disclose. The task is demanding. Assume that an institutional investor was to devote one minute to each of 400 ESG indicators. If so, it would take more than 25 years for one sustainability expert to manually screen the approximately 8000 companies for which there are ESG data today. Consider the difficulty of such manual work competing in this task with the precision (70–84 per cent of controversies correctly predicted) and speed (in seconds once the training is done) of state-of-the-art ML models. There is no lack of information regarding the growing complexity of financial reporting, but the analytical tools with which institutional investors can assess ESG issues have to be hand crafted by each investor because no available tools meet the accuracy and information content demands of institutional investors. The problem is how data are aggregated to form valid and intelligent ratings. Conventional ESG indicator aggregation methodology is more likely to measure the capacity for sustainability rhetoric in large companies than it is to measure sustainable performance (Drempetic et al., 2020).

We find a conceptual weakness that propagates into methodological weaknesses in the aggregation methodology used in ESG ratings. The conceptual weakness is that the ratings do not assess ESG relative to a performance standard. Conventional ratings assume that an average of ESG indicators accurately represents CSP. This is a misperception of the performance concept. This concept has escaped theoretical scrutiny in the CSR and CSP literature despite being established in, for example, the education and work performance literatures. Omitting the idea of performance as behavior relative to a standard causes aggregated ESG ratings to be arbitrary, because an estimator that averages the indicators conveys no information about what is sufficient or good performance on the respective indicators or about the extent to which performance on one indicator could or could not make up for deficient performance on another. The use of equally weighted averages in aggregated ESG ratings is the result of this conceptual error. The methodological problems include raters mainly using equally weighted arithmetic means to aggregate ESG indicators. This aggregation method is, as several studies of ESG rating validity show, unable to produce valid aggregated ESG ratings (e.g. Berg et al., 2019; Chatterji et al., 2016; Semenova & Hassel, 2015; Trumpp et al., 2015).

As a solution to the challenge posed by institutional investors' need for accurate SRI decision support systems, we develop a new type of ESG rating methodology. Our rating methodology uses ML to predictively model ESG. This is a data-driven solution to the problem of assessing company behaviors relative to a performance standard, because we anchor our rating methodology in CSR controversies, which are proxies for companies' compliance with compulsory corporate social *responsibilities* (CS*R*). This data-driven approach studies associations between a broad list of ESG indicators and more than 20,000 controversies with the goal of predicting controversies involving individual companies. Since the prediction of controversies is equivalent to the prediction of compliance with CS*R*, the best ML model (i.e. random forest) assesses companies' behaviors relative to CS*R*. Based on how each indicator is associated with the likelihood of a company experiencing a CS*R* controversy, state-of-the-art ML algorithms learn how to assess the weights on ESG indicators in their models. The best ML model not only produces an unequal weighting scheme that mirrors companies' compliance with CS*R*, it also produces a complex and nonlinear model of ESG that represents the abundant interactions that may occur between ESG indicators. Extrapolating the relationships between ESG indicator patterns and CS*R* controversies to all companies, this ML model can predict compliance with CS*R* for any company for which there is indicator data.

The best ML model has high predictive validity and is consistent with the information needs of institutional investors. These investors need ratings that predict CS*R* controversies so that they can avoid investing in controversy-prone companies. This is the most important feature of any ESG rating, because institutional investors underweight stocks that exhibit traits typical of controversy-prone companies (Nofsinger et al., 2019) and because the strongest motivation for institutional investors to integrate ESG concerns in their investment decisions is the reputation protection they can gain by not investing in controversy-prone companies (Krueger et al., 2020). Our best ML model provides ESG information that meets these needs.

Our various studies have several limitations. First, we use models of the temporal distribution of data. Future modelling could use longitudinal designs with the aim of predicting future controversies. Second, our models do not consider whether controversies in the environmental, social, and governance categories differ from one another, which may introduce bias in the assessment of a company's likelihood of experiencing a controversy because different controversies may have different associations with ESG indicators. Third, we do not adjust the models for unequal media attention paid to companies. Company size and media attention are likely positively correlated, resulting in size-biased ratings. Future work to refine our rating methodology could investigate how to replace CS*R* controversies with a company-level wrongdoing index, which is a scaled and filtered metric of the wrongdoing signaled by controversies (cf. Fiaschi et al., 2020).

Despite these limitations, we argue that our ratings have more merit than the conventional method of simply aggregating ESG indicators. We demonstrate how moving away from explanatory to predictive modelling and how exploiting AI can make ESG a visible target to institutional investors. Future rating research might benefit from investigating how investment managers use additional information to construct supporting arguments for decisions based on ratings. As demonstrated, the importance of individual ESG indicators to the ML ratings can be described using ML models. This is not relevant to conventional ESG rating methodology because the relative importance of indicators does not reflect the relative importance of different features of CSP.

Acknowledgements We gratefully acknowledge the financial and material support from Länsförsäkringars Forskningsfond and the Asset Management Department at Länsförsäkringar. We are particularly indebted to Lars Höglund, Kristofer Dreiman, and Alexander Elfving for their valuable input to this research project.

References

Ambec S, Lanoie P (2008) Does it pay to be green? A systematic overview. *Academy of Management Perspectives, 22*, 45–62. https://doi.org/10.5465/amp.2008.35590353

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

Aupperle, K. (1984). An empirical measure of corporate social orientation. *Research in Corporate Social Performance and Policy, 6*, 27–54.

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

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Points of significance: Statistics versus machine learning. *Nature Methods, 15*, 233–234. https://doi.org/10.1038/nmeth.4642

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*, 61–78. https://doi.org/10.1002/csr.182

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

Chatterji, A., & Levine, D. (2006). Breaking down the wall of codes: Evaluating non-financial performance measurement. *California Management Review, 48*, 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. *The Strategic Management Journal 37*, 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*, 789–804. https://doi.org/10.1111/j.1937-5956.2010.01202.x

Chiu, S. C., & Sharfman, M. (2018). Corporate social irresponsibility and executive succession: An empirical examination. *Journal of Business Ethics, 149*, 707–723. https://doi.org/10.1007/s10551-016-

Choi, D., Gao, Z., & Jiang, W. (2020). Global carbon divestment and firms' actions. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3589952

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

Cui, B, & Docherty, P. (2020). Stock price overreaction to ESG controversies. *Monash Centre for Financial Studies*. Retrieved from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3559915

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

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*, 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. *Review of Accounting Studies, 25*, 1098–1134. https://doi.org/10.1007/s11142-020-09546-9

Drempetic, S., Klein, C., & Zwergel, B. (2020). The influence of firm size on the ESG score: Corporate sustainability ratings under review. *Journal of Business Ethics, 167*, 333–360. https://doi.org/10.1007/s10551-019-04164-1

Earle, J. S., Spicer, A., & Peter, K.S. (2010). The normalization of deviant organizational practices: Wage arrears in Russia, 1991–98. *Academy of Management Journal, 53*, 218–237. https://doi.org/10.5465/amj.2010.49387426

EUROSIF. (2018). *European SRI study 2018*. Brussels, EUROSIF. Retrieved from: http://www.eurosif.org/eurosif-2018-sri-study-launch-event/

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*, 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*, 287–299. https://doi.org/10.1016/j.bushor.2019.09.004

Gond, J.-P., & Crane, A. (2010). Corporate social performance disoriented: Saving the lost paradigm? *Business & Society, 49*, 677–703. https://doi.org/10.1177/0007650308315510

Grappi, S., Romani, S., & Bagozzi, R. P. (2013). Consumer response to corporate irresponsible behavior: Moral emotions and virtues. *Journal of Business Research, 66*, 1814–1821. https://doi.org/10.1016/j.jbusres.2013.02.002

Groover, M. P. (2007). *Work systems: The methods, measurement & management of work*. Pearson Education International

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

Harmes, A. (2011). The limits of carbon disclosure: Theorizing the business case for investor environmentalism. *Glob Environ Polit, 11*, 98–119. https://doi.org/10.1162/GLEP_a_00057

Heo, J., & Yang, J. Y. (2014). AdaBoost based bankruptcy forecasting of Korean construction companies. *Applied Soft Computing, 24*, 494–499. https://doi.org/10.1016/j.asoc.2014.08.009

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

Hügle, M., Omoumi, P., van Laar, J. M., et al. (2020). Applied machine learning and artificial intelligence in rheumatology. *Rheumatology Advances in Practice, 4*. https://doi.org/10.1093/rap/rkaa005

Krueger, P., & Sautner, Z., & Starks, L. (2020). Importance of climate risks for institutional investors. *The Review of Financial Studies, 33*:1067–1111.

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

Kruppa, J., Schwarz, A., Arminger, G., & Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications, 40*, 5125–5131. https://doi.org/10.1016/j.eswa.2013.03.019

Luo, H.A., & Balvers, R. J. (2017). Social screens and systematic investor boycott risk. *Journal of Financial and Quantitative Analysis, 52*, 365–399. https://doi.org/10.1017/S0022109016000910

Margolis, J. D., Elfenbein, H. A., & Walsh, J. P. (2012). Does it pay to be good...and does it matter? A meta-analysis of the relationship between corporate social and financial performance. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1866371

Muller, A., & Kräussl, R. (2011). Doing good deeds in times of need: A strategic perspective on corporate disaster donations. *Strategic Management Journal, 32*, 911–929. https://doi.org/10.1002/smj.917

Nieri, F., & Giuliani, E. (2018). International business and corporate wrongdoing: A review and research agenda. In D. Castellan, R. Narula, Q. Nguyen et al. (Eds.), *Contemporary issues in international business* (pp. 35–53). Springer International Publishing.

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

Öhman, P., & Wallerstedt, E. (2012). Audit regulation and the development of the auditing profession: The case of Sweden. *Accounting History, 17*, 241–257. https://doi.org/10.1177/1032373211434723

Oikonomou, I., Platanakis, E., & Sutcliffe, C. (2018). Socially responsible investment portfolios: Does the optimization process matter? *The British Accounting Review, 50*, 379–401. https://doi.org/10.1016/j.bar.2017.10.003

Orlitzky, M., Schmidt, F. L., & Rynes, S. L. (2003). Corporate social and financial performance: A meta-analysis. *Organization Studies, 24*, 403–441. https://doi.org/10.1177/0170840603024003910

Pedregosa, F et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research, 12*, 2825–2810.

Renneboog, L., Ter Horst, J., & Zhang, C. (2008). The price of ethics and stakeholder governance: The performance of socially responsible mutual funds. *Journal of Corporate Finance, 14*, 302–322. https://doi.org/10.1016/j.jcorpfin.2008.03.009

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

Sharfman, M. (1996). The construct validity of the Kinder, Lydenberg & Domini social performance ratings data. *Journal of Business Ethics, 15*, 287–296. https://doi.org/10.1007/BF00382954

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

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

Statman, M. (2006). Socially responsible indexes. *The Journal of Portfolio Management, 32*, 100–109.https://doi.org/10.3905/jpm.2006.628411

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*, 549–572. https://doi.org/10.5465/amj.2010.0962

Svanberg, J., Ardeshiri, T., Samsten, I., Öhman, P., Rana, T., & Danielson, M. (2022). Prediction of environmental controversies and development of a corporate environmental performance rating methodology. *Journal of Cleaner Production, 342*, 130979. https://doi.org/10.1016/j.jclepro.2022.130979

Trumpp, C., Endrikat, J., Zopf, C., & Guenther, E. (2015). Definition, conceptualization, and measurement of corporate environmental performance: A critical examination of a multidimensional construct. *Journal of Business Ethics, 126*, 185–204. https://doi.org/10.1007/s10551-013-1931-8

USSIF. (2018). *Report on US sustainable, responsible and impact investing trends.* US SIF and US SIF Foundation. Retrieved from: https://www.ussif.org/files/Trends/Trends%202018%20executive%20summary%20FINAL.pdf

Wolfe, R. (1991). The use of content analysis to assess corporate social responsibility. *Res Corp Soc Perform Policy, 12*, 281–307.

Wood, D. J. (2010). Measuring corporate social performance: A review. *International Journal of Management Reviews, 12*, 50–84. https://doi.org/10.1111/j.1468-2370.2009.00274.x

Wokutch, R. E., & Spencer, B. A. (1987). Corporate saints and sinners: The effect of philanthropic and illegal activity on organizational performance. *Calif Manag Rev, 29*, 62–77.

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*, 1079–1101. https://doi.org/10.5465/amj.2010.0608

Jan Svanberg, PhD, is an Associate Professor of Business Administration at University of Gävle, and associated with Centre for research on Economic Relations (CER), Sweden. He earned his PhD in Business Administration at Stockholm University, Sweden, and was recently a research fellow at RMIT University, Melbourne, Australia, for a period of two years. In addition to these studies, he has a diverse education with degrees in engineering, mathematics, and physics (M.Sc.) and law (LL.M). His current research focuses on how core problems in accounting, auditing and finance can be solved by interdisciplinary approaches with tools made available by advances in data science, especially machine learning.

Tohid Ardeshiri, PhD, is a Postdoctoral Fellow at the Division of Automatic Control at the Department of Electrical Engineering at Linköping University and Research fellow at the Centre for research on Economic Relations, Sweden. His research interests are in analytical approximations for and various aspects of Bayesian learning, such as variational Bayes, maximum entropy kernels for Gaussian process regression, and mixture reduction.

Isak Samsten, PhD, is a Senior Lecturer in the Department of Computer and Systems Sciences at Stockholm University, Sweden. His main research interest includes machine learning, especially focused on classification and interpretable machine learning, and on complex and temporal data.

Peter Öhman, PhD, is a Professor and Head of Business Administration at Mid Sweden University in Sundsvall, Sweden, and director of Centre for research on Economic Relations (CER). His research focuses on behavioral issues, primarily in accounting and auditing, but also in the banking and property industries. He has published a significant number of scholarly articles in leading academic journals, for example Contemporary Accounting Research, and Behavioural Research in Accounting. He has published books, book chapters and reports, and presented papers at international conferences. He has received international awards for excellent research, is a member of Editorial Boards, and a reviewer of several well-known international journals.

Presha Neidermeyer, PhD, is a Professor of Accounting at the West Virginia University, USA. She is a proud graduate of the West Virginia University College of Business & Economics and the Eberly College of Arts and Sciences. Her research investigates the way culture impacts auditor decision-making. She is serving as an ADVANCE Fellow and is coordinating the Women's Leadership Initiative on campus. Her philanthropic work includes collaborating with various groups both domestically and internationally to assist business growth primarily with micro businesses.