International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-023-00405-9

REGULAR PAPER

Bayesian learning models to measure the relative impact of ESG factors on credit ratings

Arianna Agosto[1] · **Paola Cerchiello**[1] · **Paolo Giudici**[1]

Received: 30 March 2023 / Accepted: 12 June 2023

Abstract
Artificial intelligence methods, based on machine learning models, are rapidly changing financial services, and credit lending in particular, complementing traditional bank lending with platform lending. While financial technologies improve user experience and possibly lower costs, they may increase risks and, in particular, the model risks that derive from inaccurate credit rating assessments. In this paper, we will show how to reduce such model risks, using a S.A.F.E. statistical learning model, which improves: Sustainability, taking environmental, social and governance factors into account; Accuracy, building a model which maximises predictive accuracy; Fairness, merging ESG scores from different data providers, improving their representativeness; Explainability, clarifying the relative contribution of each ESG score to predictive accuracy.

Keywords Sustainability · Explainability · ESG scores · Credit ratings · Machine learning · Bayesian models

1 Background

Artificial intelligence methods, based on machine learning applied to the data, are rapidly changing financial services, in all areas, such as lending, asset management and payment services, transforming "finance" into "financial technologies".

While financial technologies, and peer-to-peer lending in particular, improve user experience, and possibly lower costs, they may increase risks. Among them, the risk of inaccurate estimates in credit scoring, i.e. non-proper measures of creditworthiness of the borrowers ("model risk"). The occurrence of this risk may lead to important credit losses, especially when credit is given to large companies. Indeed, incentives are rather different: while in classic bank lending the costs of wrong credit rating assessments are paid by banks themselves, in peer-to-peer lending they are paid by the borrowers.

✉ Paolo Giudici
paolo.giudici@unipv.it

Arianna Agosto
arianna.agosto@unipv.it

Paola Cerchiello
paola.cerchiello@unipv.it

[1] Department of Economics and Management, University of Pavia, Pavia, Italy

These considerations suggest, given the increased economic importance of platform lending, that regulators and supervisors should carefully supervise the model risks that arise from credit ratings and their use by lending platforms.

A first important model risk concerns Sustainability, and it arises when the model is not resilient to cyber attacks or to extreme data and, in particular, when it is affected by "external" factors, represented by Environmental, Social and Governance (ESG) factors. The problem is quite challenging. First of all, it is not clear whether ESG factors do impact on credit ratings, particularly as they refer to a long term time horizon, differently from credit ratings. The most important problem is however the lack of standardisation of ESG scores. ESG scores are currently made available by various specialised companies, including rating agencies. The presence of ESG scores in the market can push companies to improve their Corporate Social Performance (CSP) or ESG behaviour [1], but it also presents possible drawbacks. Multiple ESG ratings for a given company can differ and create opaqueness in the company's actual ESG standing or greenwashing misbehaviour A recent survey by KPMG [2] showed the existence of more than 160 ESG ratings and data providers, with multiple agencies (e.g. Bloomberg, Thomson Reuters, S&P, etc.) whose ESG ratings may however differ. [3] showed little convergence between different ESG ratings. More recently, Abhayawansa and Tyagi [4] provided evidence of the low correlation between ESG ratings issued

Published online: 01 July 2023

by different providers. The lack of standardisation of ESG metrics is a problem for both investors and borrowers. From the investors' point of view, it could be challenging to understand and choose among the ESG ratings to select the best investment opportunities. Similarly, it would be difficult for borrower companies to establish financing plans in a correct way.

We believe that taking into account ESG factors is a necessary step for a sustainable finance and, for this reason, we will consider the issue of Sustainable credit scores, through the investigation of the impact of ESG scores on credit ratings, as the main focus of our paper.

A second important model risk concerns lack of predictive accuracy. Credit scoring in peer-to-peer lending has been studied in a few recent papers, that propose network models to take into account platform risk arising from the connectivity between companies. In these papers, financial network models allow to improve the predictive accuracy of the individual probability of default by considering similarities or linkages among borrowers. This becomes crucial for peer-to-peer lending platforms, in which individuals are able to directly provide small and, in most cases, unsecured loans to small and medium enterprises, without the availability of financial and behavioural information typically leveraged by banks. A network-based scoring model built upon balance sheet similarities between P2P borrowing companies was applied by Agosto et al. [5], while Ahelegbey et al. [6] improved P2P credit scoring models by clustering SMEs based on latent risk factors, deduced from financial ratios. In [5], a network is instead built upon trade flows between the companies joining the platform, proxied by input–output data at the sector level. While network models, and similarly complex machine learning models, may seem appealing, capturing nonlinearities and, thereby, improving predictive accuracy, in some cases they can be limited by their "black-box" nature, which makes it difficult to interpret the results. Although complex machine learning models may reach high predictive accuracy, their predictions are not Explainable, in the sense that they cannot be understood, and therefore oversight, by humans.

We believe that such models may be useful when they improve model accuracy in a manner that overcompensates their lack of explainability, making the further computational burden of making them explainable affordable. This may not be the case when data are of limited quality.

Indeed, following what we already discussed, a third important model risk that may arise in machine learning credit scoring is that of data quality, whose lack may lead to unfair results, as stated, for example, in the recent European Artificial Intelligence Act [7]. The problem of data quality arises in credit scoring when some necessary information is missing or contradictory. This is the case of sustainability factors, encoded in ESG measures: they are not yet standardised, with different data providers assigning a different ESG value to the same company, and with a relatively short time series available. This lack of standardisation may lead to unfair credit ratings, which creates a distorted credit allocation.

We believe that lack of data quality is a real concern that prevents from a correct understanding of the impact of ESG factors on credit ratings. However, in line with our focus, we will employ the data available so far, trying to leverage not only the disadvantages but also the advantages of inconsistent ESG databases.

A fourth important model risk is lack of explainability of the credit scores. This is a very relevant problem for many stakeholders: for investors, who cannot rationalise their investment decisions, not knowing why some companies have a higher score than others; for borrowers, who cannot improve their scores, without knowing the drivers of their values; for regulators and supervisors, which cannot evaluate the impacts of the proposed models, particularly under stress scenarios and, therefore, may not validate them. Complex machine learning models may be highly accurate, as they can capture nonlinearities and interdependences, but are typically "black-box": they assign predictive scores without explaining their determinants, in terms of the most correlated explanatory variables, as "classic" regression models do, leading to a lack of model explainability. The recent machine learning literature has proposed methods to explain black box models, by means of further processing of the predictive output: see e.g. [8–11].

We believe that Explainable AI methods are useful, but their extra computational burden is not justified when the available data are of limited quality and/or size. In this case it would be better to build a model that is, while complex, and capable to capture nonlinearities, "explainable by design", as a simple regression model.

Sustainability, Accuracy, Fairness and Explainability are desirable characteristics of a machine learning model, which should be monitored along time for high-risk applications of AI, as stated in the recent European AI Act (and in similar proposals to regulate artificial intelligence that are being developed worldwide). The importance of these four characteristics highlights the need to build appropriate statistical metrics to measure them, currently not available.

To fill the gap, we have been working in close collaboration between academics and policymakers, within the Milano Hub of the Bank of Italy, to develop a S.A.F.E. learning model that can take Environmental, Social and Governance factors into account.

The result of the collaboration, reported in the present paper, is a credit scoring model for companies that, given the available data, is: Sustainable, as it contributes to sustainability efforts in Finance, by taking into account ESG indicators in the prediction of creditworthiness; Accurate, as it indicates that ESG factors predict to some extent credit rat-

ings, even when controlling for balance sheet information; Fair, as it "compensates" different data providers into one combined score; Explainable, as based on a mixture model whose weights indicate the importance of each ESG score in determining the credit scores.

From a methodological viewpoint, the main contribution of the paper is a data-driven model that describes how ESG scores affect credit ratings, by means of a statistical learning model that is explainable by design, as the final ESG score is a linear combination of the ESG sources, with weights that are proportional to their predictive accuracies.

We remark that the aim of the paper is not to evaluate what is the effect of ESG indicators on credit ratings but, rather, whether there is such an effect, and whether different ESG indicators (or their combined score) contribute differently to this effect (even if potentially limited). This is why we focus on the Bayesian model, which can produce a weight for each indicator, that depends on its accuracy, allowing to judge the accuracy of each ESG score for credit ratings and, furthermore, providing a way to aggregate the indicators in a combined measure that we show to improve accuracy. The weights depend on the in-sample accuracy of each ESG indicator in explaining a target variable related to the company's creditworthiness, such as the credit rating or a default binary variable. In other words, the combined ESG score will be strongly impacted by good scores and less impacted by bad scores.

The methodology proposed in this paper can be usefully addressed to different stakeholders. For data scientists, it provides assessment metrics for different ESG indicators, which is proportional to their (credit rating) predictive accuracy. For investors, it can provide an aggregate ESG indicator, more robust than single indicators, that can be used in investment decisions. For borrowers, it provides a mean to evaluate long term lending perspectives, taking ESG factors into account.

To our knowledge, this is the first data-driven model based on the relationship between credit ratings and ESG scores, by means of a statistical learning model that is explainable by design, as the final ESG score is a linear combination of the ESG sources, with weights that are proportional to their predictive accuracies.

The remainder of this paper is organised as follows: Sect. 2 presents a discussion on the main focus of our paper: the relationship between ESG factors and credit ratings; Sect. 3 introduces the proposed modelling approach; Sect. 4 presents an application of the methodology to a sample of European companies and, finally, Sect. 5 concludes.

2 ESG scores and credit rating

Corporate Social Performance (CSP) is aimed at evaluating the degree to which companies are sustainable, that is, how they perform their business activities in relation to the external stakeholders and taking into account the economic, environmental, social, and time factors [12, 13]. Environmental, Social and Governance (ESG) factors are often taken as a proxy for the sustainable behaviour of companies.

Environmental factors (E) relate to the impact on the environment deriving from the production of goods or services and include carbon emissions, preservation of the natural environment, biodiversity protection, and waste and water management [14–16]. A company that operates with less harm to the environment might reduce the probability of future scandals, legal actions, losses related to legal claims etc. and benefit from a better reputation and lower risks [17].

Social factors (S) refer to the impacts of companies on society, including issues of employee satisfaction, diversity, inequality, gender gap, protection of young and children, investment in human capital and communities, and human rights [14, 18].

Governance factors (F) measure the quality of corporate governance. Shortcomings in governance have been in the past the cause of major scandals and crises, such as the Enron crisis in the USA, Volkswagen in Germany, Parmalat in Italy, and the banking crisis of 2007–2008 [19, 20]. Improved governance settings can contribute to a more sustainable and balanced firms' growth, therefore contributing to a more sustainable economic development [21, 22].

The above factors are the basis for investment decisions and drive the choice of investors in terms of which companies to finance through equity or debt. To improve the interpretability of ESG, specialised companies (including rating agencies) have started to provide measures and proxies for ESG behaviour, publishing ESG ratings or ESG scores that convey the level of sustainability of companies and the degree of accountability of these companies on ESG aspects [23, 24].

Each rating provider collects information from different sources (company reports, news, stock exchange information, etc.) and applies proprietary methodologies to combine information and produce a summary measure of ESG behaviour. Different methodologies yield different measures, that often produce divergent results [3, 4, 25, 26], and this induces lack of standardisation.

The importance of ESG metrics is bound to grow in the future, with ESG ratings likely to affect investors' decisions, firms' ability to finance their investments and pursue a sustainable business model. It follows that understanding whether and how ESG ratings affect creditworthiness is a very important managerial and policy challenge.

To our knowledge, this is the first work to: (1) analyse the relationship between ESG scores and credit ratings through a data-driven model that predicts the company's credit rating class based on the ESG rating; (2) use the ESG scores assigned by different providers to create a combined metric

where each ESG score is weighted based on its predictive accuracy.

In the next section, we describe our proposed methodology, which is applied to real data in Sect. 4. Section 5 concludes the paper with a final discussion.

3 Proposal

In this section, we introduce our proposed Bayesian learning model, which leads to an indicator for the ESG performance of listed companies that integrates the ESG scores assigned by different providers. The indicator is obtained attributing to each available ESG score a weight that is a function of the likelihood of the observed counts of companies belonging to the different credit rating classes, under the alternative partitions generated by the ESG scores. The likelihood weights express in-sample predictive performance and are obtained through the application of Bayes' theorem.

Our model is based on the assumption that there is an effect of ESG scores and credit rating. However, our aim is not to build a model that employs ESG scores to improve credit rating predictive accuracy but, rather, to investigate the relative importance of each ESG data score. To this end, we extend to the ESG context the methodology proposed by Cerchiello and Giudici [27], who considered the case of estimating a company's probability of default using a set of explanatory financial variables. Our proposal relies on the modelling approach by Cerchiello and Giudici [27], but applies it to study the relationship between ESG indicators and credit risk, extending what proposed by Agosto et al. [28] to multiple ESG scores and to a binomial response variable.

In [27], based on the mixture of Dirichlet processes model proposed by Giudici et al. [29], it is assumed that the partition g_k generated by the k-th among K covariates is made up of $j = 1, ..., J_k$ levels and that the probability of default of company i, $Prob(Y_i = 1)$, where Y_i is a binary variable equal to 1 if company i defaults, 0 otherwise) is constant within the same j level of the covariate and equal to θ_j.

Here, we extend their work assuming that the partition g_k is generated by the values of the ESG scores assigned by the k-th data provider, and that Y_i is a binary variable which indicates whether a company rating is speculative (equal to 1) or investment grade (equal to 0). These assumptions do not imply a loss of generality: different partitions can be assumed, for example, corresponding to a combination of ESG scores, and a different binarisation of the rating can be considered to obtain Y.

Letting Y_i be a Bernoulli(θ_j) variable and the θ_j's Beta random variables with parameters α and β, which implies that, a priori, $E(\theta_j) = \frac{\alpha}{\alpha+\beta}$, the marginal likelihood contribution of level j can be obtainedas:

$$\begin{aligned} p(y\|j) &= \int_0^1 p(y\|\theta_j)p(\theta_j)d\theta_j \\ &= \int_0^1 \theta_j^{d_j}(1-\theta_j)^{n_j-d_j}\frac{1}{B(\alpha,\beta)}\theta_j^{\alpha-1}(1-\theta_j)^{\beta-1}d\theta_j \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+d_j)\Gamma(\beta+n_j-d_j)}{\Gamma(\alpha+\beta+n_j)} \end{aligned} \quad (1)$$

where $p(\theta_j)$ is the prior distribution of θ_j, d_j is the number of defaulted companies and n_j is the total number of companies sharing level j of the k covariate. Furthermore, B is the Beta function, defined by:

$$B(z_1, z_2) = \frac{\Gamma(Z_1)\Gamma(Z_2)}{\Gamma(Z_1+Z_2)},$$

where for each positive integer n:

$$\Gamma(n) = (n-1)!$$

Under the assumption that the θ_j's are independent random variables, the marginal likelihood of the partition g_k is:

$$p(y\|g_k) = \prod_{j=1}^{J_k} p(y\|j), \quad (2)$$

which determines the posterior probability of the partition:

$$p(g_k\|Y) \propto p(y\|g_k)p(g_k), \quad (3)$$

where $p(g_k)$ can be set a priori, for example, according to the uniform distribution: $p(g_k) \propto 1/M$ where M is a constant.

The expected probability of default of company i, conditional on the available set of covariates X, can then be obtained as follows:

$$E(\theta_i\|X, Y) = \sum_{k=1}^{K} E(\theta_j\|g_k, Y)p(g_k\|Y), \quad (4)$$

with $E(\theta_j\|g_k, Y) = \frac{\alpha+d_j}{\alpha+\beta+n_j}$, in which the posterior probability $p(g_k\|Y)$ acts as k-th covariate weight in determining the expected probability of the default event.

Equations (3) and (4) summarise the essence of our proposed machine learning model. It is a Sustainable model, as it allows to measure the impact of ESG factors on credit ratings; it is a Fair model, as it averages the contribution of different ESG providers, compensating their differences, due to different objectives; it is an Explainable model, as it is a linear combination of weights with posterior probabilities, which, although calculated in a nonlinear way, have a clearly interpretable meaning. In the next section, we will verify, for our available data, whether the model is also accurate, that

is, whether ESG factors have a predictive relevance for credit ratings, and what are the relative weights of each ESG factor in the model.

For the sake of comparison and completeness, we consider as benchmark model XGBOOST [30], for its well-known capability of modelling nonlinearity in a very efficient way, without imposing any distributional assumption. Moreover, together with deep neural network, they represent the state of the art, as far as the overall accuracy is concerned. Since deep neural network cannot be profitably employed in the current exercise, given the dimensions of the dataset, we resort to XGBOOST. Indeed, the latter is an ensemble model which works over the idea of combining several weak classifiers to create a strong one characterised by extremely good performance thanks to a regularised gradient boosting framework. As further term of comparison, we also consider bagging and random forest models which belong to the same ensemble approach family but exploiting different strategies [31].

4 Application

4.1 Data

In this section, we apply our proposed methodology to a sample of 1382 European companies for which we retrieve:

- the MSCI ESG Score: a continuous variable ranging from 0 (lowest sustainability) to 10 (highest sustainability);
- the Refinitiv ESG Score: a continuous variable ranging from 0 to 100. As for the MSCI ESG score, higher values indicate better sustainability profiles;
- the Standard and Poor's (S&P) Global ESG Rank: a discrete variable defined as the total sustainability percentile rank, ranging from 0 (lowest sustainability) to 100 (highest sustainability);
- the risk class assigned to the company based on the Bloomberg Issuer Default Risk model generated probability of default over the next one year: an ordinal variable whose categories in the sample range from IG1 (highest credit worthiness) to D4 (lowest credit worthiness). Specifically, classes from IG1 to IG10 identify Investment Grade bond issuers, while classes from HY to H6 and from D1 to D4 identify High Yield and Distressed bond issuers, respectively. Starting from the rating class information, we define a binary variable which is equal to 1 if the company belongs to a speculative (high-yield or distressed) class, 0 otherwise. This will be our target variable in the application of the Bayesian model presented in Sect. 3;

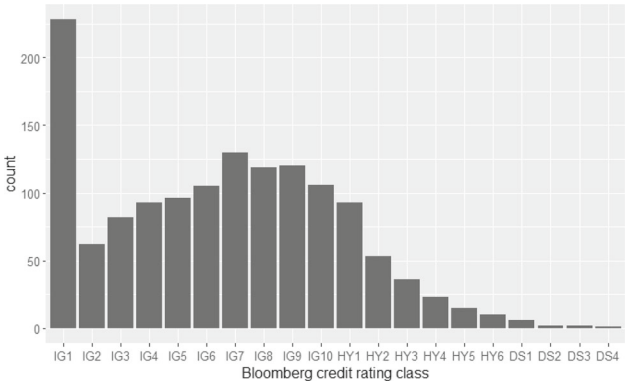

Fig. 1 Distribution of the analysed companies by Bloomberg credit rating. Source: own elaborations based on Bloomberg data

- a set of 13 financial ratios[1] which should reflect company profitability, growth and liquidity, together with the value of market capitalisation, which serves as a dimensional indicator.

To allow the comparability of the scores, the MSCI ESG score has been rescaled in the 0–100 range.

Data are the last available as of August 3, 2022, and is retrieved from various sources: MSCI ESG Research (for the MSCI ESG scores), Refinitiv LSEG business (for the Refinitiv ESG scores), Bloomberg (for the S&P Global ESG rank and the credit ratings). All Data is pre-processed so that no missing values are present in our sample. In our setting, among the European companies having an ESG rating, we only select those (1382) for which all three ESG scores are available at the considered date. The data have got a cross-sectional structure, all being referred to a single date, the 3rd of August, 2022.

The distribution of sample companies among the credit rating classes is shown in Fig. 1.

Figure 1 shows that, for the considered companies, the distribution of ratings is quite skewed to the right, and that there is a large group of companies with very high ratings (IG1). Both aspects will make it more challenging to attain a good level of predictive accuracy.

As it can be seen from Fig. 2,[2] the distribution of the three ESG scores in the analysed sample is instead left-skewed, meaning that a few number of companies have a much worst ESG evaluation than the mean one.

[1] Our balance sheet dataset includes the following indicators: Return on Equity, Return on Asset, Return on Investment, Short Term Debt on 1-year Growth, Total Debt on 1-year Growth, Free Cash Flow on 1-year Growth, Free Cash Flow on 5-year Growth, EBITDA to Interest Expenses, Long Term Debt to Total Equity, Quick Ratio, Capital Expenditure Ratio, Financial Leverage, Asset Turnover.

[2] Reproduced by permission of MSCI ESG Research LLC, copyright 2023 MSCI ESG Research LLC, All rights reserved.

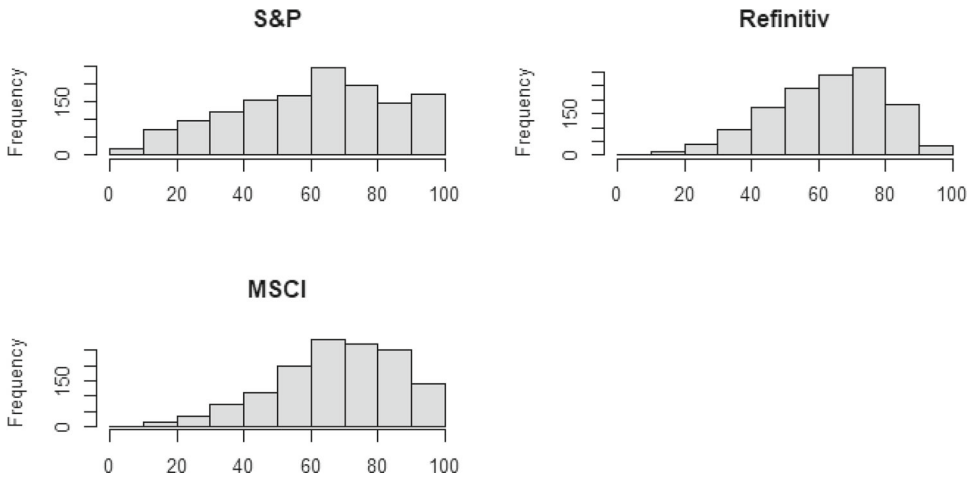

Fig. 2 Distribution of the analysed companies by ESG score. To allow the comparability of the scores, the MSCI ESG score has been rescaled in the 0–100 range. Source: own elaborations based on MSCI ESG Research, Refinitiv (LSEG business), S&P Global and Bloomberg data

Table 1 Pearson correlation between the ESG scores

	S&P	Refinitiv	MSCI
S&P	1	0.692	0.373
Refinitiv	0.692	1	0.383
MSCI	0.372	0.383	1

Source: own elaborations based on MSCI ESG Research, Refinitiv (LSEG business), S&P Global and Bloomberg data

Table 2 Spearman correlation between the ESG scores

	S&P	Refinitiv	MSCI
S&P	1	0.689	0.356
Refinitiv	0.689	1	0.376
MSCI	0.356	0.376	1

Source: own elaborations based on MSCI ESG Research, Refinitiv (LSEG business), S&P Global and Bloomberg data

Table 3 Kendall's tau correlation between the ESG scores

	S&P	Refinitiv	MSCI
S&P	1	0.501	0.246
Refinitiv	0.501	1	0.259
MSCI	0.246	0.259	1

Source: own elaborations based on MSCI ESG Research, Refinitiv (LSEG business), S&P Global and Bloomberg data

Table 4 Somers D correlation between the ESG scores

	S&P	Refinitiv	MSCI
S&P	1	0.498	0.247
Refinitiv	0.498	1	0.262
MSCI	0.247	0.262	1

Source: own elaborations based on MSCI ESG Research, Refinitiv (LSEG business), S&P Global and Bloomberg data

Concerning the concordance between the ESG scores, it can then be noticed from Tables 1, 2, 3 and 4 that correlation between the Refinitiv and the S&P ESG scores is relatively high according to the Pearson and Spearman measures, but decreases to nearly 50% when moving to rank-based concordance measures. Correlation between the MSCI ESG scores and the other two indicators is instead low, never reaching 40%. This increases the interest in reaching a sustainability metric that combines alternative ESG scores based on their capability to order the observed companies by their creditworthiness.

4.2 Results

4.2.1 In-sample analysis

The first step in our empirical analysis consists of the calculation of the posterior probability-based weights according to the methodology described in Sect. 3. Having no *a priori* reasons to assign different weights to the scores, we set the M constant in (3) equal to 3, which means that the three scores are *a priori* equally weighted. The α parameter is set equal to the ratio between the number of investment grade companies and the total number of companies in the sample, so that $\beta = 1 - \alpha$ is set to be the proportion of speculative grade companies in the sample.

The posterior weights associated with the scores are estimated on a random training sample of 829 companies (60% of the available observations) and are shown in the second column of Table 5. The third column of Table 5 reports instead the weights obtained by applying the same methodology to the residuals of stepwise linear regression models where the dependent variable is a given ESG score (MSCI, Refinitiv or S&P) and the regressors are the company's balance sheet variables and market capitalisation. This allows indeed to consider the extent to which the financial information—on which both the ESG scores and the credit ratings are supposed to be related to—influences the capability of ESG scores to

Table 5 Weights derived from the posterior probabilities associated to the ESG scores, before and after controlling for financial ratios

ESG score	Before control	After control
MSCI	0.36	0.34
Refinitiv	0.37	0.32
S&P	0.27	0.34

Source: own elaborations based on MSCI ESG Research, Refinitiv (LSEG business), S&P Global and Bloomberg data

predict the credit ratings. Coefficient estimates for the estimated linear regression models are shown in Tables 7, 8 and 9.

Table 5 shows that model weights are somewhat different, before controlling for financial ratios. But also that such difference nearly disappears, once controlling for the same ratios. This may be the effect of different attention given by the providers to the financial ratios. Once they are taken into account, however, the ESG scores have a similar importance, in determining credit worthiness. This shows that our proposed model is able to improve fairness, reducing inconsistencies among the data providers. And, by taking an equally weighted average of the ESG scores, it does not generate any bias deriving from using one rather than the other.

We also remark that the weights in Table 5 are the main output of our proposed model: a set of weights which is easy to interpret and implement in the monitoring of credit risk.

In other words, with the in-sample analysis we have shown that our proposed model is fair and explainable.

4.2.2 Out-of-sample analysis and robustness

We now provide a predictive analysis where the probability that a company belongs to a certain rating class—conditional on the ESG score—is estimated based on the methodology described in Sect. 3.

Specifically, we use the weights associated to the ESG scores estimated on the training sample (see Sect. 4.2.1) to predict the credit rating in the validation sample (40% of the available observations). According to the proposed merged scoring methodology, the weights are then used to determine, for each company, and for each provider domain (Refinitiv, Standard and Poor's, MSCI) the probability associated to each of the two considered rating categories: Investment Grade or Speculative (High Yield or Distressed) class.

Figure 3 shows the posterior probabilities associated to the different classes of the ESG score distribution, for each of the three scores considered. These probabilities are used to determine the probabilities assigned by the merged score. Indeed, for each company, the probability of belonging to a speculative rating class is calculated as the weighted mean of the probabilities assigned by the three scores, using the Bayesian likelihood-based weights.

Figure 4 shows the ROC curves of the credit rating prediction based in the ESG scores, obtained by applying the Bayesian model.

From Fig. 4 note that there is no absolute dominance of one specific ROC curve. The relationship depends strictly on the quantiles of reference. More in detail, if we compare the related AUROC measures, the two leading models are the Merged score and the MSCI ones. The Merged score model is, furthermore, more robust (more sustainable in the statistical sense) as it does better in modelling the tails of the distribution, where the more extreme financial profiles lie: companies that are either very bad or very good.

The results are confirmed after controlling for the financial ratios. We can conclude from Fig. 4 that the merged model leads to predictions that are better than those of the single ESG scores on the tails and, in particular, for high cut-off levels. This means that the merged model is resilient to extreme values (upper tail): its performance does not decrease when extreme values are considered, as indìvidual ESG models do. The proposed model is thus a sustainable credit rating model, as it shows that ESG factors are important to predict credit ratings, even when financial variables are inserted into the model. The proposed model also improves predictive accuracy, with respect to what the separate ESG scores would do.

A question that may arise, especially for the sake of comparison, is whether a different (non-Bayesian) machine learning model would improve predictive accuracy, although being not explainable. If it were so, computationally expensive explainable AI methods, such as Shapley values [8, 10] could be applied as an "add-on" to the model.

To this end, we additionally fit a competing model, which is typical expression of machine learning approaches.

As already introduced, we fit a XGBOOST by means of the package 'xgboost' of R software and by setting three tuning parameters as follows: a parameter d, which determines the depth of each boosted tree; a learning parameter η, which determines the updating rate, and a parameter B, which determines the number of boosted trees. We select the values of such parameters after a fine tuning exercise and specifically we take: $d = 1$ or 2; $\eta = 0.001$; $B = 5000$. Indeed the parameter d controls for the complexity/size of the trees in terms of considered variables and depth levels. Given the limited number of variables, we consider very small values for d: 1 and 2. The features employed are the three ESG scores, exactly as for the Bayesian model and we estimate XGBOOST on a 60% training set and we evaluate it on the remaining 40% test set (same strategy employed for the Bayesian model).

We ended up with a boosting model whose predictive performance is reported in Fig. 5. For the sake of complete-

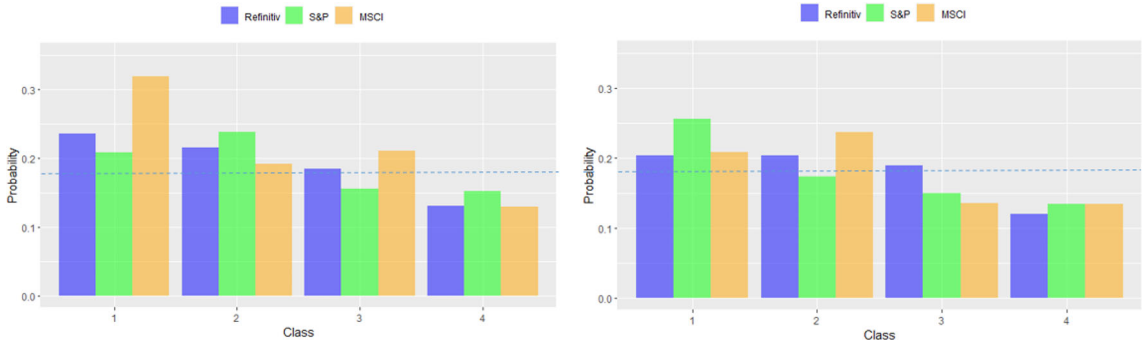

Fig. 3 Estimated probability of belonging to a High Yield or Distressed credit rating class by ESG score class, before (left) and after (right) controlling for financial and dimensional indicators. The dashed line indicates the ratio of speculative-grade rated companies in the sample. Source: own elaborations based on MSCI ESG Research, Refinitiv (LSEG business), S&P Global and Bloomberg data

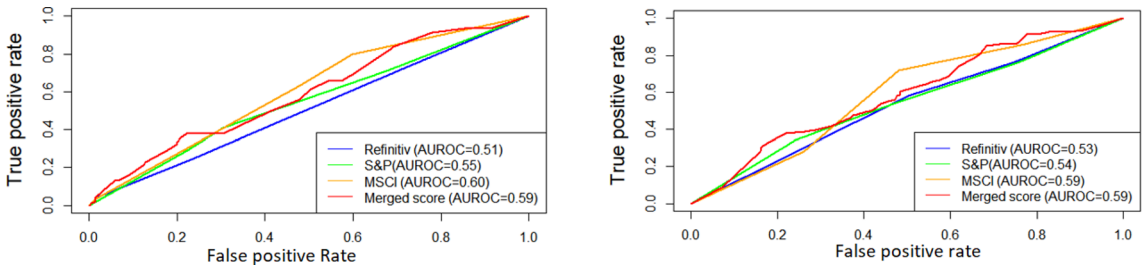

Fig. 4 ROC curve of credit rating predictions based on ESG scores, before (left) and after (right) controlling for financial and dimensional indicators, obtained through application of the proposed Bayesian model. Source: own elaborations based on MSCI ESG Research, Refinitiv (LSEG business), S&P Global and Bloomberg data

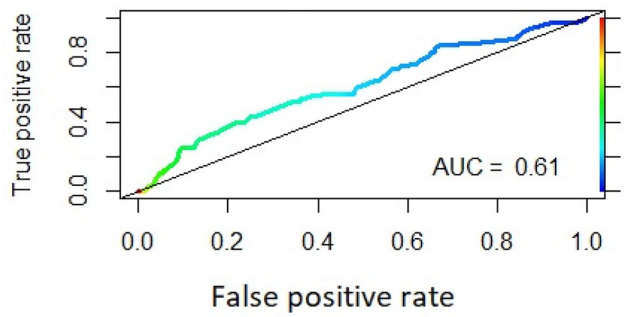

Fig. 5 ROC curve of credit rating predictions based on ESG scores obtained through the XGBOOST algorithm. Source: own elaborations based on MSCI ESG Research, Refinitiv (LSEG business), S&P Global and Bloomberg data

ness, we have also included two further classification models, namely Bagging and Random forest, which are ensemble methods still based on classification trees.

In Table 6, we report a full comparison of the different approaches.

Table 6 reports the value of AUROC (area under the ROC curve) and AUPRC (area under the precision and recall curve) for the individual ESG scores, the merged ESG score, the XGBBOST, the bagging and the random forest. AUROC accounts for the overall accuracy for each and every possible threshold, AUPRC similarly considers the areas under the precision and recall curves regardless the threshold. Such strategy allows us to produce a robust quality assessment of the competing models, without imposing any subjective assumption. From Table 6, we infer that either AUROC or AUPRC are very close to each other when considering the Bayesian model and the XGBOOST. Indeed, the slight improved accuracy of XGBOOST is limited and it is not statistically significant.

Indeed, the proposed Bayesian model does not offer an exceptional performance, especially because the effect of ESG factors on credit ratings is probably limited, but it has a clear and unavoidable advantage: it is explainable by design and it offers a system of weights that can be used in further analysis. On the other hand, the XGBOOST model, which is not explainable by design, does not lead to a gain in predictive accuracy that can justify the use of a computationally expensive AI method, such as Shapley values. The same applies to Bagging and Random Forest which show even worse performances than XGBOOST.

We remark that both XGBOOST or bagging/random forest can be made explainable (in a qualitative sense) using a variable importance plot. Unfortunately, the variable impor-

Table 6 AUROC and AUPRC of credit rating predictions based on ESG scores, obtained through application of the proposed Bayesian model and the XGBOOST algorithm

ESG score model	AUROC	AUPRC
MSCI	0.60	0.22
Refinitiv	0.51	0.18
S&P	0.55	0.19
Merged Bayesian score	**0.59**	**0.24**
XGBOOST	**0.61**	**0.25**
Bagging	0.54	0.20
Random forest	0.55	0.21

Source: own elaborations based on MSCI ESG Research, Refinitiv (LSEG business), S&P Global and Bloomberg data
Bold values refer to the performance of the two compared approaches, i.e. our merged Bayesian score vs XGBOOST

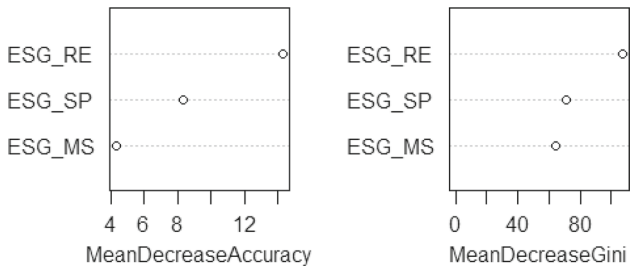

Fig. 6 Variable Importance Plot obtained upon the XGBOOST algorithm. Source: own elaborations

tance plot is not fully agnostic: we cannot use it for the Bayesian model, for example, and, thus, make comparisons. In this regard, we report in Fig. 6 the variable importance plots obtained upon the XGBOOST algorithm. Two measures are used for the ranking of the used variables: mean decrease accuracy and mean decrease Gini. Both agree on the ranking in the importance of the variables: first ESG from Refinitiv, second ESG from S&P, third ESG from MSCI. The results confirm what obtained from the Bayesian model, that is the relevance of ESG scores produced by Refinitiv. As a second important variable, the variable importance plot selects ESG scores from S&P conditionally on Refinitiv, differently from the Bayesian model which proceeds with a simultaneous selection. We remark that the weights and the importance attributed to the different ESG scores has merely a descriptive purpose within the framework of our model and it does not imply any evaluation of their inner quality.

Although a computationally expensive explainable AI method may not be justified in our context, we have tried to interpret the predictions obtained from XGBOOST with a graphical method, comparing the plots of the estimated probabilities by the three ESG scores, similar to what obtained in Fig. 3 for the Bayesian model, reported in Fig. 7.

Comparing Figs. 7 with 3 (left), note that the behaviours of the estimated probabilities are rather similar. In both cases,

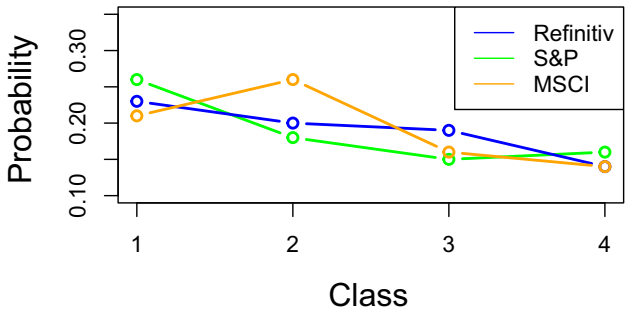

Fig. 7 Estimated probability of belonging to a high yield or distressed credit rating class by ESG score class, with the XGBOOST model. Source: own elaborations based on MSCI ESG Research, Refinitiv (LSEG business), S&P Global and Bloomberg data

there is an overall negative dependence between the ESG score class and the probability of default; moreover, the range of variation of Refinitiv is the smallest. This implies a higher weight in the Bayesian model for Refinitiv than for more discriminant scores, such as S&P. Figure 7 shows that the probabilities estimated by the XGBOOST have generally a lower variability, with respect to those from the Bayesian model. This is in line with the smoothing effect carried out by the (nonlinear) XGBOOST model.

5 Conclusions

In the paper, we have shown how credit worthiness could be measured by means of a S.A.F.E. machine learning model which reduces model risks, in line with the emerging regulations of artificial intelligence, which aim to measure the risks of artificial intelligence to promote its usage.

The model is Sustainable, as credit ratings can take Environmental, Social and Governance factors into account. The model is Accurate, as it indeed shows that ESG scores have an effect, although limited, in the prediction of credit ratings. The model is Fair, as it can level out differences between different ESG data providers, taking an averaged score. The model is Explainable as it can be easily interpreted by means of a set of normalised weights assigned to the different ESG providers, which are function of their relative predictive accuracy.

We believe that this paper is the first of this kind, and it may generate debate and impact, in the AI and in the financial community altogether.

This, in particular, because it can improve ESG standardisation, providing a solution to the problem of multiple ESG ratings. The increased attention to sustainability issues has yielded the proliferation of rating agencies and ESG scores, with multiple ESG scores on the market that are often divergent and provide different types of information. In the paper we show how to combine different ESG scores into a single

ESG one that combines the information given by different providers of ESG scores. A combined score can bring better evidence on whether ESG factors can be predictors of credit rating classes.

Our findings have many implications for the application of statistical learning and artificial intelligence methods in the financial sector. What presented can be useful for investors in financial markets, who can exploit the information provided by different ESG scores in a comprehensive setting, reducing information asymmetries on ESG company performance. It can also be useful for lenders in credit markets, as they can make a better informed use of ESG factors in determining credit worthiness, to the benefit of the best-performing companies in terms of sustainable behaviour. It can be of interest also for insurance companies, helping to assess pricing of climate and ESG related events.

Our research is also of interest for regulators and supervisors in the financial sector, as it provides a standardised metric to measure the impact of different ESG scores, along with a combined score, thereby improving the assessment of the sustainability of the company which receives ESG ratings. And, finally, it is important for ESG data providers, as they can receive feedback on the relative quality of their metrics, and, possibly, improve them.

We also remark that the scope of this paper is to provide indications to financial institutions on the relative quality of different ESG providers (in terms of their predictive accuracy).

Future research could replicate how our results, obtained on the available data, clean of missing values, can be extended, to a different and possibly larger database.

Future research should also extend our work to cover companies for which ESG scores are missing for some providers. Our approach can be easily generalised to this context assigning companies with missing scores to a distinct new category that contains all companies with missing information.

Future research should also concern the implementation of the proposed methodology to other regulated industries, such as the health care and the automotive sectors, and, possibly, to other high-risk artificial intelligence applications.

Author Contributions While Paolo Giudici supervised the work and wrote Sects. 1 and 5 of the paper, Arianna Agosto and Paola Cerchiello developed the methodology and the data analysis and wrote Sects. 2 (PC), 3 and 4 (AA).

Funding Open access funding provided by Università degli Studi di Pavia within the CRUI-CARE Agreement. The authors acknowledge the European HORIZON 2020 PERISCOPE Project (Contract Number 101016233) and the Italian MUR PRIN Project FIN4GREEN for financial support. They also acknowledge continuous support from the Milano Hub of the Bank of Italy and, in particular, Juri Marcucci for the coordination activity; Marco Fanari, Johnny Di Giampaolo and Enrico Foscolo for useful discussion and the provision of data.

Code availability Code is available upon reasonable request.

Declarations

Conflict of interest The authors declare they have no conflicts of interest.

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Appendix A

The following tables report the coefficient estimates[3] for the linear regression models of the ESG scores on balance sheet ratios and dimensional indicators (Tables 7, 8, 9).

Table 7 Results of stepwise linear regression model of the MSCI ESG score on balance sheet ratios and dimensional indicators

Variable	Estimate	*P*-value
(Intercept)	68.9	< 2e−16***
EBITDA to revenue	6.55e−03	0.029**
Free cash flow to total equity	−2.40e−03	0.007***
Quick ratio	−1.21	0.002***
Market capitalisation	9.22e−06	0.041**

Source: Own elaborations based on MSCI ESG Research and Bloomberg data

[3] In all tables, ***, **, * denote statistical significance at the 1%, 5% and 10%, respectively.

Table 8 Results of stepwise linear regression model of the Refinitiv ESG score on balance sheet ratios and dimensional indicators

Variable	Estimate	*P*-value
(Intercept)	63.8	<2e−16***
EBITDA to revenue	7.87e−03	0.003***
Short term debt on 1-year growth	−1.33e−03	0.108*
EBITDA on interest expenses	−7.13e−04	0.138
Quick ratio	−2.182	5.94 –10***
Financial leverage	0.148	0.009***
Market capitalisation	1.83e−05	6.66e−06***

Source: Own elaborations based on Refinitiv (LSEG business) and Bloomberg data

Table 9 Results of stepwise linear regression model of the S&P ESG score on balance sheet ratios and dimensional indicators

Variable	Estimate	*P*-value
(Intercept)	64.6	<2e−16***
EBITDA to revenue	5.83e−03	0.121
Long term debt to total equity	−1.60e−03	0.081*
Quick ratio	−3.18	1.59e−10***
Financial leverage	0.210	0.014**
Asset turnover	−3.382	0.003*
Market capitalisation	2.30e−05	5.52e−05***

Source: Own elaborations based on S&P Global and Bloomberg data

References

1. Zeng, G., Xu, Y.: Sustainable development and the rating effects: a strategic categorization approach. Corp. Soc. Responsib. Environ. Manag. **26**(6), 1554–1564 (2019)
2. KPMG: Sustainable investing: Fast-forwarding its evolution. Technical report (2020)
3. Dorfleitner, G., Halbritter, G., Nguyen, M.: Measuring the level and risk of corporate responsibility-an empirical comparison of different ESG rating approaches. J. Asset Manag. **16**(7), 450–466 (2015)
4. Abhayawansa, S., Tyagi, S.: Sustainable investing: the black box of environmental, social, and governance (ESG) ratings. J. Wealth Manag. **24**(1), 49–54 (2021)
5. Agosto, A., Giudici, P., Leach, T.: Spatial regression models to improve p2p credit risk management. Front. Artif. Intell. **2**, 6 (2019)
6. Ahelegbey, D.F., Giudici, P., Hadji-Misheva, B.: Latent factor models for credit scoring in p2p systems. Physica A **522**, 112–121 (2019)
7. European Commission (2022). https://artificialintelligenceact.eu/wp-content/uploads/2022/06/AIA-FRA-Consolidated-Version-15-June.pdf. Accessed 26 Dec 2023
8. Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Guyon, I., Von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017)
9. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
10. Bussmann, N., Giudici, P., Marinelli, D., Papenbrock, J.: Explainable machine learning in credit risk management. Comput. Econ. **57**(1), 203–216 (2021)
11. Giudici, P., Raffinetti, E.: Shapley Lorenz explainable artificial intelligence. Expert Syst. Appl. **167**, 114104 (2021)
12. Lozano, R.: Are companies planning their organisational changes for corporate sustainability? An analysis of three case studies on resistance to change and their strategies to overcome it. Corp. Soc. Responsib. Environ. Manag. **20**(5), 275–295 (2013)
13. Muñoz-Torres, M.J., Fernández-Izquierdo, M.Á., Rivera-Lirio, J.M., Escrig-Olmedo, E.: Can environmental, social, and governance rating agencies favor business models that promote a more sustainable development? Corp. Soc. Responsib. Environ. Manag. **26**(2), 439–452 (2019)
14. European commission: overview of sustainable finance. https://finance.ec.europa.eu/sustainable-finance/overview-sustainable-finance_en
15. Financial times: definition of ESG. http://lexicon.ft.com/term?term=esg
16. Robeco: Sustainability investing glossary: ESG definition. https://www.robeco.com/hk/en/key-strengths/sustainability-investing/glossary/esg-definition.html (n.d.)
17. Fafaliou, I., Giaka, M., Konstantios, D., Polemis, M.: Firms' ESG reputational risk and market longevity: a firm-level analysis for the united states. J. Bus. Res. **149**, 161–177 (2022)
18. Van Duuren, E., Plantinga, A., Scholtens, B.: ESG integration and the investment management process: fundamental investing reinvented. J. Bus. Ethics **138**(3), 525–533 (2016)
19. Shin, S., Lee, J., Bansal, P.: From a shareholder to stakeholder orientation: evidence from the analyses of CEO dismissal in large us firms. Strateg. Manag. J. **43**(7), 1233–1257 (2022)
20. Soltani, B.: The anatomy of corporate fraud: a comparative analysis of high profile American and European corporate scandals. J. Bus. Ethics **120**(2), 251–274 (2014)
21. Adams, R.B., Mehran, H.: Bank board structure and performance: evidence for large bank holding companies. J. Financ. Intermed. **21**(2), 243–267 (2012)
22. Esteban-Sanchez, P., de la Cuesta-Gonzalez, M., Paredes-Gazquez, J.D.: Corporate social performance and its relation with corporate financial performance: international evidence in the banking industry. J. Clean. Prod. **162**, 1102–1110 (2017)
23. Scalet, S., Kelly, T.F.: CSR rating agencies: What is their global impact? J. Bus. Ethics **94**(1), 69–88 (2010)
24. Avetisyan, E., Ferrary, M.: Dynamics of stakeholders' implications in the institutionalization of the CSR field in France and in the united states. J. Bus. Ethics **115**(1), 115–133 (2013)
25. Dimson, E., Marsh, P., Staunton, M.: Divergent ESG ratings. J. Portfo. Manag. **47**(1), 75–87 (2020)
26. Billio, M., Costola, M., Hristova, I., Latino, C., Pelizzon, L.: Inside the ESG ratings: (dis) agreement and performance. Corp. Soc. Responsib. Environ. Manag. **28**(5), 1426–1445 (2021)
27. Cerchiello, P., Giudici, P.: Bayeslan credit ratings. Commun. Stat.-Theory Methods **43**(4), 867–878 (2014)
28. Agosto, A., Giudici, P., Tanda, A.: How to combine ESG scores? A proposal based on credit rating prediction. Corp. Soc. Responsib. Environ. Manag. (2023). https://doi.org/10.1002/csr.2548

29. Giudici, P., Mezzetti, M., Muliere, P.: Mixtures of products of Dirichlet processes for variable selection in survival analysis. J. Stat. Plan. Inference **111**(1–2), 101–115 (2003)
30. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., et al.: Xgboost: extreme gradient boosting. R Pack. Vers. 0.4-2 **1**(4), 1–4 (2015)
31. Altman, N., Krzywinski, M.: Ensemble methods: bagging and random forests. Nat. Methods **14**, 933–934 (2017)