# ESG investments: filtering versus machine learning approaches*

Carmine De Franco,† Christophe Geissler,‡ Vincent Margot§ and Bruno Monnier**

## Abstract

We designed a machine learning algorithm that identifies patterns between environmental, social and governance (ESG) profiles and financial performance for companies in a large investment universe. The goal of the algorithm, which falls in the category of supervised machine learning, is to predict the (conditional) excess return of each company over the benchmark, given the specific values taken by some of its ESG indicators (the features). In other words, the algorithm identifies regions in the high-dimensional space of ESG features that are statistically related to financial outperformance or underperformance. The final aggregated predictions are transformed into scores, which allow us to design simple strategies that screen the investment universe for stocks with positive scores. By linking ESG features with financial performance in a non-linear way, our strategy is shown to be an efficient stock picking tool, outperforming classic strategies that screen stocks according to their ESG ratings, such as the popular best-in-class approach. Our paper introduces new ideas into the growing field of financial literature investigating the links between ESG behaviour and the economy. We show, indeed, that there is clearly some form of alpha in the ESG profile of a company, but that this alpha can be accessed only with powerful, non-linear techniques such as machine learning.

JEL classification: D83, G10, G11, G34.

Keywords: best-in-class approach, ESG, machine learning, portfolio construction, sustainable investments.

* Papers in this volume were prepared for a conference co-hosted by the BIS, the World Bank, the Bank of Canada and the Bank of Italy, in Rome, Italy on 22–23 October 2018. The views expressed are those of the authors and do not necessarily reflect the views of the BIS, the World Bank or the central banks represented at the meeting. Individual papers (or excerpts thereof) may be reproduced or translated with the authorisation of the authors concerned.

† Carmine De Franco PhD, Head of fundamental research at Ossiam (Paris, France), carmine.de-franco@ossiam.com.

‡ Christophe Geissler CEO at Advestis (Paris, France), cgeissler@advestis-conseil.com.

§ Vincent Margo, Analyst at Advestis (Paris, France), vmargot@advestis-conseil.com.

** Bruno Monnier, CFA, Head of operational research at Ossiam (Paris, France), bruno.monnier@ossiam.com.

# 1. Introduction

The relationship between corporate social performance (CSP) and corporate financial performance (CFP) is a fairly old theme in economic research. In its earlier stages, CSP was met with scepticism among critics: Nobel prize-winning economist Milton Friedman wrote in *The New York Times Magazine* in 1970 that "... there is one and only one social responsibility of business – to use its resources and engage in activities designed to increase its profits so long as it stays within the rules of the game, which is to say, engages in open and free competition without deception or fraud...." (Friedman (1970)).

As time went on, however, the number of studies highlighting the positive, or at least non-negative, relationship between CSP and CFP has grown significantly, probably beginning with the initial work by Bragdon and Marlin (1972) on the link between environmental virtue and financial performance. Fifty years later, the number of proponents of CSP and, more broadly, environmental, social and governance (ESG) integration in both corporate management and investors' choices has grown exponentially. As has the number of financial products, funds and exchange-traded funds that offer ESG versions of a large panel of investment strategies (mainly on equity and bonds). The current investment approach now seems in complete contrast to that of Friedman's, with the most recent empirical literature highlighting the link between ESG performance and alpha (Chong and Phillips (2016), Giese et al (2016), Zoltan et al (2016)).

Nevertheless, the question regarding the relationship between CSP and CFP remains largely unanswered. Reviews of published papers (meta-analysis) highlight that most empirical studies published on this theme report a non-negative or weak positive relationship between CSP and CFP (see eg Orlitzky et al (2003), Allouche and Laroche (2005), Wu (2006), Van Beurden and Gössling (2008), Margolis et al (2009), Friede et al (2015)). Other researchers take a more optimistic view and report either a significant relationship between CSP and CFP (Peiris and Evans (2010), Filbeck et al (2014), Indrani and Clayman (2015)) or, at the least, that CSP is not detrimental to CFP as long as one manages to build the portfolio with care, even if there is no clear value added in ESG integration (Kurtz and Di Bartolomeo (2011)).

Although we do not share the very optimistic and mostly overstated enthusiasm about the direct relationship between ESG and financial performance, we do believe a strong relationship exists between ESG and the sustainability of corporate businesses. Specifically, we believe ESG has an impact on financial performance and risks, but not linearly. We welcome the efforts that investors are undertaking to include ESG criteria in their portfolio choices, and hope this will trigger economic and cultural changes in corporate management. At the same time, however, we remain sceptical regarding the far-too flaunted capability of basic ESG ratings to act as an alpha generator in a portfolio. It remains true, however, that ESG data, reports and analysis can contain useful information related to the strengths and weaknesses of corporations. Unfortunately, ESG ratings are, by construction, a composite measure that dramatically reduces this rich set of information.

Our contribution to the growing literature on this topic is to show that, empirically, there is no value added in portfolios based on simple ESG screenings. Although it usually results in no harm to the performance, we do not find any alpha in such approaches. However, by recognising the intrinsic value of the large panel of ESG indicators that are aggregated to form the ESG ratings, we show that it is possible

to extract value from them, which, in turns, translates into real alpha. By exploring large data sets of specific ESG indicators, we are able to identify those that significantly impact corporate financial performance. In a simplified example, we can agree that for a company in the utility sector, the environmental performance can, most likely, be a discriminating criterion for financial performance; at the same time, governance can play an important role if we compare a utility company in an advanced market economy in Europe, for example, with one in an emerging market economy. Similarly, direct carbon emissions for banks are probably not as relevant to them as would the exposures of these banks, through loans, to highly polluting companies. In short, aggregate measures such as ESG ratings lose valuable information contained in the ESG indicators, which therefore lower their predictive power.

Searching for interesting patterns between specific ESG indicators and financial performance for a large set of companies remains out of reach for the standard tools available to econometricians. This search takes place in a high-dimensional space and is not oriented by previously derived information relating to these ESG features. To deal with this complexity, we developed a machine learning algorithm that allows us to identify features and patterns that are relevant to explain the link between CSP and CFP. The algorithm maps the regions in our high-dimensional space of ESG features that have been consistently associated with outperformance or underperformance. In the econometric parlance, we look at those regions for which the conditional expectation of each stock's forward return is statistically positive (or negative), given that its relevant ESG features fall in these regions. We say that these relevant ESG features "activate" the region. By observing the ESG features, we then obtain a significant signal regarding the future financial performance of the stock.

This identification is done with a set of rules that take the form of *If-Then* statements. The *If* statement identifies the region in the ESG space: ie the values that some ESG features must take in order to activate the rule. The *Then* statement produces a prediction of the excess return, over the benchmark, that we can expect from a stock whose ESG features fall in that region. The final prediction is the aggregation of the predictions made by these rules and is transformed into a score [–1, 0, +1]. We therefore focus on the sign of the prediction of excess return rather than on its value. This usually makes the estimation more robust.

The aggregation method mimics a panel of experts, each of whom specialises in an ESG feature (eg environment, independence of the board, ESG reporting verification, employee incidents etc) and makes a prediction given the ESG behaviour of the company. When the aggregated prediction is close to zero, ie the panel of experts is split between optimistic and pessimist forecasters, the final prediction is set at zero. The algorithm is regularly trained over time so that it can react and readjust to the new observed data. The algorithm is used to design a very simple strategy that screens the investment universe and selects all stocks with a positive score. The resulting portfolio is compared with a classic ESG best-in-class portfolio, which consists of all stocks in the investment universe whose ESG ratings are above a given threshold within their peer groups. Our empirical results show that the simple machine learning screened portfolio significantly outperforms the ESG best-in-class approach and the benchmark.

This is in line with the economic belief that ESG data are valuable in assessing financial performance, but also confirms that aggregated ESG ratings are not suited to distinguishing between outperformers and underperformers over the long run.

Even if a perfect distinction is out of reach, our results clearly confirm that there is alpha in the granular ESG data, but the relationship between ESG and financial performance is definitely not linear. Furthermore, the predictive power of the scores vanishes with time. We prove, indeed, that regularly training the algorithm over time and producing up-to-date sets of rules are key components of the superior performance of machine learning when it comes to stock screening.

## 2. Data

The analyses in this paper are carried out on portfolios based on the investment universe defined by the market capitalisation-weighted MSCI World Index USD, which consists of the largest corporations by market capitalisation listed in the United States, Canada, western Europe, Japan, Australia, New Zealand, Hong Kong SAR and Singapore. Portfolios are calculated in USD and net dividends are reinvested in the portfolio itself. Stock prices and dividends are taken from Thomson Reuters/Datastream. We reconstruct a proxy of the MSCI World Index by using end-of-month compositions as well as proxies for benchmarks in the United States, Europe[6] and developed economies in Asia.[7]

We also consider sector portfolios derived from the MSCI World Index and the regional benchmarks by filtering stocks that belong to the same sector: consumer staples (CS), consumer discretionary (CD), energy (EN), financials (FI), health care (HC), industrials (IN), information technology (IT), materials (MA), telecommunication services (TL) and utilities (UT).

For each company in the investment universe, we collect ESG ratings from Sustainalytics.[8] An ESG rating is a comprehensive measure based on three pillars – environment, social and governance – that assesses the strengths and weaknesses of a company along these three directions. The pillars are themselves based on a large set of specific indicators. For the purposes of this study, the composite ESG rating is the arithmetic average of the three ratings – environment (E), social (S) and governance (G) – each of which is itself the combination of roughly 50 narrower indicators. Finally, for each company, we consider its relative peer group, which consists of all companies with a similar business, hence comparable from a sustainability point of view. ESG data are available from 2009 onwards, collected with a relatively stable methodology and uniform coverage. All portfolios presented in the following sections are rebalanced on a monthly basis, at the end of every month, with a four-day lag between data extraction and a portfolio's implementation. Portfolios are benchmarked against classical cap-weighted, liquid and investable portfolios, a standard practice in the financial industry. Through the entire analysis, the term "alphas" refer to CAPM-alphas unless stated otherwise.

6 Stocks in the MSCI World Index domiciled in Austria, Belgium, Denmark, Finland, France, Germany, Ireland, Italy, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland and the United Kingdom.

7 Stocks in the MSCI World Index domiciled in Australia, Hong Kong SAR, Japan, New Zealand and Singapore.

8 One of the largest providers of ESG ratings.

## 3. The best-in-class approach

One of the most popular approaches to embed ESG criteria in the portfolio construction process is the so-called best-in-class approach.

Given a threshold *x*, one excludes the stocks whose ESG ratings belong to the lowest *x*-quantile. The exclusion is usually carried across peer groups, ie groups of stocks with very similar characteristics. The reason behind this is twofold:

- Removing stocks with low ESG ratings within peer groups ensures that the final economic mesh of the filtered universe remains similar to the initial investment universe.
- ESG ratings have a structural, sector-driven bias that usually favours specific sectors (eg IT or health care sectors) while penalising others (eg energy or utilities). Given this bias, the filtering of peer groups makes comparisons of ESG ratings independent of the sectors.

For the purpose of this study, an ESG best-in-class portfolio derived from a market capitalisation-weighted portfolio removes, within each peer group, the stocks whose ratings belong to the lowest *x*-quantile. The portfolio is finally scaled to sum up to one. This approach, quite popular among investors, should not be thought of as a way to enhance performance. As Tables 1–4 show, ESG best-in-class filters applied to standard market capitalisation-weighted indexes do not lead to outperformance.

Except for Europe and relatively low threshold levels, we find small but negative excess returns and negative information ratios for the ESG best-in-class portfolios over their benchmarks with almost unchanged risks. Although the approach does not create outperformance per se, it does not carry structural underperformance either. Optimistically, one could accept the fact that embedding ESG objectives in a portfolio does not significantly modify its risk/return profile.

Our findings are not in contradiction with the large literature that finds positive links between ESG and financial performance. But the consistency and durability over time of the ESG factor has been questioned since the very beginning. Aupperle et al (1985) finds no significant relationship between social responsibility and corporate profitability, and similar results were obtained in Capelle-Blancard and Monjon (2012) and Humphrey and Tan (2014). Griffin and Mahon (1997) report that a correlation between financial performance and social performance depends on the measure used to distinguish between high and low social performers.

Table 1: World Developed

| ESG best-in-class | Bench | 10% | 30% | 50% |
|---|---|---|---|---|
| Ann performance | 10.07% | 10.01% | 9.93% | 9.51% |
| Ann volatility | 13.34% | 13.31% | 13.44% | 13.82% |
| Sharpe ratio | 0.73 | 0.73 | 0.72 | 0.67 |
| Max drawdown | –21.91% | –21.79% | –22.02% | –22.57% |
| Information ratio | - | –0.27 | –0.25 | –0.41 |

Table 2: United States

| ESG best-in-class | Bench | 10% | 30% | 50% |
|---|---|---|---|---|
| Ann performance | 13.45% | 13.25% | 13.46% | 13.2% |
| Ann volatility | 14.61% | 14.54% | 14.49% | 14.4% |
| Sharpe ratio | 0.9 | 0.89 | 0.91 | 0.9 |
| Max drawdown | –18.99% | –18.87% | –18.71% | –18.04% |
| Information ratio | - | –0.71 | 0.02 | –0.18 |

Table 3: Europe

| ESG best-in-class | Bench | 10% | 30% | 50% |
|---|---|---|---|---|
| Ann performance | 6.37% | 6.55% | 6.47% | 6.31% |
| Ann volatility | 19.25% | 19.19% | 19.19% | 19.29% |
| Sharpe ratio | 0.32 | 0.33 | 0.32 | 0.31 |
| Max drawdown | –30.25% | –30.21% | –30.2% | –30.54% |
| Information ratio | - | 0.43 | 0.22 | –0.11 |

Table 4: Asia-Pacific

| ESG best-in-class | Bench | 10% | 30% | 50% |
|---|---|---|---|---|
| Ann performance | 6.83% | 6.71% | 6.41% | 5.75% |
| Ann volatility | 15.54% | 15.71% | 16% | 16.2% |
| Sharpe ratio | 0.42 | 0.41 | 0.38 | 0.34 |
| Max drawdown | –24.8% | –24.95% | –25.27% | –25.8 |
| Information ratio | - | –0.21 | –0.36 | –0.46 |

Key performance indicators of the MSCI World Index and three market capitalisation-weighted regional benchmarks, together with ESG best-in-class filtered portfolios with different thresholds: 10%, 30% and 50%. Data are shown in USD from August 2009 to March 2018.

Source MSCI, Datastream, Sustainalytics.

Our results are more in line with Revelli and Viviani (2015), for which "... the consideration of corporate social responsibility in stock market portfolios is neither a weakness nor a strength compared with conventional investments...". It should be noted that many fund managers and institutional investors surveys report that ESG is mostly viewed firstly as a risk mitigation tool (Van Duuren et al (2016)) and eventually as a long-term performance driver. We share the optimistic view of Nobel prize-winning economist Robert Shiller, for which both society and the financial community would find the use of socially responsible practices mutually beneficial (Shiller (2013)). At the same time, we also believe that short- to medium-term financial performance is, at best, weakly correlated to ESG ratings, at least for such broad investment universes as the MSCI World Index (which contains more than 1,600 companies). We find this to be the case for several reasons:

i. The investment universes are relatively large and the aggregated ESG ratings have too low a signal-to-noise ratio to allow for an efficient selection of outperforming stocks.
ii. ESG ratings are global metrics that embrace environmental, social and governance criteria. As such, they may be too reductive, and we may lose a significant amount of information from the single indicator to the aggregated scores.
iii. Granularity is key: as an example, it is likely that companies in specific sectors (eg energy) react differently to changes in the E score compared with the S score.
iv. In the search for a rational economic theory behind ESG, some argue that by divesting low ESG-rated companies, investors raise their cost of capital and, in turn, the return these companies have to offer to attract new investors. As such, in the short run, they may show higher performance, but over time, the level of return they have to offer becomes unsustainable. In other words, the action of divesting may take time to materialise in both investors' portfolios and low ESG-rated companies (see eg Asness (2017)).
v. The period considered in this study spans from the earlier stages of the recovery in 2009 to March 2018. Therefore, we consider key performance indicators over

a period of strong equity market, characterised by high returns and historically lower levels of volatility. This market regime can potentially affect the overall strength of ESG filtered portfolios.

To illustrate item (iii), we consider sector portfolios derived from the MSCI World Index and from the three regional benchmarks (the United States, Europe and Asia-Pacific) and we apply both ESG and single pillars E, S and G, at 30% best-in-class filtering. Tables 5–8 show the results. For the sake of simplicity, we only show annualised excess returns over the relative benchmark sector portfolios and information ratios.

Overall, it is not straightforward to detect clear patterns between excess returns and ESG metrics conditionally to the regional benchmarks. But we can definitely detect specific triplets sector/region/metric that produce significant positive excess returns. Clearly, integrating ESG criteria in the utilities sector enhances in-sample performance. But the right metric to use clearly depends on the geography: in the world developed region (Table 5) the best excess return for the utilities sector is achieved when one uses the G score only at 0.82%; in the United States (Table 6) it is better to look at the S rating under which utilities achieve 0.7%. In Europe (Table 7) it is with the E score that utilities obtain the best result with 3.25%, while in Asia-Pacific (Table 8) it is, once again, the composite ESG rating that achieves the highest excess return at 3.07%.

More generally, there is no sector nor metric for which the excess return of the best-in-class filtered sector achieves a positive excess return in all the regions. Similarly, there is no region nor sector for which all metrics produce positive excess returns. Finally, no sector achieves positive excess returns across all regions and metrics. In other words, finding performance drivers when integrating ESG criteria in a best-in-class fashion is out of reach.

From Tables 5–8, only 12 out of 40 sector/metric portfolios in the world developed region turn out to have a positive excess return, and seven of them are obtained when one considers the G score. In the United States, we find positive excess returns in 12 out of 40, with no clear indication on the best metric to use. We notice, however, that all the metrics seem to work in the utilities sector. In Europe, we count 22 out of 40 sector/metric pairs with positive excess returns. For four sectors (consumer discretionary, materials, telecommunication services and utilities) all metrics work accurately. In Asia, we have 16 out of 40 portfolios with positive excess returns, with no clear patterns between sectors and metrics, except for the energy sector, for which all metrics produce positive excess returns, even if their magnitudes are relatively small.

Table 5: World Developed

| | ESG | E | S | G |
|---|---|---|---|---|
| CD | −0.33% (−0.3) | −0.46% (−0.46) | −0.65% (−0.4) | −1.22% (−0.65) |
| CS | −0.26% (−0.26) | −0.45% (−0.47) | −0.65% (−0.59) | **0.62% (0.53)** |
| EN | −0.24% (−0.13) | −0.41% (−0.19) | −0.26% (−0.17) | **0.86% (0.31)** |
| FI | −0.51% (−0.43) | −0.46% (−0.42) | −0.87% (−0.63) | **0.19% (0.13)** |
| HC | −0.39% (−0.35) | −0.5% (−0.3) | −0.47% (−0.4) | −0.53% (−0.43) |
| IN | −0.31% (−0.28) | −0.32% (−0.27) | −0.32% (−0.31) | −0.33% (−0.19) |
| IT | −0.13% (−0.12) | −0.44% (−0.45) | −1.53% (−0.52) | **0.1% (0.05)** |
| MA | −0.09% (−0.05) | −0.09% (−0.03) | −0.17% (−0.08) | **0.32% (0.18)** |
| TL | **1.26% (0.64)** | **1.17% (0.74)** | **0.15% (0.06)** | **0.73% (0.21)** |
| UT | **0.16% (0.1)** | −0.94% (−0.39) | **0.72% (0.43)** | **0.82% (0.32)** |

Table 6: United States

| | ESG | E | S | G |
|---|---|---|---|---|
| CD | −1.95% (−0.82) | −0.48% (−0.39) | −0.91% (−0.55) | −2.92% (−0.98) |
| CS | −0.94% (−0.84) | −0.46% (−0.43) | −0.46% (−0.32) | −0.26% (−0.12) |
| EN | −0.09% (−0.06) | −0.66% (−0.44) | 0.01% (0.01) | **0.55% (0.16)** |
| FI | −0.22% (−0.19) | −0.11% (−0.11) | −0.2% (−0.17) | **1.02% (0.67)** |
| HC | −0.23% (−0.22) | **0.19% (0.15)** | −0.55% (−0.5) | −0.58% (−0.37) |
| IN | **0.4% (0.54)** | **0.25% (0.32)** | 0.01% (0.01) | −0.6% (−0.56) |
| IT | −1.36% (−0.96) | −0.7% (−0.75) | −1.76% (−0.51) | −0.37% (−0.11) |
| MA | **0.34% (0.18)** | **0.72% (0.32)** | −0.76% (−0.32) | −0.14% (−0.07) |
| TL | −0.38% (−0.37) | −0.32% (−0.38) | −0.4% (−0.2) | **1.32% (0.28)** |
| UT | **0.63% (0.56)** | **0.47% (0.38)** | **0.7% (0.7)** | **0.13% (0.12)** |

Table 7: Europe

| | ESG | E | S | G |
|---|---|---|---|---|
| CD | **0.51% (0.32)** | **0.45% (0.29)** | **0.18% (0.14)** | **1.07% (0.68)** |
| CS | **0.16% (0.14)** | **0.03% (0.03)** | −0.01% (−0.01) | **0.39% (0.32)** |
| EN | −1.2% (−0.3) | −2.01% (−0.38) | −0.5% (−0.12) | −1.05% (−0.27) |
| FI | **0.31% (0.18)** | **0.04% (0.03)** | **0.33% (0.17)** | −0.28% (−0.21) |
| HC | −0.38% (−0.49) | −0.44% (−0.55) | −0.53% (−0.67) | −0.11% (−0.18) |
| IN | 0.03% (0.03) | 0.03% (0.03) | −0.39% (−0.35) | −0.67% (−0.44) |
| IT | −0.63% (−0.29) | −1.13% (−0.49) | 0.01% (0.08) | −0.71% (−0.43) |
| MA | **0.12% (0.03)** | **0.23% (0.06)** | **0.51% (0.14)** | **0.6% (0.18)** |
| TL | **1.81% (0.67)** | **1.72% (0.63)** | **1.82% (0.62)** | **2.06% (0.71)** |
| UT | **1.79% (0.54)** | **3.25% (0.87)** | **0.09% (0.03)** | **0.41% (0.16)** |

Table 8: Asia-Pacific

| | ESG | E | S | G |
|---|---|---|---|---|
| CD | −0.29% (−0.19) | −0.11% (−0.11) | **0.09% (0.04)** | **0.75% (0.37)** |
| CS | **0.6% (0.52)** | **0.37% (0.12)** | −0.3% (−0.28) | −0.18% (−0.09) |
| EN | **0.12% (0.04)** | **0.34% (0.11)** | **0.03% (0.01)** | **0.25% (0.1)** |
| FI | −0.27% (−0.16) | −0.09% (−0.06) | −0.71% (−0.46) | −0.08% (−0.05) |
| HC | **0.29% (0.19)** | −0.67% (−0.27) | **0.14% (0.08)** | −0.03% (−0.02) |
| IN | −0.76% (−0.39) | −0.41% (−0.22) | −0.42% (−0.22) | −0.61% (−0.35) |
| IT | −1.32% (−0.59) | −0.92% (−0.4) | −1.01% (−0.43) | −0.96% (−0.28) |
| MA | −0.29% (−0.26) | **0.04% (0.03)** | −0.58% (−0.44) | −0.07% (−0.05) |
| TL | **0.17% (0.04)** | −0.18% (−0.11) | −0.61% (−0.09) | **1.83% (0.24)** |
| UT | **3.07% (0.65)** | −3.57% (−0.87) | **0.16% (0.04)** | **2.6% (0.55)** |

Annualised excess returns (information ratios) between market capitalisation-weighted sector portfolios and their ESG best-in-class filtered versions for the MSCI World Index and the derived regional benchmarks. In bold pairs sector/indicator for which the excess return is positive. Best-in-class filters are performed with the ESG rating together with the single pillars environment (E), social (S) and governance (G) ratings. Data are shown in USD from August 2009 to March 2018. Source: MSCI, Datastream and Sustainalytics. The rows correspond to the standard GICS sectors: consumer staples (CS), consumer discretionary (CD), energy (EN), financials (FI), health care (HC), industrials (IN), information technology (IT), materials (MA), telecommunication services (TL) and utilities (UT).

In conclusion, our empirical findings confirm that simple ESG filtering does not result in better performance. Rather, it behaves as a negative factor, reducing performance. Given the short period we consider, and the market regime that equity markets have experienced since 2009, we share the view that ESG best-in-class integration is, most likely, neutral to financial performance. Nevertheless, our results highlight the fact that geographies and sectors do not react to ESG criteria in the same way. But finding interesting and statistically significant patterns between ratings, pillars, their underlying narrow indicators (*features*) and financial performance, for more than 150 indicators on more than 1,600 companies in the MSCI World Index, over a roughly 10-year period, is out of reach for both human and linear statistical tools. The next section introduces other techniques that can overcome this complexity and exploit this huge set of data.

## 4. Machine learning

In this section, we introduce a deterministic, easily understandable machine-learning prediction algorithm, aimed at finding consistent and statistically significant patterns between ESG ratings and financial performance. The algorithm explores a high-dimensional data set of ESG granular indicators for all the companies in our investment universe. The goal of the algorithm, which falls in the category of supervised machine learning, is to predict the (conditional) excess return of each company over the benchmark, given the specific values taken by some of its ESG indicators (the features). In other words, the algorithm identifies regions in the high-dimensional space of ESG features that are statistically related to financial outperformance or underperformance. Features include raw and derived ESG indicators,[9] sector and country classifications, company size and controversy indicators.

The regions are characterised by *rules* in the form *If-Then*, so that the algorithm finally consists of a set of such rules. The *If* statement is a list of conditions on the features $x_t \in X = X_1 \times X_2 \ldots \times X_d$, where $X_i$ is the set of possible outcomes of the feature $i$ and $d=447$ is the total number of features.[10] Therefore, a rule defines a hyper-rectangle of $X$. The *Then* statement is the prediction of the three-month forward excess return conditional to the *If* statement. Since the rules correspond to hyper-rectangles in the feature space, we finally obtain relatively simple and understandable regions. Furthermore, to avoid overfitting, the algorithm only selects a finite number of such rules. At each time *t*, the predictions of each rule are aggregated into one prediction, $\hat{y}$, through convex combination. The algorithm is calibrated (*trained*) on the training set and the rules are used out-of-sample. The learning process works at two independent levels:

[9] For each raw indicator, as for example the E score, we also look at the derived indicator relative to the peer group and the sector. All these transformations can potentially contain useful information. On the other side, the use of both raw and derived indicators rapidly increases the dimension of the feature space *d*.

[10] We use 164 ESG raw indicators, from which we derive peer group and sector relative indicators and three valuation indicators. In total 164*3 + 3 = 495. From these indicators we remove 48 indicators for which either the sector or the peer group derived indicators are too close, or for which historical data are missing.

- At the end of year *N+1* we train the algorithm on an expanded data set of features and stock total returns that contain the data set used at the end of year *N* augmented by all the new observed data (features and stock total returns) from the end of year *N* to the end of year *N+1*. To initialise the algorithm, we train it over three years of data (from 2009 to 2012). By expanding the data set, the algorithm is able to access new data and explore new patterns, so that it can strengthen or nuance some rules that were previously discovered.
- Daily, the algorithm can update the weights used to aggregate each rule's prediction, by overweighting rules with a good prediction rate and underweighting the others. Therefore, following day predictions will benefit from the *experience* the algorithm is gaining on the rules and their predictive power. The weight of each rule can be viewed as a confidence index. Of course, this is possible because the algorithm is able to assess the goodness of its predictions by looking at the realised three-month return.

To avoid threshold effects, we transform the final prediction for each stock into a score $S \in \{-1, 0, +1\}$, where *+1* stands for significantly positive excess return prediction, *-1* for negative prediction and *0* for an uncertain prediction. The case where *S=0* is usually related to stocks for which some of their ESG indicators would eventually signal financial outperformance, while other ESG indicators rather signal potential underperformance. The picture is then nuanced, and the algorithm cannot make a precise prediction. This is a very common situation in finance, where different indicators can yield different forecasts, so that, in aggregate, the forecast turns out to be uninformative. The learning process is divided into two steps. Following Nemirovski (2000) and Tsybakov (2003), the training set $D_N$ at the end of year *N* is divided into two sub-data sets: $D_n$ the learning set and $D_t$ the aggregation set, with $t \gg n$ and $n + t = N$. The learning set $D_n$ is used to design and select the set of rules used by the algorithm to make predictions. The aggregation set is used to fit the coefficients of the convex combination, in line with the *expert aggregation theory* of Cesa-Bianchi and Lugosi (2006) and Stoltz (2010).

**Independent suitable rules.** Let $\mathrm{D_N} = \big((x_1, y_1), \cdots, (x_N, y_N)\big) \in (X \times \mathbb{R})^{\mathrm{N}}$ be the training set. Here $y_i$ denotes the three-month return for some stock and $x_i$ is the *d*-dimensional vector of its ESG features. The training set consists of a large but finite number of *(d+1)*-vectors spanning all stocks in the investment universe and all available dates. The training set $D_n \subseteq D_N$ includes the first *n* data points in $D_N$ and $\mathrm{D_t} = \big((x_{n+1}, y_{n+1}), \cdots, (x_N, y_N)\big)$ the order being induced by the time.

**Definition 4.1.** *For any set $E \subset X$, we define*

$$\mu(E, D_n) \coloneqq \frac{\sum_{i=1}^{n} y_i \mathbf{1}_{x_i \in E}}{\sum_{i=1}^{n} \mathbf{1}_{x_i \in E}}$$

*where, by convention,* $0/0 = 0$.

The set-valued map $\mu$ represents the conditional excess return of a stock over the benchmark, given that its ESG features *x* belong to *E*.

**Definition 4.2.** *Let $\boldsymbol{r}$ be a hyper-rectangle on X: $\boldsymbol{r} = \prod_{k=1}^{d} \boldsymbol{I}_k$ where each $\boldsymbol{I}_k$ is an interval of $\boldsymbol{X}_k$. A rule f is a function defined on $\boldsymbol{r} \times (X \times R)^N$ as*

$$f(x, D_n) =: \mu(\boldsymbol{r}, D_n), \forall\, x \in \boldsymbol{r} \tag{4.1}$$

*The hyper-rectangle* ***r*** *is called the* ***condition*** *and* $\mu(\boldsymbol{r}, D_n)$ *is called the* ***prediction*** *of the rule f. The event* $\{x \in \boldsymbol{r}\}$ *is called the* ***activation conditions*** *of the rule f.*

A rule *f* is completely defined by its condition ***r***. So, with an abuse of notation, we do not distinguish between a rule and its condition. We define two crucial numbers for a rule:

**Definition 4.3.** *Let f be a* rule as in Definition 4.2 defined on $\boldsymbol{r} = \prod_{k=1}^{d} \boldsymbol{I}_k$.

*a. The* ***number of activations*** *of f in the sample* $D_n$ *is*

$$n(\boldsymbol{r}, D_n) \coloneqq \sum_{i=1}^{n} \mathbf{1}_{x_i \in E}$$

*b. The* ***complexity*** *of f is*

$$cp(\boldsymbol{r}) \coloneqq d - \#\{i \leq k \leq d \mid I_k = X_k\}$$

The algorithm does not consider all the possible rules, but only those with a given *coverage* and *significance*. We call these rules *suitable,* and their definition is given below.

**Definition 4.4.** *A rule f, defined on* ***r****, is a* ***suitable rule*** *for the training set* $D_n$ *if and only if it satisfies the two following conditions:*

***a. Coverage condition***

$$C_{min} \leq \frac{n(\boldsymbol{r}, D_n)}{n} \leq C_{max} \tag{4.2}$$

*with* $C_{min}$ *and* $C_{max}$ *suitably chosen in the calibration step.*

***b. Significance condition***

$$|\mu(\boldsymbol{r}, D_n) - \mu(X, D_n)| \geq z(\boldsymbol{r}, D_n, \alpha) \tag{4.3}$$

*for a chosen* $\alpha \in [0, 1]$ *and a function z.*

The coverage condition (4.2) excludes rules that are activated only on small sets (ie with a low coverage rate, $C_{min}$) and rules that are too obvious (ie with a high coverage rate, $C_{max}$). The threshold in the significance condition (4.3) is set such that the probability of falsely rejecting the null hypothesis $\mu(\boldsymbol{r}, D_n) = \mu(X, D_n)$ is less than $\alpha$. The parameter $\alpha$ permits to control the number of suitable rules. The higher $\alpha$, the higher the number of suitable rules. In what follows, we generate rules of complexity $c \geq 2$ by a *suitable intersection* of rules of complexity *1* and rule of complexity *c-1*.

**Definition 4.5.** *Two rules $f_i$ and $f_j$ defined on $\boldsymbol{r}_i$ and $\boldsymbol{r}_i$ respectively, form a* ***suitable intersection*** *if and only if they satisfy the two following conditions*:

***a. Intersection condition***

$$r_i \cap r_j \neq \emptyset,$$

$$n(\boldsymbol{r}_i \cap \boldsymbol{r}_j, D_n) \neq n(\boldsymbol{r}_i, D_n), \tag{4.4}$$

$$n(\boldsymbol{r} \cap \boldsymbol{r}_j, D_n) \neq n(\boldsymbol{r_j}, D_n)$$

***b. Complexity condition***

$$cp(\boldsymbol{r}_i \cap \boldsymbol{r}_j) = cp(\boldsymbol{r}_i) + cp(\boldsymbol{r}_j) \tag{4.5}$$

The intersection condition (4.4) avoids adding a useless condition for a rule. In other words, to define a suitable intersection, $\boldsymbol{r}_i$ and $\boldsymbol{r}_i$must not be satisfied by the same points in $D_n$. The complexity condition (4.5) means that $\boldsymbol{r}_i$ and $\boldsymbol{r}_i$have no marginal index in common.

**Designing suitable rules.** The design of suitable rules is made recursively on their complexity. It stops at a complexity *c* if no rule is suitable or if the maximal complexity $c = cp_{max}$ is achieved.

*Complexity 1*: The first step is to find suitable rules of complexity *1*. First notice that the complexity of evaluating all rules of complexity *1* is $O(ndm^2)$. Rules of complexity *1* are the base of the algorithm search heuristic. So, all rules are considered, and only suitable ones are kept, ie rules that satisfied the coverage condition (4.2) and the significance condition (4.3). Since rules are considered independently, the search can be parallelised.

*Complexity c:* Among the suitable rules of complexity *1* and *c-1*, we select *M* rules of each complexity (*1* and *c-1*) according to a chosen criterion. Then it generates rules of complexity *c* by pairwise *suitable intersection* according to Definition 4.5. The complexity of evaluating all rules of complexity *c*, obtained from their intersections, is $O(nM^2)$. Here again, since rules are considered independently, the evaluation can be parallelised. The parameter *M* helps to control the computing time.

**Selecting suitable rules.** We select a subset $\boldsymbol{S}$ from all suitable rules which maximises the gains expected from rule in $D_n$ and such as their conditions form a covering of $X$.

**Algorithm.** The calibration of the algorithm is structured in two parts: in the first one, it finds all suitable rules, and in the second one it retains only an optimal subset of it. To avoid threshold effects, overfitting and to manage the numerical complexity, we discretise each feature in $X$ into *m* classes with empirical quantiles (modalities).[11] Thus, each modality of each variable covers about *100/m* percent of the sample. In practice, *m* must be inversely related to *d*: The higher the dimension of the problem, the smaller the number of modalities.

[11] Of course, such procedure is performed only on real-valued features with more than $m$ different values. Categorical features are left unchanged.

The parameters of the algorithm are:

- $m$, the sharpness of the discretisation;
- $\alpha \in [0, 1]$, which specifies the false rejecting rate of the test;
- $z$, the significance function of the test;
- $C_{max}$ and $C_{min}$ the coverage bounds;
- $cp_{max}$ the maximal complexity of a rule; and
  $M > 0$, the number of rules of complexity *1* and *c-1* used to define the rules of complexity. *c*.

**Aggregation.** Let $\mathrm{D_t} = \big((x_{n+1}, y_{n+1}), \cdots, (x_N, y_N)\big) \in (X \times \mathbb{R})^{\mathrm{N}}$, where $n + t = N$ be the aggregation set and let $\boldsymbol{S}$ be the set of *R* rules selected by the algorithm. At each time *t*, the predictions of each rule are aggregated into one prediction $\widehat{y_t}$ as follows:

$$\widehat{y_t} = \frac{\sum_{i=1}^{R} \pi_{i,t}\, f_i\,(x_t, D_n)}{\sum_{i=1}^{R} \pi_{i,t} \mathbf{1}_{x_t \in \boldsymbol{r_i}}} \tag{4.6}$$

with $\pi_{i,t} = 1/R$. When the realised value $y_t$ is known, the weights $\pi_{.,t+1}$are updated with the following formula:

$$\pi_{i,t+1} = \pi_{i,t} \frac{\exp\big(-\eta\, l(f_i\,(x_t, D_n), y_t)\big)}{\sum_{k=1}^{R} \pi_{k,t} \exp\big(-\eta\, l(f_k\,(x_t, D_n), y_t)\big)} \tag{4.7}$$

with $\eta > 0$ and $l$ a convex loss function.
Remark 4.6. One can notice that $f_i\,(x_t, D_n)$ is not defined if $x_t \notin \boldsymbol{r}_i$. In (4.6), $\widehat{y_t}$ is well defined for all $t$, since the set $\boldsymbol{S}$ is a covering of $X$. In (4.7) we follow the methodology of the **sleeping expert aggregation** from Devaine et al (2013). Once trained, the machine learning algorithm produces predictions of the excess returns, which are transformed into a scores $S \in \{-1, 0, +1\}$, given the out-of-sample ESG features $x_t$ for each company. Table 9 shows some examples of rules taken from the learning process of the algorithm. The table lists three rules associated with positive predictions (opportunities) and five rules with negative predictions.

Opportunity rules: positive excess return — Table 9

| Feature | Relative to | Activation set | Rule description |
|---|---|---|---|
| Business ethics incidents | Sector | [5.9] | WHEN business ethics incidents is high relative to sector AND board remuneration disclosure is high relative to sector THEN opportunity |
| Board remuneration disclosure | Sector | [5.9] | |
| Board independence | All | [9.9] | WHEN board independence is at the maximum AND board remuneration disclosure is high relative to sector THEN opportunity |
| Board remuneration disclosure | Sector | [5.9] | |
| Board independence | All | [9.9] | WHEN board independence is at the maximum AND business ethics incidents is high relative to sector THEN opportunity |
| Business ethics incidents | Sector | [5.9] | |

Risk rules: negative excess return

| Feature | Relative to | Activation set | Rule description |
|---|---|---|---|
| Verification of ESG reporting | Sector | [0.7] | WHEN verification of ESG reporting is not high relative to sector AND board remuneration disclosure is low relative to sector THEN risk |
| Board remuneration disclosur | Sector | [0.4] | |
| Quantitative performance | All | [5.9] | WHEN quantitative performance score is high AND board remuneration disclosure is low relative to sector THEN risk |
| Board remuneration disclosur | Sector | [0.4] | |
| Verification of ESG reporting | All | [0.6] | WHEN verification of ESG reporting is not high AND quantitative performance score is high THEN risk |
| Quantitative performance | All | [6.9] | |
| Gender diversity of board | Peer group | [0.8] | WHEN gender diversity of board is not high relative to peer group AND employee incidents is very low relative to peer group THEN risk |
| Employee incidents | Peer group | [0.2] | |
| Green logistics programmes | Delta score | [0.2] | WHEN green logistics programmes delta score is ve low AND qualitative performance delta score is very low THEN risk |
| Qualitative performance | Delta score | [0.2] | |

Some rules from the learning process of the algorithm at end 2012, 2013 and 2016. All features are discretised over 10 modalities (zero to nine) except for qualitative performance, which is discretised over six modalities (zero to five). High values for the features correspond to good ESG performance.

Each rule consists of two features and two intervals. The "relative to" properties indicate whether the feature must be calculated over all stocks in the universe (all), over a sector, over a peer group, or whether we should look at the variations of the feature over time (delta score).

Whenever the values taken by the features for a given company fall in the given intervals (we say that the stock activates the rule) the algorithm makes a prediction on its excess return. It is important to remark that we aggregate all the predictions, and we transform the final aggregated prediction into a score $S \in \{-1, 0, +1\}$, so that in the end we mainly look at the sign of the prediction rather than at its magnitude. We also remark that, while the set of rules remains unchanged for one year (until the next learning process), the output of the rules can change over time, because raw indicators can change and because the aggregated weights of the rules change over time.

Finally, the use of granular, rich ESG data are a key element of the power of the machine learning algorithm: indeed, it works quite poorly if one only considers aggregated E, S and G scores.

# 5. Machine learning application

We now compare the predictive power of the machine learning algorithm developed in Section 4 with the classical best-in-class approach. More precisely, we try to assess whether filtering stocks over scores derived from the algorithm outperforms the standard filtering over ESG ratings (best-in-class).

For the sake of simplicity, we only present the "world developed" universe and, among the strategies presented in Section 3, we only consider the 30% best-in-class, as it is very close to what investors look at for their ESG portfolios.

We recall that this strategy excludes, at each monthly review, the stocks whose ESG ratings are in the lower tercile within each peer group, and finally scale the weights so that their sum is one. To insure replicability of the portfolio, the ESG ratings are taken four days before the review date (which is end-of-month). At the monthly review, we also build three portfolios based on the scores calculated with the machine learning algorithm, with the rules calculated at the end of the year that precedes the review:

**Positive ML screening:** The portfolio selects all stocks in the investment universe whose scores are *+1*. The weights are finally scaled to sum up to one (maintaining the market capitalisation-weighting scheme of the benchmark)

**Positive ML screening sector-matched:** Same selection as for the positive ML screening portfolio, but the scaling of the weights is done in such a way that the final sector breakdown of the portfolio is matched to the benchmark's one.

**Negative ML screening:** The portfolio selects all stocks in the investment universe whose scores are *-1*. The weights are finally scaled to sum up to one (maintaining the market capitalisation-weighting scheme of the benchmark)

As before, the scores are taken four days before the review date. We consider the sector-matched portfolio because the absolute screening usually introduces significant sector deviations with respect to the benchmark.

It should be noticed that market capitalisation-weighted portfolios have some drawbacks: they are trend-following and show sector concentrations. However, this is relatively limited in our case as the benchmark is a large and relatively well diversified portfolio, where even large cap stocks rarely exceed 3% of the index. We do not report results relative to the equally-weighting scheme (1/N) as they are very similar to the cap-weighted scheme in relative terms.

Alternatively, one could use the standard mean-variance approach. We do not consider it in the research process because for this approach one should design a covariance estimation procedure, which may introduce noise and subjectivity in the portfolio construction, and define a procedure to calculate expected return, either model-based, or estimated from past data or again in a Black-Littermann fashion. In both cases, the sensitivity of the strategy to the set of procedures (and corresponding parameters) will play a key role in the outcome. And associated turnover will be very high when compared to cap-weighted selections. Our choice of a market capitalisation-weighted scheme is therefore the best way to assess the power of our ML algorithm as it is simple, stable and produces low-turnover portfolios.

Table 10 collects the main results for these portfolios using data since January 2013. The sample length is driven by the availability of good and uniform quality ESG

data (since 2009) and the initial three years of data needed for the first training of the ML algorithm. Although we recognise that the period over which we can test the machine learning algorithm is relatively short (five years and three months), the results we obtain contain some interesting insights.

Table 10

| | | Machine learning screening | | | ESG best-in-class |
|---|---|---|---|---|---|
| | Bench | Positive | Positive sect matched | Negative | 30% |
| Ann performance | 10.32% | 13.07% | 11.66% | 8.31% | 10.13% |
| Ann volatility | 10.50% | 11.14% | 10.96 | 10.95% | 10.57% |
| Sharpe ratio | 0.94 | 1.14 | 1.03 | 0.72 | 0.92 |
| Max drawdown | –18.07% | –14.99% | –16.46% | –22.47% | –17.91% |
| Information ratio | - | 1.01 | 0.58 | -0.54 | –0.32 |
| Ann CAPM alpha | - | 2.47% | 1.15% | –1.81% | –0.24% |

Key performance indicators of the MSCI World Index (bench), the market capitalisation-weighted selection filtered over positive scores from the ML algorithm, the one with the sector allocation matched to the benchmark, the one screened over negative scores and the 30% ESG best-in-class filtered portfolios. Data are shown in USD from January 2013 to March 2018.

Source: MSCI, Datastream, Sustainalytics.

First, the positive ML screening outperforms all the other portfolios: the benchmark on an annualised basis by 2.76%, the ESG best-in-class portfolio by 2.94% and the negative ML screening by 4.77%. And while the realised annual volatilities remain in the range 10.50% to 11.14%, there are significant differences in the realised maximum drawdowns: the negative ML screening shows a –22.47% loss from its peak, while the positive ML screening loss from its peak accounts for –14.99%.

These two combined results show that the machine learning algorithm is clearly able to distinguish between opportunity stocks (the ones with positive scores) from risky stocks (negative scores). Figure 1 shows the historical behaviour of these two portfolios and the benchmark. We notice that the positive ML screening outperforms the negative ML screening over time, with the benchmark in between.

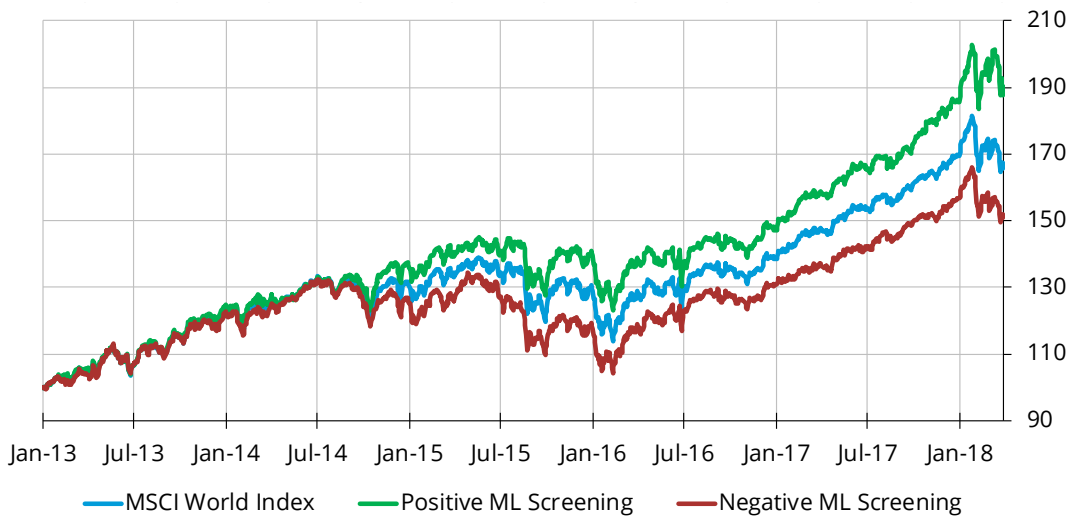
Furthermore, in years when the benchmark shows very high performance with very low volatility, typically in bull market regimes, the differences between the two strategies are less pronounced. On the contrary, when the market is in a bear regime or has no clear trend, the positive ML screening clearly outperforms its negative counterpart, as shown in Table 11.

In years when the benchmark performance is very significant (2013 or 2017), the positive ML screening is still able to achieve some outperformance, but the spread with the negative ML screening is somehow lower in years when the market performance is negative or low (in 2014, 2015 and, most recently, 2018).

Simulated strategy levels for the benchmark MSCI World Index, the selection filtered over positive scores and the selection filtered over negative scores

Figure 1

Base level = 100




Simulated strategy levels for the benchmark MSCI World Index, the market capitalisation-weighted selection filtered over positive scores from the ML algorithm (positive ML screening) and the one screened over negative scores (negative ML screening). Data in USD from January 2013 to March 2018.

Source: MSCI, Datastream, Sustainalytics.

Interestingly, the excess return of the sector-matched version is also positive, even if lower in magnitude when compared to the positive ML screening. By neutralising the sector component (because their weights are the same in the benchmark), the outperformance essentially comes from the stock-picking.

Table 11

| | | Excess return | | | |
|---|---|---|---|---|---|
| | | Machine learning screening | | | ESG best-in-class |
| Year | Bench | Positive | Positive sect matched | Negative | 30% |
| 2013 | 23.95% | 0.72% | 0.12% | –1.39% | –0.14% |
| 2014 | 4.97% | 3.88% | 3.65% | –2.75% | –0.17% |
| 2015 | –0.89% | 3.79% | 1.89% | –5.30% | 0.07% |
| 2016 | 7.40% | –2.14% | –1.91% | 3.73% | –0.44% |
| 2017 | 22.44% | 3.82% | 0.61% | –2.57% | –0.02% |
| 2018 | –1.37% | 3.92% | 2.24% | –1.63% | –0.24% |

Calendar year performance for the MSCI World Index (bench) and the excess returns for the positive, positive sector-matched and negative ML screening as well as for the ESG 30% best-in-class portfolio. Data are shown in USD from January 2013 to March 2018.

Source: MSCI, Datastream, Sustainalytics.

For the negative ML screening, the excess return is always negative except for 2016. Finally, the best-in-class portfolio shows almost systematically small but negative excess returns, except in 2015 when it managed to outperform by 0.07%. Once again, our findings confirm that for very large and diversified universes, the

simple ESG filtering does not bring alpha, although it does not significantly reduce the performance with the best-in-class approach.

Table 12 collects the results of standard factor regressions of the portfolios over the classic Fama and French four-factor model. The positive ML strategy delivers strong, positive and statistically significant alpha, with no specific exposure to size or momentum factors, while it is slightly exposed to the growth factor and a market beta close to one. The same behaviour exists for the sector-matched version, although the resulting alpha is now smaller. $R^2$ are close to one for all specifications. Evidence from these factor exposures point to the strong picking ability of the ML algorithm, as the outperformance is not coming from unintended factor bets.

Table 12

| | Machine learning screening | | | ESG best-in-class |
|---|---|---|---|---|
| | Positive | Positive sect matched | Negative | 30% |
| Ann alpha | 2.041%** | 1.128%* | –0.142% | –1.088% |
| Market beta | 1.006*** | 0.989*** | 1.003*** | 0.99*** |
| Size | –0.052 | 0.043 | –0.029** | 0.095 |
| Value | –0.148*** | –0.096** | 0.008 | 0.163** |
| Momentum | 0.042 | 0.009 | –0.002 | –0.078 |
| $R^2$ | 96.64% | 97.77% | 99.83% | 93.32% |

Four-factor model regressions. Returns are sampled at a monthly frequency. We have replaced Kenneth French's Market Factor (MKT) with the MSCI World Index, to force a more intuitive market factor for these strategies. Stars refer to statistical significance: *** = 99% significant, ** 95% significant, * = 90% significant, no star = not significant. Data from January 2013 to March 2018.

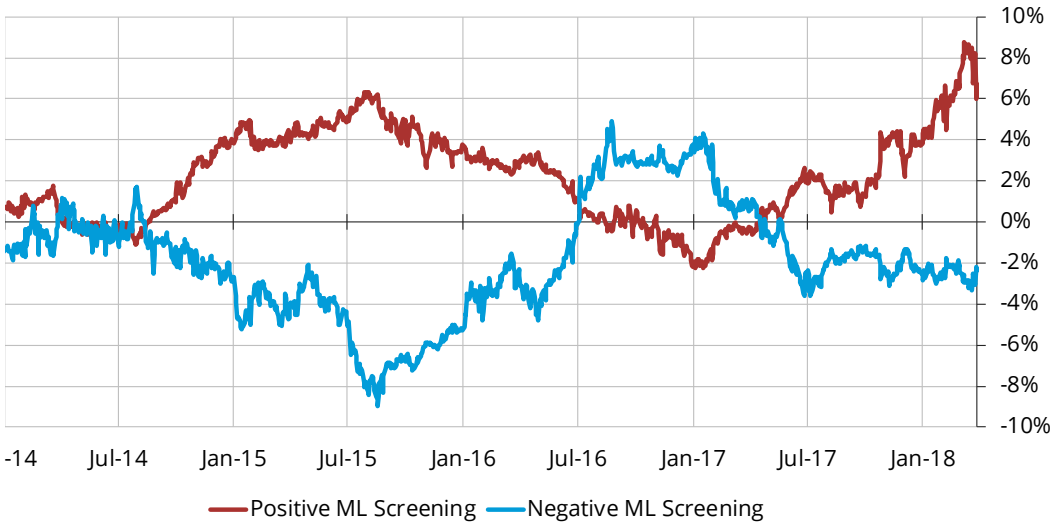
Source: MSCI, Datastream, Sustainalytics, Kenneth French's website.

Finally, Figure 2 shows the one-year excess return of the ML strategies against the benchmark. The positive ML screening delivers consistent positive excess return over time, relatively regular, excepted between Q2 and Q4 2016. At the same time, the negative ML screening shows consistently negative and highly volatile excess returns over time. Said otherwise, the ML algorithm achieves its objective of identifying stocks, from their ESG profile, that are indeed able to deliver superior returns.

Simulated one-year rolling excess returns of positive ML screening and the negative ML screening over the benchmark MSCI World Index

Figure 2

Percentage in USD




Data in USD from January 2014 to March 2018.

Source: MSCI, Datastream, Sustainalytics.

**The effects of learning.** The machine learning algorithm is initially trained over three years of data and then updated yearly. During these regular updates, the algorithm learns from the new flow of data it can access: it can test its rules to confirm, nuance or remove some of them, and selects new rules linked to statistically significant patterns. This learning process is key in the final performance of the model (and for the positive ML screening portfolio built upon it). To measure this effect, we form four portfolios named LEARNING Y, where Y = 2012, 2013, 2014, 2015 as follows:

- For each year Y, we consider the set of rules related to the learning at the end of the year Y.
- We calculate the scores for all stocks in the universe from the end of year Y to March 2018 with this set of rules.
- LEARNING Y is built as positive ML screening, except that the underlying scores are calculated with the same, not updated set of rules calibrated at the end of year Y.

LEARNING Y uses a unique, static set of rules that is never updated (no learning).
By construction, the portfolios positive ML screening and LEARNING Y coincide over the period *1 January, Y+1* to *31 December Y+1*, because, over this period, they use the same set of rules (hence the same scores) to screen the investment universe. Figure 3 shows the calendar excess returns of these portfolios together with the positive ML screening portfolio over the benchmark MSCI World Index.
Since we only show out-of-sample results, the time frame of each LEARNING Y portfolio is different. In most cases, we see that positive ML screening outperforms the LEARNING Y portfolios after the first year (since they are the same on the first year). Indeed, the excess return for the LEARNING Y portfolios usually shrinks to zero and becomes even negative over time. In other words, the predictive power of the scores vanishes over time, so that it is important to train the algorithm on the new observed data to update the set of rules.

Calendar excess returns of the positive ML screening and the four portfolios

Figure 3

Percentage in USD


Calendar excess returns of the positive ML screening and the four portfolios LEARNING 2012, LEARNING 2013, LEARNING 2014 and LEARNING 2015 over the MSCI World Index. Data are shown in USD from January 2013 to March 2018.

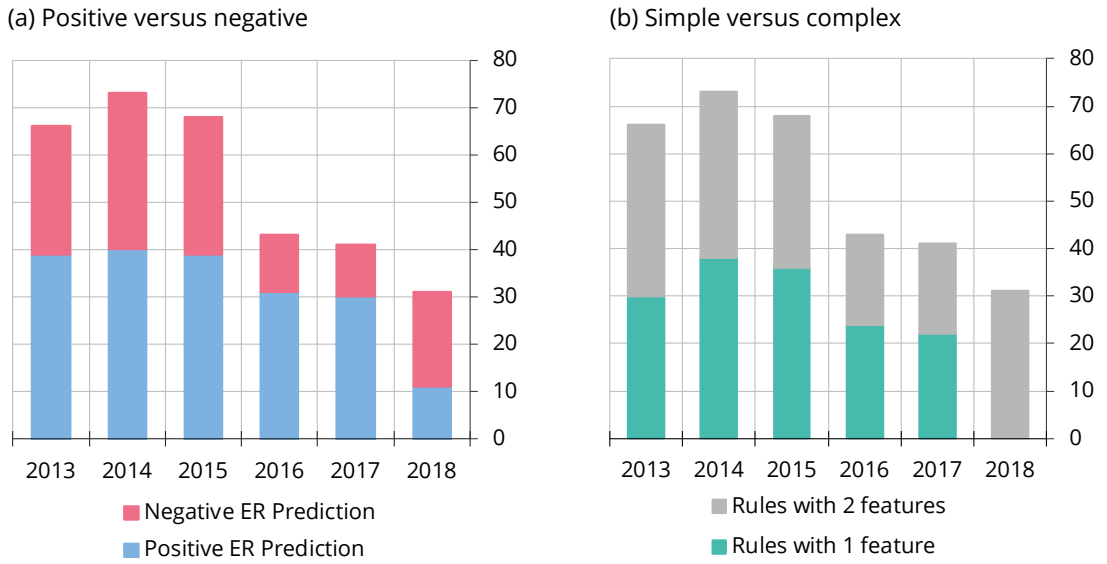
Source: MSCI, Datastream, Sustainalytics.

The number of rules used by the algorithm changes over time: as shown in Panel (a) of Figure 4, this number evolves in the range [31,73] with the split between positive rules (ie rules related to positive predictions of the excess return) and negative ones also changing over time. Interestingly, the number of rules related to negative excess return increased from 12 in 2016 to 20 in the latest 2018 learning. Panel (b) of Figure 4 shows the same number of rules split between simple rules (ie those that only make use of one feature) and complex rules (ie those that use two features, as the examples shown in Table 9). Both Figures 3–4 suggest that, to extract alpha from the ESG

features, one needs to regularly update the algorithm and consider a newly created set of rules to detect patterns between ESG profiles and financial performance.

Number of rules at each update of the algorithm

Figure 4




Number of rules at each update of the algorithm: (a) the split bsetween rules that predict positive or negative excess returns (ER); (b) the split between rules that make use of one feature (simple) or two features (complex).

## 6. Conclusion

The last few years have seen increasing interest in ESG investing and the integration of socially responsible principles at the portfolio construction level. Managers and investors are asked to complement pure financial objectives with non-traditional financial ones.

Our study brings some new ideas and insights into the way investors could achieve ESG objectives in their investments. The literature on the theme is mixed: initial studies were mostly sceptical of the benefit of ESG integration into the portfolio. Over time the mindset has evolved, and several studies have empirically proved that ESG integration in the portfolio does not lower performance. Most recently, the financial literature has gone one step further and claim that, indeed, ESG integration is a way to extract alpha or, at least, to reduce risks. We recognise the need for serious integration of ESG objectives alongside classic financial ones, and the existence of an economic link between the ESG profile of a company and its financial performance over the long run. Nevertheless, we tend to agree with the pioneers of ESG research, which assert that, at best, ESG integration does not significantly degrade financial performance, especially for large and diversified investment universes.

ESG profiles can impact financial performance in a non-linear way, and the impact can depend on the sector, the country or other specific characteristics of each company. Thus, we designed and implemented a sophisticated machine learning algorithm that identifies patterns between ESG profiles and performance, and is statistically robust across the universe and over time.

The algorithm produces a set of rules, each of which identifies a region in the high-dimensional space of the ESG features, conditionally on which we can make a prediction on the stock's excess return. All the predictions are finally aggregated and transformed into a score taking values in $\{-1, 0, +1\}$, so that in the end we can effectively look at the sign of the excess return rather than its magnitude.

With this algorithm, trained over time to remain updated, we empirically proved that the link between ESG profiles and financial performance exists, but that it can only be accessed with non-linear techniques. Indeed, a simple strategy that selects stocks whose scores are positive significantly outperforms the well known ESG best-in-class approach.

# References


Allouche, J and P Laroche (2005): "A meta-analytic investigation of the relationship between corporate social and financial performance", *Revue de gestion des ressources humaines*, no 57, pp 18.

Asness, C (2017): "Virtue is its own reward: or, one man's ceiling is another man's floor", *AQR Blog,* https://www.aqr.com/Insights/Perspectives/Virtue-is-its-Own-Reward-Or-One-Mans-Ceiling-is-Another-Mans-Floor.

Aupperle, K, A Carroll and J Hatfield (1985): "An empirical examination of the relationship between corporate social responsibility and profitability", *Academy of Management Journal*, vol 28, no 2, pp 446–63.

Bragdon, J and J Marlin (1972): "Is pollution profitable?", *Risk Management*, vol 19, no 4, pp 9–18.

Capelle-Blancard, G and S Monjon (2012): "Trends in the literature on socially responsible investment: looking for the keys under the lamppost", *Business Ethics: A European Review*, vol 21, no 3, pp 239–50.

Cesa-Bianchi, N and G Lugosi (2006): *Prediction, learning and games*, Cambridge University Press.

Chong, J and G Phillips (2016): "ESG investing: a simple approach", *The Journal of Wealth Management*, vol 19, no 2, fall, pp 73–88.

Devaine, M, P Gaillard, Y Goude and G Stoltz (2013): "Forecasting electricity consumption by aggregating specialized experts", *Machine Learning*, vol 90, no 2, pp 231–60.

Filbeck, G, H Holzhauer and X Zhao (2014): "Using social responsibility ratings to outperform the market: evidence from long-only and active-extension investment strategies", *The Journal of Investing*, vol 23, no 1, spring, pp 79–96.

Friede, G, T Bush and A Bassen (2015): "ESG and financial performance: aggregated evidence from more than 2000 empirical studies", *Journal of Sustainable Finance & Investment*, vol 5, no 4, pp 210–33.

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

Giese, G, A Ossen and S Bacon (2016): "ESG as a performance factor for smart beta indexes", *The Journal of Index Investing*, vol 7, no 3, winter, pp 7–20.

Griffin, J and J Mahon (1997): "The corporate social performance and corporate financial performance debate: twenty-five years of incomparable research", *Business & Society*, vol 36, no 1, pp 5–31.

Humphrey, J and D Tan (2014): "Does it really hurt to be responsible?", *Journal of Business Ethics*, vol 122, no 3, pp 375–86.

Indrani, D and M Clayman (2015): "The benefits of socially responsible investing: an active manager's perspective", *The Journal of Investing*, vol 24, no 4, winter, pp 49–72.

Kurtz, L and D Di Bartolomeo (2011): "The long-term performance of a social investment universe", *The Journal of Investing*, vol 20, no 3, fall, pp 95–102.

Margolis, J, H Elfenbein and J Walsh (2009): "Does it pay to be good and does it matter? A meta-analysis of the relationship between corporate social and financial performance", https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1866371.

Peiris, D and J Evans (2010): "The relationship between environmental social governance factors and US stock performance", *The Journal of Investing*, vol 19, no 3, fall, pp 104–12.

Nemirovski, A (2000): "Topics in nonparametric", *Ecole d'Eté de Probabilités de Saint-Flour*, vol 28, pp 85.

Orlitzky, M, F Schmidt, S and Rynes (2003): "Corporate social and financial performance: a meta-analysis", *Organization Studies*, vol 24, no 3, pp 403–41.

Revelli, C and J Viviani (2015): "Financial performance of socially responsible investing (SRI): What have we learned? A meta-analysis", *Business Ethics: A European Review*, vol 24, no 2, pp 158–85.

Shiller, R (2013): "Capitalism and financial innovation", *Financial Analysts Journal*, vol 69, no 1.

Stoltz, G (2010): "Agrégation séquentielle de prédicteurs: méthodologie générale et applications à la prévision de la qualité de l'air et celle de la consommation électrique," *Journal de la Société Française de Statistique*, vol 151, no 2, pp 66–106.

Tsybakov, A (2003): "Optimal rates of aggregation", in B Schölkopf and M Warmuth (eds) *Learning theory and kernel machines*, pp 303–13.Van Beurden, P and T Gössling (2008): "The worth of values – A literature review on the relation between corporate social and financial performance", *Journal of Business Ethics*, vol 82, no 2, pp 407–24.

Van Duuren, E, A Plantinga and B Scholtens (2016): "ESG integration and the investment management process: fundamental investing reinvented," *Journal of Business Ethics*, vol 138, no 3, pp 525–33.

Wu, M L (2006): "Corporate social performance, corporate financial performance, and firm size: a meta-analysis", *Journal of American Academy of Business*, vol 8, no 1, pp 163–71.

Zoltan, N, A Kassam and L E Lee (2016): "Can ESG add alpha? An analysis of ESG tilt and momentum strategies", *The Journal of Investing*, vol 25, no 2, summer, pp 113-24.