Must social performance ratings be idiosyncratic? An exploration of social performance ratings with predictive validity

Jan Svanberg
University of Gavle, Gävle, Sweden

Tohid Ardeshiri
RISE Research Institutes of Sweden AB, Mölndal, Sweden

Isak Samsten
Department of Computer and Systems Sciences, Stockholm University, Kista, Sweden

Peter Öhman
Department of Economics, Geography, Law and Tourism, Centre for Research on Economic Relations, Mid Sweden University – Sundsvall Campus, Sundsvall, Sweden

Presha E. Neidermeyer
John Chambers College of Business and Economics, West Virginia University, Morgantown, West Virginia, USA

Tarek Rana and Frank Maisano
School of Accounting, Information Systems and Supply Chain, RMIT University, Melbourne, Australia, and

Mats Danielson
Department of Computer and Systems Sciences, Stockholm University, Kista, Sweden and International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria

Received 11 March 2022
Revised 8 August 2022
10 January 2023
14 March 2023
Accepted 14 April 2023

Abstract

Purpose – The purpose of this study is to develop a method to assess social performance. Traditionally, environment, social and governance (ESG) rating providers use subjectively weighted arithmetic averages to combine a set of social performance (SP) indicators into one single rating. To overcome this problem, this study investigates the preconditions for a new methodology for rating the SP component of the ESG by applying machine learning (ML) and artificial intelligence (AI) anchored to social controversies.

The authors gratefully acknowledge the financial support from Länsförsäkringars Forskningsfond.

Sustainability Accounting, Management and Policy Journal
Vol. 14 No. 7, 2023
pp. 313-348
Emerald Publishing Limited
2040-8021
DOI 10.1108/SAMPJ-03-2022-0127

Design/methodology/approach – This study proposes the use of a data-driven rating methodology that derives the relative importance of SP features from their contribution to the prediction of social controversies. The authors use the proposed methodology to solve the weighting problem with overall ESG ratings and further investigate whether prediction is possible.

Findings – The authors find that ML models are able to predict controversies with high predictive performance and validity. The findings indicate that the weighting problem with the ESG ratings can be addressed with a data-driven approach. The decisive prerequisite, however, for the proposed rating methodology is that social controversies are predicted by a broad set of SP indicators. The results also suggest that predictively valid ratings can be developed with this ML-based AI method.

Practical implications – This study offers practical solutions to ESG rating problems that have implications for investors, ESG raters and socially responsible investments.

Social implications – The proposed ML-based AI method can help to achieve better ESG ratings, which will in turn help to improve SP, which has implications for organizations and societies through sustainable development.

Originality/value – To the best of the authors' knowledge, this research is one of the first studies that offers a unique method to address the ESG rating problem and improve sustainability by focusing on SP indicators.

Keywords Artificial intelligence, Machine learning, ESG, Social performance indicators, Weighting problem, Social controversies, Socially responsible investment

Paper type Research paper

1. Introduction

Environmental, social and governance (ESG) reporting has emerged as a driving force for institutional investors making portfolio investment decisions; however, organizational impediments exist in achieving progress with sustainability (Adams and McNicholas, 2007). It appears that the lack of consistent sustainability metrics constrains the progress of identifying, and thus making, sustainable investments. The sustainability of potential investments is measured on a broad range of scales known as ESG ratings. These scores are meant to represent the extent to which social *responsibilities* are addressed by the reporting company (Chen and Delmas, 2011; Nitsche and Schröder, 2018), or more specifically, corporate social responsibilities (CSRs), as Chatterji *et al.* (2016) express the construct targeted by ratings. ESG ratings are problematic conceptually because different raters use inconsistent idiosyncratic definitions for these responsibilities, thus it is uncertain what the ratings represent, and empirically because evidence indicates that ESG ratings have low validity (Chatterji *et al.*, 2016; Christensen *et al.*, 2022), with some research even suggesting these ratings are a proxy for company size (Drempetic *et al.*, 2020). Studies have shown that rater inconsistency is considerable; for example, Christensen *et al.* (2022) found a 30% divergence in the ESG ratings of the same company by three major raters (MSCI, Sustainalytics and Refinitiv). Such assessment differences can play into the hands of institutional investors, who may purposely select information that supports their desire to pursue a policy that improves their public image by favoring "sustainable" investments (Adams, 2002). Despite previous findings that legitimacy depends on the amount of sustainability disclosures (de Villiers and van Staden, 2006), better data availability does not make ESG ratings better – rather it increases rater inconsistency (Christensen *et al.*, 2022). The latter effect suggests that it is the way that rating models aggregate the data, i.e. the importance attributed to indicators of various kinds, that contributes substantially to rater inconsistency. Chatterji *et al.* (2016) argued that the subjectivity with which raters select indicators and assign weights to indicators plays a key role in causing the differences.

This study is concerned with how data is aggregated into holistic ratings. The problem with aggregating indicators into one metric is how to determine the importance attributed to each indicator. Chatterji *et al.* (2016) described differences in theorizing that exist between raters – for example, some analysts rate firms according to their products' safety, while others do not, and some raters place more weight than other raters on, for example, the social component relative to the environmental and governance components by representing it with more indicators and assigning more weight to it. Several indicators used by some raters are not used at all by others (equivalent to weight = 0), and the emphasis given to them may also differ, for example, by using one metric for employee issues versus using several different metrics for employee issues, such as employees' health and safety, training programs and labor relations. The relative weights of the latter three when forming the subcategory "employee issues" may also differ, and this may be affected by raters' manual adjustments, for example, by regarding health and safety as relevant only to some industries while irrelevant to others. The discretionary approach to all such issues by traditional raters leads to rater inconsistency or rater idiosyncrasy.

We see two problems with this situation. The first problem is that ratings are idiosyncratic and therefore subjective in the sense that reasons for both including and excluding indicators and for assigning weights to indicators are not determined by reference to a holistic standard that can be justified conceptually, nor are these reasons disclosed. Although ESG ratings are supposed to measure how companies perform on responsibilities (Chatterji *et al.*, 2016), raters provide no evidence that their inclusion and exclusion of indicators and the weighting schemes with which indicators are aggregated to a rating score represent anything other than their own preferences. The research by Chatterji *et al.* (2016) therefore indicates that there is a need for a holistic standard for companies' performance on CSRs. We address this research gap by exploring how data on companies' compliance with social responsibilities can be used to solve the weighting-scheme problem with a model that predicts compliance with social responsibilities.

A second problem is that raters provide little or no evidence that their ESG ratings are valid. The idiosyncratic weighting schemes make ratings unanchored and floating, therefore having low convergent validity (Chatterji *et al.*, 2016; Christensen *et al.*, 2022). Weighting schemes tend to be justified by referring to the idea that weights reflect financial materiality, but a recent review found little evidence in the literature that financial materiality can be estimated (Christensen *et al.*, 2021). Chatterji *et al.* (2016) found large differences between ESG theorizations, and to our knowledge, there is no evidence that ratings are valid measures of how companies perform on CSRs. We demonstrate how the adoption of compliance with social responsibilities as a measure for companies' performance on social responsibilities enables predictively valid social performance (SP) ratings.

Our focus on SP is primarily motivated by the rich availability (more than ten times greater) of social controversies compared with the availability of environmental and governance controversies. Furthermore, the differences between the company behaviors included in the three topics E, S and G are so distinct that it is necessary to study each sustainability topic separately. The topic differences have been confirmed by findings in many studies (cf. Lee and Suh, 2022), for example, one regarding the difference between the importance attributed to the topics E, S and G for investment decisions made by large institutional investors (Krueger *et al.*, 2020) and another concerning their different relationships with financial outcomes (Barnett and Salomon, 2006; Zhang *et al.*, 2022). To the best of the authors' knowledge, the literature describes neither how the problem with subjective or arbitrary ESG model weights can be solved, nor how an ESG rating methodology with predictive validity can be designed.

We propose solutions to these problems by investigating *the overall research question whether social controversies are predicted by SP indicators,* i.e. *whether companies' SP can be characterized by patterns of SP indicators typical of non-compliance with social responsibilities?* The answer to this question is a prerequisite for the contributions of this study: a positive finding indicates that a weighting scheme can be calculated that is not discretionarily determined by raters but instead estimated from data and represents the significance of SP feature indicators for predicting companies' compliance with social responsibilities. This also suggests that there is evidence that an ESG rating methodology can be constructed with predictive validity – responding to the call for predictive validity tests by Chatterji *et al.* (2016, p. 1608).

2. Prior studies and theoretical framework

2.1 Background

We refer to traditional ESG ratings as those that are aggregations of the information in accounting data, i.e. sustainability reporting, and, as such, claim to measure CSR or CSP. There are a multitude of such metrics, and the screening methods investors refer to when considering ESG in portfolio composition are a spectrum of overlapping approaches (Latinovic and Obradovic, 2013), making conceptual clarification both important and problematic. Despite the considerable biases that may be the effect of financial incentives, most CSR or CSP metrics are produced by commercial companies, not by financially independent institutions (Dahlsrud, 2008). Such ratings aspire to represent a holistic view of a company's sustainable performance, defined as CSR or CSP. The holistic view is achieved through the aggregation of many feature-specific indicators, which involves gauging the importance of one against the other because all features and behaviors of a company cannot be equally important for the holistic assessment. The ESG industry seems not to have derived the ratings from a theoretical concept but has jumped immediately to measurement as if the validity of the ingredients would guarantee the validity of the compound metric. In our view, problems with traditional ESG ratings discussed in section 2.2 are caused by a contradiction between the idea of CSR activities or CSP as voluntary acts of "doing good" and the concept of "responsibility." We therefore outline conceptual nuances as they have gradually emerged in the development of CSR and CSP since the 1950s before discussing validity problems with traditional ESG ratings.

Commercial raters propose the ESG ratings as measures of the extent to which companies act or perform CSR. The ratings are the offspring of a long conceptual evolution since Bowen published his seminal book *Social Responsibilities of the Business Man* (Bowen, 1953). Back then, three CSR themes dominated: the manager as public trustee, balancing competing claims to corporate resources and corporate philanthropy (Frederick, 1994). Formal definitions of CSR appeared in the 1970s with an emphasis on performance (Carroll, 1999). Accepting responsibility was not enough, and responding to responsibility and showing responsiveness became the keys. CSP was an attempt to relate the nuances of responsibilities and responsiveness, the CSR_1 and CSR_2, and represented a desire to emphasize the outcome of socially responsible initiatives (Carroll, 1979; Wartick and Cochran, 1985; Wood, 1991). The link between CSR and corporate financial performance (CFP) became popular in the 80s (Lee, 2008) and has since continued to be the dominant research issue and the main concern for industry practitioners.

Before we discuss the CSR–CFP link, there are a few conceptual distinctions that need to be clarified because they have implications for measurement. A conventional way of thinking about CSR (and CSP) is to assume that it includes certain features and behaviors of companies. The definition then becomes an issue of defining the features and acts by

naming them in categories that represent types of CSR. A definition of CSR developed by researchers for research and used for longer than three decades is provided by Carroll' *s* (1979) four categories of normative expectations: economic, legal, ethical and discretionary. The glue that provides internal consistency to this definition is that companies need to *comply* with societal responsibilities.

Two of the categories in Carroll's concept do not, however, constitute normative expectations, or norms in a strict conceptual sense. In jurisprudence, which is a subject that specializes in the specification of the meaning of norms, the 'ought to' refers to a statement's legal validity (Dworkin, 1985; Kelsen and Knight, 1967; Summers, 1963), which means that a statement is uttered as part of the legal system and obtains its normative meaning from its legal-institutional context. Similarly, formalization in organizations has been described in the sociological literature as the outcome of a process that associates behavioral expectations with organizational membership (Luhmann, 1995), which constructs and maintains the meaning difference between individual and organizational perception. These differences in meaning refer to Galtung's sociological explanation of the difference between normative and cognitive expectations (Galtung, 1959). The sociological explanation of normative expectations is that they are formed by social decision-making processes associated with a social entity, e.g. a group or organization, establishing expectations that become criteria for membership in the entity. Normative expectations therefore differ from cognitive expectations in the way they can be changed in the face of non-compliance: individuals can become disappointed or upset when observing non-compliance with a norm (which they cannot themselves adjust to facts) but they can easily adjust their cognitive expectations if they are not consistent with reality.

None of these theoretical distinctions are noticeable in Carroll's CSR concept. On the contrary, it mixes cognitive and normative expectations as if there were no difference between them. Carroll's inclusion of economic responsibilities is inconsistent with how financial performance is viewed in practice – a company's profit is expected by stakeholders as a fact, not as a norm. CSR defines the responsibilities companies need to comply with to earn the right to prosper financially, but society does not demand that companies or their owners prosper (Kang and Wood, 1995). This interplay between the responsibilities, rights and voluntary aspects of sustainable performances suggests a difference between CSR, which semantically refers to responsibilities and has a tradition of emphasizing compliance, and CSP, which emphasizes performance often in terms of discretionarily selected acts of "doing good" (Wood, 2010). The differences in focus and emphasis have important implications for operationalization. In Carroll's terms, there is a difference in the degree of precision or determinacy between, on the one end of the spectrum, legal responsibilities, ethical or moral obligations and, on the other end, discretionary responsibilities. The most precisely determined responsibilities are the legal, for which there may be financial or even punitive repercussions for non-compliance. The moral obligations are less distinct than the law and are not institutionally clarified, e.g. through judicial procedure. There is arguably a moral responsibility for businesses to reduce CO_2 emissions, but it would be difficult for any business to know whether they have done enough or not. Carroll's discretionary responsibilities express the expectation that any company should invest efforts in doing good, but neither the nature of the act nor its object are specified. The problem with this notion is that a responsibility to act in a manner the company finds "good" cannot be the norm because the difference between whether the company's acts are compliant or not must be decided by the company. No one else knows what the company discretionarily regards as a good deed. Non-compliance is therefore undeterminable, and a CSR concept defined as

compliance can therefore neither include financial performance nor the discretionary aspect of "doing good."

For these reasons, the aspect of CSP that targets "doing good" is less suitable for operationalization than CSR, defined as the societal normative requirements on businesses. While "doing good" makes sustainability a subjective, highly observer-dependent construct, the compliance-oriented construct is anchored on societal preferences regarding business externalities. Furthermore, the doing good perspective is difficult to reconcile with the idea of a social contract and legitimacy theory because breaches of a social contract cannot be determined unless the terms of the contracts are fairly clear, and loss of legitimacy cannot be argued unless a company is at fault. The idea of companies deciding for themselves what they find to be sustainable business is, in our view, logically inconsistent with legitimacy theory, which emphasizes that companies earn their right to do business by complying with CSR (Deegan, 2002). In the following, we consider these conceptual differences, but we follow the practice of Christensen *et al.* (2022, 2021) and use CSR and CSP interchangeably, both referring to the extent that companies comply with societally defined environmental and social responsibilities. We treat ESG ratings as a practice-focused operationalization of various meanings of CSR and CSP, often vaguely defined and having come to include the governance aspect of a business. ESG ratings are provided by commercial firms that sell both underlying data and assessments to banks, insurance companies and funds. Before turning to the research that specifically addresses the validity of ESG ratings we briefly discuss findings about the relationship between companies' CSR activities and their financial performance because it is considered evidence of the validity of CSR metrics.

In the traditional view, managers should only engage in activities that increase shareholder value (Friedman, 1970), but CSR can present companies and their owners with financial incentives. In contrast, it is conceivable that CSR may be carried out in the interest of shareholders even if they have a negative net present value. Activities could be in the interest of shareholders if they have preferences for CSR (Fama and French, 2007; Friedman and Heinle, 2016; Hart and Zingales, 2017), and the company's management would then maximize shareholder welfare but not shareholder financial value with their CSR investments (Hart and Zingales, 2017). Other non-financial motives to engage in CSR activities are that it satisfies the preferences of broader stakeholder groups than the investors or because managers pursue personal goals (Adams and McNicholas, 2007) at the expense of the owners (Masulis and Reza, 2015), providing managers with a "warm glove effect" (Barnea and Rubin, 2010). In doing so, firms may sacrifice profits (Roberts, 1992; Benabou and Tirole, 2010).

Studies find significant associations for various CSR activities or policy measures regarding the relationship between firm value and CSR. However, the sign of the association differs between results. There is no consensus on whether there is a positive or negative relationship between CSR and financial performance or the firm's value (Christensen *et al.*, 2021). Because much CSR is measured as voluntary activities, there is a selection problem involved – there is often a positive association between voluntary CSR activities and firm value, but when companies are forced to implement CSR, the valuation effect can be negative (Manchiraju and Rajgopal, 2017). The literature also identifies several mediating factors that alter the sign and strength of the relationship between CSR and firm value. The value of companies is affected if CSR is related to positive media coverage (Cahan *et al.*, 2015), to good products that cause satisfied customers and innovation (Luo and Bhattacharya, 2006), to customer awareness of CSR (Servaes and Tamayo, 2013) and to investors' affective reactions to CSR performance (Elliott *et al.*, 2014). The presence and absence of such mediators can cause conflicting results.

The conflicting results regarding the relationship between ESG and firm value are reflected in research findings about the results of investing in high/low ESG portfolios and individual high/low ESG stocks. A good example of the many contingencies that appear to interfere with the prospects of drawing straightforward conclusions is the findings of Zhang *et al.* (2022), who examined ESG using Bloomberg's ratings of Chinese companies. For portfolios, they found a non-linear relationship between ESG and portfolio returns, with the highest and lowest ESG-ranking portfolios earning significantly positive abnormal returns compared with the other portfolios, and for individual stocks, they found pillar differences: good governance plays a distinguishingly positive role in increasing stock returns, while responsible social behavior has a negative impact and environmental performance has an ambiguous effect. They concluded that "investors cannot obtain excess returns by simply holding high-ESG profile stocks" (Zhang *et al.*, 2022). They also found that ESG valuation is sector-sensitive. This study overall reflects the complexity involved in the CSR–CFP relationship and the selection issue discussed by Christensen *et al.* (2021), which means that the weak CSR–CFP or return on investment relationships are sensitive to sample composition because so many aspects of a business potentially impact returns that they cannot be controlled for.

The effects of CSR should appear in the companies' accounts, but as with firm value, there are many inconsistent conclusions about whether it pays off for a company to invest in CSR. Furthermore, the evidence is normally not causal and is based on voluntary firm choices. Firms with strong performance may be prone to invest in CSR, and the relationship could therefore go from financial performance to CSR (Margolis *et al.*, 2009). Combining studies in a meta-analysis does not address this underlying selection problem in CSR activities. A positive relationship between CSR and profitability has been found by Simpson and Kohers (2002), Flammer (2015), Cornett *et al.* (2016). A literature review by Kitzmueller and Shimshack (2012) finds weak support for a positive effect of CSR on firm profitability. Likewise, Margolis *et al.* (2009) perform a meta-analysis of 251 studies and find a small magnitude positive correlation between CSR and financial performance, but other meta-studies find a more robust positive relation between CSR and financial performance (Orlitzky *et al.*, 2003; Busch and Friede, 2018; Atz *et al.*, 2023). The results may be less reliable than they seem, however, because they aggregate results across studies of different aspects of CSR. Moreover, several studies suggest that the relation between CSR and financial performance is mediated by firm-level innovation and industry-level differentiation (Hull and Rothenberg, 2008), a firm's intangible resources (Surroca *et al.*, 2010), reputation and competitive advantage (Saeidi *et al.*, 2015; Busch and Friede, 2018) and how a firm implements its CSR strategy (Tang *et al.*, 2012). Finally, the relationship could even be U-shaped (Brammer and Millington, 2008).

With the many contingencies that can impact the relationship between CSR activities and companies' financial performance, and thus the returns of investing in the companies, it is reasonable to have doubts about causes and effects. The literature appears to conclude that there might be a positive financial outcome from CSR activities, but that the evidence is inconsistent, either because of the complex nature of the relationship or because of the difficulties with measuring the extent to which companies engage in CSR. Some recent findings indicate that markets consider CSR-disclosures relevant as they value companies; for example, Grewal *et al.* (2020) found that sustainability reporting according to the materiality standards developed by the Sustainability Accounting Standards Board resulted in greater price informativeness for some companies. The financial effects of CSR activities on the reporting companies are supposed to be measurable through markets' reactions to CSR disclosures. Such reactions would indicate financial materiality in the sense provided

by referring to a US Supreme Court decision [1] that defined information as financially material if its disclosure would impact a reasonable investor's decisions. The recent finding of Christensen *et al.* (2022) that ESG ratings, which focus on capturing ESG primarily in the financial materiality sense, differ as much as 30% in the assessment of an average CSR activity company supports the interpretation that the inconsistency among the cited findings can be because of differences in CSR metrics. In Section 2.2, we discuss the weaknesses of the commonly used ESG ratings as CSR metrics, and in Section 2.3, we pose our research question and outline how we address it by proposing a rating methodology that has features that should mitigate the problems we identify with traditional ESG ratings.

2.2 Problems with traditional environment, social and governance ratings

The inconsistent evidence regarding the CSR–CFP relationship suggests that ESG ratings may have validity problems. The standard method with which to assess ESG in research and practice is to adopt an off-the-shelf social-component ESG rating or to reconstruct such a rating from a selection of indicators (Chen and Delmas, 2011; Drempetic *et al.*, 2020; Nitsche and Schröder, 2018; Oikonomou *et al.*, 2018). Traditional ESG ratings compute the arithmetic average of ESG indicators (Chen and Delmas, 2011). The use of such simplistic models to estimate the arguably complex ESG concept has several constraints that contribute to the absence of convergent validity of such operationalizations found in previous research (Berg *et al.*, 2022; Chatterji *et al.*, 2016; Christensen *et al.*, 2022). ESG ratings with zero or low validity are a great problem for the finance industry, researchers and corporate managers because the ratings are the most popular navigation instrument for assessing companies' CSR, or ESG, among practitioners. We discuss the weaknesses of the traditional ESG rating methodology that, in our view, are the most likely to cause the low validity of traditional ESG ratings.

The first problem is the predominant use of *discretionary indicator weights*. The rating industry has no answer to questions such as: How should we weigh the indicator "Total Injury Rate Employees" against, for example, the indicator "Gender Pay Gap Percentage"? Neither theory nor empirical findings offer an objective answer to this type of question (Chatterji and Levine, 2006; Hillman and Keim, 2001), but weighing the relative importance of these two indicators against each other is necessary for an aggregated, holistic ESG rating – and entirely determines how companies are rated. For example, the rating industry does not know whether 20 workplace accidents for a particular corporation should be viewed as good or bad (Kotsantonis and Serafeim, 2019). Traditional ESG ratings therefore aggregate indicators discretionarily.

This core problem may be seen as though raters assume the meaning of ESG to be inherent in the indicators. Determining the weights of indicators, however, affects the measure's ability to be a valid representation of ESG as a concept, be it CSR or CSP. What is actually measured with ESG ratings can be questioned because the inconsistency between leading raters is so great that they disagree about which should be the three most significant ESG categories (Berg *et al.*, 2022). The inconsistency between raters' assessments of the same company may be partly explained by the voluntary disclosure of ESG data by reporting companies (Orlitzky, 2013) and inconsistent collection of ESG feature data by raters (Berg *et al.*, 2022), but Christensen *et al.* (2022) find that data availability and quality do not drive rater inconsistency as much as does the use of inconsistent aggregation methodologies. As data availability increases from the 25 percentiles of companies reporting the least ESG data to the top 25% of ESG reporting companies, the difference between ratings of the same company increases on average by more than 30%. Their finding is clear evidence that the main constraint of the current ESG rating methodology is the weighting

schemes and associated (manual) procedures with which ESG ratings are computed from raw data. The use of arbitrary weighting schemes per definition produces arbitrary ratings, and it is not surprising that those different agencies produce inconsistent results compared with other raters.

Arbitrary, discretionary weighting postulates that the relative importance of features to the holistic construct is known by the rater, which is obviously not true (Callan and Thomas, 2009). For example, ESG ratings may overestimate the meaning of programs and policies for reducing a business' impact on society relative to substantial outcomes (Delmas and Burbano, 2011). Discretionary weighting may dilute and distort the impact of the most substantial outcome indicators because it does not rationalize the actual weight differences. Dilution of the impact of the arguably most important indicators would be a hazard to construct validity because there is no guarantee that neglect of substantial outcomes would be consistent with society's conception of acceptable ESG business practice. It follows from logic that subjective weighting schemes can explain the lack of validity of leading raters' ESG ratings found in previous research.

In addition, the use of linear models with manually determined weights creates issues when it comes to representing nonlinearity in the underlying construct. It is likely that ESG indicators have nonlinear relationships with the holistic ESG construct because a policy to increase workplace safety must be more important the more injuries occur in a corporation. This feature interaction is inevitable for a broad range of ESG features. An arithmetic average of indicators-type rating models does not capture nonlinearity (Ding *et al.*, 2020). There is potential nonlinearity in the relationships between individual ESG features and total ESG that is most likely caused by interaction between ESG features. For example, the importance of "Employee Health and Safety Training Hours" depends on the number of employee accidents and the injury rate in the company. The significance is near zero if the company has no issues with accidents or injuries. Furthermore, if CSR or ESG is defined as compliance with the responsibilities defined by society, the construct should have the form of a step function (sharp non-linearity) because companies are either compliant or non-compliant on each topic. Traditional rating methodology does not represent this high-degree of non-linearity.

In conclusion, the literature identifies serious deficiencies in traditional ESG ratings. In relation to this research, Barnett and Solomon (2006) found that the relationship between ESG and investment returns depends on the number of ESG screens the investor uses for screening. Their finding indicates that the screens may not measure ESG or measure it with substantial variation and error. Given these findings, responsible investors are appropriately suspicious of the information the ratings contain. Wong and Petroy (2020) find in their annual survey of major institutional investors' perception and use of ESG ratings that asset managers are more and more dissatisfied with ESG ratings the more they analyze the ratings' inner logic. Furthermore, the discretionary weighting scheme for ESG ratings appears not to be suitable for institutional investors' ESG information preferences (Nofsinger *et al.*, 2019). Institutional investors are indifferent to whether companies have ESG features that are not responsibilities, i.e. features not required of companies to have, but institutional investors avoid investing in companies with ESG features that suggest non-compliance with CSR in the responsibilities sense. Traditional ESG ratings do not satisfy such information needs because there is no evidence that these ratings would measure whether companies comply with environmental, social or governance responsibilities. Nofsinger *et al.* (2019) found that this asymmetric information preference is driven by the fact that, on the one hand, the potential benefits of companies' CSR performance on topics not required of companies are offset by the costs of the CSR activities, while, on the

other hand, the financial damages caused by companies' non-compliance with ESG responsibilities are higher than the cost savings on non-compliance. Controversies caused by non-compliance, e.g. customer boycotts and strikes, trigger unsurmountable costs of non-compliance, and institutional investors therefore pay close attention to this aspect of CSR. There is no indication in the literature that traditional ESG ratings would satisfy the core needs of long-term institutional investors. On the contrary, Utz (2019) found that the Refinitiv ESG ratings do not predict relevant controversies and thus cannot help institutional investors avoid investment in non-compliant companies.

2.3 Social controversies as proxies for holistic social performance

The main problem with the traditional approach to ESG ratings appears to be that the relative importance of the many features that need to be addressed by the ratings cannot be determined without a considerable amount of subjective discretion. This is a problem that, for obvious reasons, calls for urgent exploration of potential solutions. We observe that the leading ESG raters, e.g. MSCI and Sustainalytics, predominantly focus on the financial materiality approach to ESG assessment, thus emphasizing the effects of sustainability issues on the reporting company and its owners rather than the effects of sustainability issues on society (cf. Grewal *et al.*, 2020; Pizzi *et al.*, 2022). This should not be seen as an expression of homogeneity, although some findings suggest that raters are affected by a global homogeneity trend (Saadaoui and Soobaroyen, 2018) and ESG practices can be affected by political changes (Aboud and Diab, 2019), because there is evidence of substantial differences between raters because of theorizing idiosyncrasies (Chatterji *et al.*, 2016). Our research question addresses one potential solution by exploring whether social controversies are predicted by SP indicators, i.e. whether companies' SP can be characterized by patterns of SP indicators typical of non-compliance with social responsibilities? We emphasize the relevance of using the societal critique of companies' SP because social responsibilities are norms placed on companies by society, not by any specific stakeholder groups such as investors or individuals. The issues addressed by such responsibilities are of general concern and relevance, e.g. justice at work, basic human rights, gender and race discrimination etc. We agree with the approaches of certain sustainability information providers, e.g. Ravenpack (www.ravenpack.com) and RepRisk (www.reprisk.com), and the use of controversies in the Refinitiv ESG controversy score (www.refinitiv.com), in that controversies are a relevant label of social responsibility non-compliance. These three use controversies as a characterization of companies based on the actual track records of controversies in the manner of a wrongdoing index (Fiaschi *et al.*, 2020). Wrongdoing indices are descriptive in the sense that they describe the amount and character of the critique each company has been given, but a wrongdoing index says nothing of a company that has received little or no media scrutiny. It is therefore not a rating of companies but rather a record of past misdemeanors. Our approach is a fundamentally different method of using the information contained in controversies. We learn to predict the likelihood that any company complies with social responsibilities based on their SP indicator pattern, regardless of the past record of controversies for a specific company. We use machine learning (ML) to build a model that characterizes companies based on how their SP indicator patterns are typically the target of media scrutiny. Our ML models learn to predict the likelihood of social controversies by looking only at the SP indicators, and the controversies are only used in our models to learn the typical SP features for a company that is criticized for social responsibility non-compliance. Thus, we use controversies not simply to label the companies on their historical track records, but to learn a model that discerns which SP feature patterns are likely predictors of such track records, whether or not a

specific company has had controversies. The relevance of characterizing companies' compliance with or non-compliance with social responsibilities from patterns of SP indicators is that the model that learns to recognize companies that have a high risk of non-compliance can function as a rating model with ratings anchored on the likelihood of compliance with social responsibilities.

The main advantage of a model that estimates the relative importance of SP features in the process of estimating the likelihood of companies' compliance with social responsibilities is that this type of model relies less on the rater's subjective discretion than traditional ESG ratings. The predictive methodology would anchor its assessments of the relative impact of ESG indicators on the estimation of compliance likelihood and would reflect society's preferences for acceptable ESG practice. Because ESG includes a diverse and disparate set of topics (Oikonomou *et al.*, 2018), we simplify the task by focusing on what raters and institutional investors refer to as the social category of ESG. A reason for our choice to focus on the social aspect of CSR and exclude environmental and governance issues is that the social aspect is much more frequently assessed by the media than the two other aspects. The number of controversies available from e.g. the Refinitiv database is more than 10 times as large for the social controversies compared to environmental and governance controversies. The societal feedback mechanism provided by social controversies is therefore a much richer source of information than the other types of controversies.

In addition to its consistency with the information needs of institutional investors (Nofsinger *et al.*, 2019), this approach is also supported by the findings of Nirino *et al.* (2021) that corporations involved in controversies have lower returns on assets than other corporations. Moreover, Aouadi and Marsat (2018) found an effect of controversies on financial performance for high-attention corporations. In addition, corporations receiving recent "bad press" tend to suffer from high volatility in their stock price and market value, often experiencing steep stock price declines (Cui and Docherty, 2020; Muller and Kräussl, 2011). Such effects may, however, depend on the extent of corporate ESG disclosures prior to the negative exposure (Beelitz *et al.*, 2021). However, these findings suggest that compliance is of interest to investors and that an ESG rating methodology anchored on compliance would provide ESG information consistent with institutional investors' preferences. Our focus is therefore on the compliance aspect of CSR, while we ignore the voluntary "doing good" perspective. This distinction has also been discussed as the difference between CSP as "doing good" (Kothari *et al.*, 2009; Mackey *et al.*, 2007) versus social irresponsibility (Arora and Dharwadkar, 2011), with the "doing good" side receiving almost all the attention.

Based on these observations reported in previous literature, we investigate the possibility of developing a rating methodology based on a measure of comprehensive compliance with social responsibilities. Compliance can be measured indirectly by the absence of non-compliance, which is represented by the situation in which a company has not been criticized for non-compliance with important social responsibilities. Social controversies are indicators of non-compliance with social topic responsibilities because controversies occur when companies are perceived to be breaching such responsibilities (Nieri and Giuliani, 2018). We emphasize the distinction between describing social features (i.e. behavior and structures), which is the basis of traditional ESG ratings, and using social feature indicators for assessing a company's compliance with social responsibilities, which is the goal of our investigated rating methodology. In our view, which contrasts with the view of traditional ESG rating methodology, a company that performs within the boundaries of its responsibilities should not be punished for being at fault because the company is in fact not at fault. The traditional ESG ratings, which do not use the responsibilities laid out by society for companies to follow as their performance standard

and reference point for ratings, may conclude that a company that is acting consistent with societally established preferences is given a low rating because of the subjective preference methodology used in traditional ratings. All ratings that do not aspire to reflect societal standards are, in our view, inconsistent with the idea of CSR or ESG as a societal-level navigation system to guide the practice of socially responsible investment and the efforts of all business managers in their efforts to improve the SP of their businesses. We therefore view social controversies as signals that a company's pattern of SP indicators is considered socially illegitimate from either a moral, soft law (international standards) or legal point of view.

The importance of investigating a rating methodology that can mirror societal responsibilities motivates our examination of the extent to which social controversies can be predicted with a company's social feature indicators. From the compliance perspective, controversies provide information different from that provided by SP indicators. Social controversies are society's reactions to social category behavior in the sense that behaviors are labeled inappropriate. Controversies thus label a particular company, which would, without the controversy, be represented only by a pattern of SP indicator values. Social controversies contribute information about the performance standards set by society that companies need to live up to (the social responsibilities). The controversies indicate that there may be problems with the transgressor's structures and processes because, as signaled by controversy, those have not been sufficient to prevent the critique (Nieri and Giuliani, 2018).

Whereas traditional ESG ratings are anchored on the subjective preferences of the rater, the legitimacy of which must be questionable in the eyes of institutional investors, the non-compliance accusation communicated by controversy is put forth by a democratic institution, i.e. the media, but the critique is possible only if it argues that the critiqued company has violated an important responsibility toward society or significant stakeholders. Defining SP as compliance with social responsibilities therefore makes the development of rater-indifferent SP ratings possible. The role of social controversies in the development of such a rating methodology is to label companies, or rather their individual patterns of SP indicators, as compliant or non-compliant, referring to each company's historical record of controversies. Social controversies concerning, for example, product harm (Cai *et al.*, 2012; Klein and Dawar, 2004), human rights violations and breaches of labor law (Aouadi and Marsat, 2018) evoke detailed criticism of instances when corporations are non-compliant with their social responsibilities. In fact, controversies operate as a labeling process (Faulkner, 2011), through which non-compliant corporations risk legitimacy loss (Deegan, 2002).

Controversies obviously differ in origin from the self-reported SP indicators that originate from annual reports. This difference in the origin of the data makes the controversies "out-of-sample" data in relation to the SP indicators. This extra-indicator information assigns compliance relevance, or weights, to individual SP indicators and to indicator patterns based on whether the indicators and patterns are typical of companies exhibiting non-compliance. Our use of controversies to label patterns of SP indicators for each company enables assessment of the relative importance of individual indicators to a holistic SP construct because the relevance of each indicator can be defined as its contribution to the likelihood that a corporation is (non-) compliant with its social topic responsibilities. According to our argument, this is the most compelling reason for investigating social topic controversies. Controversies provide the constitutive information for the definition and measurement of a compliance-oriented and rater-indifferent SP rating.

Social controversies are, in addition to being the decisive component in a rating methodology anchored on societal social responsibilities, events of interest in themselves, and the critique they express is of such a nature and significance that institutional investors typically strive to avoid investing in companies that attract controversy in their effort to protect their reputation as responsible investors (Krueger *et al.*, 2020; Nofsinger *et al.*, 2019) because investment in controversial corporations may cause the withdrawal of funds (Grappi *et al.*, 2013).

Controversies also reflect a lack of policies, structures and routines necessary for compliance with society's expectations (Nieri and Giuliani, 2018). Controversies signal systematic breaches of moral or legal norms in corporations striving to earn profits and gain market share (Fiaschi *et al.*, 2017, 2020; Giuliani *et al.*, 2015; Surroca *et al.*, 2013). A controversy indicates that the misbehaving corporation lacks governance structures or processes that would, if present and operational, have prevented the corporation from engaging in reprehensible behavior (Nieri and Giuliani, 2018), for example, inappropriate management-team or supply-chain procedures (Chiu and Sharfman, 2018). For SP, there is an abundance of such potential structures and processes that may, if failing, cause social controversies relating to anti-competitive behavior, business ethics in general, patents and intellectual property, a lack of respect for human rights, tax fraud, child labor, inappropriate executive management salary schemes, etc. If social controversies can be predicted by using the abundantly available SP indicators, this would suggest that the inappropriate behaviors that have become normal in misbehaving corporations (Earle *et al.*, 2010) are present in the information provided by these corporations' accounting disclosures. Our research may therefore, in addition to examining a precondition for controversy-anchored SP ratings, be understood as investigating the possibility of detecting such weaknesses in the at-fault corporations' accounting.

3. Method and data

3.1 Research design and measures

We adopted a cross-sectional design with predictive modeling performed through ML experiments. To develop an SP rating, it is appropriate to sacrifice the theoretical accuracy of explanatory modeling for the considerably higher empirical precision associated with predictive modeling (Collopy *et al.*, 1994). A rating that cannot distinguish between corporations with high precision is of little use to an investor, as discussed by Shmueli (2010) and Bzdok *et al.* (2018). The differences between methodologies indicate that predictive modeling is better than explanatory modeling for developing a diagnostic method, such as an SP rating.

We used eight ML algorithms that learn to predict the social controversies of corporations by examining their SP indicators over a 10-year window. The five performance measures used in evaluating the predictive performance were precision, recall, *F*-measure, area under the receiver operating characteristic (ROC) curve (hereafter, AUC) and precision recall curve (PRC).

Data were obtained from Refinitiv Eikon, which compiles more than 400 ESG indicators, of which 129 are SP indicators enumerated in Appendix 1, and a summary description of the inputs and outputs of the ML models is presented in Figure 1 (regarding the algorithms in the figure, see Sections 3.2 and 3.3). SP indicators can be divided into substantial outcomes, for example, salary gap and employee satisfaction and policies to improve them (Delmas and Blass, 2010). We are not primarily interested in understanding the individual features that drive SP in general but in predicting SP using the collective information contained in the large number of SP indicators. Therefore, we use the full set of SP indicators provided by

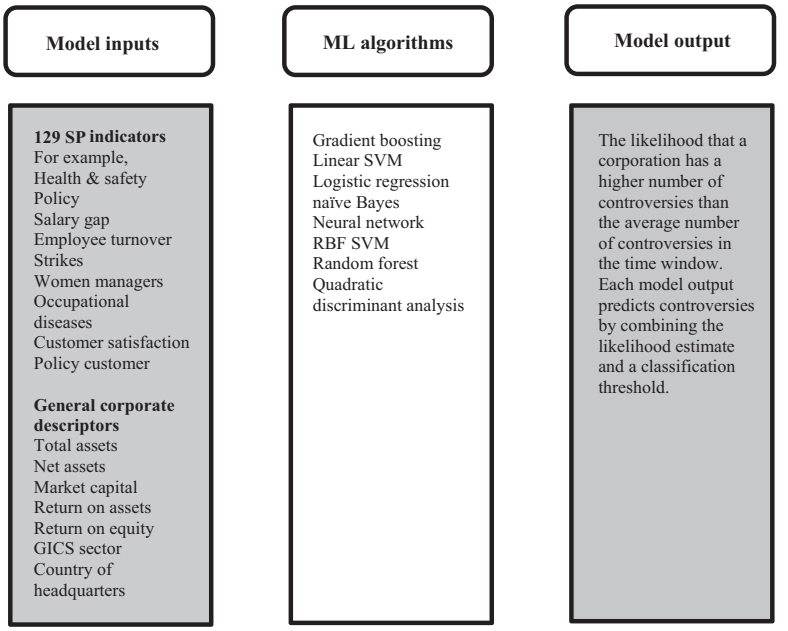


Figure 1.
Summary of inputs, algorithms and the output variable

Source: Created by the authors

Refinitiv Eikon, together with their definitions of the indicators. The relative importance of the indicators is determined by the data, not the researcher. In addition to the SP indicators, we included seven corporate descriptors that do not target SP: total assets, net assets, market capital, return on assets, return on equity, the Global Industry Classification Standard sector and country of headquarters.

The data set covers 2,517 corporations over the 2009–2018 period, during which about 80% of the corporations experienced a social controversy. Table 1 describes the sample, which included all available corporations with an overall ESG rating from Refinitiv Eikon for all studied years. We chose this sample because we wanted to make sure that we selected corporations for which SP indicators were richly represented. Even in this ESG reporting-oriented sample, there were many corporations for which data on certain indicators are lacking. Regarding the country of headquarters, our sample includes corporations based in 61 countries. Most corporations were based in the USA (707); four other countries, i.e. Japan, the UK, Canada and Australia, had more than 100 corporations in the sample.

The corporations were classified into disjoint categories, one with high SP and one with low SP. Social controversies were frequent, with the average number of controversies per corporation during the 10-year window being 5.17. We defined the disjoint classes as corporations having fewer (high SP) and more (low SP) than the average number of controversies. We called the corporations "negative cases" in the former class and "positive cases" in the latter. The number of positive cases, each having six or more social controversies, was 561.

The goal was not to model the risk of future controversy based on a corporation's past SP indicators but to assess the likelihood that an indicator pattern is associated with a corporation's risk of having a controversy in a cross-sectional sense. The longitudinal aspect

Industry	No. corporations	No. controversies	No. corporations with controversies	No. controversies when > 0					Assets in US $ billion	Assets in US$ billion per corporation				
				Min	Max	Mean	SD	Skewness	Kurtosis		Min	Max	Mean	SD	Skewness	Kurtosis
Communication services	161	1566	143	1	131	9.33	16.6	4.45	25.13	3415	0.06	284	21	40	3	14
Consumer discretionary	297	2508	225	1	92	8.43	13.5	3.36	13.06	4153	0.11	331	13	35	6	46
Consumer staples	156	1364	119	1	107	9.24	14	3.8	20.49	2377	0.29	160	15	23	3	15
Energy	183	1137	135	1	65	6.53	9.87	3.36	13.45	4581	0.07	291	25	51	3	12
Financial	398	3640	270	1	115	8.77	16.8	3.9	16.86	79534	0.24	2509	199	393	3	11
Health care	141	1733	111	1	71	10.89	14.2	2.01	3.91	2257	0.05	144	16	24	2	7
Industrial	420	2005	295	1	46	5.92	7.7	2.69	7.95	5770	0.25	457	13	35	10	122
Information technology	201	1566	143	1	127	7.82	15.1	5.03	32.24	2387	0.16	182	11	24	4	24
Materials	277	1068	181	1	53	4.93	7.1	4.13	21.36	2703	0.04	96	9	15	3	14
Real estate	155	138	65	1	12	2	1.87	3.07	12.36	1667	0.42	78	10	12	2	7
Utilities	124	505	87	1	29	5.22	5.69	2.24	5.54	3494	0.13	259	28	36	3	15

Source: Created by the authors

Table 1.
Sample corporations included in the data set described by industry, number of controversies and average total assets per corporation

of the data was captured by using a simple method in which the indicators are averaged if numerical or encoded with dummy variables if binary.

3.2 Machine learning algorithms

The models for assessing SP were developed in ML experiments in which one algorithm at a time extracted information from SP indicators by associating patterns of indicators with social controversies. The eight ML algorithms have different functions and therefore capture different aspects of this learning task. The idea is not to optimize the prediction or find the best method, but to demonstrate whether prediction is possible with a library of algorithms. The broad set of algorithms illustrates how various aspects of the data affect predictability; for example, the impact of nonlinearity can be understood by looking at the difference between a linear and a nonlinear algorithm. We summarize the algorithms in alphabetical order below. All experiments are implemented in Python using scikit-learn and xgboost for fitting estimators and using Pandas for data management and manipulation.

Gradient boosting (GB) is an ML technique for regression and classification problems that produces a model comprising a prediction-model ensemble typically organized as decision trees. The model was built stage-wise to minimize a loss function. Multiple weak models were combined to form a strong ensemble model by reweighting the training data of the SP indicators to focus the learning on those corporations that the algorithm cannot predict correctly. GB defined a loss function and used the gradient of the loss function to reweight the firms to focus on misclassifications using the logistic loss. It has been shown to be a strong classifier for a wide range of tasks (Babajide Mustapha and Saeed, 2016; Sigrist and Hirnschall, 2019).

Linear support vector machine (linear SVM) is a multidimensional linear method that separates two classes of data by drawing an $(n - 1)$-dimensional plane, where n is the number of indicators in the feature space, and separating instances of data lying on one side of the plane from instances on the other side. Introduced by Vapnik (1995), this method is said to be an effective approach to pattern recognition and prediction. For example, prediction of stock price movements and bankruptcy can be addressed if the data are linearly separable (Xu *et al.*, 2009).

Logistic regression (LR) estimates the coefficients of a regression equation and classifies data according to the output of this equation (Menard, 2011). With LR, the independent variables in the present case were the SP indicators, and the dependent variable was social controversies. Each feature value was multiplied by a weight and then summed. LR related controversies to the indicators with the goal of finding the best-fit set of a linear combination of indicators to distinguish between positive and negative cases. Each feature was multiplied by a weight, and then all were summed. The result was transformed by a sigmoid function, producing the binary output, and the supervised learning generated the coefficients predicting a logit transformation of the probability.

Given the class, naïve *Bayes* (NB) classifiers build on the assumption that all feature values are independent. These methods are probabilistic classifiers developed from Bayes' theorem. Here, the NB classifier computed the conditional probability of a controversy from the pattern of SP indicators. Under the naïve assumption that these indicators are independent, the classifier was the conditional probability of a controversy multiplied by the product of the conditional probability of each SP indicator given a controversy (Gulo *et al.*, 2015).

Neural networks (NN) are multiple layers of functions, called artificial neurons, with an output layer consisting of a logistic, softmax or linear regression model. When the layers relate to one another via an activation function, NNs can reproduce any linear or nonlinear

function. With their artificial neurons, NNs mimic the functioning of the human brain's interconnected neurons. The links between neurons have numeric weights that form the long-term memory of the NN. NNs have many good features, such as generalization capability, the ability to learn highly nonlinear relationships and not requiring any assumptions about data distribution. They are, however, of limited use in applications requiring interpretation because their operation is opaque. NNs have better forecasting abilities than statistical regression models and work well on fuzzy or complex data (Ghritlahre and Prasad, 2018). They are nonlinear and nonparametric models that capture unknown interactions (DeTienne *et al.*, 2003; Safa and Samarasinghe, 2011; Sözen, 2009).

Radial basis function support vector machine (RBF SVM) is a version of linear SVM that uses a radial basis function to potentially improve data classification when the data are not linearly separable with an $(n - 1)$-dimensional hyperplane (Ring and Eskofier, 2016). The RBF SVM is called a kernel, which is a window for mapping the nonlinearity in the original space to a higher order space in which the data are linearly separable.

Random forest (RF) has become a popular ML algorithm and is considered state of the art in many applications. Random samples of training data are used to create decision trees, which are then combined to form an ensemble model (Breiman, 2001). At each node, a limited number of SP indicators are sampled to construct each tree separately, which increases variability. A majority vote is taken among the trees, here representing the social controversy prediction. RF works well with outliers and noise in the training set (Yeh *et al.*, 2014), which can be expected with SP data because of "greenwashing" (i.e. information disseminated to create a false appearance of environmental friendliness) and the absence of binding standards for CSR disclosure, both of which are prone to causing data errors. Overfitting is avoided, and precise forecasts are obtained by using the majority vote among trees (Breiman, 2001).

Quadratic discriminant analysis (QDA) is a method that can estimate nonlinear dependencies between complex indicator patterns, such as those in the SP indicators and social controversies. Unlike linear discriminant analysis, it can capture such dependencies by not assuming that the covariance of each class of data is identical (Anagnostopoulos *et al.*, 2012; Ou and Wang, 2009; Yuan *et al.*, 2017). QDA generates a model developed from conditional data densities by constructing a quadratic decision boundary.

3.3 Experiments

Hyper-parameters for the algorithms are presented in Table 2 and in full in Appendix 2. These parameters ensure that our study can be reproduced and provide additional background to its interpretation because variations in settings affect the performance of algorithms. For simplicity, we used the default settings for scikit-learn, version 0.22.

A partition method that achieves reliable estimations of the generalization performance of the ML algorithms and economizes on scarce data is k-fold cross-validation. The cross-validation partitions the data set into k disjoint folds and allows training to be conducted iteratively on $k - 1$ fold, with one fold spared for testing. Repeating this procedure in a circular manner uses the entire data set for both training and testing, ensuring that the predictive performance is evaluated from k different viewpoints. The result of this procedure is k performance measures, the mean of which represents a more reliable estimate of the generalization performance than if the predictive performance were evaluated using one simple partition. In this study, we used stratified 10-fold cross-validation to ensure an equal number of positive and negative cases in each test set.

The performance of our ML models was estimated using several measures, each measuring different aspects of learning ability (Alpaydin, 2010). The measures are summarized in Table 3.

Classifier	Description	Notes
GB	Gradient boosting	Learning rate of 0.1
Linear SVM	Linear support vector machine	Linear kernel with $C = 0.025$
LR	Logistic regression	Ridge regularization with $C = 1$
NB	Naïve Bayes	No hyperparameters
NN	Neural network	Four hidden layers of size 100 using the RELU activation function
RBF SVM	Radial basis function support vector machine	RBS kernel with $C = 0.025$
RF	Random forest	100 trees
QDA	Quadratic discriminant analysis	No hyperparameters

Table 2.
Hyperparameters

Note: Scikit-learn log likelihood ratio (tol) of 0.0001 for the QDA
Source: Created by the authors

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

Table 3.
Basic measures

Source: Created by the authors

Precision measures the sensitivity of the classifier, i.e. its accuracy in predicting the controversy, and recall is considered the ability to find as large a fraction as possible of the controversy-affected corporations in the whole data set. Precision [equation (4.1)] and recall [equation (4.2)] are conflicting measures, while the F-measure [equation (4.3)] captures the trade-off between the two measures.

The AUC measure is the area under the curve defined as a plot of $True_{positive}$ [equation (4.4)] versus $False_{positive}$ [equation (4.5)]. It estimates the probability of a classifier ranking a true-positive case ahead of a false-positive case and is therefore a measure of the model's ranking performance. Similarly, the PRC estimates the mean precision for multiple thresholds of recall and is used to measure the trade-off between precision and recall. It is defined as the area under the plot of precision versus recall. The main benefit of both AUC and PRC is that they are insensitive to the class distribution of the training and testing data, as opposed to the accuracy.

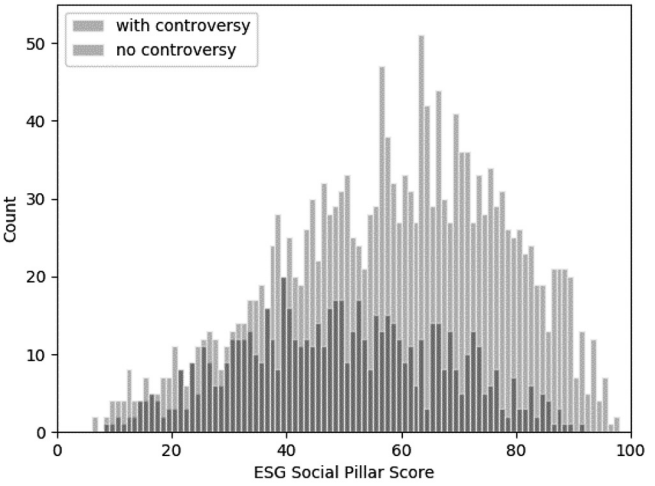
4. Results

4.1 Predictive performance of different learning approaches

We first investigate whether the social component of the ESG rating from Refinitiv Eikon would provide appropriate data for classifying corporations with and without

controversy. Figure 2 displays a histogram where the grey distribution shows corporations with controversy and the black distribution shows corporations without controversy. Visual inspection reveals that the distributions are similar, with no social-component ESG rating being a suitable point on the *x*-axis for defining the two classes. The social-component ESG rating is particularly insensitive to the likelihood of controversies in the 50–70 range, where most of the corporations are found. In this range, the two distributions do not provide any support for a classification. This evidence is consistent with the findings of Utz (2019), who found that Refinitiv ESG ratings do not predict controversies at all.

Analyzing the predictive models, we first discuss the ability of the ML algorithms to learn the SP estimate. Table 4 depicts the performance of the ML algorithms. The static



Notes: Corporations experiencing at least one controversy in the 10-year window are classified as "with controversy" on the y-axis and corporations experiencing no controversies are classified as "without controversy"
Source: Created by the authors

Figure 2. Distribution of social controversies over corporations and the social-component ESG rating from Refinitiv Eikon

Algorithm	Precision	Recall	*F*-measure	AUC	PRC
Gradient boosting	0.7935	0.6414	*0.7088*	*0.9248*	*0.8295*
Linear SVM	0.5541	0.7378	0.6320	0.8674	0.7276
Logistic regression	0.6132	*0.7788*	0.6852	0.8982	0.7898
Naïve Bayes	0.6803	0.4632	0.5468	0.8335	0.5976
Neural network	0.7837	0.5665	0.6533	0.8920	0.7797
RBF SVM	0.4527	0.7216	0.5554	0.8172	0.6280
Random forest	*0.8417*	0.5756	0.6823	0.9165	0.8104
Quadratic discriminant analysis	0.7750	0.0195	0.0378	0.5149	0.3096

Notes: Linear classifiers are linear SVM, logistic regression and naïve Bayes. The highest value in each column is italicized
Source: Created by the authors

Table 4. Predictive and learning performance

measures precision, recall and *F*-measure primarily reveal the predictive performance of each algorithm at a set certainty threshold in predicting a positive case, whereas the dynamic measures AUC and PRC address performance trade-offs. While balanced performance is preferable, an emphasis on precision is inevitable because high precision is a key to portfolio composition. The precision column reveals that RF is the most precise (84.2%), followed by GB (79.4%), NN (78.4%) and QDA (77.5%). The bottom four clearly lag in performance.

RF's high precision can be linked to its learning capabilities because it produces models that have uncalibrated probability estimates, which require the model to have high confidence in predictions. Table 4 also shows that GB has the highest levels in three of the five categories (i.e. *F*-measure, AUC and PRC), but the difference from RF in these three scores is small. It can also be seen that the top performers are not the same in all measures, suggesting that the choice of algorithm depends on which performance measure is the most important. Logistic regression has high recall and good AUC and PRC, suggesting that a linear model performs well in these aspects of the prediction despite the arguments that the underlying construct is complex and nonlinear.

In evaluating the ranking and predictive performance of the algorithms, we also compute the AUC and PRC in Figures 3 and 4. The AUC measures discrimination, equivalent to the probability of a randomly chosen positive instance being ranked higher than a randomly chosen negative instance, i.e. it is equivalent to the two-sample Wilcoxon rank-sum statistic. The AUC should be as close to the top left-hand corner as possible, and a curve below the dashed line is no better than a random guess. The plot in Figure 3 is consistent with Table 4, i.e. RF, GB, NN and logistic regression perform more strongly than the other algorithms. This is partly contrary to our expectation that logistic regression would be less suitable as a predictor of controversies because of the nonlinearity of the underlying construct.

The PRC is a dynamic measure of the predictive performance of the algorithms, and Figure 4 shows how much precision needs to be sacrificed to obtain a certain level of recall. Ideally, a curve should reside in the top-right corner, as far as possible from the bottom-left corner. As long as the algorithms need to predict with absolute certainty which corporations

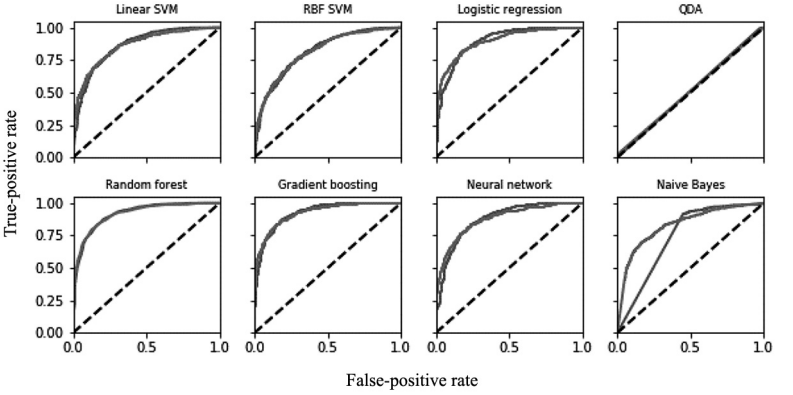
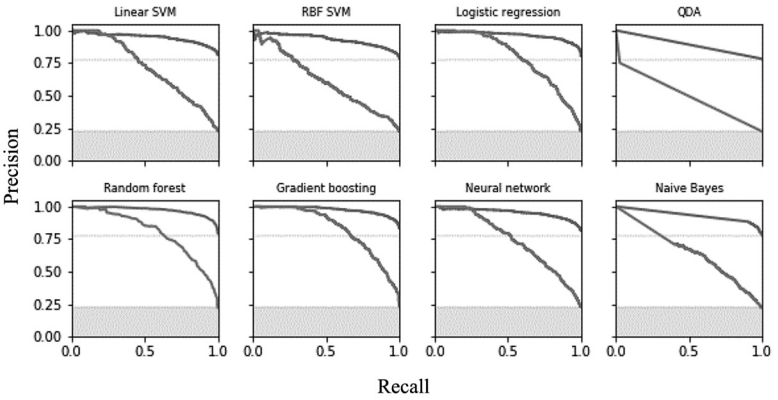


Figure 3.
AUC for the eight learning algorithms

Notes: The black and grey lines represent the AUC for predicting non-controversy and controversy, respectively. Linear classifiers are linear SVM, logistic regression and naïve Bayes
Source: Created by the authors



Notes: The black and grey lines represent the PRC for predicting non-controversy and controversy, respectively. The PRC shows the precision of a classifier as the recall increases. The top region (defined by the dashed grey line) is where a predictor performs better than random guessing for the non-controversy cases, and the region between the bottom and top regions shows the region where a predictor performs better than random guessing for the controversy cases. Linear classifiers are linear SVM, logistic regression and naïve Bayes

Source: Created by the authors

Figure 4.
PRC for the eight learning algorithms

have controversies, they maintain high precision; however, because they are required to find a larger proportion of the total number of corporations with controversies, the algorithms are, as the graphs show, forced to sacrifice certainty to identify more corporations with controversies. Visual inspection confirms the calculations in Table 4. The black (i.e. prediction of non-controversy) and grey (i.e. prediction of controversy) curves are situated in the top right-hand corner for the same group of algorithms that performed well in terms of AUC. RF, GB, NNs and logistic regression can better accomplish the trade-off between precision and recall than can the other algorithms. The grey area at the bottom of the graphs represents the class distribution. Several algorithms offer high precision of around 0.8 at a recall well above 0.5.

4.2 Controversy prediction as social performance

An SP rating anchored to social controversies should be negatively correlated with the number of controversies, but the ML ratings do not automatically assign low ratings to all corporations that have been involved in many controversies. This is because the rating is estimated solely from SP indicators once the rating model is learned. When the ratings are calculated, the predictors do not know which corporations have had controversies and how many. Despite enduring more controversies than average, some corporations deviate from having social-component behavioral indicator patterns indicative of compliance with social norms, i.e. they look good from the rating model's point of view. Such corporations are awarded high ratings. Other corporations have had no or fewer than an average number of controversies, although they exhibit the typical pattern of traits of corporations that are controversy prone according to our best-performing models. They are therefore awarded

low SP ratings. Nevertheless, on average, corporations involved in more controversies should have lower ratings than corporations involved in fewer than average or no social controversies.

Figure 5 reveals a negative correlation between all ML ratings and the number of controversies corporations have. In this graph, the *x*-axis shows the number of controversies per corporation and the *y*-axis shows the ratings. The methods with the highest predictive performance all have smooth distributions. As can be seen, the complex models tend to use what could be called more discretion than do the linear models in identifying exceptions to the many-controversies/low-rating pattern. This would not be the case if the models were overtrained.

As expected, there is a difference between logistic regression and, for example, RF in how the ratings can depart from strict adherence to assigning a low rating to all corporations with many controversies. There is a considerably wider distribution in the ratings for corporations in the range of 20–40 controversies for RF than for logistic regression. This suggests that RF may give weight to more complex SP indicator patterns than does logistic regression, allowing the former model to detect exceptions to the many-controversies/low-rating rule when determining the SP rating. The ability to balance between such exceptions and high performance in predicting controversies is a virtue of a good SP rating, because some corporations that have had many controversies may have traits and behaviors typical of high SP performers.

5. Discussion and conclusions

A prerequisite for socially responsible investment is the availability of holistic ESG ratings built with legitimate and valid methodology. Legitimacy and validity may have different meanings depending on the goal and beliefs of the investor. Institutional investors, believing that a company's harmful impacts on society are of interest only if they also harm the company itself financially, would ask for ratings measuring financial materiality.

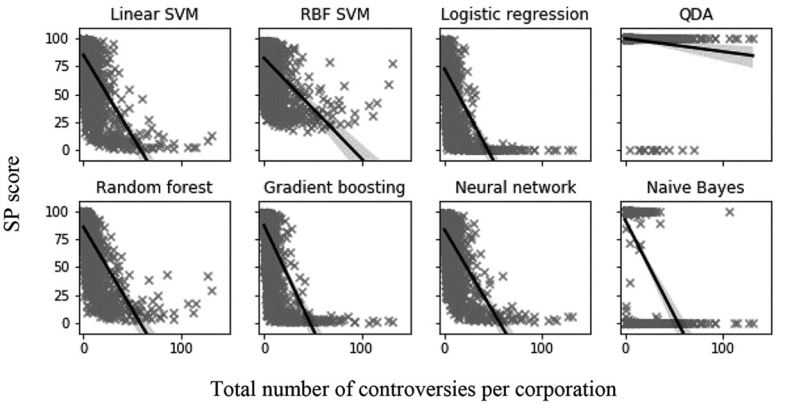


Figure 5.
Correlation between the SP score produced by the eight ML algorithms and the number of years in which corporations experience controversies

Notes: Linear classifiers are linear SVM, logistic regression and naïve Bayes. The SP score refers to the "likelihood estimate" that each model outputs prior to classification. In a fully developed SP rating method, this score would be the rating

Source: Created by the authors

The financial materiality perspective falls short, however, on a number of issues compared with perspectives on materiality that include externalities as described by Adams *et al.* (2021), and many stakeholders, therefore, would be interested in sustainability information on company performance that covers topics that are material according to an impact materiality perspective. Impact materiality is judged by, broadly speaking, the company's impacts on people and the environment, regardless of financial impacts on the company. It has an emphasis on the recognition of broad stakeholder impacts, non-financial impacts and long-term cumulative impacts (Cooper and Michelon, 2022). The upcoming EU Corporate Sustainability Reporting Directive and the EU Sustainability Reporting Standards are the EU's response to this need. The legislation mandates EU companies to include impact materiality as a main principle for identifying topics to report. This revolutionary shift in focus on what companies should report ought to be accompanied by a corresponding shift in focus on what should be assessed by a rating. Current practice among leading ESG raters, however, continues to argue that their ratings should measure the contribution of ESG according to various features' financial materiality.

One possibility to develop a rating methodology that targets impact materiality from the point of view of society (externalities) is to anchor the ratings on a societal-level assessment of companies' compliance with ESG responsibilities. Such an approach should not come as a surprise or stand out as remarkably different from current popular approaches, given that leading ESG ratings are supposed to measure how companies perform on CSRs (Chatterji *et al.*, 2016). A prerequisite for this approach is that ESG feature indicators contain information about company behaviors indicative of non-compliance with ESG responsibilities.

We address this issue for the social component of ESG by investigating whether SP indicators can predict social controversies. We view evidence of such a prediction as evidence that information on compliance with social responsibilities is contained in the SP indicator patterns. Our results indicate that the likelihood of a company's non-compliance with social responsibilities can be predicted accurately with a range of ML models, even without tuning the parameter settings of the algorithms. We also illustrate, as a contrast to our model's goal and in contrast with what Chatterji *et al.* (2009, 2016) consider an evident tacit goal of all ESG ratings, that the Refinitiv ESG ratings *do not* represent information useful for predicting non-compliance with social responsibilities. It appears that Refinitiv ESG ratings have no ability to identify companies that perform poorly on CSRs. A technical aspect of our findings is that the controversy prediction models that account for feature interaction offer a slightly better prediction of the controversies than other models. For example, the effects on a company's finances of SP may depend on what combination of features a company has, and the relationship with the financial effects as a function of the number of SP screens applied for portfolio composition may accordingly be very complex, e.g. U-shaped (Barnett and Salomon, 2006). This is important because feature interaction between, for example, policy indicators and substantial outcome indicators are an inherent aspect of SP data because of the interplay between such policies and substantial outcomes. Our finding that compliance with social responsibilities can be predicted using the information contained in SP indicators has two main implications.

The first implication of our findings is that our proposed estimation method offers a solution to the weighting-scheme problem with ESG ratings that researchers have struggled with for more than 25 years. For example, Mitchell *et al.* (1997) noted that there is no universal ranking of CSR issues and therefore no way of knowing the relative importance of the issues for a holistic ESG metric. Kennelly (2000) argued that surveys such as those used by Waddock and Graves (1997) do not solve the weighting-scheme problem because response rates are low and respondents have inadequate knowledge. In contrast, anchoring

ratings on the prediction of compliance with social responsibilities is a way of systematically extracting the information that Waddock and Graves (1997) saw as legitimizing their ESG metric by reducing the idiosyncrasy of their indicator weights. Anchoring ratings on the prediction of social controversies provides certainty that SP ratings are developed with the importance attributed to SP issues by the entirety of all media scrutiny of all companies' non-compliance with social responsibilities. From a legitimacy point of view, obtaining importance attributions from the bulk of media coverage of thousands of cases in which companies have failed to comply with social responsibilities should be preferable to relying on the discretionary weights chosen by individual raters.

While the importance attributions developed by a ML model that predicts companies' non-compliance with social responsibilities mirror societal preferences for company behaviors insofar as they are expressed in the company's conformation to social expectations, which is subject to media scrutiny, the weighting schemes provided by leading raters such as Morgan Stanley, Bloomberg or Refinitiv lack such a legitimacy guarantee. Furthermore, our use of a data-driven model that rates companies' performance on social responsibilities is consistent with the generally accepted criterion of reproducibility of scientific data and results because ratings that reflect the likelihood of compliance with social responsibilities can be reproduced by anyone who has access to the controversy data, the ESG feature data and the standard settings of, e.g. the RF ML algorithm. Tradition ESG ratings do not fulfill this fundamental quality criterion for scientific research because they are proprietary assessments, often manual and undisclosed, about which the researchers or investors using them know very little. Our proposed approach makes ESG ratings completely reproduceable public property. Full transparency regarding how the model assigns importance to features for its assessment, omitted in this study, can be extracted from the model, for example, by using the variable importance assessment method called SHAP by Lundberg and Lee (2017). Importance attribution can be determined both locally, i.e. how important features are for the assessment of a particular company, and globally, i.e. feature importance for a sector or the whole universe of companies.

The second implication of the finding that compliance with social responsibilities can be predicted is that predictive modeling offers a way to assess the validity of ESG ratings. Previously, researchers had to rely on assessing convergent validity, which is simply a comparison of ratings with other ratings. However, if other ratings are wrong, they will serve as poor references for validation; thus, Chatterji *et al.* (2016) suggested that approaches testing the validity of ratings with predictive modeling should be explored. Our research shows that predictive validity is a viable criterion for evaluating ESG ratings. Predictive modeling with supervised learning has a built-in logic for assessing validity because it requires a holistic measure to represent the predicted construct. The unequivocal evidence in several previous studies of the low convergent validity of leading ESG ratings indicates that predictive validity may shed new light on whether ratings capture how companies perform in their CSRs.

There is no evidence in the literature indicating that the leading traditional raters measure externalities relating to CSR or that they are successful in measuring non-compliance with responsibilities (cf. Utz, 2019), and we demonstrate that the weighting scheme of Refinitiv's ESG ratings appears unsuitable to predict compliance with social responsibilities. Further, the information services on the market providing controversy track-records, e.g. RepRisk, Ravenpack and Refinitiv controversy scores, which are wrongdoing indices (cf. Fiaschi *et al.*, 2020; Nieri and Giuliani, 2018), provide no solution to the problem of discretionary weighting schemes because controversy track-records are

descriptive – a wrongdoing index does not constitute an assessment of the companies that have not been criticized for non-compliance.

This study has some limitations and provides opportunities for future research. First, we did not model the time structure of the data, meaning that we did not use all the information in the SP indicators. Future modeling to improve predictive performance should use a longitudinal design. Second, our prediction models treated all social controversies as if they were identical, which may have introduced a bias in assessing a corporation's likelihood of experiencing a controversy. Future research might consider a model investigating the likelihood of controversies and assigning this to the assessment. Third, we did not adjust the models for corporation size and unequal media coverage between countries. Finally, we used a 10-year window in which we observed the differences between companies' SP indicator patterns and controversies record. The long-time window may introduce endogeneity problems. The endogenous effects would typically decrease the size of the effects in the study because controversial companies strive to avoid being targeted by more criticism in the future by implementing more policies and adjusting their behaviors (Utz, 2019). Future refinement of our rating methodology should investigate how to replace social controversies with a corporate-level wrongdoing index, a scaled and filtered metric of the wrongdoing signaled by controversies (Fiaschi *et al.*, 2020). Nevertheless, this study indicates how supervised ML promises to make social responsibility performance ratings less dependent on raters' preferences than is the case with today's idiosyncratic ratings, reproduceable with standard ML methods and easy to validate with generally accepted measures of predictive validity.

Note

1. TSC Industries v. Northway, Inc., 426 U.S. 438, 449 (1976). See also Basic, Inc. v. Levinson, 485 U. S. 224 (1988).

References

Aboud, A. and Diab, A. (2019), "The financial and market consequences of environmental, social and governance ratings: the implications of recent political volatility in Egypt", *Sustainability Accounting, Management and Policy Journal*, Vol. 10 No. 3, pp. 498-520, doi: 10.1108/SAMPJ-06-2018-0167.

Adams, C. (2002), "Internal organisational factors influencing corporate social and ethical reporting: beyond current theorising", *Accounting, Auditing and Accountability Journal*, Vol. 15 No. 2, pp. 223-250.

Adams, C.A. and McNicholas, P. (2007), "Making a difference: sustainability reporting, accountability and organisational change", *Accounting, Auditing and Accountability Journal*, Vol. 20 No. 3, pp. 382-402, doi: 10.1108/09513570710748553.

Adams, C.A., Alhamood, A., He, X., Tian, J., Wang, L. and Wang, Y. (2021), "The double-materiality concept: application and issues".

Alpaydin, E. (2010), *Introduction to Machine Learning*, 2nd ed., The MIT Press, Cambridge, MA.

Anagnostopoulos, C., Tasoulis, D.K., Adams, N.M., Pavlidis, N.G. and Hand, D.J. (2012), "Online linear and quadratic discriminant analysis with adaptive forgetting for streaming classification", *Statistical Analysis and Data Mining: The Asa Data Science Journal*, Vol. 5 No. 2, pp. 139-166, doi: 10.1002/sam.10151.

Aouadi, A. and Marsat, S. (2018), "Do ESG controversies matter for firm value? Evidence from international data", *Journal of Business Ethics*, Vol. 151 No. 4, pp. 1027-1047, doi: 10.1007/S10551-016-3213-8/TABLES/11.

Atz, U., Van Holt, T., Liu, Z.Z. and Bruno, C.C. (2023), "Does sustainability generate better financial performance? Review, meta-analysis, and propositions", *Journal of Sustainable Finance & Investment*, Vol. 13 No. 1, pp. 802-825, doi: 10.1080/20430795.2022.2106934.

Arora, P. and Dharwadkar, R. (2011), "Corporate governance and corporate social responsibility (CSR): the moderating roles of attainment discrepancy and organization slack", *Corporate Governance: An International Review*, Vol. 19 No. 2, pp. 136-152, doi: 10.1111/J.1467-8683.2010.00843.X.

Babajide Mustapha, I. and Saeed, F. (2016), "Bioactive molecule prediction using extreme gradient boosting", *Molecules (Molecules)*, Vol. 21 No. 8, p. 983, doi: 10.3390/molecules21080983.

Brammer, S. and Millington, A. (2008), "Does it pay to be different? An analysis of the relationship between corporate social and financial performance", *Strategic Management Journal*, Vol. 29 No. 12, pp. 1325-1343, doi: 10.1002/SMJ.714.

Barnea, A. and Rubin, A. (2010), "Corporate social responsibility as a conflict between shareholders", *Journal of Business Ethics*, Vol. 97 No. 1, pp. 71-86, doi: 10.1007/s10551-010-0496-z.

Barnett, M.L. and Salomon, R.M. (2006), "Beyond dichotomy: the curvilinear relationship between social responsibility and financial performance", *Strategic Management Journal*, Vol. 27 No. 11, pp. 1101-1122, doi: 10.1002/SMJ.557.

Beelitz, A., Cho, C.H., Michelon, G. and Patten, D.M. (2021), "Measuring CSR disclosure when assessing stock market effects", *Accounting and the Public Interest*, Vol. 21 No. 1, pp. 1-22, doi: 10.2308/API-2020-017.

Benabou, R. and Tirole, J. (2010), "Individual and corporate social responsibility", *Economica*, Vol. 77 No. 305, pp. 1-19, doi: 10.1111/J.1468-0335.2009.00843.X.

Berg, F., Kölbel, J.F. and Rigobon, R. (2022), "Aggregate confusion: the divergence of ESG ratings", *Review of Finance*, Vol. 26 No. 6, pp. 1315-1344, doi: 10.1093/ROF/RFAC033.

Bowen, H. (1953), *Social Responsibility of the Businessman*, Harper and Row, New York, NY, (Issue 4).

Breiman, L. (2001), "Random forests", *Machine Learning*, Vol. 45 No. 1, pp. 5-32, doi: 10.1023/A:1010933404324.

Busch, T. and Friede, G. (2018), "The robustness of the corporate social and financial performance relation: a second-order meta-analysis", *Corporate Social Responsibility and Environmental Management*, Vol. 25 No. 4, pp. 583-608, doi: 10.1002/csr.1480.

Bzdok, D., Altman, N. and Krzywinski, M. (2018), "Points of significance: statistics versus machine learning", *Nature Methods*, Vol. 15 No. 4, pp. 233-234, doi: 10.1038/nmeth.4642.

Cahan, S.F., Chen, C., Chen, L. and Nguyen, N.H. (2015), "Corporate social responsibility and media coverage", *Journal of Banking and Finance*, Vol. 59, pp. 409-422, doi: 10.1016/j.jbankfin.2015.07.004.

Cai, Y., Jo, H. and Pan, C. (2012), "Doing well while doing bad? CSR in controversial industry sectors", *Journal of Business Ethics*, Vol. 108 No. 4, pp. 467-480, doi: 10.1007/s10551-011-1103-7.

Callan, S.J. and Thomas, J.M. (2009), "Corporate financial performance and corporate social performance: an update and reinvestigation", *Corporate Social Responsibility and Environmental Management*, Vol. 16 No. 2, pp. 61-78, doi: 10.1002/csr.182.

Carroll, A. (1979), "A three-dimensional conceptual model of corporate performance", *The Academy of Management Review*, Vol. 4 No. 4, pp. 497-505.

Carroll, A.B. (1999), "Corporate social responsibility: evolution of a definitional construct", *Business and Society*, Vol. 38 No. 3, pp. 268-295, doi: 10.1177/000765039903800303.

Chatterji, A. and Levine, D. (2006), "Breaking down the wall of codes: evaluating non-financial performance measurement", *California Management Review*, Vol. 48 No. 2, pp. 29-51, doi: 10.2307/41166337.

Chatterji, A.K., Levine, D.I. and Toffel, M.W. (2009), "How well do social ratings actually measure corporate social responsibility?", *Journal of Economics and Management Strategy*, Vol. 18 No. 1, pp. 125-169, doi: 10.1111/J.1530-9134.2009.00210.X.

Chatterji, A.K., Durand, R., Levine, D.I. and Touboul, S. (2016), "Do ratings of firms converge? Implications for managers, investors and strategy researchers", *Strategic Management Journal*, Vol. 37 No. 8, pp. 1597-1614, doi: 10.1002/smj.2407.

Chen, C.M. and Delmas, M. (2011), "Measuring corporate social performance: an efficiency perspective", *Production and Operations Management*, Vol. 20 No. 6, pp. 789-804, doi: 10.1111/j.1937-5956.2010.01202.x.

Chiu, S.C. and Sharfman, M. (2018), "Corporate social irresponsibility and executive succession: an empirical examination", *Journal of Business Ethics*, Vol. 149 No. 3, pp. 707-723, doi: 10.1007/s10551-016-3089-7.

Christensen, H.B., Hail, L. and Leuz, C. (2021), "Mandatory CSR and sustainability reporting: economic analysis and literature review", *Review of Accounting Studies*, Vol. 26 No. 3, pp. 1176-1248, doi: 10.1007/S11142-021-09609-5.

Christensen, D., Serafeim, G. and Sikochi, A. (2022), "Why is corporate virtue in the eye of the beholder?", *The Accounting Review*, Vol. 97 No. 1, pp. 147-175.

Collopy, F., Adya, M. and Armstrong, J.S. (1994), "Principles for examining predictive validity: the case of information systems spending forecasts", *Information Systems Research*, Vol. 5 No. 2, pp. 170-179, doi: 10.1287/isre.5.2.170.

Cornett, M.M., Erhemjamts, O. and Tehranian, H. (2016), "Greed or good deeds: an examination of the relation between corporate social responsibility and the financial performance of U.S. commercial banks around the financial crisis", *Journal of Banking and Finance*, Vol. 70, pp. 137-159, doi: 10.1016/J.JBANKFIN.2016.04.024.

Cooper, S. and Michelon, G. (2022), "Conceptions of materiality in sustainability reporting frameworks: commonalities, differences and possibilities", *Handbook of Accounting and Sustainability*, Edward Elgar Publishing, Cheltenham, UK, pp. 44-66.

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

Dahlsrud, A. (2008), "How corporate social responsibility is defined: an analysis of 37 definitions", *Corporate Social Responsibility and Environmental Management*, Vol. 15 No. 1, pp. 1-13, doi: 10.1002/csr.132.

de Villiers, C. and van Staden, C.J. (2006), "Can less environmental disclosure have a legitimising effect? Evidence from Africa", *Accounting, Organizations and Society*, Vol. 31 No. 8, pp. 763-781, doi: 10.1016/j.aos.2006.03.001.

Deegan, C. (2002), "Introduction: the legitimising effect of social and environmental disclosures", – *Accounting, Auditing and Accountability Journal*, Vol. 15 No. 3, pp. 282-311, doi: 10.1108/09513570210435852.

Delmas, M. and Blass, V.D. (2010), "Measuring corporate environmental performance: the trade-offs of sustainability ratings", *Business Strategy and the Environment*, Vol. 19 No. 4, pp. 245-260, doi: 10.1002/bse.676.

Delmas, M.A. and Burbano, V.C. (2011), "The drivers of greenwashing", *California Management Review*, Vol. 54 No. 1, pp. 64-87, doi: 10.1525/cmr.2011.54.1.64.

DeTienne, K.B., DeTienne, D.H. and Joshi, S.A. (2003), "Neural networks as statistical tools for business researchers", *Organizational Research Methods*, Vol. 6 No. 2, pp. 236-265, doi: 10.1177/1094428103251907.

Ding, K., Lev, B., Peng, X., Sun, T. and Vasarhelyi, M.A. (2020), "Machine learning improves accounting estimates: evidence from insurance payments", *Review of Accounting Studies*, Vol. 25 No. 3, pp. 1098-1134, doi: 10.1007/s11142-020-09546-9.

Drempetic, S., Klein, C. and Zwergel, B. (2020), "The influence of firm size on the ESG score: corporate sustainability ratings under review", *Journal of Business Ethics*, Vol. 167 No. 2, pp. 333-360, doi: 10.1007/s10551-019-04164-1.

Dworkin, R.M. (1985), *Matter of Principle*, Oxford University Press, Oxford, available at: www.books.google.com/books/about/A_Matter_of_Principle.html?hl=sv&id=FUz9VVTIQakC

Earle, J.S., Spicer, A. and Peter, K.S. (2010), "The normalization of deviant organizational practices: wage arrears in Russia, 1991-98", *Academy of Management Journal*, Vol. 53 No. 2, pp. 218-237, doi: 10.5465/amj.2010.49387426.

Elliott, W.B., Jackson, K.E., Peecher, M.E. and White, B.J. (2014), "The unintended effect of corporate social responsibility performance on investors' estimates of fundamental value", *The Accounting Review*, Vol. 89 No. 1, pp. 275-302, doi: 10.2308/ACCR-50577.

Fama, E.F. and French, K.R. (2015), "A five-factor asset pricing model", *Journal of Financial Economics*, Vol. 116 No. 1, pp. 1-22, doi: 10.1016/J.JFINECO.2014.10.010.

Faulkner, R.R. (2011), *Corporate Wrongdoing and the Art of the Accusation*, Anthem Publishers, London, doi: 10.7135/UPO9780857284204.

Fiaschi, D., Giuliani, E. and 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*, Vol. 52 No. 4, pp. 546-563, doi: 10.1016/j.jwb.2016.09.001.

Fiaschi, D., Giuliani, E., Nieri, F. and Salvati, N. (2020), "How bad is your company? Measuring corporate wrongdoing beyond the magic of ESG metrics", *Business Horizons*, Vol. 63 No. 3, pp. 287-299, doi: 10.1016/j.bushor.2019.09.004.

Flammer, C. (2015), "Does corporate social responsibility lead to superior financial performance? A regression discontinuity approach", *Management Science*, Vol. 61 No. 11, pp. 2549-2568, doi: 10.1287/MNSC.2014.2038.

Frederick, W.C. (1994), "From CSR1 to CSR2: the maturing of business-and-society thought", *Business and Society*, Vol. 33 No. 2, pp. 150-164, doi: 10.1177/000765039403300202.

Friedman, M. (1970), "The social responsibility of business is to maximise its profits", *New York Times Magazine*, September, 13.

Friedman, H.L. and Heinle, M.S. (2016), "Lobbying and uniform disclosure regulation", *Journal of Accounting Research*, Vol. 54 No. 3, doi: 10.1111/1475-679X.12118.

Galtung, J. (1959), "Expectations and interaction processes", *Inquiry (Inquiry)*, Vol. 2 Nos 1/4, pp. 213-234, doi: 10.1080/00201745908601296.

Ghritlahre, H.K. and Prasad, R.K. (2018), "Application of ANN technique to predict the performance of solar collector systems – a review", *Renewable and Sustainable Energy Reviews*, Vol. 84, pp. 75-88, doi: 10.1016/j.rser.2018.01.001.

Giuliani, E., Nieri, F. and Fiaschi, D. (2015), "BRIC companies seeking legitimacy through corporate social responsibility", *Transnational Corporations*, Vol. 22 No. 3, pp. 5-42, doi: 10.18356/e13d5a2e-en.

Grappi, S., Romani, S. and Bagozzi, R.P. (2013), "Consumer response to corporate irresponsible behavior: moral emotions and virtues", *Journal of Business Research*, Vol. 66 No. 10, pp. 1814-1821, doi: 10.1016/j.jbusres.2013.02.002.

Grewal, J., Hauotmann, C. and Serafeim, G. (2020), "Material sustainability information and stock price informativeness", *Journal of Business Ethics*, Vol. 171 No. 3, pp. 513-544.

Gulo, C.A.S.J., Rúbio, T.R.P.M., Tabassum, S. and Prado, S.G.D. (2015), "Mining scientific articles powered by machine learning techniques", *OpenAccess Series in Informatics*, Vol. 49, pp. 21-28, doi: 10.4230/OASIcs.ICCSW.2015.21.

Hart, O. and Zingales, L. (2017), "Companies should maximize shareholder welfare not market value", *Journal of Law, Finance, and Accounting*, Vol. 2 No. 2, doi: 10.1561/108.00000022.

Hillman, A.J. and Keim, G.D. (2001), "Shareholder value, stakeholder management, and social issues: what's the bottom line?", *Strategic Management Journal*, Vol. 22 No. 2, pp. 125-139, doi: 10.1002/1097-0266(200101)22:2<125::AID-SMJ150>3.0.CO;2-H.

Hull, C.E. and Rothenberg, S. (2008), "Firm performance: the interactions of corporate social performance with innovation and industry differentiation", *Strategic Management Journal*, Vol. 29 No. 7, pp. 781-789, doi: 10.1002/smj.675.

Kang, Y.C. and Wood, D. (1995), "Before-profit social responsibility: turning the economic paradigm upside down", in Nigh, D. and Collins, D. (Eds), *Proceedings of the International Association for Business and Society*, IABS, Vol. 6, pp. 809-829, doi: 10.5840/iabsproc1995672.

Kelsen, H. and Knight, M. (1967), "Pure theory of law", 356.

Kennelly, J.J. (2000), "Institutional ownership and multinational firms: relationships to social and environmental performance", *Institutional Ownership and Multinational Firms: Relationships to Social and Environmental Performance*, Taylor and Francis, New York, doi: 10.4324/9781315053325.

Kitzmueller, M. and Shimshack, J. (2012), "Economic perspectives on corporate social responsibility", *Journal of Economic Literature*, Vol. 50 No. 1, doi: 10.1257/jel.50.1.51.

Klein, J. and Dawar, N. (2004), "Corporate social responsibility and consumers' attributions and brand evaluations in a product-harm crisis", *International Journal of Research in Marketing*, Vol. 21 No. 3, pp. 203-217, doi: 10.1016/j.ijresmar.2003.12.003.

Kothari, S., Shu, S. and Wysocki, P.D. (2009), "Do managers withhold bad news?", *Journal of Accounting Research*, Vol. 47 No. 1, pp. 241-276, doi: 10.1111/j.1475-679X.2008.00318.x.

Kotsantonis, S. and Serafeim, G. (2019), "Four things no one will tell you about ESG data", *Journal of Applied Corporate Finance*, Vol. 31 No. 2, pp. 50-58, doi: 10.1111/JACF.12346.

Krueger, P., Sautner, Z. and Starks, L. (2020), "Importance of climate risks for institutional investors", *The Review of Financial Studies*, Vol. 33 No. 3, pp. 1067-1111, available at: www.academic.oup.com/rfs/article-abstract/33/3/1067/5735302

Latinovic, M. and Obradovic, T. (2013), "The performance of socially responsible investments", *Entrepreneurial Business and Economics Review*, Vol. 1 No. 2, pp. 29-40, doi: 10.15678/EBER.2013.010203.

Lee, M.D.P. (2008), "A review of the theories of corporate social responsibility: its evolutionary path and the road ahead", *International Journal of Management Reviews*, Vol. 10 No. 1, pp. 53-73, doi: 10.1111/J.1468-2370.2007.00226.X.

Lee, M.T. and Suh, I. (2022), "Understanding the effects of environment, social, and governance conduct on financial performance: arguments for a process and integrated modelling approach", *Sustainable Technology and Entrepreneurship*, Vol. 1 No. 1, p. 100004, doi: 10.1016/J.STAE.2022.100004.

Luhmann, N. (1995), "Funktionen und Folgen formaler Organisation: mit einem Epilog 1994", *Schriftenreihe der Hochschule Speyer (Issue Bd 20)*, Duncker and Humblot, Berlin.

Luo, X. and Bhattacharya, C.B. (2006), "Corporate social responsibility, customer satisfaction, and market value", *Journal of Marketing*, Vol. 70 No. 4, pp. 1-18, doi: 10.1509/jmkg.70.4.001.

Lundberg, S.M. and Lee, S.I. (2017), "A unified approach to interpreting model predictions", *Advances in Neural Information Processing Systems*, 2017-December, pp. 4766-4775.

Mackey, A., Mackey, T.B. and Barney, J.B. (2007), "Corporate social responsibility and firm performance: investor preferences and corporate strategies", *Academy of Management Review*, Vol. 32 No. 3, pp. 817-835, doi: 10.5465/AMR.2007.25275676.

Manchiraju, H. and Rajgopal, S. (2017), "Does corporate social responsibility (CSR) create shareholder value? Evidence from the Indian companies act 2013", *Journal of Accounting Research*, Vol. 55 No. 5, doi: 10.1111/1475-679X.12174.

Margolis, J.D., Elfenbein, H.A. and Walsh, J.P. (2009), "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*, doi: 10.2139/SSRN.1866371.

Masulis, R.W. and Reza, S.W. (2015), "Agency problems of corporate philanthropy", *Review of Financial Studies*, Vol. 28 No. 2, pp. 592-636, doi: 10.1093/RFS/HHU082.

Mitchell, R.K., Agle, B.R. and 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*, Vol. 22 No. 4, p. 853, doi: 10.2307/259247.

Muller, A. and Kräussl, R. (2011), "Doing good deeds in times of need: a strategic perspective on corporate disaster donations", *Strategic Management Journal*, Vol. 32 No. 9, pp. 911-929, doi: 10.1002/smj.917.

Nieri, F. and Giuliani, E. (2018), "International business and corporate wrongdoing: a review and research agenda", in Castellan, D., Narula, R., Nguyen, Q., Surdu, I. and Walker, J. (Eds), *Contemporary Issues in International Business*, Springer International Publishing, London, pp. 35-53, doi: 10.1007/978-3-319-70220-9_3.

Nirino, N., Santoro, G., Miglietta, N. and Quaglia, R. (2021), "Corporate controversies and company's financial performance: exploring the moderating role of ESG practices", *Technological Forecasting and Social Change*, Vol. 162, p. 120341, doi: 10.1016/J.TECHFORE.2020.120341.

Nitsche, C. and Schröder, M. (2018), "Are SRI funds conventional funds in disguise or do they live up to their name?", In Boubaker, S., Cumming, D. and Nguyen, D. (Eds), *Research Handbook of Investing in the Triple Bottom Line: Finance, Society and the Environment*, Edward Elgar Publishing, Cheltenham, UK, pp. 414-446, doi: 10.4337/9781788110006.00028.

Nofsinger, J.R., Sulaeman, J. and Varma, A. (2019), "Institutional investors and corporate social responsibility", *Journal of Corporate Finance*, Vol. 58, pp. 700-725, doi: 10.1016/j.jcorpfin.2019.07.012.

Oikonomou, I., Platanakis, E. and Sutcliffe, C. (2018), "Socially responsible investment portfolios: does the optimization process matter?", *The British Accounting Review*, Vol. 50 No. 4, pp. 379-401, doi: 10.1016/j.bar.2017.10.003.

Orlitzky, M., Schmidt, F.L. and Rynes, S.L. (2003), "Corporate social and financial performance: a meta-analysis", *Organization Studies*, Vol. 24 No. 3, pp. 403-441, doi: 10.1177/0170840603024003910.

Orlitzky, M. (2013), "Corporate social responsibility, noise, and stock market volatility", *Academy of Management Perspectives*, Vol. 27 No. 3, pp. 238-254, doi: 10.5465/AMP.2012.0097.

Ou, P. and Wang, H. (2009), "Prediction of stock market index movement by ten data mining techniques", *Modern Applied Science*, Vol. 3 No. 12, pp. 28-42, doi: 10.5539/mas.v3n12p28.

Pizzi, S., Principale, S. and de Nuccio, E. (2022), "Material sustainability information and reporting standards: exploring the differences between GRI and SASB", *Meditari Accountancy Research*, doi: 10.1108/MEDAR-11-2021-1486.

Ring, M. and Eskofier, B.M. (2016), "An approximation of the Gaussian RBF kernel for efficient classification with SVMs", *Pattern Recognition Letters*, Vol. 84, pp. 107-113, doi: 10.1016/J.PATREC.2016.08.013.

Saeidi, S.P., Sofian, S., Saeidi, P., Saeidi, S.P. and Saaeidi, S.A. (2015), "How does corporate social responsibility contribute to firm financial performance? The mediating role of competitive advantage, reputation, and customer satisfaction", *Journal of Business Research*, Vol. 68 No. 2, pp. 341-350, doi: 10.1016/J.JBUSRES.2014.06.024.

Saadaoui, K. and Soobaroyen, T. (2018), "An analysis of the methodologies adopted by CSR rating agencies", *Sustainability Accounting, Management and Policy Journal*, Vol. 9 No. 1, pp. 43-62, doi: 10.1108/SAMPJ-06-2016-0031.

Safa, M. and Samarasinghe, S. (2011), "Determination and modelling of energy consumption in wheat production using neural networks: a case study in Canterbury province", *Energy*, Vol. 36 No. 8, pp. 5140-5147, doi: 10.1016/j.energy.2011.06.016.

Servaes, H. and Tamayo, A. (2013), "The impact of corporate social responsibility on firm value: the role of customer awareness", *Management Science*, Vol. 59 No. 5, pp. 1045-1061, doi: 10.1287/MNSC.1120.1630.

Shmueli, G. (2010), "To explain or to predict?", *Statistical Science*, Vol. 25 No. 3, pp. 289-310, doi: 10.1214/10-STS330.

Sigrist, F. and Hirnschall, C. (2019), "Grabit: gradient tree-boosted Tobit models for default prediction", *Journal of Banking and Finance*, Vol. 102, pp. 177-192, doi: 10.1016/j.jbankfin.2019.03.004.

Simpson, W.G. and Kohers, T. (2002), "The link between corporate social and financial performance: evidence from the banking industry", *Journal of Business Ethics*, Vol. 35 No. 2, pp. 97-109, doi: 10.1023/A:1013082525900/METRICS.

Sözen, A. (2009), "Future projection of the energy dependency of Turkey using artificial neural network", *Energy Policy*, Vol. 37 No. 11, pp. 4827-4833, doi: 10.1016/j.enpol.2009.06.040.

Summers, R.S. (1963), "Professor H. L. A. Hart's 'concept of law'", *Duke Law Journal*, Vol. 1963 No. 4, p. 629, doi: 10.2307/1371248.

Surroca, J., Tribó, J.A. and Waddock, S. (2010), "Corporate responsibility and financial performance: the role of intangible resources", *Strategic Management Journal*, Vol. 31 No. 5, pp. 463-490, doi: 10.1002/SMJ.820.

Surroca, J., Tribó, J.A. and Zahra, S.A. (2013), "Stakeholder pressure on MNEs and the transfer of socially irresponsible practices to subsidiaries", *Academy of Management Journal*, Vol. 56 No. 2, pp. 549-572, doi: 10.5465/amj.2010.0962.

Tang, Z., Hull, C.E. and Rothenberg, S. (2012), "How corporate social responsibility engagement strategy moderates the CSR-financial performance relationship", *Journal of Management Studies*, Vol. 49 No. 7, pp. 1274-1303, doi: 10.1111/J.1467-6486.2012.01068.X.

Utz, S. (2019), "Corporate scandals and the reliability of ESG assessments: evidence from an international sample", *Review of Managerial Science*, Vol. 13 No. 2, pp. 483-511, doi: 10.1007/s11846-017-0256-x.

Vapnik, V.N. (1995), *The Nature of Statistical Learning Theory*, Springer, Berlin, doi: 10.1007/978-1-4757-2440-0.

Waddock, S.A. and Graves, S.B. (1997), "The corporate social performance-financial performance link", *Strategic Management Journal*, Vol. 18 No. 4, pp. 303-319, doi: 10.1002/(SICI)1097-0266(199704)18:4<303::AID-SMJ869>3.0.CO;2-G.

Wartick, S.L. and Cochran, P.L. (1985), "The evolution of the corporate social performance model", *Academy of Management Review*, Vol. 10 No. 4, pp. 758-769, doi: 10.5465/AMR.1985.4279099.

Wood, D.J. (1991), "Corporate social performance revisited", *The Academy of Management Review*, Vol. 16 No. 4, pp. 691-718, doi: 10.2307/258977.

Wood, D.J. (2010), "Measuring corporate social performance: a review", *International Journal of Management Reviews*, Vol. 12 No. 1, pp. 50-84, doi: 10.1111/J.1468-2370.2009.00274.X.

Wong, C. and Petroy, E. (2020), *Rate the Raters 2020: Investor Survey and Interview Results*, Sustainability/ERM Group.

Xu, X., Zhou, C. and Wang, Z. (2009), "Credit scoring algorithm based on link analysis ranking with support vector machine", *Expert Systems with Applications*, Vol. 36 No. 2, pp. 2625-2632, doi: 10.1016/j.eswa.2008.01.024.

Yeh, C.C., Chi, D.J. and Lin, Y.R. (2014), "Going-concern prediction using hybrid random forests and rough set approach", *Information Sciences*, Vol. 254, pp. 98-110, doi: 10.1016/j.ins.2013.07.011.

Yuan, L., Yong, F., Wei, Z. and Shan, K. (2017), "Using quadratic discriminant analysis to predict protein secondary structure based on chemical shifts", *Current Bioinformatics*, Vol. 12 No. 1, pp. 52-56, doi: 10.2174/1574893611666160628074537.

Zhang, X., Zhao, X. and He, Y. (2022), "Does it pay to be responsible? The performance of ESG investing in China", *Emerging Markets Finance and Trade*, Vol. 58 No. 11, pp. 3048-3075, doi: 10.1080/1540496X.2022.2026768.

Appendix 1

Table A1. Standards for standardized logistic regression coefficients

SP indicators	Scale
Health and safety policy	1/0
Policy employee health and safety	1/0
Policy supply chain health and safety	1/0
Training and development policy	1/0
Policy skills training	1/0
Policy career development	1/0
Policy diversity and opportunity	1/0
Targets diversity and opportunity	1/0
Employees health and safety team	1/0
Health and safety training	1/0
Supply chain health and safety training	1/0
Supply chain health and safety improvements	1/0
Employees health and safety OHSAS 18001	1/0
Employee satisfaction	Numerical
Salary gap	Numerical
Salaries and wages from CSR reporting	Numerical
Net employment creation	Numerical
Number of employees from CSR reporting	Numerical
Trade union representation	1/0
Turnover of employees	Numerical
Announced layoffs to total employees	Numerical
Announced layoffs	Numerical
Management departures	Numerical
Strikes	1/0
Women employees	Percent
New women employees	Percent
Women managers	Percent
HRC corporate equality index	Numerical
Flexible working hours	1/0
Day care services	1/0
Employees with disabilities	Percent
Employee health and safety training hours	Numerical
Injuries to million hours	Numerical
Total injury rate total	Numerical
Total injury rate contractors	Numerical
Total injury rate employees	Numerical
Accidents total	Numerical
Contractor accidents	Numerical
Employee accidents	Numerical
Occupational diseases	Numerical
Employee fatalities	Numerical
Contractor fatalities	Numerical
Lost days to total days	Numerical
Lost time injury rate total	Numerical
Lost time injury rate contractors	Numerical
Lost time injury rate employees	Numerical
Lost working days	Numerical
Employee lost working days	Numerical
Contractor lost working days	Numerical
HIV-AIDS program	1/0

(*continued*)

SP indicators	Scale
Average training hours	Numerical
Training hours total	Numerical
Training costs total	Numerical
Training costs per employee	Numerical
Internal promotion	1/0
Management training	1/0
Supplier ESG training	1/0
Employee resource groups	1/0
BBBEE level	Numerical
Gender pay gap percentage	Percent
Voluntary turnover of employees	Numerical
Involuntary turnover of employees	Numerical
HSMS certified percent	Percent
Human rights policy	1/0
Policy freedom of association	1/0
Policy child labor	1/0
Policy forced labor	1/0
Policy human rights	1/0
Fundamental human rights ILO UN	1/0
Human rights contractor	1/0
Ethical trading initiative (ETI)	1/0
Human rights breaches contractor	1/0
Policy fair competition	1/0
Policy bribery and corruption	1/0
Policy Business Ethics	1/0
Policy community involvement	1/0
Improvement tools business ethics	1/0
Whistle blower protection	1/0
OECD guidelines for multinational enterprises	1/0
Extractive industries transparency initiative	1/0
Total donations to revenues	Numerical
Donations total	Numerical
Community lending and investments	Numerical
Political contributions	Numerical
Lobbying contribution amount	Numerical
Employee engagement voluntary work	1/0
Corporate responsibility awards	1/0
Product sales at discount to emerging markets	1/0
Diseases of the developing world	1/0
Crisis management systems	1/0
Policy customer health and safety	1/0
Policy data privacy	1/0
Policy responsible marketing	1/0
Policy fair trade	1/0
Product responsibility monitoring	1/0
Quality management systems	1/0
ISO 9000	1/0
Six sigma and quality management systems	1/0
QMS certified percent	1/0
Customer satisfaction	1/0
Product access low price	1/0
Healthy food or products	1/0

(*continued*)

Table A1.

SP indicators	Scale
Embryonic stem cell research	1/0
Retailing responsibility	1/0
Alcohol	1/0
Alcohol revenues	Numerical
Alcohol 5% revenues	1/0
Gambling	1/0
Gambling revenues	Numerical
Gambling 5% revenues	1/0
Tobacco	1/0
Tobacco revenues	Numerical
Tobacco 5% revenues	1/0
Armaments	1/0
Armaments revenues	Numerical
Armaments 5% revenues	1/0
Nuclear 5% revenues	1/0
Pornography	1/0
Contraceptive	1/0
Obesity risk	1/0
Cluster bombs	1/0
Anti-personnel landmines	1/0
Abortifacients	1/0
Drug delay	1/0
FDA warning letters	1/0
Product delays	1/0
Not approved drug	1/0
Product recall	1/0
Recent FDA warning letters	1/0
Firearms	1/0

Table A1. **Source:** Created by the authors

Appendix 2

Description	Full list
Gradient boosting	"ccp_alpha: 0.0, criterion: friedman_mse, init: None, learning_rate: 0.1, loss: deviance, max_depth: 3, max_features: None, max_leaf_nodes: None, min_impurity_decrease: 0.0, min_samples_leaf: 1, min_samples_split: 2, min_weight_fraction_leaf: 0.0, *n*_estimators: 100, *n*_iter_no_change: None, random_state: 123, subsample: 1.0, tol: 0.0001, validation_fraction: 0.1, verbose: 0, warm_start: False"
Linear support vector machine	"C: 0.025, break_ties: False, cache_size: 200, class_weight: balanced, coef0: 0.0, decision_function_shape: ovr, degree: 3, gamma: scale, kernel: linear, max_iter: −1, probability: True, random_state: 123, shrinking: True, tol: 0.001, verbose: False"
Logistic regression	"C: 1.0, class_weight: balanced, dual: False, fit_intercept: True, intercept_scaling: 1, l1_ratio: None, max_iter: 100, multi_class: auto, *n*_jobs: None, penalty: l2, random_state: 123, solver: lbfgs, tol: 0.0001, verbose: 0, warm_start: False"
Naïve Bayes	"priors: None, var_smoothing: 1e-09"
Neural network	"activation: relu, alpha: 1, batch_size: auto, beta_1: 0.9, beta_2: 0.999, early_stopping: False, epsilon: 1e-08, hidden_layer_sizes: (100,), learning_rate: constant, learning_rate_init: 0.001, max_fun: 15000, max_iter: 1000, momentum: 0.9, *n*_iter_no_change: 10, nesterovs_momentum: True, power_*t*: 0.5, random_state: 123, shuffle: True, solver: adam, tol: 0.0001, validation_fraction: 0.1, verbose: False, warm_start: False"
Radial basis function support vector machine	"C: 0.025, break_ties: False, cache_size: 200, class_weight: balanced, coef0: 0.0, decision_function_shape: ovr, degree: 3, gamma: scale, kernel: rbf, max_iter: −1, probability: True, random_state: 123, shrinking: True, tol: 0.001, verbose: False"
Random forest	"Bootstrap: True, ccp_alpha: 0.0, class_weight: balanced_subsample, criterion: gini, max_depth: None, max_features: auto, max_leaf_nodes: None, max_samples: None, min_impurity_decrease: 0.0, min_samples_leaf: 1, min_samples_split: 2, min_weight_fraction_leaf: 0.0, *n*_estimators: 101, *n*_jobs: None, oob_score: False, random_state: 123, verbose: 0, warm_start: False"
Quadratic discriminant analysis	"copy: True, norm: l2"

Source: Created by the authors

Table A2. Hyperparameters

About the authors

Jan Svanberg is an Associate Professor (PhD) of Business Administration at the University of Gävle and the Centre for Research on Economic Relations (CER). His research interests are beavioral issues, primary in accounting and auditing. Jan Svanberg is the corresponding author and can be contacted at: jan.svanberg@hig.se

Tohid Ardeshiri has a Doctor of Philosophy degree in Statistical Signal Processing from Linköping University. He conducts applied machine learning research on financial applications at RISE, Research Institutes of Sweden.

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

Peter Öhman is Professor of Accounting at the Centre for Research on Economic Relations (CER) at Mid Sweden University. His research interests are primarily in accounting and auditing.

Presha E. Neidermeyer is Professor of Accounting at West Virginia University, USA. Her research work focuses on gender issues and international auditor behavior and has been recognized by Emerald Insight publishers.

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

Frank Maisano has a Doctor of Philosophy degree in Accounting and is Lecturer at The Royal Melbourne Institute of Technology. He conducts research on auditor independence and analytics.

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