Rule Extraction from Large Numerical Datasets: Construction of ESG Signals

December 2016

About Sustainalytics

Sustainalytics is an independent ESG and corporate governance research, ratings and analysis firm supporting investors around the world with the development and implementation of responsible investment strategies. With 14 offices globally, Sustainalytics partners with institutional investors who integrate environmental, social and governance information and assessments into their investment processes. Today, the firm has more than 300 staff members, including 170 analysts with varied multidisciplinary expertise of more than 40 sectors. Through the IRRI survey, investors selected Sustainalytics as the best independent responsible investment research firm for three consecutive years, 2012 through 2014 and in 2015, Sustainalytics was named among the top three firms for both ESG and Corporate Governance research. For more information, visit www.sustainalytics.com.

About Advestis

Advestis is a Paris-based FinTech that specializes in machine learning and big data techniques for asset management firms. Founded in 2011 by Christopher Geissler, Advestis employs four full-time professionals and is backed by three senior members of its Scientific Advisory board. Geissler is a financial data scientist with more than 30 years of experience in quantitative finance and machine learning. The firm invests more than 75 percent of its revenues in research and development, and has been awarded the 'Innovating Fintech' label by Finance Innovation for its work with Sustainalytics on ESG Signals. Advestis' capital is owned primarily by the founder, members of its Scientific Advisory board, and Quinten, a Paris-based data science company operating primarily in the healthcare and insurance sectors. For more information, visit www.advestis.com/en/.

Table of Contents

1. MOTIVATIONS FOR THE PROJECT

The investment industry has not waited for the Big Data era to exploit price patterns hidden in historical data. However, the explosion of computational power over the past decade, along with an intense focus on machine learning, has created a completely new framework for data analysis. Today, the proliferation of data sources has become an opportunity for organizations.

At the beginning of the 21st century, machine learning as well as environmental, social, and governance (ESG) research was starting to be used more frequently by asset managers worldwide. However, synergies between the two areas had not been effectively explored. To-date, ESG research has primarily been used as part of a qualitative process for risk mitigation, and big data techniques have not been applied to large sets of ESG information. Yet, over the last few years, asset managers have been seeking ways to integrate ESG research into their quantitative models to extract additional alpha and beta sources.

Sustainalytics, a leading global provider of ESG and corporate governance research and ratings, recognized a few years ago the promise of machine learning. The firm realized that by analyzing thousands of correlations between variables over time, machine learning could be applied to ESG research to extract meaningful and unexplored insights. Seeking advancement and innovation in ESG-related investment strategies, Sustainalytics began working with ***Advestis***, a FinTech company that specializes in machine learning and big data techniques for asset management firms.

In 2014, the two firms began working on a proof of concept project restricted to the oil sector. Advestis applied its machine learning technology to the historical ESG data sets provided by Sustainalytics. The early results were considered promising enough to justify a larger scale project. In fact, in June 2016, Advestis earned the label 'Innovating Project' from Finance Innovation, a French competition that boosts research and innovation by bringing the academic and financial fields together.

In 2015, Sustainalytics and Advestis collaborated on a larger project to calculate systematic daily signals on 1,600 companies from developed markets. The firms exploited big data techniques to analyze the interaction between ESG, financial and trading variables, and thus the ESG Signals product emerged.

ESG Signals is a Sustainalytics' service powered by Advestis' technology that applies state-of-the art machine-learning algorithms to a large historical set of ESG scores, incident levels and market data. The main features of this innovative approach are:

- the production of explicit rules associated with out- or under-performance biases,
- the capacity to handle a large number (500+) of explanatory variables,
- the capacity to detect interactions or synergies between variables,
- the capacity to uncover the most influent variables, and
- Ultimately, the production of signals qualifying expected positive ('Opportunity Signal') or negative ('Risk Signal') performance bias for each stock.

2. RULES EXTRACTION PROCESS

2.1 Brief overview of statistical learning methods

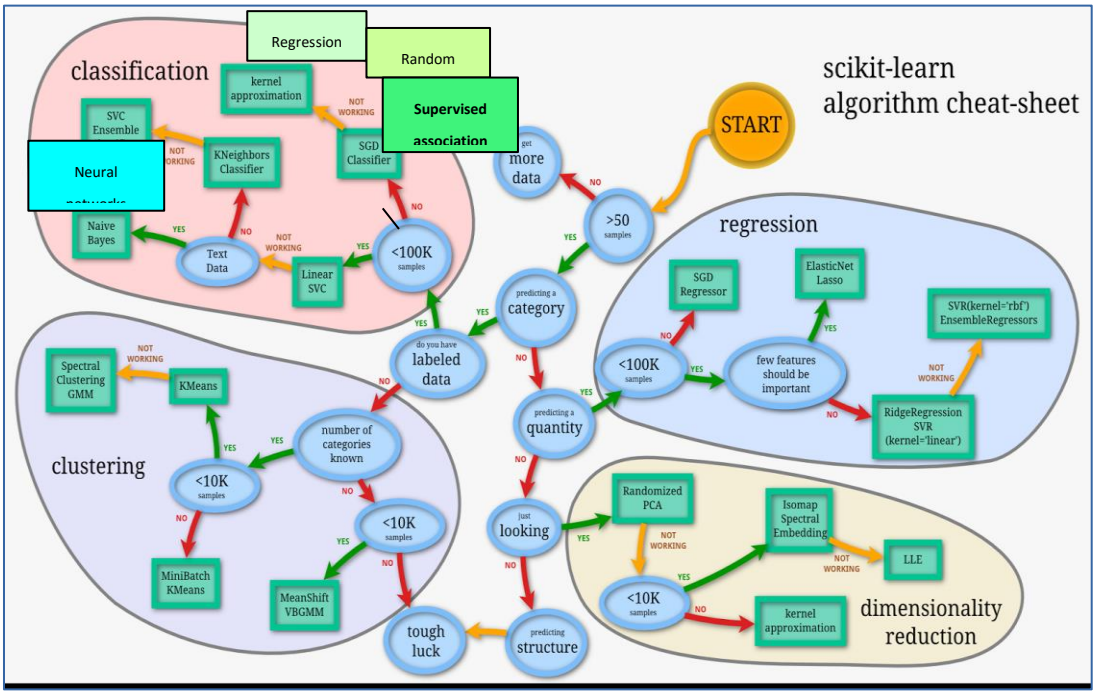
The purpose of statistical learning is to find a relationship between a random quantity Y_t to be forecasted, and a set of observable variables X^1_t, …, X^n_t, based on a sample of historical data. For example, Y_t is the future return of a stock on a three-month period, and the X^i_t are a set of variables observable at time t for this stock: price, financial metrics, ESG scores.

The oldest and best known among statistical learning methods is the linear regression model, in which the forecasted variable is expressed as a linear combination of the observable variables, plus a noise that one tries to minimize.

It is well-known, and mathematically justified, that linear models do not work very well in large dimensions, i.e. when the number of explanatory variables is too large, typically above 20. In other words, it is hopeless to search for a robust linear model in the presence of hundreds of variables. Important progresses have been made towards reducing the dimension (ridge or 'lasso' methods) when a linear dependence between the quantity Y_t to forecast and the explanatory variables can be assumed.

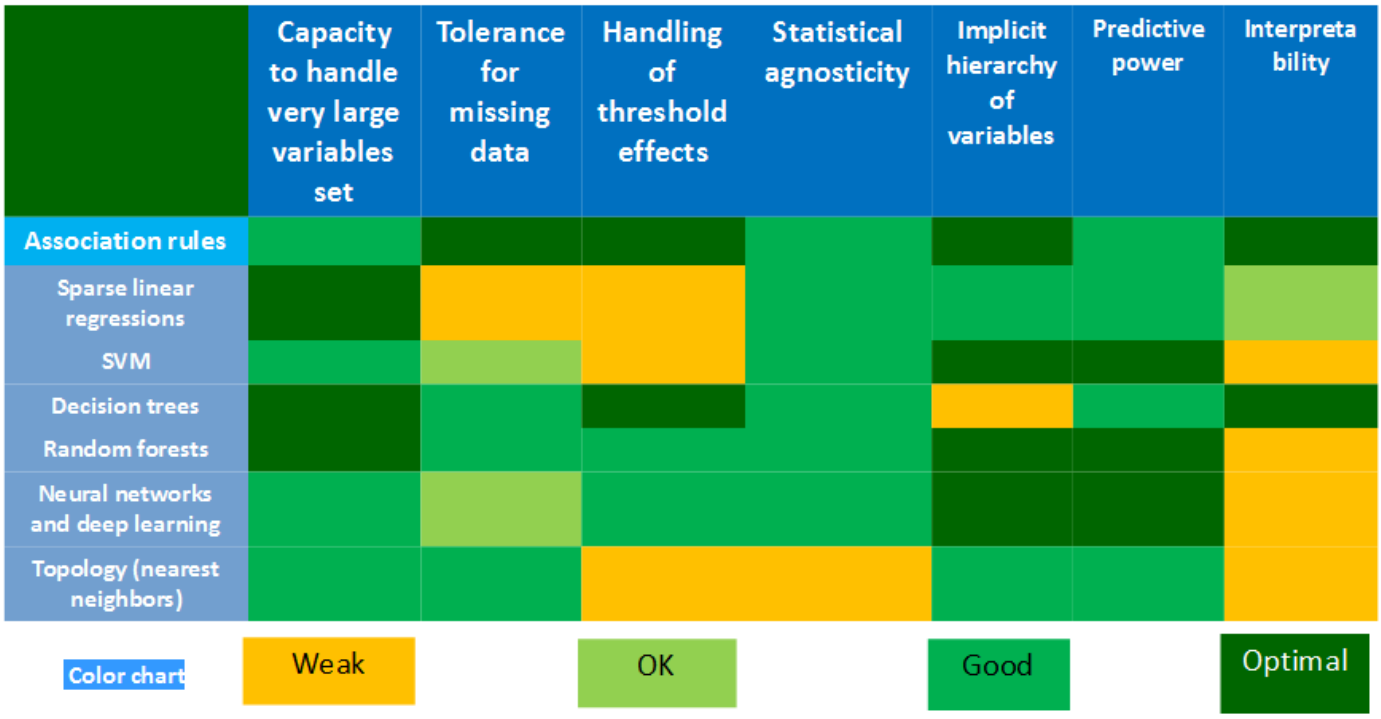
In the early 1990's, even before the so-called age of Big Data, characterized by Volume, Variety and Velocity of data, the limitations of plain linear models led to the emergence of many data-driven learning methods, among which are neural networks and regression trees, which are just two of the most popular techniques. These methods generalize regressions in the sense that they do not rely on a linear dependence between the observed phenomenon and the variables. This generalization is probably more realistic in the field of economy, where dependencies are not linear by nature (e.g. threshold effects).

The following chart (courtesy of Scikit-learn.org) gives a visual and partial summary of the current techniques. Classification and regression are two sides of the same coin depending on the categorical or numerical nature of the explained phenomenon, and constitute the family of supervised learning methods.



2.2 The choice of association rules method

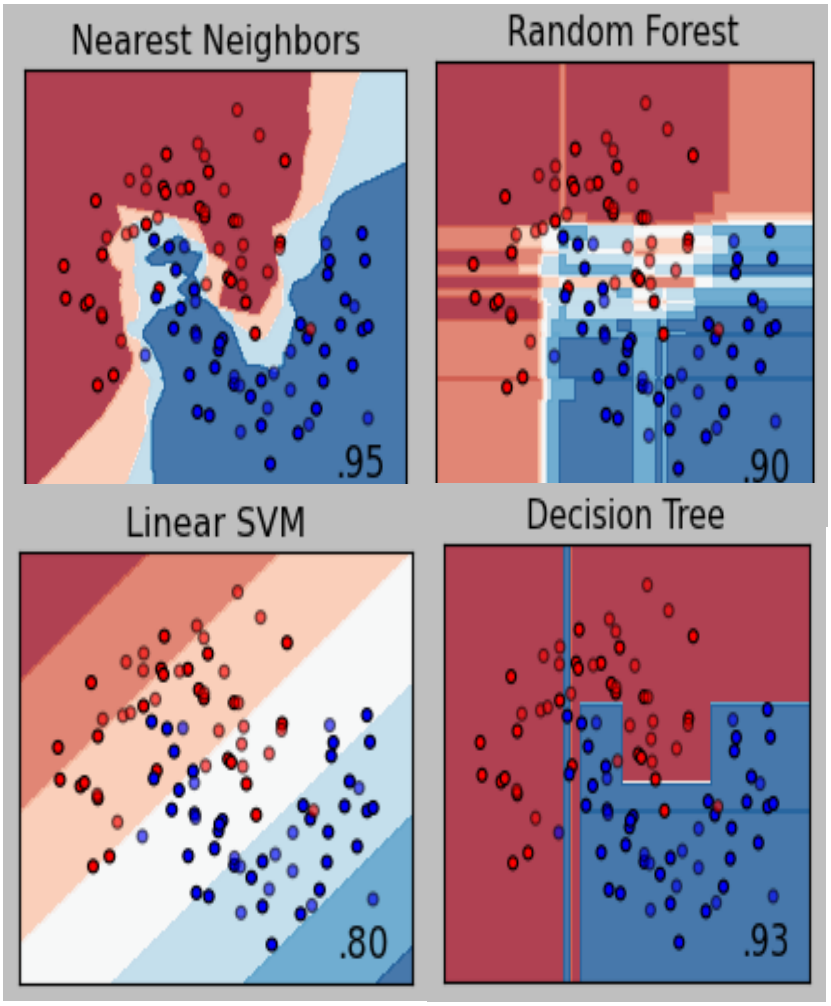
The task of extracting future price information for a large set of co-variables must meet a series of constraints. Here is a benchmark of several techniques with respect to the most important criteria.



Among these criteria, **interpretability** deserves a special mention. It qualifies a 'non-black-box' type of approach, in which the prediction can always be justified by a very limited number of influent variables depending on the context.

The bias toward an interpretable method has been taken from the beginning of ESG Signals project, thus eliminating directly some powerful but 'black-box' methods like random forests or neural networks. This bias is also required in most healthcare applications by the biologists who want to keep a control over the parameters involved in a diagnosis. This is also the case more generally in any domain where decisions (medical treatment, marketing action, investment…) have to be taken based on a small set of variables.

As an illustration, here is the difference of interpretability among four frequently used methods: nearest neighbors, linear support vector machines, classification tree and random forests (courtesy of Scikit-learn.org).



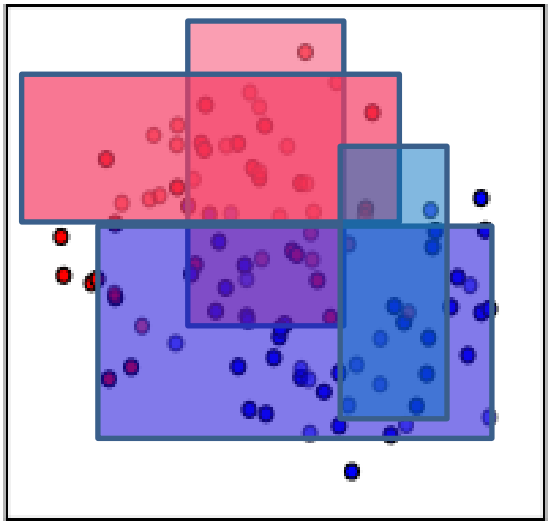
This 'toy problem' is here to correctly predict the location of red and blue dots. The shaded areas indicate the direction and strength of the prediction. The figure in the right lower corner is the rate of good predictions out of sample. The interpretability can be visually understood as the simplicity of the border of shaded areas.

As expected, linear separators are easily interpretable but give poor performances. Conversely, the nearest neighbours method (predicting like close past observations) shows the best results but cannot be described with simple conditions on the variables.

The problem is general in machine-learning: finding a good compromise between parsimony of description and predictive power. That is one of the purposes of association rules.

2.3 Outline of AdLearn ® algorithm

AdLearn ® is a supervised learning algorithm belonging to the family of supervised association rules.
A chart built from the same toy set will give a good understanding of the method:



The algorithm searches for rectangles (or hyper-rectangles in larger dimension) containing a significantly high proportion of either color. The level of significance is measured in terms of p-value, or probability of false discovery, i.e. the probability of attributing a non-zero mean to the distribution of a subgroup of observations, while the actual mean is in fact null.

The main differences with a decision tree are:

- some points can be left unpredicted (associated with no decision)
- some points can belong to two or more rectangles of the same modality.
- some points can be ambiguous in that they belong to rectangles of opposite modalities.

These features are closer to the complex reality of economy, where given phenomenon (e.g. the rise of a stock price) cannot in general be attributed with certainty to a single cause.

Together with this flexibility, the description of a prediction stays interpretable if:

- the dimension of a rectangle, or the number of variables involved in a rectangle, is bounded: for ESG Signals, this dimension has a maximum of three.
- the number of rectangles covering each point: this parameter is controlled by the algorithm.

2.3.1 Parameters settings

The AdLearn ® algorithm has a limited set of input parameters.
The most important ones are:

- The historical period of the training set, i.e. the data used to extract the rules.
- The set of X variables retained as potentially explanatory.
- The choice of the variable Y to predict, and the horizon of prediction (here three months).
- The minimum number of points covered by a rectangle or equivalently formulated, explained by a rule. This floor is generally set at 1% of the total size of the training sample.
- The complexity or maximum number of variables defining a rule: it is set to two or three.

2.3.2 Execution

AdLearn ® is written in Python and uses standard packages like Numpy, Scipy and Pandas.
It runs on Advestis' servers from the latest available data provided by Sustainalytics.
The execution is parallelized and takes between 24h and 48h on a single machine equipped with Intel ® Core i7 4790K and 32 GO of RAM.

None of the software modules making the algorithm includes any randomization. As a consequence, the output is fully deterministic. It is also insensitive to linear scale changes in the data.

The rule sets issued by the AdLearn ® are non-crypted text files. They are installed on the server running the calculation of ESG Signals.

	RuleId	Var1	Min1	Max1
R0(1)+0114		Esg_ANAll!S.4.3 Society & Community Incidents-Weighted Score_M6	3	4
R1(1)+0114		Esg_DeltaTime!G.2.12 Compensation Committee Independence-Weighted Score_M10	8	8
R2(1)+0114		Esg_ANSector!G.2.9 Board Independence-Weighted Score_M10	9	9
R3(1)+0114		Esg_ANPeerG!G.2.9 Board Independence-Weighted Score_M10	9	9
R4(1)+0114		Esg_Delta!G.2.5 ESG Governance-Weighted Score_M9	8	8
R5(1)+0114		Esg_ANSector!S.1.2 Discrimination Policy-Weighted Score_M10	6	6
R6(1)+0114		Esg_ANSector!G.2.10 Audit Committee Independence-Weighted Score_M10	7	7
R7(1)+0114		Esg_ANAll!S.3.3 Customer Incidents-Weighted Score_M7	6	6
R8(1)+0114		Esg_ANPeerG!S.3.2.1 QMS Certifications-Weighted Score_M10	4	4
R9(1)+0114		Esg_ANAll!E.1.12 Operations Incidents-Weighted Score_M9	4	4
R10(1)+0114		Esg_ANPeerG!G.1.5 Business Ethics Incidents-Weighted Score_M10	6	6
R11(1)+0114		Esg_ANSector!S.4.1 Activities in Sensitive Countries-Weighted Score_M10	7	7
R12(1)+0114		Esg_ANSector!G.1.4.3 Animal Testing Policy-Weighted Score_M7	4	5
R13(1)+0114		Esg_Delta!E.1.7.0 GHG Reduction Programme-Weighted Score_M9	6	6
R14(1)+0114		Esg_Delta!E.3.2 Product & Service Incidents-Weighted Score_M6	3	3

3. THE DATA SETS

The set of data used for the learning is relative to the largest 1600 companies in developed countries ('the Universe').

In this section, we give a detailed overview of the primary and derived variables related to ESG Signals.

3.1 Primary data

Primary data are collected as issued by their provider: Sustainalytics for the ESG scores, and Capital IQ for financial data.

3.1.1 ESG data

The data provided by Sustainalytics are the following:

- 147 ESG scores compiled since September 2009 for all companies in the Universe. These scores are typically reviewed on a yearly basis.
- 10 incidents-weighted scores recalculated daily for the whole Universe.
- Carbon data: on a yearly basis, Scope 1- and 2- GHG emissions, Carbon Intensity.

3.1.2 Fundamental data

- History of weekly consensus of earnings per share estimated on a forward 12-month rolling period and converted in USD.
- History of monthly total market capitalization in USD.
- Sector and Industry group.

3.1.3 Market data

- History of daily stock prices adjusted for corporate actions (split, dividends) and converted in USD.
- History of daily prices for 10 sector ETFs corresponding to the 10 GICS level one sectors.
- History of monthly interest rates (10y government bond) of developed countries.

3.2 Derived data

The purpose of creating new variables is to increase the possibilities of discrimination between companies across time. A total of about 800 derived data set is created. Each dataset keeps the daily history of a derived variable since September 2009 for all the stocks in the universe, hence a 1800 days x 1600 stocks matrix. 800 such matrices are stored and submitted to the rule extraction algorithm.
We introduce here the various types of transformations or derivations that are performed on primary datasets.

3.2.1 Normalizing data

As a first step, the level of a given score for a given stock is compared to the levels of the same score for three different groups:

- the whole Universe,
- the companies in the same sector,
- the companies in the same industry (or peer) group

For a stock S, let be G(S) the set of stocks belonging to the same sector as S (resp. to the same peer group or universe).
Let I be an indicator and I(S, t) its level for stock S at time t.
We denote by $m_I(G,t)$ (resp. $\sigma_I(G,t)$)) the average (resp. the standard deviation) of the levels I (S, t) for all stock S belonging to the group G.
The normalized value of indicator I for stock S at time t is thus:
$I_{norm}(S,t) = (I(S,t) - m_I(t,G(S)) / \sigma_I(t,G(S)$

For example, the variable:
Esg_ANSector!S.4.2.6 Equitable Pricing and Availability-Weighted Score designates the Sector-normalized value of the score **'S.4.2.6 Equitable Pricing and Availability-Weighted Score'** in Sustainalytics' terminology of scores.

The '**Esg'** prefix classifies the variable as being linked to ESG scores.
The '**ANSector'** prefix reminds the fact that the value is normalized with respect the stocks of the same sector.

The purpose of normalizing the data with respect to a subgroup of companies is the ability to compare companies belonging to different sectors by getting rid of the biases affecting these sectors. For example, banks enjoy generally better environmental scores than mining companies. Having said this, one is not yet able to perform a relevant comparison between a bank and a mining company on environmental criteria.

3.2.2 Time variations

- For each indicator, the following derived variables are calculated:
- The time elapsed since the last indicator change: the corresponding variables are prefixed with 'DeltaTime',
- The variation of the indicator over the last year: the corresponding variables are prefixed with 'Delta'.

3.2.3 Financial derived variables

Financial inputs include price and estimated forward earnings. The following transformations are performed to discriminate between stocks.

- The **annualized volatility σ(S,t)** of daily returns for each stock.
  The higher the volatility, the riskier the stock.
- The **yield Y(S,t)** equal for each stock S to the ratio of the estimated earnings at time t, to the price at time t.
  The higher the yield, the cheaper is the stock compared to its earnings perspectives.

- The **excess return XR (S,t)** equal to the difference Y(S,t) - RF (Currency(S), t) between the yield and the risk-free rate in the denomination currency of S.
  The higher the excess return, the cheaper is the stock regardless of its currency.
- The **risk premium RP (S, t)** equal for each stock S to the ratio XR(S,t) / σ(S,t)
  where σ(S,t) is the volatility of stock S at time t.
  The higher the risk premium, the better the risk remunerated by the expected return.
- The reduced momentum or **price Z-Score**
  Let:
  MAvg(S,t, 6m)
  be the six-month moving average of price P(S,t),

  R (S,t) = P(S,t) - MAvg (S,t,6m)
  be the price residue, or the difference between the price P(S,t) and its six-month moving average.

  The Z-Score is thus defined by: Z(S,t) = R(S,t) / Stdev(R(S,t))
  This quantity is dimensionless. It expresses the distance between the market price and its average in number of standard deviations.

All derived variables are submitted to the normalization described above.

3.2.4 Sector proxies

The 10 major sectors are represented by their respective price index $PSect_i(t)$ for i = 1,...,10.
For each stock S, 10 correlations are calculated at every date t:

$Corr_i(S,t) = Corr(\delta P(S,t) , \delta PSect_i(t))$

where 'δ' operator means daily variation, and Corr the rolling six-month correlation operator.
The higher the correlation with a sector, the closer the stock is 'viewed by the market' to this sector.
This numerical measure of the degree of 'belonging-ness' to a sector allows for more flexibility when analyzing a company, rather than a mere category like 'Energy' or 'Consumer Staples'.

3.2.5 Making the variables discrete

To reduce artificial mining and risk of overfitting, all the variables are discretized.
The values of a variable (e.g. a score) are divided in a maximum of k = 10 'buckets' collecting each approximately 1/k of the total sample, with no bucket having less than 2% of the population.
If the distribution of the variable has less than 10 modalities (e.g. Neutral / Medium / High), the variable will be discretized with the greatest possible number of bins.
The name of a discretized variable has the number of modalities as a suffix, e.g.:

Esg_Delta!S.3.1.7 Conflict of Interest Policy-Weighted Score_M7,

the value of this score being spread across seven different modalities between **0** (lowest end of the range) and **6** (highest end of the range).

4. RESULTS

4.1 Rule sets overview

The output of the algorithm consists in a **set of rules**, or conditions relating to one, two or three variables. Each rule is associated with a positive or negative sign corresponding to its behaviour on the training set. A positive rule is thus a condition associated with a performance, which is positive on average for those stocks meeting the condition expressed by the rule.

Each rule is interesting in that it expresses a **stylized fact** observed on a significant period of time and for a large number of stocks.

Example:
Negative rule of complexity 1 (univariate) taken from the rule set trained on the data between September 2009 and December 2013.

The natural language description of this rule is:
WHEN Governance Incidents relative to universe IS EQUAL OR UNDER 2.0/6.0 THEN: SIGNAL IS ON -RISK-

Its mathematical definition is as follows:

For all stock S and date t, **if the following condition is met:**
Esg_ANAll!G.2.13 Governance Incidents-Weighted Score_M6 (S,t) <= 1
Then:
The expectation of return for the stock S over the next 3 months is **negative.**

The 'equal or under 2/6' condition means 'in the lowest 33% of the distribution' of this particular variable. The convention for scores is that 0 is associated with the worst practice. This rule qualifies companies having their Governance incidents score among the worst 33%. The statistical under-performance associated with this rule is coherent with intuitive economic expectations.

4.2 Rule reports

Once the rules have been extracted for a given historical period, they remain unchanged until the next training.
It is possible to illustrate the repartition of the stocks activating the rule (i.e. meeting at certain dates the conditions defining the rules). The repartition is done on several numerical (market capitalization, volatility, ESG scores) or categorical (country, sector) criteria.

Here is an example of a report issued about a single rule:

Rules Reporting

July 11, 2016

SUSTAINALYTICS advestis

WHEN Operations Incidents relative to sector IS EQUAL OR UNDER 5.0/10.0 AND Business Ethics Incidents relative to sector IS EQUAL OR UNDER 6.0/10.0 THEN: SIGNAL IS ON -RISK-

	Rule	% of total	Index	Rule vs index
Under 4438.12	1.03 %	3.07 %	3.46 %	-11.44 %
4438.12 to 7229.98	1.79 %	5.33 %	6.04 %	-11.69 %
7229.98 to 11848.25	2.98 %	8.87 %	9.56 %	-7.22 %
11848.25 to 24280.15	5.56 %	16.56 %	17.29 %	-4.26 %
Over 24280.15	22.23 %	66.17 %	63.65 %	3.97 %
Total	34.00 %	100.00 %	100.00 %	-31.00 %

Table 1: MarketCap

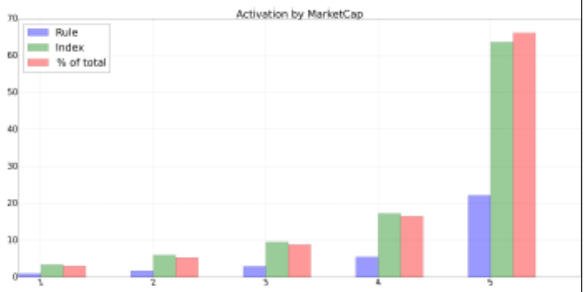
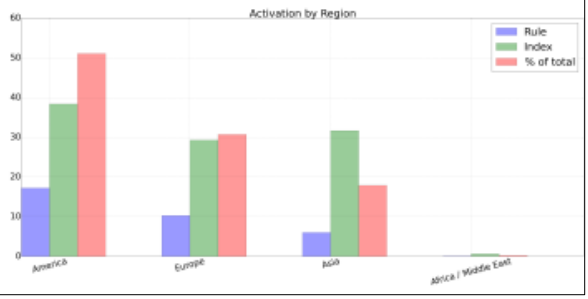


Table 2: MarketCap

	Rule	% of total	Index	Rule vs index
America	17.19 %	51.08 %	38.36 %	33.15 %
Europe	10.33 %	30.70 %	29.29 %	4.81 %
Asia	6.03 %	17.92 %	31.68 %	-43.43 %
Africa / Middle East	0.10 %	0.30 %	0.66 %	-55.10 %
Total	34.00 %	100.00 %	100.00 %	-61.00 %

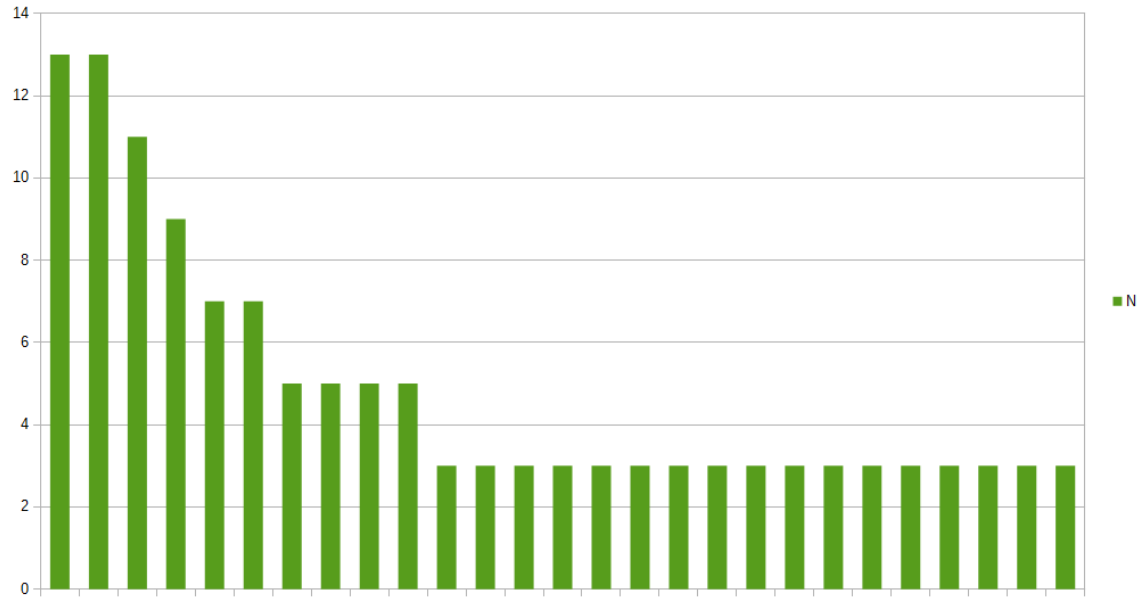


Information is also conveyed by the variables involved in a set of rules.
Here is an example, taken from the training set completed in December 2014.

Esg_ANAll!G.2.9 Board Independence-Weighted Score_M6,13
Esg_ANAll!G.1.5 Business Ethics Incidents-Weighted Score_M6,13
Esg_ANSector!E.1.12 Operations Incidents-Weighted Score_M10,11
Esg_ANPeerG!G.2.12 Compensation Committee Independence-Weighted Score_M10,9
Esg_ANAll!Quantitative Performance Score_M10,7
Esg_ANPeerG!G.2.13 Governance Incidents-Weighted Score_M10,7
Esg_ANSector!G.3.1 Political Involvement Policy-Weighted Score_M10,5
Esg_ANAll!S.4.3 Society & Community Incidents-Weighted Score_M6,5
Fin_Corr!MSCI World Information Technology Index UCITS ETF _M10,5
Esg_ANSector!G.2.9 Board Independence-Weighted Score_M10,5
Esg_ANPeerG!G.1.5 Business Ethics Incidents-Weighted Score_M10,3
Esg_DeltaTime!E.1.12 Operations Incidents-Weighted Score_M10,3
Esg_ANAll!E.1.12 Operations Incidents-Weighted Score_M9,3
Esg_ANSector!S.1.5 Employee Turnover Rate-Weighted Score_M9,3
Esg_ANPeerG!Quantitative Performance Score_M10,3
Esg_ANSector!G.1.4.3 Animal Testing Policy-Weighted Score_M7,3
Esg_ANPeerG!G.2.10 Audit Committee Independence-Weighted Score_M10,3
Esg_ANPeerG!G.1.4.3 Animal Testing Policy-Weighted Score_M9,3

Esg_ANPeerG!S.1.7 Employee Incidents-Weighted Score_M10,3
Esg_ANSector!S.4.2.11 Community Development Programmes-Weighted Score_M10,3
Esg_ANAll!E.2.2 Environmental Supply Chain Incidents-Weighted Score_M5,3
Esg_ANSector!G.2.3 Board Remuneration Disclosure-Weighted Score_M10,3
Esg_ANSector!Quantitative Performance Score_M10,3
Esg_ANPeerG!S.4.2.11 Community Development Programmes-Weighted Score_M10,3
Fin_Vol!NDevAll_M10,3
Esg_ANSector!S.1.1 Freedom of Association Policy-Weighted Score_M9,3
Esg_ANSector!S.3.3 Customer Incidents-Weighted Score_M10,3

The integer shown at the right of each variable is the total number of occurrences of each variable across the various rules.



Out of the initial set of 800 variables, only the 27 listed above have finally been selected as locally influent on forward returns. Out of this list, 25 variables are related to ESG scores, only two to financial data.

4.3 Back-test of the method

A frequent and justified objection against statistical learning is the possible instability over time, and non-reproducibility of historical findings.

The adopted strategy to objectively assess the stability of statistical rules is the out-of-sample simulation of their systematic application in the management of a stocks portfolio. Advestis has used its financial expertise to produce reliable simulations of several investment strategies derived from the rules.

4.3.1 Training and test periods

Every set of rules is associated with a training period. To prevent 'forward-trading', any simulation using rules must not start before the end of the training period.

The duration of validity of a rule set is an essentially empirical measure. Being based on past information, findings represented by rules cannot be considered eternal. The time to peremption is linked to changes in the perception by the market of the facts underlying the explanatory variables.

For ESG variables, reassessing their influence by renewing the learnings every quarter is a prudent attitude in the situation of real portfolio management.

In the simulations, trainings have been renewed on a yearly basis.
The following training generations have thus been created:

Training start	**Training end**	**Simulation start**	**Test end**
2009 Sep 15th	2012 Dec 31st	2013 Jan 1st	2013 Dec 1st
2009 Sep 15th	2013 Dec 31st	2014 Jan 1st	2014 Dec 1st
2009 Sep 15th	2014 Dec 31st	2015 Jan 1st	2015 Dec 1st
2009 Sep 15th	2015 Dec 31st	2016 Jan 1st	2016 Oct 30th

4.3.2 Signal calculation

Let us use the following notations:

S is a stock,
t is a date,
R_i is a rule,
e_i is a constant equal to the expectation of excess return conditional to the activation of rule R_i, estimated on the training set.

The individual signal associated to this rule is defined by:

$$Sig_i(S,t) = sign(e_i) = 1 \text{ if } e_i > 0 \text{ else } -1 \quad \text{if S satisfies } R_i \text{ at date t}$$
$$= 0 \quad \text{if S does not satisfy } R_i \text{ at date t.}$$

For a rule set $\{R_1, \ldots, R_p\}$, the global signal is the sign of the sum of individual signals.

$$Sig(S,t) = Sign\left(\sum_{1 <= i <= p} Sig_i(S,t) \right)$$

4.3.3 Applications of signals to a portfolio strategy

A rule set can be viewed as a signal emitter functioning for every stock and at every date.

The signals can be injected in a portfolio strategy to 'twist' the stock weighting scheme: the signals will generally lead to overweight stocks having positive signals, underweight the stocks having negative signals.

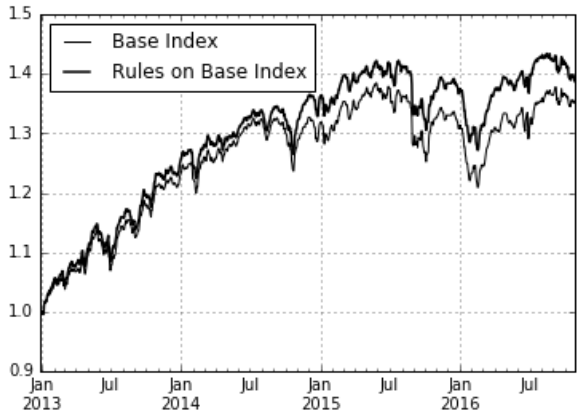
The situation can be more complex in the presence of constraints like minimum diversification, sector weighting, maximum volatility, etc.

It would be out of the scope of this paper to scan even a fraction of the strategies mainly used by fund managers.

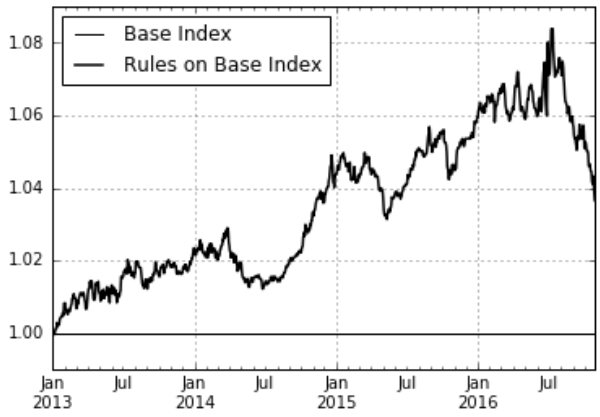
We deliberately chose to test the signals by taking as a benchmark portfolio the market capitalization-weighted set of all stocks in the investable universe.

The strategy used here is :
- doubling the weight in case of a positive signal,
- setting the weight at 0 in case of a negative signal,
- leaving the weight unchanged in case of a neutral signal.

The portfolio rebalancing takes place at the end of each quarter.
The fees used for the simulated transaction costs are set at 0.10% of the volume traded.



Comparison of results with and without application of signals



Excess return of attributable to signals

The risk and return metrics are as follows on the period between 2013 Jan 1st and 2016 Aug 30th:

	Base return	Excess return	Excess volatility	Daily additional turnover	Information ratio	Excess return / excess turnover
Rules on Base Index	9.01%	1.13%	1.40%	0.19%	0.8	2.2%

The excess return chart shows a rather well-distributed performance creation over time.

The figure of excess return as a fraction of the induced turnover is also of interest. Across simulations, excess returns ranging between 1% and 5% of the volume are generally observed.

5. CONCLUSIONS

The use of non-linear, non-global search methods applied to large ESG datasets seem to make possible the extraction of signals having a life expectation compatible with the horizon of long-term fund managers.

The systematic application of these signals to a standard portfolio strategy would have significantly increased the return over the studied period, without altering the risk profile and without creating excessive turnover.

This approach enjoys a high degree of scalability thanks to its low turnover.

By exploiting new sources of information, whose variety cannot be apprehended as a whole by individual investors, this systematic allocation method brings an additional source of performance that is very likely to be uncorrelated with most standard, price-based approaches.

Current research aims at applying clustering techniques both to the stocks and to the set of scores to reduce noise and increase the predictive power of the rules.