Credit and Capital Markets, Volume 53, Issue 4, pp. 513–554
Scientific Papers

Corporate Social Irresponsibility and Credit Risk Prediction: A Machine Learning Approach

Daniel V. Fauser* and Andreas Gruener**

Abstract

This paper examines the prediction accuracy of various machine learning (ML) algorithms for firm credit risk. It marks the first attempt to leverage data on corporate social irresponsibility (CSI) to better predict credit risk in an ML context. Even though the literature on default and credit risk is vast, the potential explanatory power of CSI for firm credit risk prediction remains unexplored. Previous research has shown that CSI may jeopardize firm survival and thus potentially comes into play in predicting credit risk. We find that prediction accuracy varies considerably between algorithms, with advanced machine learning algorithms (e. g. random forests) outperforming traditional ones (e. g. linear regression). Random forest regression achieves an out-of-sample prediction accuracy of 89.75 % for adjusted R^2 due to the ability of capturing non-linearity and complex interaction effects in the data. We further show that including information on CSI in firm credit risk prediction does not consistently increase prediction accuracy. One possible interpretation of this result is that CSI does not (yet) seem to be systematically reflected in credit ratings, despite prior literature indicating that CSI increases credit risk. Our study contributes to improving firm credit risk predictions using a machine learning design and to exploring how CSI is reflected in credit risk ratings.

Keywords: Corporate Social Irresponsibility, Credit Risk, Machine Learning, Random Forests, Regression, Classification, ESG

JEL Classification: G10, G15, G17, G32

I. Introduction

This study investigates the prediction accuracy of machine learning (ML) algorithms for firm credit risk using data on corporate social irresponsibility

* Dr. des. Daniel V. Fauser, School of Finance, University of St. Gallen, Guisanstrasse 1a, 9010 St. Gallen, E-Mail: daniel.fauser@unisg.ch (corresponding author).

** Prof. Dr. Andreas Gruener, School of Finance, University of St. Gallen, Switzerland.

The authors are grateful to Tami Dinh and Sebastian Utz for their very valuable feedback. Special thanks go to Jeroen Derwall for his advice and continuous effort to strengthen this paper. We also thank an anonymous reviewer for providing feedback that considerably improved this paper.

(CSI). We aim to leverage the power of both advanced ML algorithms and data on irresponsible firm behavior with regard to ESG aspects (i. e. CSI) to improve the accuracy of credit risk prediction. Our study is motivated by research that points to the increasingly important role of environmental, social, and governance (ESG) aspects for firm credit risk. ESG performance has been linked to lower firm and credit risk (*Stellner* et al. 2015; *Jo/Na* 2012; *Sassen* et al. 2016) and to higher credit ratings (*Jiraporn* et al. 2014; *Kiesel/Lücke* 2019; *Dorfleitner* et al. 2020; *Oikonomou* et al. 2014).

Common to these previous studies is that they have predominantly used ESG performance data to examine firms' corporate social performance (CSP). One exception, however, is *Kölbel* et al.'s (2017) study on the exposure to negative ESG performance, based on the idea of corporate social irresponsibility (CSI) (*Strike* et al. 2006). They showed that CSI leads to higher credit risk as priced by the market through increased spreads of credit default swaps (CDS). However, it is still unclear whether CSI is also systematically reflected in credit ratings and thus informs credit risk predictions. Therefore, our study uses a comprehensive machine learning regression design to test how consistently CSI has been considered in the credit risk analyses of major credit rating agencies. It is important to note that our study differs from previous research on ESG performance and credit risk (e. g. *Stellner* et al. 2015; *Dorfleitner* et al. 2020; *Sassen* et al. 2016) as we focus on negative ESG performance as opposed to positive ESG performance.

Our study also has both a practical, "supply-driven" motivation and a practical, "demand-driven" motivation. The suppliers of credit ratings, so-called credit rating agencies, have recognized that ESG aspects are no longer solely a matter of philanthropy but may deeply impact firm performance and risk. The documentation by major credit rating agencies (such as S&P Global) indicates that ESG aspects are to some extent reflected in credit ratings. For example, S&P Global claims that it "has long considered Environmental, Social, and Governance (ESG) factors in its credit ratings […], [capturing] ESG factors in many areas of [its] methodology." According to S&P Global, ESG aspects are considered in credit risk analysis along three dimensions: business risk (e. g. competitive position), financial risk (e. g. cash flow/leverage assessment), and management and governance.[1] On November 21, 2019, S&P Global issued a press release on its most recent step toward ESG integration by acquiring RobecoSAM's ESG rating business.

As far as the "demand" side is concerned, the Principles Responsible Investment (PRI) Association conducted an investor survey in 2017 and 2018 among

[1] For S&P Global's claims about ESG integration in credit risk analysis, please see: https://www.spglobal.com/ratings/en/products-benefits/products/sustainable-finance and https://www.spglobal.com/ratings/en/products-benefits/products/esg-in-credit-ratings#.

credit analysts, portfolio managers, and ESG analysts about the integration of ESG aspects into credit risk analysis and ratings (PRI 2018). The survey found that 62.9 % of participants agreed that valuation models are adjusted as a result of ESG considerations in their organization. According to participants, the governance dimension of ESG most frequently impacts credit risk analysis, followed by the environmental and social dimensions. The survey identified the three most common reasons for including ESG aspects in credit risk analyses as "Risk management," "Client demand," and "Fiduciary duty." The results of this survey underpin the constantly increasing demand to integrate ESG aspects into credit risk analysis.

Regarding methodology, our decision to use ML algorithms was motivated by research in finance that has found evidence for the superior predictive power of ML algorithms compared to more traditional algorithms (e. g. *Barboza* et al. 2017; *Gu* et al. 2020). ML has proven highly successful in addressing many problems ranging from early-day applications (e. g. spam detection) to more recent fields of application (e. g. financial forecasting and portfolio building). After long playing merely a minor role in finance, the potential of ML in financial prediction and modeling has started to gain momentum. Using ML in bankruptcy prediction (*Barboza* et al. 2017), empirical asset pricing and portfolio construction (*Gu* et al. 2020), stock selection (*Rasekhschaffe/Jones* 2019), and sovereign credit ratings prediction (*Bennell* et al. 2006) are just a few examples. ML techniques are particularly powerful in solving prediction and classification tasks due to their ability to closely capture non-linearity, complex relationships, and interactions in the data. Common issues of traditional linear algorithms such as multi-collinearity or low signal-to-noise ratios are not a serious concern for ML techniques, as we show below. Accordingly, our study tests a comprehensive set of ML algorithms for predicting credit risk as measured by credit ratings.

We use an extensive international sample of 43,476 firm-quarter observations with a time-series of over 12 years (March 2007 to September 2019). To address the potential impact of CSI on credit risk, we used data from RepRisk, which specializes in ESG and business conduct risk research. Data on our measure for credit risk, long-term issuer credit ratings, were primarily obtained from S&P Global. We also used data from the two other major rating agencies, Moody's and Fitch, in an additional analysis. Our study uses algorithms ranging from "plain vanilla" linear regression to more sophisticated machine learning techniques such as Multi-layer Perceptron regression. In line with previous research (e. g. *Barboza* et al. 2017), we used sklearn's[2] package default settings for all algorithms to ensure a high degree of replicability and applicability. We assessed

[2] sklearn is the short version for scikit learn. The application programming interface (API) reference can be found on https://scikit-learn.org/stable/index.html.

prediction accuracy with the three most common measures in machine learning: R^2, adjusted R^2, and root mean squared error (RMSE). Using default settings also aims to alleviate the frequently expressed criticism that ML techniques are a "blackbox". Default hyper-parameters enable precisely replicating our research. To ensure transparency, all algorithms, hyper-parameters, and settings can be found in Table A.7 and are also available in the sklearn API reference.

Our study has two main findings. First, using advanced ML algorithms considerably increases prediction accuracy for credit risk by accounting for non-linearity and complex interaction effects in the data. Random Forest regression offers the highest out-of-sample prediction accuracy of 89.75 % for adjusted R^2, indicating that this algorithm reliably predicts credit risk two quarters into the future. This result is in line with previous research findings on the similar superiority of Random Forest algorithms in default prediction (*Barboza* et al. 2017) and time-series forecasting (*Gu* et al. 2020). Second, CSI does not (yet) seem to be consistently and systematically reflected in credit risk as measured by credit rating data. Incorporating information about CSI into our models does not unambiguously increase prediction accuracy for credit risk. This finding at least holds for both the data and the comprehensive set of ML algorithms used in this study. One possible interpretation is that CSI is not (yet) incorporated into the credit risk analyses of rating agencies to such a degree that this would result in considerably higher prediction accuracy, as is the case for profitability or liquidity, for example. Another interpretation is that rating agencies use a significantly different operationalization of ESG aspects, which is not available to the public.

Our study makes several contributions to the literature and has several implications for market participants. First, to our best knowledge, our study is the first of its kind to synthesize credit risk prediction with state-of-the-art machine learning algorithms using credit rating data. In a credit risk context, our results support the findings of previous research, among others, that advanced ML algorithms are superior to traditional algorithms, particularly in financial prediction tasks (e. g. *Barboza* et al. 2017; *Khandani* et al. 2010; *Gu* et al. 2020). Our finding of high prediction accuracy for advanced ML algorithms is based on high out-of-sample adjusted R^2 and low RMSE, which are robust compared to alternative specifications of our ML models. High prediction accuracy allows market participants to more reliably assess future credit risk. Our trained models are ready to use to predict the credit risk of any firm for which relevant data are available to the predictors used in our study. Our result points to the growing importance of machine learning models in the paradigmatic fintech industry. However, our predictive method fundamentally differs from methods of statistical inference. In particular, machine learning models do not tell us anything about the underlying mechanisms (economic and theoretical) between the pre-

dictor and the target variables (e. g. the reasons for a certain relationship between predictors and targets).[3]

Second, to our best knowledge, our study is the first of its kind to investigate whether including CSI into ML models increases the prediction accuracy of credit risk. In doing so, we also test the extent to which CSI is reflected in the credit risk analyses of rating agencies. Even though previous research has found a positive relationship between positive ESG performance and credit risk (e. g. *Dorfleitner* et al. 2020; *Stellner* et al. 2015; *Sassen* et al. 2016), we explicitly explore negative ESG performance in a novel machine learning design. Previous research has shown that market participants account for CSI in evaluating credit risk, as evidenced by increased spreads of CDS (*Kölbel* et al. 2017). However, using credit ratings as a proxy for credit risk, our results indicate that higher CSI is not (yet) reflected to a larger extent in the credit risk analyses of major rating agencies. Therefore, market participants should be aware that common credit ratings might not entirely reflect the environmental and societal business conduct risks increasingly confronting firms. Examples of increased business conduct risk include weaknesses in corporate governance, low product safety, or compromised employee well-being. All of these factors may entail substantial future expenses (e. g. compensation payments) and reputational damages. The remaining five sections of this paper are structured as follows. Section II discusses related ML applications and empirical literature on ESG aspects and credit risk. Section III describes the research design of this study. Section IV provides some descriptive statistics for our sample. Section V displays the results while Sections VI and VII consider the implications of our findings and offer concluding remarks.

II. Literature Review

This section considers related ML applications and their potentially superior explanatory power in credit risk prediction. It also discusses the empirical literature on the relationship between ESG aspects and credit risk.

1. Machine Learning Applications

The application spectrum of ML algorithms is vast and growing. ML applications range from cancer prognosis and prediction (*Kourou* et al. 2015), genetics and genomics (*Zou* et al. 2019), and healthcare (*Jiang* et al. 2017) to automated

[3] In this paper, we use the terms "predictor" and "target," which are more common in a machine learning context, yet roughly similar to the terms "independent variable" and "dependent variable," respectively.

text classification (*Sebastiani* 2002). ML algorithms have also been shown to better solve financial prediction and classification problems than traditional financial models. Therefore, ML algorithms potentially add value to many finance-related fields of application. For instance, machine learning can improve forecasting consumer credit risk and thus cuts costs from credit losses (*Khandani* et al. 2010). Machine learning may also help to solve financial prediction problems such as security pricing, portfolio construction, financial time series prediction, or risk management (*Gu* et al. 2020; *Rasekhschaffe/Jones* 2019). Additional ML applications include sovereign credit rating prediction (*Bennell* et al. 2006), forecasting interest rates, corporate bond ratings, loan approvals, and identifying suspicious transactions (*Bose/Mahapatra* 2001).

Most noticeably, predicting firm default has gained quite some momentum over the last 25 years with ever-evolving ML algorithms and exponentially increasing computing power and data availability. Early studies on corporate default using "intelligent" techniques such as neural networks include *Altman* et al. (1994) and *Boritz/Kennedy* (1995). Most of these studies offered partially inconclusive results, possibly due to the comparably low computational power, data availability, or pure immaturity of the subject and algorithm implementations.

In a more recent study, *Barboza* et al. (2017) compared a wide range of more advanced ML algorithms to traditional ones of default prediction, such as logistic regression. Default prediction is a typical classification problem (i. e. 1 for default, 0 otherwise). Using US data, they showed that the majority of more advanced ML algorithms are significantly superior in default prediction accuracy than traditional techniques. The authors were also able to increase the total estimated prediction accuracy by about 11 percentage points using advanced ML algorithms (e. g. random forests).[4] This increased prediction accuracy is very likely due to ML algorithms' special treatment of non-linearity, complex interaction effects, and lower sensitivity to multi-collinearity. Building on these findings, we expect that ML techniques will also add considerable explanatory power to help predict credit risk in a regression context.

2. Corporate Social Irresponsibility and Credit Risk

Similar to default risk, predicting credit risk is important for both researchers and practitioners for numerous reasons. Credit risk is the probability of a debtor not repaying the principal and interest on the debt in part or in full (*Vassalou/Xing* 2004). Being unable or unwilling to repay all debts is equivalent to a de-

[4] Some of the work in predicting default risk even considers advanced hybrid ML algorithms as a superior method (*Yeh* et al. 2014). However, we will not discuss these studies due to their specificity and restricted applicability.

fault. Thus, credit risk is also an ex-ante evaluation of firm default risk. The first comprehensive theory and model of the pricing of corporate debt (i. e. credit risk) was developed by *Merton* (1973, 1974), whose work was later advanced by *Leland* (1994). These models not only enabled pricing any corporate liability but were also used to predict bankruptcy (e. g. *Bharath/Shumway* 2008).

Merton's (1974) and *Leland's* (1994) models have also contributed to continually developing the credit risk analyses of credit rating agencies. For market participants, these rating agencies became an interesting source of credit risk assessments as they provided information about the downside risk of firms no longer being able to repay their debts in part or in full. Besides their importance for market participants, credit ratings are also one of the most important drivers of capital structure choice by executives (*Graham/Harvey* 2001). Changes in credit ratings may result in very concrete costs such as changes in coupon rates, losing a contract, or triggering the repurchase of bonds (*Kisgen* 2006). Similar to previous research (e. g. *Sun/Cui*, 2014; *Hsu/Chen* 2015; *Kealhofer* 2003), our study uses these credit ratings as a proxy for credit risk. Credit risk is economically relevant in various dimensions. Firms with higher distress and credit risk offer significantly lower financial returns (*Dichev* 1998; *Campbell* et al. 2008). Moreover, credit risk has been linked to market illiquidity and increased yield spreads (*Ericsson/Renault* 2006). Ultimately, higher credit risk hurts firm profitability due to higher refinancing costs. Therefore, credit risk is key not only to the decision-making processes of lenders but also to those of investors.

In addition to the importance of credit risk, the impact of CSP on credit risk has attracted quite some attention. CSP is multidimensional and thus covers firm behavior in the environmental (e. g. pollution control), social (e. g. diversity), and corporate governance (e. g. stakeholder strategy) dimensions (*Waddock/Graves* 1997). CSP has been shown to reduce the risk of falling into default (e. g. *Sun/Cui* 2014) and firm risk (e. g. *Jo/Na* 2012; *Lee/Faff* 2009). The aforementioned research suggests that the risk-reducing effect of CSP is also reflected in lower credit risk. In fact, rating agencies tend to reward socially and environmentally responsible firms with higher credit ratings (*Hsu/Chen* 2015; *Jiraporn* et al. 2014; *Kiesel/Lücke* 2019). In particular, high performance on ESG aspects concerning primary stakeholders (i. e. employees, customers, etc.) plays a major role in higher credit ratings (*Ashbaugh-Skaife* et al. 2006; *Attig* et al. 2013). Anecdotal evidence exists for cases in which credit ratings were alternated due to changes in the CSP of firms such as Wells Fargo or Toshiba in the period 2016 – 2018 (*Henisz/McGlinch* 2019).

Another stream of research has investigated the spreads of bonds and bank loans (henceforth, credit spreads) as a measure of credit risk. Several studies have shown that firms with high (low) CSP are rewarded (penalized) by the corporate bond market with lower spreads (*Oikonomou* et al. 2014; *Goss/Roberts*

2011; *Chava* 2014; *Stellner* et al. 2015; *Drago* et al. 2019). On the one hand, less responsible firms pay between 7 and 18 basis points higher bond risk premia (*Goss/Roberts* 2011). Moreover, firms that have environmental concerns are charged significantly higher interest rates on bank loans (*Chava* 2014). On the other hand, performing well on ESG aspects such as "support for local communities, higher levels of marketed product safety and quality characteristics, and avoidance of controversies regarding the firm's workforce" positively impacts bond risk premia (*Oikonomou* et al. 2014). Moreover, high CSP is linked to reduced zero-volatility spreads (*Stellner* et al. 2015). The announcement of ESG performance ratings has also been shown to decrease CDS spreads, pointing to the credit risk-decreasing role of ESG performance (*Drago* et al. 2019).

However, the evidence on CSP and credit risk is rather mixed. The same study that found reduced zero-volatility spreads due to high CSP did not find a statistically significant effect of high CSP on credit ratings (*Stellner* et al. 2015). In terms of credit spreads, another study found that more responsible firms have to pay higher bond risk premia (*Menz* 2010). However, due to the lack of statistical significance, *Menz* (2010) concluded that environmental and social factors have not yet been considered in corporate bond pricing. Moreover, both *Stellner* et al. (2015) and *Menz* (2010) focused on the European market, which might explain their findings.

While the relationship between positive CSP (i. e. ESG performance) and credit risk is well explored, a more recent study has shifted the focus from CSP to CSI (i. e. negative ESG performance). CSI can be defined as a "set of corporate actions that negatively affects an identifiable social stakeholders legitimate claims" (*Strike* et al. 2006, p. 852). It comprises behavior that is explicitly irresponsible, while the absence of this behavior is not necessarily responsible (*Strike* et al. 2006). For example, while the violation of human rights is generally perceived as irresponsible, not violating human rights, however, should be a matter of course. This observation also explains why corporate socially responsible behavior is hardly reported in the news, whereas socially irresponsible behavior receives high media coverage (*Kölbel* et al. 2017).

Kölbel et al. (2017) investigated whether CSI is linked to higher financial risk. Contrary to the idea of a value-enhancing and reputation-building effect of CSP (i. e. ESG performance), they instead examined the risk-generating effects of CSI, based on research by *Strike* et al. (2006). To measure CSI, they used data from RepRisk. The link between CSI and financial risk is promising as both concern the risk dimension of firms. *Kölbel* et al. (2017) found that firms facing CSI also experience significantly higher CDS spreads and that this effect is primarily driven by the governance dimension.

Our investigation is based primarily on *Kölbel* et al. (2017)'s findings on CSI and CDS spreads. Their results lead us to expect that information on CSI should

be relevant to predicting credit risk and thus be priced in credit ratings. In contrast to previous studies, we are interested in robust empirical evidence as to whether CSI (i. e. compared to ESG performance) informs credit risk predictions as reflected in credit ratings (i. e. compared to CDS spreads). Similar to *Kölbel* et al. (2017), we used data on CSI from RepRisk. Moreover, we are also interested in the predictive ability that a machine learning regression design can add to this kind of prediction problem.

III. Research Design

This section discusses CSI predictors, additional market- and accounting-based predictors, the measure of credit risk (i. e. the target), and the sample collection process. The data part is followed by explaining the ML algorithms used in this study.

1. Predictors for Corporate Social Irresponsibility

We used data from RepRisk to measure the CSI of firms. RepRisk specializes in ESG and business conduct risk research by leveraging artificial intelligence and big data to derive relevant risk metrics. Therefore, RepRisk adheres to the term "ESG risk." RepRisk captures material ESG risks and violations of international standards, which can impact the reputation and financial performance of firms. They screen over 90,000 public sources and stakeholders in 20 different languages on a daily basis. Thus, risk metrics are also updated daily. Their screening universe comprised 130k public and private firms (cross-sector and cross-market) when we conducted our study. The historical data on RepRisk dates back to January 2007. The research scope covers 28 broad, comprehensive, and mutually-exclusive ESG issues. For instance, "waste issues," "local pollution," "forced labor," "human rights," and "fraud." Every risk incident in the RepRisk database is linked to one of these 28 ESG issues. RepRisk identifies any company that is associated with a negative ESG incident.

RepRisk analyzes each risk incident according to its severity (harshness), reach (influence), and novelty. Finally, using a proprietary algorithm, RepRisk quantifies a firm's ESG risk with its "RepRisk Index (RRI)." The RRI captures and quantifies reputational risk exposure related to the 28 ESG issues. It ranges from 0 to 100, and higher values indicate higher ESG risk exposure. RepRisk provides three different RRI values. First, Current RRI, which measures the current ESG risk exposure (i. e. quarterly in our case). Second, Peak RRI, which captures the highest level of ESG risk exposure over the two previous years. While the Current RRI is instead a snapshot of a firm's exposure to ESG risks, Peak RRI accounts for the time-delayed manifestations of overall ESG risk.

Third, RRI Change or Trend, which is the past 30-day increase or decrease of the RRI. Since we used quarterly observations, we only considered Current RRI and Peak RRI as predictors of credit risk. We further used a vector that contains the number of incidents for the 28 single ESG issues multiplied by the respective news count as an alternative measure of CSI.[5]

One of the advantages of RepRisk compared to, for instance, MSCI ESG or Refinitiv is that the RRI is based on third-party disclosed information (i. e. news media) rather than on self-disclosed information. We are, of course, aware that this fact does not necessarily imply an entirely objective measure of CSI. This concern is reinforced by the fact that the algorithm calculating the RRI is proprietary and thus not available for external scrutiny. Moreover, only a few studies have so far used RepRisk data (e. g. *Kölbel* et al. 2017). We believe this presents a chance but must also be kept in mind as a caveat. Ultimately, we find that RepRisk provides a reliable estimate of CSI (i. e. not ESG performance), one that is readily available to researchers and market participants.

2. Additional Accounting and Market Related Predictors

In addition to CSI, we controlled for several important factors of credit risk, which are well-founded in the previous literature. We obtained our additional credit risk predictors from Compustat/CRSP for all our sample firms. We follow Agarwal and *Taffler's* (2008) three main arguments for an accounting-based model: First, credit risk does not suddenly arise but evolves slowly and thus is best reflected in fundamental values. Second, "window dressing" very likely only marginally affects predictions when using a sophisticated set of accounting variables. Third, loan covenants tend to be based on fundamental values, as reflected in accounting-based numbers.

Most importantly, we used the common accounting-based variables for corporate distress as evidenced by *Altman* (1968) and *Ohlson* (1980): Working capital/Total assets (LIQU), Retained earnings/Total assets (PROF), Earnings before interest and taxes (Ebit)/Total assets (OPEF), Market value/Long-term debt (ME), and Sales/Total assets (AT). Moreover, we used predictors that account for short-term effects on financial performance and credit risk (*Barboza* et al. 2017). These variables include growth in assets (ASSETS), growth in sales (SALES), and percentage change in the price-to-book ratio (P/B). Since firms with higher leverage are also more likely to have difficulties in repaying their debts, we accounted for the predictive ability of long-term debt/common equity

[5] More information on our ESG Issues measure can be found in Table A.6. For more information on the RepRisk methodology, please see https://www.reprisk.com/content/static/reprisk-methodology-overview.pdf.

(LEV) (*Oikonomou* et al. 2014; *Attig* et al. 2013; *Sun/Cui* 2014). More mature firms are usually less likely to face default, which is why we included firm age (AGE) as a predictor (*Sun/Cui* 2014). Firm age is calculated as the current year minus the year in which the company first appeared in the Compustat/CRSP database. Smaller firms usually experience much higher credit risk than larger firms (*Vassalou/Xing* 2004). Therefore, we included firm size (SIZE) as a predictor.

Following *Merton* (1974) and *Ashbaugh-Skaife* et al. (2006), we added operating income before depreciation/interest expense (interest coverage; INTCOV) as an early indicator of the ability to repay debt and net PPE/total assets (CAPINT).[6] Based on the Global Industry Classification Standard, we also included sector dummies into our models to account for the impact of sector specificity on credit risk. All our variables are calculated as ratios at the end of every quarter, which allowed us to investigate a large international sample without facing currency conversion issues. However, in an additional analysis, we repeated our main analyses with a more homogeneous sample by using only US firms. All predictors were measured at two quarters before the credit rating (i. e. with a lag of two quarters, i. e. at $t - 2$) to measure the ex-ante predictive ability of our models.

3. *Proxy for Credit Risk*

Regarding the target variable, our study differs substantially from previous research on default or bankruptcy prediction, which has commonly used actual firm defaults, as shown in *Alaka* et al. (2018). However, we are interested in firm credit risk that ultimately informs *ex-ante* about firm default probability. We measured credit risk by S&P long-term issuer credit ratings (henceforth S&P ratings) as these ratings constitute an overall and forward-looking evaluation of a debtor's creditworthiness (i. e. the capacity and willingness to repay the obligations due).[7] S&P ratings have been widely used as a proxy for firm credit risk (e. g. *Sun/Cui* 2014; *Hsu/Chen* 2015; *Kealhofer* 2003).

In line with *Hsu/Chen* (2015) and *Avramov* et al. (2009), we coded the S&P ratings as a numerical variable ranging from 1 to 22. 1 indicates a AAA rating (i. e. the highest possible) and 22 reflects a D rating (i. e. the lowest possible). Thus, higher values for S&P ratings indicate higher credit and default risk, and

[6] In contrast to *Ashbaugh-Skaife* et al. (2006), we used net values for Property, Plant, & Equipment (PPE) as these were available for both US and international firms on Compustat/CRSP.

[7] For detailed information on S&P ratings, please see https://www.standardandpoors.com/en US/web/guest/article/-/view/sourceId/504352.

vice versa. It is important to note that S&P claims that their rating process considers corporate sustainability and social responsibility activities.[8] They also claim that ESG aspects are considered along three dimensions in their rating process: business risk (e. g. competitive position), financial risk (e. g. cash flow/leverage assessment), and management and governance. Therefore, when assuming that ESG aspects are reflected in credit ratings, using information about CSI in our models should inform credit risk predictions and thus increase prediction accuracy.

4. Sample Development

We began collecting data for CSI from RepRisk by retrieving quarterly observations between March 31st, 2007 (i. e. the inception of the RepRisk database), and September 30th, 2019 (i. e. the latest available data), for all firms in the RepRisk universe. This procedure yielded 805,617 firm-quarter observations with available data. However, the RepRisk universe also contains many private (i. e. unlisted) firms whose accounting- and market-based values are available only to very limited extent. In fact, after obtaining data from Compustat/CRSP on the additional predictors and merging the two datasets, our sample shrank to 502,090 firm-quarter observations. Next, we obtained quarterly S&P long-term issuer credit ratings from Compustat-Capital-IQ for the same period. If a rating was revised within the same quarter, we used the rating that was closer to the end of the quarter and thus also more up-to-date. After merging credit risk data with RepRisk and Compustat/CRSP data and after dropping observations with missing values, we ended up with an unbalanced final sample of 43,476 firm-quarter observations (representing 1,850 firms). In a machine learning context, it is vital to further split one's sample into a training sample and a test (or validation) sample. More specifically, our ML models were only trained on the training sample and not on the test sample. Accordingly, we used the test sample to make out-of-sample predictions. For the training sample, we used 80 % of our data (i. e. 34,780 firm-quarter observations). The test sample consisted of the remaining 20 % (i. e. 8,696 firm-quarter observations). We developed a stratified sampling approach, ensuring that the target variable's labels (1 for AAA to 22 for D) were first shuffled and then approximately evenly distributed among the training and test samples.[9] Without stratifying, we would have

[8] For more information on ESG considerations in S&P credit ratings, please see https://www.spglobal.com/ratings/en/products-benefits/products/sustainable-finance and https://www.spglobal.com/ratings/en/products-benefits/products/esg-in-credit-ratings#.

[9] We stratified our sample by applying the allclose function of NumPy (https://docs.scipy.org/doc/numpy/reference/generated/numpy.allclose.html). The exact code can be found in our Github repository, to which access can be granted on request.

risked biasing our results by uneven distribution or by the order of the target variable labels due to a purely random split between training and test sample (i. e. sample bias).

5. Prediction Using a Selection of Machine Learning Algorithms

We used Python 3.7.4 and a set of algorithms implemented by the widespread ML library sklearn to train all models, except for one, and to make predictions with new and unseen data (i. e. out-of-sample). We used the following algorithms: Linear regression, Elastic Net, Ridge, Ordinal Ridge, Support Vector Regression (SVR) with a linear kernel, SVR with a radial basis function (RBF) kernel, Decision Tree regression (DTR), Random Forest regression (RFR), AdaBoost regression (AdaBoostR), Gradient Boosting regression (GBR), K-nearest Neighbors regression (KNNR), and Multi-layer Perceptron regression (MLPR). While the first eleven algorithms belong to the typical ML repertoire, the last algorithm, MLPR, is a deep learning technique. Algorithms that do not explicitly account for non-linearity in the data (i. e. Linear regression and SVR with a linear kernel) or simply apply a regularization term (i. e. Elastic Net, Ridge, and Ordinal Ridge) are considered "traditional" algorithms in this study. The remaining algorithms are considered some sort of "advanced" ML algorithms. Below, we describe the applied algorithms in more detail.

a) Linear Regression

Our first algorithm is "plain vanilla" linear regression (or ordinary least squares (OLS) regression). By fitting a linear model's coefficients to the training data, we minimized the residual sum of squares between the actual output and the approximated output.

b) Elastic Net

Elastic Net is a regularization technique that adds a cost term to linear regression. Elastic Nets can be useful when only a few predictors are relevant or when the predictors are highly correlated. Elastic Net simply shrinks the weights of the irrelevant predictors. Since default and credit risk is usually determined by a considerable number of potential factors (e. g. accounting- and market-related data as well as non-financial data), we ran an Elastic Net algorithm on our extensive sample. This approach allowed us to test whether only a subset of predictors already yields reasonable results. Elastic Net is a common regularization technique that has also been applied in finance (e. g. *Wu/Yang*, 2014).

c) Ridge

Ridge regression is another regularization technique that applies a penalty to the size of coefficients, usually increasing their robustness against collinearity and avoiding over-fitting on the training data.

d) Ordinal Ridge

Similar to Ridge regression, Ordinal Ridge applies the C2 penalty for regularization. However, the main difference to Ridge regression is that the output value is ordinal (i. e. a scale of categories where the relative ordering matters). Ordinal regression models explicitly account for data ordinality (*McCullagh* 1980). Ordinal Ridge seems to be a promising algorithm for our type of optimization problem since we are measuring credit risk continuously, yet on a predetermined scale from 1 to 22.[10]

e) Support Vector Regression

Support Vector Regression (SVR) or Support Vector Machines (SVM) are another very common ML algorithm with various possible applications, for instance, bankruptcy prediction (*Chen* 2011). sklearn implements *t:*-SVR based on the LIBSVM library from *Chang/Lin* (2011) and the LIBLINEAR library from *Fan* et al. (2008).[11] SVR uses kernel functions to solve the optimization problem. We used two SVR algorithms: one with a linear kernel (LinearSVR), another with a radial basis function (RBF) kernel (RbfSVR). In general, SVR fits a hyperplane on the training data, which maximizes the margin (i. e. as many training instances within the margin, restricted by the margin of tolerance, defined as *t:*) and thus minimizes the error.

[10] We used the "mord" package to implement Ordinal Ridge since sklearn has no ordinal ridge (or even simple ordinal regression) implementation. For more information, please see https://pythonhosted.org/mord/. More details on the scale of our target variable can be found in Section 3.

[11] For reasons of consistency, we adhered to the LIBSVM implementation in sklearn for RbfSVR and LinearSVR. However, it is very likely that using the LIBLINEAR implementation for LinearSVR increases computing efficiency and decreases training time compared to the LIBSVM implementation (*Fan* et al. 2008).

f) Decision Tree Regression

In contrast to, for instance, linear regression, Decision Tree regression (DTR) is a non-parametric algorithm. However, it also leverages supervised learning[12], similar to all the other algorithms, except for the only artificial neural network (ANN) method in our study, Multi-layer Perceptron Regression (MLPR). DTR fits a model to the training data that predicts an output value by learning from simple decision rules based on the predictor's data. DTs are easy to understand and interpret. They also require almost no data adjustments (e. g. in contrast to MLPR). sklearn uses an optimized version of the Classification And Regression Tree (CART) algorithm, which was first introduced by *Breiman* et al. (1984).

g) Random Forest Regression

DTR typically over-fits the training data and thus exhibits high variance when making predictions based on new and unseen data (i. e. the test data). Random Forest regression (RFR) is an ensemble approach. It makes predictions by combining several base estimators (e. g. decision tree regressors) through averaging (also called "bootstrap aggregating" or "bagging"). The combined and averaged prediction obtained by bagging is usually superior to the prediction of a single estimator (i. e. higher accuracy, robustness, and generalizability) (*Breiman* 2001). This superiority originates from two sources of randomness in every single estimator: bootstrapped sampling and random subsets of predictors (*Breiman* 2001). Bootstrapped sampling ensures that every estimator (e. g. decision tree regressor) is trained on a different draw (with replacement) from the sample. The increased randomness implicates more decoupled prediction errors between the single estimators; thus, trading decreased variance for slightly increased bias.[13] Ultimately, bagging reduces over-fitting the model on the training data (i. e. higher generalizability).

h) AdaBoost Regression

In contrast to the bagging algorithm RFR, AdaBoost regression (AdaBoostR) uses a technique called "boosting." In boosting algorithms, the single estimators are built sequentially rather than independently or in parallel (as in bagging algorithms). Therefore, boosting fits a series of so-called "weak learners" (e. g. small decision tree regressors) on repeatedly adjusted training data (*Hastie* et al.

[12] For a detailed explanation of supervised vs. unsupervised learning, please see *Basheer/Hajmeer* (2000).

[13] For a discussion on the bias-variance trade-off, please see *Géron* (2017, p. 129).

2009). At each iteration, the boosting algorithm learns from the errors made by the previous weak learner and improves the prediction. Thus, samples that are difficult to predict become increasingly influential with every iteration (*Hastie* et al. 2009). The general idea of AdaBoostR is that the weights of the various instances are adjusted at every iteration (*Géron* 2017).

i) Gradient Boost Regression

In addition to AdaBoostR, we decided to use Gradient Boosting regression (GBR) as an alternative algorithm using the boosting technique. In contrast to AdaBoostR, at every iteration, GBR fits a new predictor on the residual errors of the previous predictor.

j) K-neaerest Neighbors Regression

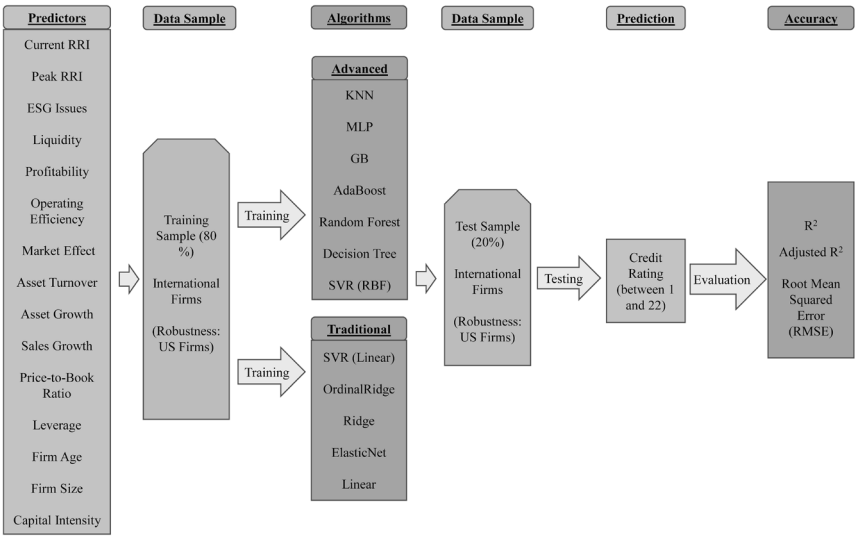
K-nearest Neighbors regression (KNNR) is another very intuitive and efficient regression algorithm. It makes predictions based on new data by relying on the similarity of predictors (or features) of data points from the training data. KNNR predicts a value based on how closely it can be resembled with the points in the training data. If k equals 5, KNNR uses the average of the five closest neighboring data points as a final prediction.

k) Multi-layer Perceptron Regression

The last algorithm is Multi-layer Perceptron regression (MLPR), which is a DL technique. We used the widespread and well-proven algorithm implementation for first-order optimization problems called Adam, developed by Kingma and Ba (2015). Adam uses stochastic optimization only requiring first-order gradients. Adam is very memory-efficient, and parameter updates occur independently of rescaling the gradient (*Kingma/Ba* 2015).

Most importantly, and similar to previous research (e. g. *Barboza* et al. 2017), we used default settings for all algorithms. Training and prediction time did not exceed one hour in total, given our tidy data set and an efficient algorithm implementation with sklearn. Similar to *Barboza* et al. (2017), none of the algorithms needed more than six minutes for training and prediction. More information on the algorithm settings can be found in Table A.7 in the appendix.

We followed previous research and explicitly refrained from data adjustments such as standardization, normalization, or the removal of outliers (e. g. *Barboza* et al. 2017; *Cleofas-Sánchez* et al. 2016). In addition to using default package settings of ML algorithms, this approach not only increases replicability but also



Note(s): This figure shows the training and testing process using all relevant predictors and algorithms. The model specifications are not part of this illustration as the same training and testing procedure was used for all specifications.

Figure 1: Graphical Approach

enables drawing conclusions about practical applicability. However, it is very likely that the predictive ability of some of our models considerably suffered under this approach. The two exceptions were SVR (linear and RBF kernel) and MLPR, for which we used standardized data (i. e. removed the mean and scaled to unit-variance). Previous research has shown that predictor scaling (i. e. standardization) is a crucial requirement for algorithms such as SVR and MLPR (e. g. *Shanker* et al. 1996). Unreported results have shown that using non-standardized predictors in SVR and MLPR may lead to highly biased and arbitrary results. Moreover, training time is much higher with non-standardized predictors, which ultimately led us to retain the standardized version as replication needs less computing power. Figure 1 illustrates the process of training and testing in our machine learning setting.

We solved the credit risk prediction task by using the three aforementioned distinct predictors (Current RRI, Peak RRI, and the vector of ESG issues) and by leveraging all of the explained ML algorithms. In each of the three specifications, we included a vector of the additional accounting- and market-related predictors and a vector of sector dummies. We benchmarked the three specifications against a baseline specification that only consisted of the accounting- and market-related predictors and sector dummies. We tested for model accuracy using the three most common measures for regression designs: R^2, adjusted R^2, and RMSE.

Similar to Gu et al. (2020), we also tested the pairwise differences in prediction accuracy for statistical significance according to the *Diebold/Mariano* (1995) test (henceforth, DM test).[14] We made two types of comparison: First, we compared differences in accuracy between the predictions of algorithms (e. g. Linear regression vs. Random Forest regression) but within our baseline model (i. e. without considering CSI). Second, we compared differences in accuracy between the predictions of our models (e. g. baseline vs. Current RRI) but within a certain algorithm (e. g. Linear regression). In both cases, we defined the test statistic between two prediction samples $DM_{12} = \frac{\bar{d}_{12}}{\sigma_{\bar{d}_{12}}}$, where

$$\bar{d}_{12} = \frac{1}{T-1}\sum_{t=1}^{T}\left[\left(e^{(1)}\right)^2 - \left(e^{(2)}\right)^2\right] \tag{1}$$

$\left(e^{(1)}\right)^2$ and $\left(e^{(2)}\right)^2$ are the prediction errors of both tested methods and $\sigma_{\bar{d}_{12}}$ denotes the standard deviation.

IV. Descriptive Statistics

Table 1 shows the mean, median, standard deviation, minimum, and maximum of the target variable and the predictors used in our models. According to RepRisk, the RRI of most international firms is located somewhere between 26 and 49 (medium risk exposure) due to their global footprint, and only very few firms reach an RRI between 75 and 100 (extremely high-risk exposure). In our sample, the mean and median of Current RRI and Peak RRI are below 25, which is the threshold for "medium ESG risk exposure" according to RepRisk. However, standard deviation is comparably high (13.9 and 18.0 points, respectively). The minimum and maximum values for Current RRI and Peak RRI are 0 and 79, respectively. Zero represents firm-quarter observations with "no ESG risk exposure" while 79 represents observations with "extremely high ESG risk exposure." Therefore, the descriptives indicate adequate variation in our data and thus the potentially good explanatory and discriminatory power of our two main predictors. The remaining predictor values are comparable with other studies (e. g. *Barboza* et al. 2017; *Oikonomou* et al. 2014). However, as mentioned, we refrained from data adjustments (e. g. removing outliers or scaling), as reflected in the summary statistics.

Table 2 shows the summary statistics for all 28 ESG issues covered by RepRisk. The three issues "Corruption, bribery, extortion and money laundery," "Fraud," and "Violation of national legislation" had the highest single news count in at least one firm-quarter observation (30, 22, and 28, respectively).

[14] Our DM test includes the modifications suggested by *Harvey* et al. (1997).

Table 1

Summary Statistics – Predictors and Targets

	Mean	*SD*	*Min*	*50 %*	*Max*
Credit Risk	10.681	3.193	1.000	10.000	22.000
Current RRI	13.466	13.856	0.000	13.000	79.000
Peak RRI	24.539	18.048	0.000	27.000	79.000
LIQU	0.100	0.148	−4.756	0.080	0.862
PROF	0.119	0.451	−9.649	0.136	2.184
OPEF	0.022	0.020	−0.360	0.019	0.458
ME	0.844	2.161	−12.189	1.001	16.801
AT	0.208	0.163	−0.127	0.166	2.245
ASSETS	0.233	42.616	−1.000	0.006	8,885.494
SALES	0.244	41.644	−5.480	0.010	8,682.722
P/B	0.248	59.139	−3,384.727	−0.011	8,303.042
LEV	6.143	746.36	7,871.500	0.689	109,761.545
AGE	27.661	17.999	0.000	22.000	69.000
SIZE	9.289	1.896	3.957	8.952	19.694
INTCOV	1.993	1.113	−6.444	1.961	9.588
CAPINT	0.352	0.256	0.000	0.300	0.989

Notes: This table reports time-series averages of cross-sectional means and standard deviations (SD), the minimum, the 50th percentiles (Median) and the maximum for all variables. The number of observations (N) is 43,476.

A value of 30 for "Corruption, bribery, extortion and money laundery" indicates that this specific firm was involved in or linked to 30 different risk incidents during the quarter. The same applies to the other ESG issues. Of course, these values are extremes, and the mean and median show that the majority of firms experienced only very few incidents around all ESG issues. However, these extremes show that single firms may be massively exposed to CSI, which in turn may considerably jeopardize firm reputation, firm performance, and solvency. Apart from that, the sample firms' incidents seem to be quite evenly distributed among the remaining ESG issues.

Table 2

Summary Statistics – ESG Issues

	Mean	*SD*	*Min*	*50 %*	*Max*
Animal Mistreatment	0.003	0.055	0.0	0.0	2.0
Anti-Competitive Practices	0.041	0.268	0.0	0.0	7.0
Child Labor	0.008	0.106	0.0	0.0	4.0
Climate Change, GHG Emissions, and Global Pollution	0.024	0.201	0.0	0.0	8.0
Controversial Products and Services	0.024	0.240	0.0	0.0	13.0
Corruption, Bribery, Extortion and Money Laundry	0.048	0.463	0.0	0.0	30.0
Discrimination in Employment	0.009	0.108	0.0	0.0	3.0
Executive Compensation Issues	0.006	0.085	0.0	0.0	3.0
Forced Labor	0.009	0.118	0.0	0.0	7.0
Fraud	0.041	0.332	0.0	0.0	22.0
Freedom of Association and Collective Bargaining	0.011	0.117	0.0	0.0	3.0
Human Rights Abuses and Corporate Complicity	0.049	0.312	0.0	0.0	18.0
Impacts on Communities	0.084	0.449	0.0	0.0	11.0
Impacts on Landscapes, Ecosystems and Biodiversity	0.076	0.460	0.0	0.0	27.0
Local Participation Issues	0.014	0.144	0.0	0.0	4.0
Local Pollution	0.061	0.404	0.0	0.0	25.0
Misleading Communication	0.014	0.140	0.0	0.0	4.0
Occupational Health and Safety Issues	0.032	0.227	0.0	0.0	10.0
Other Environmental Issues	0.000	0.005	0.0	0.0	1.0
Other Issues	0.000	0.011	0.0	0.0	1.0
Other Social Issues	0.000	0.005	0.0	0.0	1.0
Overuse and Wasting of Resources	0.005	0.081	0.0	0.0	4.0
Poor Employment Conditions	0.034	0.240	0.0	0.0	10.0
Products (Health and Environmental Issues)	0.045	0.374	0.0	0.0	17.0
Social Discrimination	0.003	0.056	0.0	0.0	2.0
Supply Chain Issues	0.043	0.322	0.0	0.0	12.0
Tax Evasion	0.006	0.091	0.0	0.0	4.0
Tax Optimization	0.005	0.095	0.0	0.0	7.0
Violation of International Standards	0.007	0.090	0.0	0.0	3.0
Violation of National Legislation	0.167	0.741	0.0	0.0	28.0
Waste Issues	0.017	0.148	0.0	0.0	5.0

Notes: This table reports time-series averages of cross-sectional means and standard deviations (SD), the minimum, the 50th percentiles (Median) and the maximum for all ESG issues. The number of observations (N) is 43,476.

V. Results

This section presents the main results of our study, the results of our series of additional analyses, and some additional analysis and practical implications.

1. Main Results

Table 3 shows the results for all machine learning algorithms for our baseline specification (i. e. without including CSI exposure). Unsurprisingly, the linear regression algorithm reached only a comparably low out-of-sample prediction accuracy of 56.31 % for adj. R^2. Regularization, i. e. ElasticNet, Ridge, and OrdinalRidge, did not increase prediction accuracy after all. Shrinking the weights of supposedly irrelevant predictors, such as in ElasticNet, even impaired prediction accuracy (reducing adj. R^2 from 56.31 % to 44.92 %). This finding indicates that many of the well-documented predictors used in this study are useful for predicting credit risk and should therefore be considered in credit risk models. While LinearSVR is supposedly unable to capture non-linearity in the data (55.36 % for adj. R^2), the comparably poor result for AdaBoostR is surprising (57.07 % for adj. R^2). Moreover, our results show that more advanced ML algorithms clearly outperformed linear regression. SVR with a radial basis function kernel and gradient boosting regression were superior to linear regression and the regularization techniques. Out-of-sample adj. R^2 increased to 76.79 % for RbfSVR and 73.39 % for GBR.

Four of the more advanced ML algorithms achieved adj. R^2s greater than 80 %: Decision Tree regression, Random Forest regression, K-nearest Neighbors regression, and Multi-layer Perceptron regression. Random Forest regression achieved the highest out-of-sample prediction accuracy of all algorithms: 89.65 % for adj. R^2 and 1.02 for RMSE. This outstanding result is most likely due to the averaging technique of the Random Forest algorithm. Random Forest regression combines the prediction of several base estimators with an ensemble prediction, which is less likely to over-fit the noise in the training data and thus to generalize better to (unseen) test data. In contrast, Decision Tree regression largely over-fitted the training data (adj. R^2 of 99.90 % on the training data compared to 80.72 % on the test data). The over-fitting tendency of Decision Tree regression and Random Forest regression has been documented in many previous studies (e. g. *Barboza* et al. 2017; *Piramuthu* 2006). However, Random Forest regression still seems to generalize best to unseen data (i. e. out-of-sample).

K-nearest Neighbors regression and Multi-layer Perceptron regression also seem to be suitable algorithms for predicting credit risk by using the similarities between different firms. After all, they reached an adj. R^2 of 85.05 % and 80.35 %, respectively. Whereas *Barboza* et al. (2017) showed that boosting and bagging

Table 3

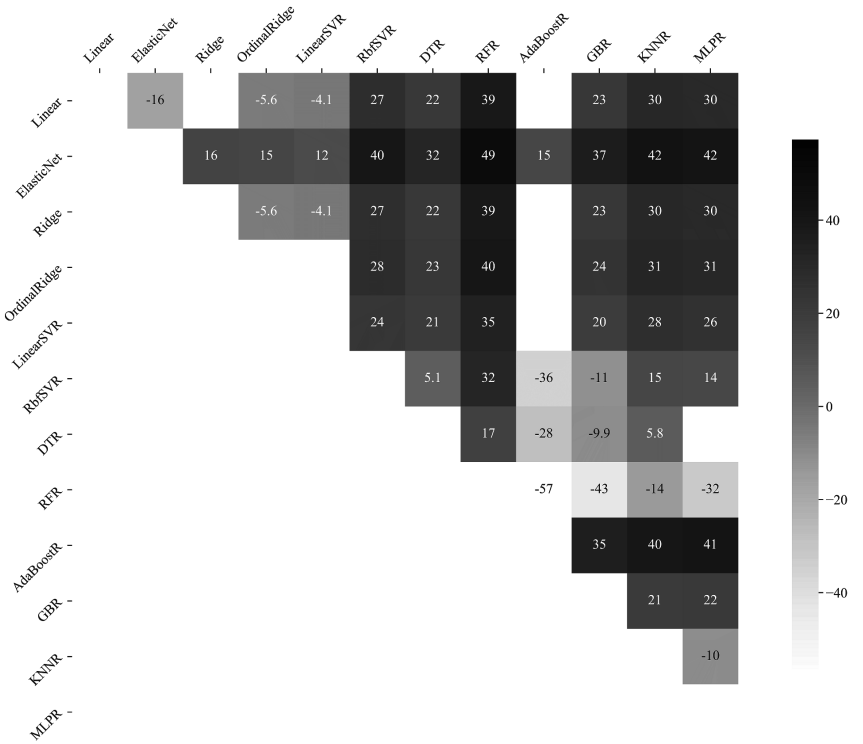
Prediction Accuracy of Machine Learning Algorithms – Baseline Model

Training Sample				*Test Sample*			
Algorithm	R^2	*adj.* R^2	*RMSE*	*Algorithm*	R^2	*adj.* R^2	*RMSE*
Linear	0.5601	0.5598	2.1186	*Linear*	0.5644	0.5631	2.1018
ElasticNet	0.4416	0.4412	2.3868	*ElasticNet*	0.4508	0.4492	2.3600
Ridge	0.5601	0.5598	2.1186	*Ridge*	0.5644	0.5631	2.1018
OrdinalRidge	0.5526	0.5522	2.1366	*OrdinalRidge*	0.5562	0.5549	2.1216
LinearSVR	0.5528	0.5524	2.1361	*LinearSVR*	0.5548	0.5536	2.1247
RbfSVR	0.7658	0.7657	1.5457	*RbfSVR*	0.7686	0.7679	1.5320
DTR	0.9990	0.9990	0.1031	*DTR*	0.8078	0.8072	1.3963
RFR	0.9787	0.9787	0.4659	*RFR*	0.8968	0.8965	1.0230
AdaBoostR	0.5831	0.5828	2.0624	*AdaBoostR*	0.5719	0.5707	2.0836
GBR	0.7356	0.7354	1.6425	*GBR*	0.7347	0.7339	1.6403
KNNR	0.9093	0.9092	0.9621	*KNNR*	0.8510	0.8505	1.2294
MLPR	0.7923	0.7921	1.4558	*MLPR*	0.8041	0.8035	1.4095

algorithms[15] have a reasonable performance in classification-based bankruptcy prediction, we only found evidence for a high prediction accuracy of bagging algorithms in the case of regression-based credit risk prediction. However, consistent with the findings of *Barboza* et al. (2017) for Random Forest classification, we found that Random Forest also provides the highest prediction accuracy in a regression and credit risk design. Other studies have found similar evidence for Random Forests, yet in different contexts than bankruptcy or credit risk prediction (e. g. *Yeh* et al. 2014; *Svetnik* et al. 2003).

Using the DM test, we further examined whether the prediction accuracy of one algorithm was significantly superior to that of the others (i. e. pairwise comparisons). Figure 2 displays the results for our baseline specification. Positive values indicate that the column algorithm outperforms the row algorithm. Only statistically significant values ($p < 0.01$) are displayed. Unsurprisingly, the regularization techniques (ElasticNet, Ridge, and OrdinalRidge) did not outperform

[15] For a brief description of "boosting" and "bagging," please see the relevant methods section above and *Barboza* et al. (2017).



Note: This figure reports Diebold-Mariano test statistics comparing the out-of-sample predictions between all algo- rithms for our baseline specification (i. e. no CSI). Positive values indicate outperformance of the column algorithm compared to the row algorithm. Only differences statistically significant at the 1 % level or better are displayed.

Figure 2: Statistical Significance of Prediction Accuracy – Baseline Model

"plain vanilla" linear regression, as already suggested by the lower adj. R^2 ElasticNet was even outperformed by any other algorithm. However, the more advanced ML algorithms – except for AdaBoostR – yielded predictions that were significantly superior to those of linear regression and the regularization techniques. Again, Random Forest regression turned out to offer the highest prediction accuracy in pairwise comparisons. Therefore, the findings are in line with our results in Table 3, that advanced algorithms are superior to traditional ones.

Turning to the inclusion of CSI predictors, Panel A of Table 4 displays the prediction accuracy of all algorithms testing the model that includes our first CSI predictor (i. e. Current RRI). The prediction accuracy of some of the algorithms only slightly increased when considering CSI in predicting credit risk. Panel B and Panel C of Table 4 support this finding of slightly increasing prediction accuracy when using Peak RRI and ESG Issues as alternative measures of CSI. Due

to the already high prediction accuracy of our baseline specification, most of the explanatory power seems to stem from the accounting- and market-related predictors. Even though previous research has shown that CSI can heavily impact firm performance, firm survival, and even credit risk, we did not find evidence that incorporating CSI into credit risk prediction considerably increases prediction accuracy.

Similar to the pairwise comparison of algorithms for our baseline model specification, we also conducted DM tests for those alternative specifications that include CSI (i. e. Current RRI, Peak RRI, ESG Issues). The two last columns of Table 4 contain the results for the DM tests, which compare the predictions of the respective alternative specifications to the baseline specification (e. g. Current RRI vs. baseline), holding the algorithm constant. The results were rather mixed. On the one hand, the more traditional ML algorithms (i. e. Linear, ElasticNet, Ridge, and OrdinalRidge) mostly yielded superior predictions when using the alternative specifications compared to the baseline specification (positive t-stat and mostly $p < 0.01$). In other words, predictions improved, yet only marginally. On the other hand, no clear trend could be identified for the more advanced ML algorithms as to whether predictions improved when using the alternative specifications compared to the baseline specifications (e. g. t-stat of −0.36 for RFR using Current RRI and t-stat of 3.85 for RFR using Peak RRI). The most consistent specification turned out to be Peak RRI, which yielded superior predictions compared to the baseline specification in the case of all algorithms – except when using DTR and GBR.

One possible explanation for our results is that our measure of credit risk (i. e. S&P long-term issuer credit ratings) does not yet reflect CSI exposure to a larger extent. Another possible explanation is that higher CSI increases the cost of capital, as suggested by *Kölbel* et al. (2017), instead of being reflected in credit ratings. Therefore, our result does not indicate that CSI is consistently and systematically considered in the credit risk analyses of rating agencies. For example, S&P Global claims that its credit risk analyses consider ESG factors along three dimensions: business risk (e. g. competitive position), financial risk (e. g. cash flow/leverage assessment), and management and governance. However, we cannot rule out that differently operationalizing ESG aspects between RepRisk and S&P Global leads to our results.[16] For example, this difference in operationalization might be due to a possibly higher weighting of the corporate governance dimension, whereas the environmental and social dimensions are underweighted.

[16] For S&P Global's claims about integrating ESG in credit risk analysis, please see: https://www.spglobal.com/ratings/en/products-benefits/products/sustainable-finance and https://www.spglobal.com/ratings/en/products-benefits/products/esg-in-credit-ratings#.

Table 4

Prediction Accuracy of Machine Learning Algorithms – Baseline Model

Panel A: Current RRI

Training Sample				*Test Sample*				*DM Test*	
Algorithm	R^2	*adj.* R^2	*RMSE*	*Algorithm*	R^2	*adj.* R^2	*RMSE*	*t-stat*	*p-value*
Linear	0.5601	0.5598	2.1186	*Linear*	0.5644	0.5631	2.1018	6.71	0.0000
ElasticNet	0.4416	0.4412	2.3868	*ElasticNet*	0.4508	0.4492	2.3600	3.20	0.0014
Ridge	0.5601	0.5598	2.1186	*Ridge*	0.5644	0.5631	2.1018	6.71	0.0000
OrdinalRidge	0.5526	0.5522	2.1366	*OrdinalRidge*	0.5562	0.5549	2.1216	5.75	0.0000
LinearSVR	0.5528	0.5524	2.1361	*LinearSVR*	0.5548	0.5536	2.1247	8.08	0.0000
RbfSVR	0.7658	0.7657	1.5457	*RbfSVR*	0.7686	0.7679	1.5320	6.18	0.0000
DTR	0.9990	0.9990	0.1031	*DTR*	0.8078	0.8072	1.3963	−1.33	0.1850
RFR	0.9787	0.9787	0.4659	*RFR*	0.8968	0.8965	1.0230	−0.36	0.7202
AdaBoostR	0.5831	0.5828	2.0624	*AdaBoostR*	0.5719	0.5707	2.0836	8.95	0.0000
GBR	0.7356	0.7354	1.6425	*GBR*	0.7347	0.7339	1.6403	0.51	0.6129
KNNR	0.9093	0.9092	0.9621	*KNNR*	0.8510	0.8505	1.2294	−25.29	0.0000
MLPR	0.7923	0.7921	1.4558	*MLPR*	0.8041	0.8035	1.4095	7.37	0.0000

Panel B: Peak RRI

Training Sample				*Test Sample*				*DM Test*	
Algorithm	R^2	*adj.* R^2	*RMSE*	*Algorithm*	R^2	*adj.* R^2	*RMSE*	*t-stat*	*p-value*
Linear	0.5696	0.5693	2.0956	*Linear*	0.5734	0.5722	2.0799	5.78	0.0000
ElasticNet	0.4487	0.4483	2.3717	*ElasticNet*	0.4562	0.4545	2.3484	2.22	0.0266
Ridge	0.5696	0.5693	2.0956	*Ridge*	0.5734	0.5722	2.0799	5.78	0.0000
OrdinalRidge	0.5613	0.5610	2.1156	*OrdinalRidge*	0.5661	0.5648	2.0977	4.23	0.0000
LinearSVR	0.5632	0.5629	2.1109	*LinearSVR*	0.5656	0.5643	2.0990	7.06	0.0000
RbfSVR	0.7741	0.7739	1.5182	*RbfSVR*	0.7770	0.7763	1.5038	7.58	0.0000
DTR	0.9998	0.9998	0.0489	*DTR*	0.8115	0.8109	1.3826	0.68	0.4983
RFR	0.9872	0.9872	0.3612	*RFR*	0.9138	0.9136	0.9349	3.85	0.0001
AdaBoostR	0.5833	0.5830	2.0618	*AdaBoostR*	0.5768	0.5755	2.0717	2.51	0.0119
GBR	0.7356	0.7354	1.6426	*GBR*	0.7357	0.7349	1.6372	1.06	0.2889
KNNR	0.8816	0.8815	1.0990	*KNNR*	0.8086	0.8080	1.3932	−7.93	0.0000
MLPR	0.8142	0.8141	1.3939	*MLPR*	0.8171	0.8166	1.3618	6.12	0.0000

(continue next page)

(Table 4 continued)

Panel C: ESG Issues

Training Sample				*Test Sample*				*DM Test*	
Algorithm	R^2	*adj.* R^2	*RMSE*	*Algorithm*	R^2	*adj.* R^2	*RMSE*	*t-stat*	*p-value*
Linear	0.5681	0.5674	2.0991	*Linear*	0.5681	0.5674	2.0991	6.99	0.0000
ElasticNet	0.4416	0.4407	2.3868	*ElasticNet*	0.4416	0.4407	2.3868	–	–
Ridge	0.5681	0.5674	2.0991	*Ridge*	0.5681	0.5674	2.0991	6.99	0.0000
OrdinalRidge	0.5610	0.5603	2.1164	*OrdinalRidge*	0.5610	0.5603	2.1164	5.83	0.0000
LinearSVR	0.5608	0.5601	2.1168	*LinearSVR*	0.5608	0.5601	2.1168	7.53	0.0000
RbfSVR	0.7157	0.7152	1.7021	*RbfSVR*	0.7157	0.7152	1.7021	−18.39	0.0000
DTR	0.9990	0.9990	0.0988	*DTR*	0.9990	0.9990	0.0988	−1.18	0.2362
RFR	0.9865	0.9865	0.3715	*RFR*	0.9865	0.9865	0.3715	−2.21	0.0269
AdaBoostR	0.5831	0.5824	2.0624	*AdaBoostR*	0.5831	0.5824	2.0624	–	–
GBR	0.7364	0.7359	1.6401	*GBR*	0.7364	0.7359	1.6401	1.30	0.1945
KNNR	0.9000	0.8998	1.0103	*KNNR*	0.9000	0.8998	1.0103	−7.14	0.0000
MLPR	0.8146	0.8143	1.3753	*MLPR*	0.8146	0.8143	1.3753	−5.98	0.0000

Notes: This table reports the three most common measures of prediction accuracy (R^2, adj. R^2, and RMSE) for the models including the CSI predictors. In Panel A, we used Current RRI as a predictor to capture CSI exposure. In Panel B, we used Peak RRI as a predictor to capture CSI exposure. In Panel C, we used the vector of ESG issues as a predictor to capture CSI exposure. The vector of additional accounting- and market-related variables and the vector of sector dummies are included in all models. The statistics from the Diebold-Mariano test are displayed in the last two columns. They test the statistical significance of differences in the accuracy of predictions compared to the baseline specification (i. e., the model does not capture CSI; see Table 3 and Figure 2). Positive t-stat values indicate outperformance compared to the baseline model. "NaN" values indicate that the difference between the predictions was zero, i. e., the models yielded the same predictions. This finding is not surprising since regularization techniques, such as ElasticNet, shrink the weights from irrelevant predictors, which can render an equivalent or highly similar model to the baseline specification. For more information on algorithm specifications, please see Table A.7 in the appendix.

Figure 3 illustrates and compares the prediction accuracy of all our algorithms in the training and test samples.[17]

2. *Additional Analyses – Regression Design*

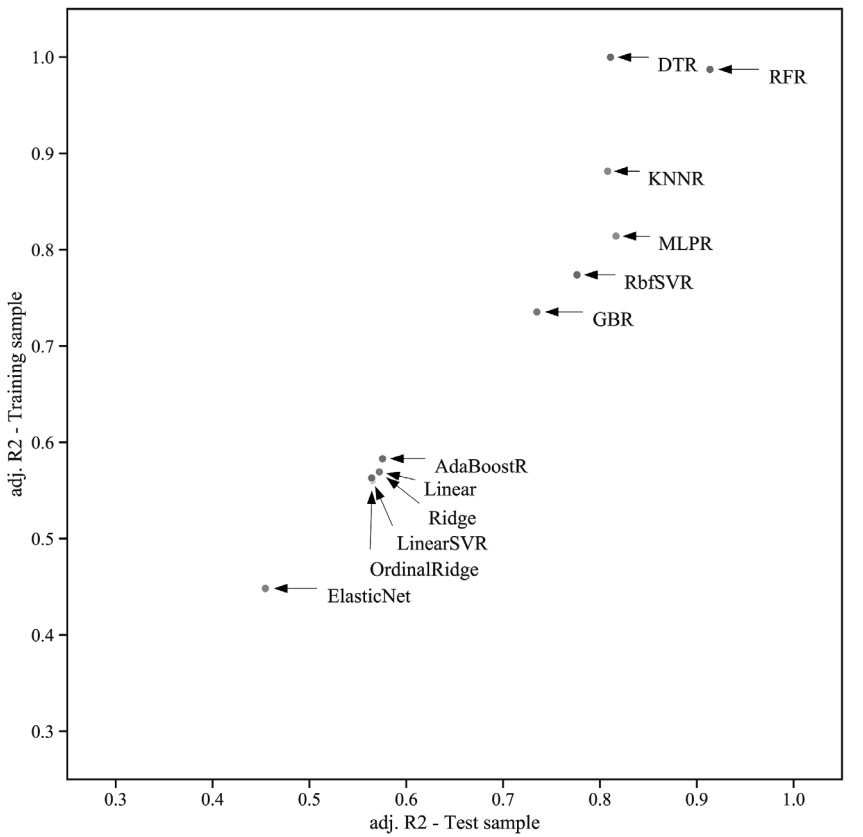
We performed a series of additional analyses.[18] First, we used only US data instead of global data. The results show that using US data increases the overall

[17] We decided only to illustrate the model using Peak RRI rather than the alternative CSI measures, as the model using Peak RRI achieved the highest out-of-sample prediction accuracy overall.

[18] The results of the additional analyses are unreported but available on request from the authors.

prediction accuracy of most ML algorithms. Using Peak RRI, Random Forest regression again yields the highest out-of-sample prediction accuracy with 92.53 % for adjusted R^2 and 0.8682 for RMSE. The higher overall prediction accuracy is presumably owed to the lower noise and the higher homogeneity in US data compared to global data.

Second, we tried to reduce the noise in the CSI data caused by negligible negative ESG incidents with no few (economic) consequences for the firm. More specifically, we followed *Kölbel* et al. (2017) and controlled for the reach of the news source in which the particular negative ESG incident was reported and for the severity (or harshness) of the negative ESG incident. A higher reach of the news source (e. g. The Financial Times vs. the Nassau Herald) and stronger se-



Note: This figure shows the prediction accuracy (adj. R^2) of all ML algorithms using Peak RRI as a predictor. The vector of additional accounting- and market-related variables and the vector of sector dummies are included. The y-axis displays adj. R^2 for the training sample and the x-axis shows adj. R^2 for the test sample.

Figure 3: Model Prediction Accuracy – Peak RRI

verity (e. g. large corruption scandal spanning several geographies vs. minor local pollution incident) are more likely to affect credit risk and thus be reflected in credit ratings. Therefore, we calculated interaction terms for our vector of ESG issues. Only negative ESG incidents that appeared in high-reach news sources and with high severity were considered in our ESG Issues predictor. All other negative ESG incidents were weighted zero. As the results show, this alternative specification of ESG Issues did not considerably change our results. We conclude that negative ESG incidents, which appeared in influential news sources and which had a severely negative impact, do not increase prediction accuracy either and thus are not systematically reflected in credit risk ratings.

Third, we used a time lag of four quarters instead of two quarters for our predictor variables to account for the possibility that it takes longer for CSI to materialize in credit ratings. However, a longer time lag did not qualitatively change our results.

Fourth, we applied a subsample analysis only using firm-quarter observations from the year 2015 and later, based on the observation that credit rating agencies have strengthened their ESG considerations in recent years. Therefore, it is more likely that CSI informs credit risk prediction as reflected in credit ratings during this subperiod. Similar to the results for the full sample, we found that the prediction accuracy of some algorithms slightly improves (e. g. DTR, AdaBoostR, LinearSVR, OrdinalRidge, Ridge, ElasticNet, Linear). In contrast, that of others even slightly declines (e. g. KNNR) due to including CSI. However, in absolute terms, the difference in prediction accuracy between models using CSI predictors and those which do not is only marginal. The effect on prediction accuracy is slightly stronger when only using data since 2015 compared to the full sample. This finding is in line with the recent efforts of credit rating agencies to incorporate CSI more strongly into their credit risk analyses.

Finally, we retrieved both Moody's and Fitch credit risk data from Thomson Reuters to rule out that our result is driven by the use of S&P Global data. Similar to S&P ratings, Moody's and Fitch ratings were converted to numerical values according to the respective scale. Using Moody's and Fitch rating data considerably reduced our sample to 8,202 and 4,261 firm-quarter observations, respectively. Again, the two samples were split into a training (80 %) and a test (20 %) sample.

The findings indicate that the prediction accuracy of traditional algorithms considerably declines, whereas the more advanced algorithms largely maintain their prediction accuracy scores. Adjusted R^2s decrease considerably to only slightly above 20 % for most algorithms (except when using ElasticNet, which is even well-below 20 %). While most more advanced ML algorithms maintain their comparably high out-of-sample prediction accuracy, only Random Forest regression, K-nearest Neighbor regression, and Multi-layer Perceptron regres-

sion achieve outstanding out-of-sample prediction accuracy close to or above 80 % for adj. R^2. This finding is somewhat surprising since algorithms like Random Forest regression usually work best with sample sizes well-beyond 10,000. In our case, and despite the considerably decreased sample size, the mentioned algorithms maintain their prediction accuracy scores. Overall, S&P ratings are best explained by our models, followed by Fitch ratings, followed by Moody's ratings.

3. Predicting Investment Grade vs. Non-Investment Grade – Classification Design

So far, we have used a regression design to predict a credit risk continuum between 0 and 22, which possibly lacks discriminatory power in terms of distinguishing observations with different CSI. To alleviate this concern, we used an alternative classification design to predict whether a firm will be rated investment grade or non-investment grade in the future. In contrast to a regression design, classification predicts whether a particular status applies or not (i. e. a categorical decision between 1 and 0). The most common classification algorithm is Logistic regression.

From both an investor's and an analyst's perspective, it would be very interesting to investigate the prediction accuracy of ML algorithms for classifying a firm into investment grade or non-investment grade. The non-investment grade status of an asset is an essential piece of information for investors and analysts alike. For example, credit rating decisions such as a non-investment grade status have signaling value for investors and may significantly impact the capital structure of firms (*Graham/Harvey* 2001; *Kisgen* 2006). Moreover, while CSI did not inform the prediction of credit risk as reflected in credit ratings in a regression design, it potentially informs a credit risk agency's decision to classify a firm as non-investment grade. Therefore, we conducted an additional analysis to test whether our models can reliably predict the non-investment grade status of a firm two quarters into the future and whether data on CSI informs these predictions.

Usually, a firm with a rating of BBB – or lower has non-investment grade status. Accordingly, we classified each firm-quarter observation with a rating of BBB – or lower as non-investment grade (i. e. a value of 1). All other firm-quarter observations with ratings above BBB – were classified as investment grade (i. e. a value of 0). As for our traditional algorithms, we used Logistic regression (Logit) and Support Vector Classification (SVC) with a linear kernel (LinearSVC) as our benchmark classification algorithms. As for our more advanced algorithms, we used SVC with an RBF kernel (RbfSVC), Decision Tree classification (DTC), Random Forest classification (RFC), AdaBoost classification (AdaBoostC), Gradient Boosting classification (GBC), K-nearest Neighbors classifi-

cation (KNNC), and Multi-layer Perceptron classification (MLPC).[19] Again, we refrained from data adjustments such as standardization, normalization, or the removal of outliers, except for SVC and MLPC. We used four common measures for prediction accuracy in a machine learning classification context (Precision, Recall, F1-Score, and Accuracy), which are explained in detail in Table 5.

Table 5

Additional Analysis – Investment Grade vs Non-Investment Grade

Panel A: Baseline – Without CSI Exposure

Training Sample					*Test Sample*				
Algorithm	*Precision*	*Recall*	*F1-Score*	*Accuracy*	*Algorithm*	*Precision*	*Recall*	*F1-Score*	*Accuracy*
Logit	0.7931	0.6782	0.7312	0.7509	*Logit*	0.7854	0.6848	0.7317	0.7570
LinearSVC	0.8244	0.8003	0.8122	0.8152	*LinearSVC*	0.8188	0.8025	0.8106	0.8185
RbfSVC	0.9007	0.8740	0.8872	0.8890	*RbfSVC*	0.8998	0.8728	0.8861	0.8914
DTC	1.0000	0.9981	0.9990	0.9991	*DTC*	0.9240	0.9216	0.9228	0.9254
RFC	0.9982	0.9952	0.9967	0.9967	*RFC*	0.9634	0.9330	0.9480	0.9504
AdaBoostC	0.8481	0.8238	0.8358	0.8383	*AdaBoostC*	0.8406	0.8210	0.8307	0.8381
GBC	0.8869	0.8625	0.8745	0.8764	*GBC*	0.8741	0.8550	0.8645	0.8703
KNNC	0.9655	0.9596	0.9626	0.9627	*KNNC*	0.9281	0.9296	0.9289	0.9311
MLPC	0.9710	0.9596	0.9653	0.9655	*MLPC*	0.9468	0.9437	0.9452	0.9471

Panel B: Current RRI

Training Sample					*Test Sample*				
Algorithm	*Precision*	*Recall*	*F1-Score*	*Accuracy*	*Algorithm*	*Precision*	*Recall*	*F1-Score*	*Accuracy*
Logit	0.7859	0.6686	0.7225	0.7435	*Logit*	0.7768	0.6677	0.7181	0.7464
LinearSVC	0.8221	0.8022	0.8120	0.8145	*LinearSVC*	0.8206	0.8046	0.8125	0.8204
RbfSVC	0.9026	0.8772	0.8897	0.8914	*RbfSVC*	0.8996	0.8750	0.8871	0.8922
DTC	1.0000	0.9993	0.9996	0.9996	*DTC*	0.9196	0.9242	0.9219	0.9242
RFC	0.9988	0.9953	0.9971	0.9971	*RFC*	0.9611	0.9277	0.9441	0.9469
AdaBoostC	0.8452	0.8241	0.8345	0.8367	*AdaBoostC*	0.8406	0.8251	0.8328	0.8397
GBC	0.8860	0.8601	0.8728	0.8748	*GBC*	0.8733	0.8519	0.8625	0.8686
KNNC	0.9334	0.8898	0.9111	0.9133	*KNNC*	0.8799	0.8338	0.8562	0.8645
MLPC	0.9731	0.9638	0.9684	0.9686	*MLPC*	0.9528	0.9460	0.9494	0.9512

[19] Due to the similarity of most classification algorithms with the regression algorithms, we refrain from further explanation. However, detailed explanations and blueprints of the algorithms can be found at https://scikit-learn.org/stable/supervised_learning.html.

Panel C: Peak RRI

Training Sample					*Test Sample*				
Algorithm	*Precision*	*Recall*	*F1-Score*	*Accuracy*	*Algorithm*	*Precision*	*Recall*	*F1-Score*	*Accuracy*
Logit	0.7733	0.6067	0.6799	0.7147	*Logit*	0.7611	0.6133	0.6792	0.7198
LinearSVC	0.8235	0.8009	0.8120	0.8148	*LinearSVC*	0.8201	0.8053	0.8127	0.8204
RbfSVC	0.9068	0.8785	0.8924	0.8942	*RbfSVC*	0.9053	0.8769	0.8908	0.8960
DTC	1.0000	0.9992	0.9996	0.9996	*DTC*	0.9173	0.9230	0.9201	0.9225
RFC	0.9987	0.9967	0.9977	0.9977	*RFC*	0.9655	0.9311	0.9480	0.9506
AdaBoostC	0.8462	0.8279	0.8369	0.8389	*AdaBoostC*	0.8381	0.8281	0.8331	0.8395
GBC	0.8857	0.8607	0.8731	0.8750	*GBC*	0.8725	0.8526	0.8625	0.8684
KNNC	0.9532	0.9440	0.9486	0.9489	*KNNC*	0.9179	0.9085	0.9132	0.9164
MLPC	0.9726	0.9591	0.9658	0.9661	*MLPC*	0.9499	0.9415	0.9457	0.9477

Panel D: Peak RRI

Training Sample					*Test Sample*				
Algorithm	*Precision*	*Recall*	*F1-Score*	*Accuracy*	*Algorithm*	*Precision*	*Recall*	*F1-Score*	*Accuracy*
Logit	0.7768	0.6831	0.7269	0.7437	*Logit*	0.7709	0.6896	0.7280	0.7507
LinearSVC	0.8229	0.8028	0.8127	0.8152	*LinearSVC*	0.8192	0.8032	0.8111	0.8190
RbfSVC	0.8817	0.8616	0.8715	0.8731	*RbfSVC*	0.8705	0.8436	0.8568	0.8636
DTC	1.0000	0.9983	0.9992	0.9992	*DTC*	0.9211	0.9180	0.9195	0.9223
RFC	0.9984	0.9955	0.9969	0.9970	*RFC*	0.9594	0.9144	0.9364	0.9399
AdaBoostC	0.8492	0.8240	0.8364	0.8390	*AdaBoostC*	0.8385	0.8222	0.8303	0.8374
GBC	0.8873	0.8633	0.8752	0.8770	*GBC*	0.8753	0.8529	0.8640	0.8701
KNNC	0.9608	0.9556	0.9582	0.9583	*KNNC*	0.9198	0.9242	0.9220	0.9243
MLPC	0.9721	0.9763	0.9742	0.9742	*MLPC*	0.9264	0.9342	0.9303	0.9323

Notes: This table reports four common measures of prediction accuracy for all classification algorithms: Precision, Recall, F1-Score, and Accuracy. Precision is the accuracy of the positive predictions, i. e. the number of true positives (TPs) divided by the number of TPs plus the number of false positives (FPs). In this context, "the number of …" always refers to "the number of observations classified as … by the algorithm." For brevity, we use the shorter form. Recall is the ratio of positive observations that are correctly detected, i. e. the number of TPs divided by the number of TPs plus the number of false negatives (FNs). The F1-Score is the harmonic mean of precision and recall, i. e. the number of TPs divided by the number of TPs plus the number of FNs and FPs divided by two. Accuracy measures the fraction of correct predictions. In Panel A, we used the baseline specification and thus no CSI predictor. In Panel B, we used Current RRI as a predictor to capture CSI exposure. In Panel C, we used Peak RRI as a predictor to capture CSI exposure. In Panel D, we used the vector of ESG issues as a predictor to capture CSI exposure. The vector of additional accounting- and market-related variables and the vector of sector dummies are included in all models.

The results of our additional analysis are displayed in Table 5. In general, the results of the classification design are in line with those of the regression design. The two traditional ML algorithms, Logit and LinearSVC, exhibit reasonable out-of-sample prediction accuracy of 75.70 % and 81.85 %, respectively. However, the more advanced ML algorithms exhibit superior prediction accuracy. For example, DTC and RFC have a prediction accuracy of 92.54 % and 95.04 %, respectively. This result is in line with results for default prediction using a machine learning classification design (*Barboza* et al. 2017). Overall, our algorithms show higher Precision, i. e. the accuracy of positive predictions, than Recall, i. e. the fraction of positive observations that are correctly detected by the classifier. The deep learning method MLPC has the highest Recall of 94.37 %, which indicates that this method reliably identifies almost 95 % of all non-investment grade firms (out-of-sample).

Moreover, we do not find that including CSI consistently increases prediction accuracy when classifying firms into investment grade or non-investment grade. The prediction accuracy of most algorithms either remains approximately the same (e. g. DTC or GBC) or even slightly declines (e. g. Logit or KNNC). Overall, Random Forest classification again dominates the other algorithms with prediction accuracy between 93.99 % and 95.06 %. These results are in line with our main analysis. Random Forest algorithms provide the highest prediction accuracy for credit risk as measured by credit ratings and for classifying firms into non-investment grade. Moreover, including CSI in models does not considerably increase either the accuracy of credit risk prediction or the accuracy of classifying firms into non-investment grade.

VI. Contributions to Research and Practical Implications

Our research contributes to the literature and has practical implications in at least two ways. First, whereas previous research has relied on statistical inference, we used a comprehensive machine learning design to train models for out-of-sample prediction of credit risk (as measured by credit ratings). Previous studies have primarily investigated the impact of ESG performance (e. g. *Oikonomou* et al. 2014; *Stellner* et al. 2015; *Hsu/Chen* 2015) and of CSI (e. g. *Kölbel* et al. 2017) on credit risk. However, they have neglected the overall prediction of credit risk using data on CSI and the potential of ML algorithms. The only exception is *Dorfleitner* et al. (2020), from whom our study differs by using data on CSI (i. e. not CSP) and a comprehensive series of machine learning algorithms. Predictive models that use advanced machine learning algorithms potentially add value by capturing the inherent non-linearity and complex interaction effects in the data. Credit risk is potentially influenced by a vast number of factors, financial and non-financial, which need to be considered in modeling. For investors and analysts, it is vital to incorporate these factors into credit risk

analysis. ML techniques can help obtain the most accurate credit risk prediction, especially if large amounts of data with complex relationships and interaction effects are available. The resource intensiveness of credit risk evaluation and the importance of firm solvency for financial institutions support the data-driven machine learning approach.

Second, previous research has shown that ESG performance is relevant to credit risk (e. g. *Oikonomou* et al. 2014; *Stellner* et al. 2015; *Drago* et al. 2019). At the same time, major rating agencies claim that they account for ESG aspects to some extent in their ratings.[20] However, we found that including data on CSI does not systematically increase the prediction accuracy of credit risk as measured by credit ratings. Therefore, our finding that CSI does not inform credit risk prediction is somewhat surprising. It also contrasts slightly with previous studies, according to which more environmentally and socially responsible firms are rewarded with higher credit ratings (*Jiraporn* et al. 2014; *Attig* et al. 2013; *Kiesel/Lücke* 2019). We do not find support for the anecdotal evidence that changes in the ESG performance of certain firms during the period 2016 – 2018 led to changes in credit risk ratings on a systematic basis (*Henisz/McGlinch* 2019). Our result is more in line with research on the European market, which has found that higher ESG performance is not significantly related to higher credit ratings (*Stellner* et al. 2015). However, our finding holds for both a global and a US sample and is related to CSI as compared to CSP (i. e. ESG performance).

A possible explanation for the diverging results is that rating agencies tend to account for positive ESG performance – at least outside of Europe – but not for the downside risk of negative ESG incidents (i. e. CSI). Another possible explanation is that the operationalization of ESG aspects by credit rating agencies significantly differs from our operationalization (i. e. CSI as measured by data from RepRisk). Our finding is also somewhat surprising in the light of results by *Kölbel* et al. (2017), who found that CSI impacts credit risk through increased CDS spreads. In contrast, our results indicate that higher credit risk due to CSI is not reflected in credit ratings. It seems that while the market already accounts for CSI, CSI does not (yet) play a major role in credit ratings. Therefore, a potential implication for market participants is that common credit ratings might not entirely reflect the environmental and societal business conduct risks (e. g. weaknesses in corporate governance, low product safety, or compromised employee well-being) that firms are increasingly facing. These business conduct risks potentially entail substantial future expenses (e. g. compensation payments) and reputational damages.

Ultimately, then, Peak RRI adds most to prediction accuracy across algorithms. In absolute terms, however, the additional explanatory power is small

[20] For a detailed discussion of this inclusion, please see above.

compared to specifications only using accounting- and market-related predictors. However, this finding likely depends on our measure of credit risk, i. e., credit rating data from S&P Global, Moody's, and Fitch. In light of our results, and considering the potential importance of ESG issues for credit risk, it seems reasonable to more strongly incorporate CSI into credit risk analysis. Major credit rating agencies such as Fitch have recently stated that they are intensifying their efforts to integrate ESG into credit risk analysis and to be more transparent about integration.[21]

VII. Concluding Remarks

We have investigated whether using advanced ML algorithms and data on CSI improves firm credit risk prediction. We therefore also tested the extent to which increased credit risk due to higher CSI is reflected in credit ratings. We found that the use of advanced ML algorithms considerably improves prediction accuracy for credit risk as measured by credit ratings. However, we found no evidence that CSI is systematically reflected in credit ratings. A long series of additional analyses and robustness checks corroborated our results. In an alternative classification design, we investigated whether advanced ML algorithms and data on CSI inform the prediction of a firm's non-investment grade status (i. e. a credit rating of BBB– or lower). Whereas advanced ML algorithms again considerably increase prediction accuracy using this classification design, data on CSI are not consistently informative as to whether a firm is predicted to have non-investment grade status.

The results of our study are limited to some extent. First, the time-series availability of CSI and credit rating data is limited. As CSI data availability increases, training ML models will most likely yield more robust systems. These systems include ones that are more robust when considering time trends.

Second, studies like ours face the trade-off between using standardized and non-standardized predictors. On the one hand, regularizing and preprocessing data potentially impedes replicating and interpreting the results. On the other, refraining from adjustments such as standardization may yield biased and arbitrary results for certain algorithms. For instance, SVR and MLPR algorithms are highly sensitive to unscaled data. Therefore, we decided to apply standardization where using non-standardized values would have disproportionately biased our results (i. e. for SVR and MLPR) but refrained from standardization where possible (i. e. all the other algorithms).

[21] For the respective announcement by Fitch, please see: https://www.fitchratings.com/site/pr/10058528.

Third, even though our trained models are ready for use, the actual prediction output and prediction accuracy seem to vary to some extent among different sources of credit risk data (i. e. S&P Global, Moody's, or Fitch). In machine learning, this limitation is well-known as the model dependency on the structure of learning data and further adjustments to the data or algorithms. While we refrained from specific data adjustments and used default algorithm settings, prediction accuracy will most likely vary when using alternative measures for credit risk.

The last limitation leads us to some avenues for future research. While CSI does not (yet) seem to be systematically considered in credit risk evaluations by rating agencies, *Kölbel* et al.'s (2017) findings indicated that the impact of CSI on credit risk is already priced in by market participants. This suggestion is supported by market-based studies, which have found a bond yield spreads-increasing effect of poor ESG performance (e. g. *Goss/Roberts* 2011; *Chava* 2014). Therefore, future research could investigate whether ML-based credit risk models can be significantly improved when using an alternative, market-based measure of credit risk (e. g. CDS spreads). While we refrained from algorithm hyper-parameter tuning, future research could also test whether prediction accuracy might be further increased when using k-fold cross-validation and hyper-parameter tuning.

References

Agarwal, V./*Taffler,* R. (2008): Comparing the performance of market-based and accounting-based bankruptcy prediction models. Journal of Banking and Finance, Vol. 32(8): 1541 – 1551.

Alaka, H. A./*Oyedele,* L. O./*Owolabi,* H. A./*Kumar,* V./*Ajayi,* S. O./*Akinade,* O. O./*Bilal,* M. (2018): Systematic review of bankruptcy prediction models: Towards a framework for tool selection. Expert Systems with Applications, Vol. 94, 164 – 184.

Altman, E. I. (1968): Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. The Journal of Finance, Vol. 23(4), 589 – 609.

Altman, E. I./*Marco,* G./*Varetto,* F. (1994): Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience). Journal of Banking & Finance, Vol. 18(3), 505 – 529.

Ashbaugh-Skaife, H./*Collins,* D. W./*LaFond,* R. (2006): The effects of corporate governance on firms' credit ratings. Journal of Accounting and Economics, Vol. 42(1 – 2), 203 – 243.

Attig, N./*Ghoul,* S. E./*Guedhami,* O./*Suh,* J. (2013): Corporate Social Responsibility and Credit Ratings. Journal of Business Ethics, Vol. 117(4), 679 – 694.

Avramov, D./*Chordia,* T./*Jostova,* G./*Philipov,* A. (2009): Credit ratings and the cross-section of stock returns. Journal of Financial Markets, Vol. 12(3), 469 – 499.

Barboza, F./*Kimura,* H./*Altman,* E. (2017): Machine learning models and bankruptcy prediction. Expert Systems with Applications, Vol. 83, 405 – 417.

Basheer, I. A./*Hajmeer,* M. (2000): Artificial neural networks: Fundamentals, computing, design, and application. Journal of Microbiological Methods, Vol. 43(1), 3 – 31.

Bennell, J. A./*Crabbe,* D./*Thomas,* S./*Gwilym,* O. A. (2006): Modelling sovereign credit ratings: Neural networks versus ordered probit. Expert Systems with Applications, Vol. 30(3), 415 – 425.

Bharath, S. T./*Shumway,* T. (2008): Forecasting Default with the Merton Distance to Default Model. The Review of Financial Studies, Vol. 21(3), 1339 – 1369.

Boritz, J. E./*Kennedy,* D. B. (1995): Effectiveness of neural network types for prediction of business failure. Expert Systems With Applications, Vol. 9(4), 503 – 512.

Bose, I./*Mahapatra,* R. K. (2001): Business data mining – A machine learning perspective. Information and Management, Vol. 39(3), 211 – 225.

Breiman, L. (2001): Random forests. Machine Learning, Vol. 45(1), 5 – 32.

Breiman, L./*Friedman* J. H./*Olshen,* R. A./*Stone,* C. J. (1984): Classification and regression trees. Wadsworth International Group, Monterey, CA.

Campbell, J. Y./*Hilscher,* J./*Szilagyi,* J. (2008): In Search of Distress Risk. The Journal of Finance, Vol. 63(6), 2899 – 2939.

Chang, C.-C./*Lin,* C.-J. (2011): LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, Vol. 2(3), 1 – 27.

Chava, S. (2014): Environmental externalities and cost of capital. Management Science, Vol. 60(9), 2223 – 2247.

Chen, M.-Y. (2011): Bankruptcy prediction in firms with statistical and intelligent techniques and a comparison of evolutionary computation approaches. Computers & Mathematics with Applications, Vol. 62(12), 4514 – 4524.

Cleofas-Sánchez, L./*García,* V./*Marqués,* A. I./*Sánchez,* J. S. (2016): Financial distress prediction using the hybrid associative memory with translation. Applied Soft Computing Journal, Vol. 44, 144 – 152.

Dichev, I. D. (1998): Is the Risk of Bankruptcy a Systematic Risk? The Journal of Finance, 53(3):1131 – 1147. Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. Journal of Business and Economic Statistics, Vol. 13(3), 253 – 263.

Dorfleitner, G./*Grebler,* J./*Utz,* S. (2020): The Impact of Corporate Social and Environmental Performance on Credit Rating Prediction: North America versus Europe. Journal of Risk, forthcoming.

Drago, D./*Carnevale,* C./*Gallo,* R. (2019): Do corporate social responsibility ratings affect credit default swap spreads? Corporate Social Responsibility and Environmental Management, Vol. 26(3), 644 – 652.

Ericsson, J./*Renault,* O. (2006): Liquidity and credit risk. The Journal of Finance, Vol. 61(5), 2219 – 2250.

Fan, R.-E./*Chang,* K.-W./*Hsieh,* C.-J./*Wang,* X.-R./*Lin,* C.-J. (2008). LIBLINEAR: A Library for Large Linear Classification. The Journal of Machine Learning Research, Vol. 9, 1871 – 1874.

Géron, A. (2017): Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media, Inc., Sebastopol, CA, 1st edition.

Goss, A./*Roberts,* G. S. (2011): The impact of corporate social responsibility on the cost of bank loans. Journal of Banking and Finance, Vol. 35(7), 1794 – 1810.

Graham, J. R./*Harvey,* C. R. (2001): The theory and practice of corporate finance: Evidence from the field. Journal of Financial Economics, Vol. 60(2 – 3), 187 – 243.

Gu, S./*Kelly,* B./*Xiu,* D. (2020): Empirical Asset Pricing via Machine Learning. The Review of Financial Studies, Vol. 33(5), 2223 – 2273.

Harvey, D./*Leybourne,* S./*Newbold,* P. (1997): Testing the equality of prediction mean squared errors. International Journal of Forecasting, Vol. 13(2), 281 – 291.

Hastie, T./*Tibshirani,* R./*Friedman,* J. (2009): The Elements of Statistical Learning. Springer-Verlag, New York, 2nd edition.

Henisz, W. J./*McGlinch,* J. (2019): ESG, Material Credit Events, and Credit Risk. Journal of Applied Corporate Finance, Vol. 31(2), 105 – 117.

Hsu, F. J./*Chen,* Y.-C. (2015): Is a firms financial risk associated with corporate social responsibility? Management Decision, Vol. 53(9), 2175 – 2199.

Jiang, F./*Jiang,* Y./*Zhi,* H./*Dong,* Y./*Li,* H./*Ma,* S./*Wang,* Y./*Dong,* Q./*Shen,* H./*Wang,* Y. (2017): Artificial intelligence in healthcare: Past, present and future. Stroke and Vascular Neurology, Vol. 2(4), 230 – 243.

Jiraporn, P./*Jiraporn,* N./*Boeprasert,* A./*Chang,* K. (2014): Does corporate social responsibility (CSR) improve credit ratings? Evidence from geographic identification. Financial Management, Vol. 43(3), 505 – 531.

Jo, H./*Na,* H. (2012): Does CSR Reduce Firm Risk? Evidence from Controversial Industry Sectors. Journal of Business Ethics, Vol. 110(4), 441 – 456.

Kealhofer, S. (2003): Quantifying credit risk I: Default prediction. Financial Analysts Journal, Vol. 59(1), 30 – 44.

Khandani, A. E./*Kim,* A. J./*Lo,* A. W. (2010): Consumer credit-risk models via machine-learning algorithms. Journal of Banking and Finance, Vol. 34(11), 2767 – 2787.

Kiesel, F./*Lücke,* F. (2019): ESG in credit ratings and the impact on financial markets. Financial Markets, Institutions & Instruments, Vol. 28(3), 263 – 290.

Kingma, D. P./*Ba,* J. L. (2015): Adam: A method for stochastic optimization. In 3rd International Conference on Learning Representations, ICLR 2015.

Kisgen, D. J. (2006): Credit ratings and capital structure. Journal of Finance, Vol. 61(3), 1035 – 1072.

Kölbel, J. F./*Busch,* T./*Jancso,* L. M. (2017): How Media Coverage of Corporate Social Irresponsibility Increases Financial Risk. Strategic Management Journal, Vol. 38(11), 2266 – 2284.

Kourou, K./*Exarchos,* T. P./*Exarchos,* K. P./*Karamouzis,* M. V./*Fotiadis,* D. I. (2015): Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal, 13, 8 – 17.

Lee, D. D./*Faff,* R. W. (2009): Corporate Sustainability Performance and Idiosyncratic Risk: A Global Perspective. Financial Review, Vol. 44(2), 213 – 237.

Leland, H. E. (1994): Corporate Debt Value, Bond Covenants, and Optimal Capital Structure. The Journal of Finance, Vol. 49(4), 1213 – 1252.

McCullagh, P. (1980): Regression Models for Ordinal Data. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 42(2), 109 – 142.

Menz, K.-M. M. (2010): Corporate Social Responsibility: Is it Rewarded by the Corporate Bond Market? A Critical Note. Journal of Business Ethics, Vol. 96(1), 117 – 134.

Merton, R. C. (1973): Theory of Rational Option Pricing. The Bell Journal of Economics and Management Science, Vol. 4(1), 141 – 183.

Merton, R. C. (1974): On the Pricing of Corporate Debt: the Risk Structure of Interest Rates. The Journal of Finance, Vol. 29(2), 449 – 470.

Ohlson, J. A. (1980): Financial Ratios and the Probabilistic Prediction of Bankruptcy. Journal of Accounting Research, Vol. 18(1), 109 – 131.

Oikonomou, I./*Brooks, C./Pavelin, S.* (2014): The Effects of Corporate Social Performance on the Cost of Corporate Debt and Credit Ratings. The Financial Review, Vol. 49(1), 49 – 75.

Piramuthu, S. (2006): On preprocessing data for financial credit risk evaluation. Expert Systems with Applications, Vol. 30(3), 489 – 497.

PRI (2018): ESG in Credit Risk and Ratings Forums: Investor Survey Results. Technical report, PRI Association, London.

Rasekhschaffe, K. C./*Jones*, R. C. (2019): Machine Learning for Stock Selection. Financial Analysts Journal, Vol. 75(3), 70 – 88.

Sassen, R./*Hinze,* A.-K./*Hardeck,* I. (2016): Impact of ESG Factors on Firm Risk in Europe. Journal of Business Economics, Vol. 86(8),867 – 904.

Sebastiani, F. (2002): Machine Learning in Automated Text Categorization. ACM Computing Surveys, Vol. 34(1), 1 – 47.

Shanker, M. S./*Hu,* M. Y./*Hung,* M. S. (1996): Effect of data standardization on neural network training. Omega, Vol. 24(4), 385 – 397.

Stellner, C./*Klein,* C./*Zwergel,* B. (2015): Corporate social responsibility and Eurozone corporate bonds: The moderating role of country sustainability. Journal of Banking and Finance, 59, 538 – 549.

Strike, V. M./*Gao,* J./Bansal, P. (2006): Being good while being bad: Social responsibility and the international diversification of US firms. Journal of International Business Studies, Vol. 37(6), 850 – 862.

Sun, W./*Cui,* K. (2014): Linking corporate social responsibility to firm default risk. European Management Journal, Vol. 32(2), 275 – 287.

Svetnik, V./*Liaw,* A./*Tong,* C./*Culberson,* J. C./*Sheridan,* R. P./*Feuston,* B. P. (2003): Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. Journal of Chemical Information and Computer Sciences, Vol. 43(6), 1947 – 1958.

Vassalou, M./*Xing,* Y. (2004): Default Risk in Equity Returns. The Journal of Finance, Vol. 59(2), 831 – 868.

Waddock, S. A./*Graves*, S. B. (1997): The corporate social performance-financial performance link. Strategic Management Journal, Vol. 18(4), 303 – 319.

Wu, L./*Yang,* Y. (2014): Nonnegative Elastic Net and application in index tracking. Applied Mathematics and Computation, 227, 541 – 552.

Yeh, C. C./*Chi,* D. J./*Lin,* Y. R. (2014): Going-concern prediction using hybrid random forests and rough set approach. Information Sciences, 254, 98 – 110.

Zou, J., *Huss/*M., *Abid,* A./*Mohammadi,* P./*Torkamani,* A./*Telenti,* A. (2019): A primer on deep learning in genomics. Nature Genetics, Vol. 51(1), 12 – 18.

Appendix

Table A.1

Variable Definitions

	Type	*SD*
Credit Risk	[1; 22]	measures credit risk by S&P's long-term issuer credit ratings. The original ratings, which ranged from AAA to D, were converted into a numerical and continuous variable (1 for AAA, 2 for AA, 3 for A, 4 for BBB, etc.). Thus, the lowest possible rating (i. e. D) resulted in a value of 22.
Current RRI	[0; 100]	measures current exposure to CSI, based on a proprietary RepRisk algorithm. This paper uses quarterly observations. Higher (lower) values indicate higher (lower) CSI exposure compared to the firm's peers.
Peak RRI	[0; 100]	measures the highest level of CSI exposure over the last two years, based on a proprietary RepRisk algorithm. Higher (lower) values indicate higher (lower) CSI exposure compared to the firm's peers.
ESG Issues	[0; ∞]	is a vector containing all 28 ESG issues from RepRisk that ultimately constitute the RRI. This vector is multiplied with the news count for every issue and firm-quarter observation. For instance, if a firm experienced a forced labor incident and the news count was 8 in that respective firm-quarter, the ESG issue "Forced labor" took the value of $1 \; x \; 8 = 8$. Summary statistics can be found in Table 2. For more information on the definition of RepRisk's ESG issues, please see: https://www.reprisk.com/content/static/reprisk-esg-issues-definitions.pdf.
LIQU	[∞; 1]	measures liquidity. Calculated as working capital divided by total assets.

(continue next page)

(Table A.1 continued)

PROF	$[-\infty; \infty]$	measures profitability. Calculated as retained earnings divided by total assets.
OPEF	$[-\infty; \infty]$	measures operating efficiency. Calculated as earnings before interest and taxes (EBIT) divided by total assets.
ME	$[-\infty; \infty]$	accounts for a market effect on credit risk. Calculated by taking the natural logarithm of market value divided by long-term debt.
AT	$[0; \infty]$	measures asset turnover. Calculated as sales divided by total assets.
ASSETS	$[-1; \infty]$	measures asset growth. Calculated as the percentage change in total assets between the current and last quarter.
SALES	$[-1; \infty]$	measures sales growth. Calculated as the percentage change in sales between the current and last quarter.
P/B	$[-1; \infty]$	measures the change in the price-to-book-value ratio. Calculated as the percentage change in the price-to-book-value ratio between the current and last quarter.
LEV	$[-\infty; \infty]$	measures leverage. Calculated as long-term debt divided by stockholders equity.
AGE	$[0; \infty]$	measures firm age. Calculated as the current year minus the year in which the firm appeared in the Compustat/SRSP database for the first time.
SIZE	$[0; \infty]$	measures firm size. Calculated as the natural logarithm of total assets
INTCOV	$[-\infty; \infty]$	measures interest coverage. Calculated by taking the natural logarithm of operating income before extraordinary items divided by interest expense.
CAPINT	$[0; 1]$	measures capital intensity. Calculated as net property, plant, and equipment (PPE) divided by total assets.

Table A.2

Algorithm Hyper-Parameters

	Type	*SD*
Linear	sklearn.linear_model	*class* LinearRegression (fit_intercept=True, normalize=False, copy_X=True, n_jobs=None)
ElasticNet	sklearn.linear_model	*class* ElasticNet (alpha=1.0, l1_ratio=0.5, fit_intercept= True, normalize=False, precompute=False, max_iter=1000, copy_X=True, tol=0.0001, warm_start=False, positive= False, random_state=target_seed_global, selection='cyclic')
Ridge	sklearn.linear_model	*class* Ridge (alpha=1.0, fit_intercept=True, normalize=False, copy_X=True, max_iter=None, tol=0.001, solver='auto', random_state=target_seed_global)
OrginalRidge	mord	*class* OrdinalRidge (alpha=1.0, fit_intercept=True, normalize=False, copy_X=True, max_iter=None, tol=0.001, solver='auto')
LinearSVR	sklearn.svm	*class* SVR (kernel='linear', degree=3, gamma='scale', coef0=0.0, tol=0.001, C=1.0, epsilon=0.1, shrinking=True, cache_size=200, verbose=False, max_iter=−1)
RbfSVR	sklearn.svm	*class* SVR (kernel='rbf', degree=3, gamma='auto', coef0=0.0, tol=0.001, C=1.0, epsilon=0.1, shrinking=True, cache_size=200, verbose=False, max_iter=−1)
DTR	sklearn.tree	*class* DecisionTreeRegressor (criterion='mse', splitter='best', max depth=None, min_samples_split=2, min_samples_ leaf=1, min_weight fraction_leaf=0.0, max_features=None, random_state=target_seed_global, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity split=None, presort='deprecated', ccp_alpha=0.0)
RFR	sklearn.ensemble	*class* RandomForestRegressor (n_estimators=10, criterion='mse', max depth=None, min_samples split=2, min_samples leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf nodes=None, min_ impurity_decrease=0.0, min_impurity split=None, bootstrap=True, oob_score=False, n_jobs=None, random_ state=target_seed_global, verbose=0, warm_start=False, ccp alpha=0.0, max_samples=None)
AdaBoostR	sklearn.ensemble	*class* AdaBoostRegressor (base_estimator=None, n_estima tors=50, learning_rate=1.0, dom_state=target_seed_global)
GBR	sklearn.ensemble	*class* GradientBoostingRegressor (loss='ls', learning_ rate=0.1, n_estimators=100, subsample=1.0, min_samples_ split=2, min_samples_leaf=1, min_weight_fraction_ leaf=0.0, max_depth=3, init=None, random_state=target_ seed_global, max_features=None, alpha=0.9, verbose=0, max_leaf_nodes=None, warm_start=False)

(continue next page)

(Table A.2 continued)

	Type	*SD*
MLPR	sklearn.neural_ network	*class* MLPRegressor (hidden_layer_sizes=(100,), activation='relu', solver=àdam', alpha=0.0001, batch_size='auto', learning rate='constant', power t=0.5, max iter=1000000,learning rate init=0.001, shuffle=True, random state=target seed global, tol=0.0001, verbose=False, warm start=False, momentum=0.9, nesterovs_momentum=True, early stopping=False, validation_fraction=0.1, beta 1=0.9, beta 2=0.999, epsilon=1e−08, n_iter_no_change=10, max fun=15000)
KNNR	sklearn.neigbors	*class* KNeighborsRegressor (n neighbors=5, weights='uniform', algorithm='auto', leaf size=30, p=2, metric='minkowski, metric_params=None, n-jobs=None,**kwargs)