The Application of Machine Learning to Sustainable Finance

Erik Allen, PhD
Erik Allen is the Founder and Chief Scientific Officer of Rho AI, which develops targeted, industry-specific software driven by artificial intelligence. He has received a PhD from the Massachusetts Institute of Technology.

Kevin Lyons
Kevin Lyons is Senior Vice President, Data Science, Digital Technology at Nielson Marketing Cloud.

Rodrigo Tavares, PhD
Rodrigo Tavares is the Founder and President of the Granito Group, a global holding company devoted to advancing the impact economy. He received a PhD from the University of Gothenburg, School of Global Studies, and conducted postdoctoral research at Columbia University, School of International and Public Affairs.
Abstract
This article explores the various ways machine learning (ML), one of the applications of artificial intelligence, can be applied to sustainable finance. In the first part of the article we describe the crucial role that ML plays in the financial ecosystem, from managing assets to assessing risks. We also highlight the growing relevance of the global environmental, social, and governance (ESG) market. The second part describes how ML, in combination with big data, can present a robust instrument for ESG data to be assessed in an efficient, standardized, and objective fashion. We argue that ML will eventually enable the identification of materially relevant ESG indicators from the large universe of current ESG metrics, automatically identifying variations by industry, geography, and firm size. However, doing so will require additional standardization of reporting. ML has a role to play in the near-term by providing a set of tools to enable real-time monitoring of sustainability performance.
Introduction

Computer-driven assessment of financial data and human-led assessment of non-financial information are two preponderant trends in financial markets. The first is related to the use of machine learning (ML) in managing financial assets and assessing risks. The latter embodies the integration of environmental, social, and governance (ESG) data into asset management. Both trends started in the 1980s and are showing similar levels of prevalence and acceptance in the industry, but up until now their growth has been independent. We argue that this will not be the case in the future.

Machine learning plays a fundamental role in the financial ecosystem. Fully automated wealth management services (robot-advising) and algorithmic trading or its subtype high-frequency trading (HFT), for instance, are used by financial institutions to optimize financial decisions. Advocates argue that humans come hard-wired with cognitive biases that often lead to suboptimal financial decisions (Kramer, 2016). Unsurprisingly, forecasts indicate that assets under management by robo-advisors will grow to 10 percent of total global assets by 2020 (Kocinaski, 2016). HFT, also a type of algorithmic trading powered by ML, which involves automated quick moves in and out of securities, with the holding period sometimes less than a second, accounted for nearly half of all equity trading in the United States in 2016. This grew from little more than 20 percent in 2006 (Market Watch, 2017). Millennials\(^1\) helped drive the application of artificial intelligence and computerized analysis to finance given their stated preferences for automated interfaces over human interaction, along with their innate trust of online engagements. While algorithmic-driven portfolio management strategies are nothing new, “the merging of these strategies with newer, mobile-friendly, state-of-the-art platforms and consumer-based applications and tools makes robo-advisors innovative and appealing” (Roche, 2017). ML can be used both to extrapolate data (e.g., inferring missing data by detecting patterns in an incomplete data set) and to predict future financial performance (e.g., predicting future financial ratios or relative stock performance based on underlying company and industry metrics). The application of machine learning in financial activities is a solid trend in capital markets.

Responsible investing, or the integration of ESG data into asset management, is also a solid trend. Its origins can be traced to 19th-century Methodist and Quaker opposition to investment in particular assets like tobacco, or anything related to war or the slave trade. In the second half of the 20th century, the war in Vietnam and South Africa’s apartheid sparked a renewed interest in responsible investing. Money managers avoided the allocation of capital in companies that could profit from war or from regimes with a poor human rights record.

---

\(^1\) Born between 1980 and 2000.
ESG investing began to play a larger role in mainstream investment after the 1980s, when the first ESG-driven asset managers and industry associations appeared. (These include Trillium Asset Management, The Forum for Sustainable and Responsible Investment (US SIF), and Ceres.) Globally, there are now US$22.89 trillion of assets professionally managed under responsible investment strategies, an increase of 25 percent since 2014 (Global Sustainable Investment Alliance, 2017). This is approximately one fourth of the US$71.4 trillion currently under management (BCG, 2016). Such staggering growth derives from three interrelated phenomena:

1. Correlation between ESG and increased returns is fairly well established (Deutsche Bank, 2012; Eccles, Ioannou, and Serafeim, 2012; Allianz Global Investors, 2015; Cambridge Associates and the Global Impact Investing Network, 2017; Mercer and LGT Capital Partners, 2015; Morgan Stanley, 2015; Eccles, Verheyden, and Feiner, 2016; Khan, Serafeim, and Yoon, 2015). A recent study demonstrated that listed companies with good ESG practices show lower stock return volatility in comparison to reference companies – 28.6 percent lower, on average. Meanwhile, the average positive effect on equity return is 6.1 percent higher for companies with good ESG practices (Kumar et al., 2016). Assessing the relationship between ESG and performance in over 2,200 studies from 1970 through 2014, Friede, Busch, and Bassen show that 90 percent demonstrate a positive correlation (2015). Market feedback goes in the same direction: 68 percent of asset owners say the integration of an ESG strategy has significantly improved returns, according to a State Street Global Advisors study (2017).

2. ESG analysis is becoming more refined, expert, and widespread. What started as straightforward negative screening of certain assets is now a robust qualitative and quantitative assessment of equities. This includes the examination of company reports or data provided by third-party ESG/CSR service providers. ESG can also be integrated in fixed-income portfolios by conducting ESG credit rating analysis and assessing issuer-level ESG data. Asset owners and asset managers can integrate ESG data using a plethora of benchmarks, studies, and guidelines.

3. Similar to ML, pressure from younger asset owners is driving ESG integration. A handful of studies show that millennials, more than any previous generation, are prioritizing social and environmental issues when choosing an investing strategy (United States Trust Company, 2017; World Economic Forum, 2013; Deloitte, 2017). For almost half (49 percent) of millennials with a net worth of more than US$1 million, social responsibility is a factor in choosing an investment. This compares to 43 percent of Gen Xers, 34 percent of Baby Boomers, and 27 percent of seniors (Spectrem Group, 2015).

Despite the growth of the ESG industry over the last decade, however, we claim that it will soon reach a plateau if it does not incorporate the same level of automation as
traditional finance. This automation will come when ML, in combination with big data, assesses ESG data in an efficient, standardized, and objective fashion.

**Machine Learning Applied to ESG Investing**

In this section, we discuss the current state of machine learning within ESG investing, the challenges that limit its current use, and the path forward.

**Machine Learning and ESG: Current State and Limitations on Adoption**

Machine learning and ESG investing can be combined to great effect, and some funds have already done this. For example, the Richmond Global Compass Fund utilizes ESG metrics and machine learning as part of a global macro hedge fund. But this kind of combined application of ML and ESG has not achieved widespread adoption. Why not?

The challenge of combining ML and ESG within an investing framework relates to the salience and completeness of underlying data. When compared to traditional financial accounting metrics, ESG reporting is non-standard: Companies voluntarily self-report information that they deem materially relevant. Under this system, cross-company or cross-industry comparisons are difficult, as is finding or developing a consistent data set with which to assess materiality. Machine-learning algorithms that extract complex patterns and signals require complete and historically rich data for research – precisely what is now unavailable in the ESG literature.

The application of ML nonetheless has a vital role to play in ESG-oriented investment. Its foundational impact, though, will not initially be in automated investment decisions, but rather in generating the inputs that allow investment firms to monitor ESG performance with improved granularity and completeness. In fact, we argue that ML represents an appealing way to provide structure to the unstructured ESG data set.

ML, then, can be applied in a number of major ways that build upon each other. In order of development, these applications include:

1) Using natural language processing, speech recognition, and image processing tools based on ML technologies to infer crucial ESG information ahead of, or not available through, data providers;

2) Imputing ESG data for currently unmeasured companies by detecting patterns within the full set of ESG metrics; and

3) In the longer-term, utilizing existing/available high-quality ESG data, together with asset managers’ specific material financial data (often also collected through machine learning), to augment and optimize investment decision making by identifying materially relevant metrics and predicting future financial performance.
Related to this, ML will utilize the same input (ESG and financial data) to model investment risk. Each of these topics is considered below.

**Machine Learning and AI Tools for Real-time Monitoring**

Foundational applications of machine learning and artificial intelligence in recent years have focused on performing automated tasks that are time-consuming, costly, or unpleasant. These applications include, for example, automatic photo classification and object detection (Joel Janai, 2017). Such technologies are well developed and can be used to supplement self-reported ESG data.

**Natural Language Processing.** Natural Language Processing (NLP) comprises tools used to automatically extract information from unstructured text. NLP algorithms can organize and structure knowledge to perform tasks such as: automatic summarization (Mohak Sukhwani, 2017); translation (Jiacheng Zhang, 2017); named-entity recognition (Franck Dernoncourt, 2017); relationship extraction (Makoto Miwa, 2016); sentiment analysis (Liu, 2017); speech recognition (Takaaki Hori, 2017); and topic segmentation (Shafiq Rayhan Joty, 2014).

NLP tools can be used to extract and summarize text from public company 10-Ks that reference ESG metrics and topics. For example, the Sustainability Accounting Standards Board (SASB) tracks company disclosure history leveraging NLP tools to automatically identify text relevant to SASB ESG metric disclosures within financial documents.

Sentiment analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer’s attitude towards a particular topic, product, etc., is positive, negative, or neutral. Sentiment analysis can be utilized to monitor public opinion on companies, specifically related to ESG metrics.

Sentiment analysis works by assigning point values (positive, negative, or indifferent) to words, phrases, sentences, and paragraphs based upon a training corpus. For example, TruValue Labs, a provider of real-time sustainability data analytics, analyzes publicly available texts about companies (including articles, blog posts, and Twitter) to establish real-time sentiment specifically tuned to ESG metrics.

Pluribus Labs provides sentiment analysis for U.S. equities by developing custom dictionaries of search terms. This technology could be extended to focus on ESG data.

**Video and Image Analysis.** Video and image analysis tools focus on two main types of tasks. The first takes account of the entire image, and classifies the scene. The second focuses on extracting objects from within the scene. ClarifAI provides automatic object recognition tools applicable to both images and video. Users submit image and video and
receive responses back documenting the objects identified within the image. To improve the recognition accuracy of specific objects, users can also submit labeled images to train their own object recognition tools.

Enhancing the feasibility of image and video analysis is the concept of transfer learning (Lei Xiao, 2017). Transfer learning enables a developer to utilize a model previously trained to detect one object type, and retrain it to detect something else. Transfer learning generally requires much smaller data sets than the original learning task, and so can be used in situations in which the cost of training data development is high.

Object recognition tools could be used for a variety of ESG estimation tactics. For example, satellite imagery related to production facilities could be analyzed to track objects relevant to environmental goals (trucks entering and exiting a facility, color and size of tailing affected runoff, or production of pollutants from smokestacks). Such tools would supplement quarterly or annual updates from company statements, providing real-time insight into current performance.

Audio Analysis. Audio analysis performs many of the same functions of NLP, but the underlying text is vocalized (Jonathan Malmaud, 2015). Speech-to-text tools like Google Voice can generate a transcript of any spoken words. This text can then be analyzed by the same NLP tools described above, providing an additional source of unstructured information.

ESG information derived from C-level speeches, investor meetings, and interviews could be leveraged in the same way as written text, providing an additional source of unstructured information.

Impute Missing ESG Data

The SASB provides a codified ESG reporting standard of variables that they consider materially relevant. These standards vary across sectors and industries, as defined by the Sustainable Industry Classification System. SASB tracks a company’s disclosure history by categorizing disclosure quality: “No Disclosure,” “Boilerplate,” “Company Tailored,” or “Metrics.” The last of these indicates disclosure consistent with the SASB standard. Current sparse coverage makes subsequent analysis of metrics challenging.

Machine learning can be utilized to impute missing values for companies with incomplete reporting. All of these techniques infer ESG metric values of non-reporting companies from two sources: the known ESG metrics of reporting companies and data that defines the similarity between companies. Fundamentally, imputed values should rely heavily on similar companies – those within the same industry or that use similar ESG-related language. The data to define similarity between two companies can be drawn from:
• Standard industry classifications;
• Financial metrics, including those defining size (e.g. revenue) and profitability (e.g. net margin);
• Unstructured data sources like website text, standard financial disclosure text, and curated company profiles (e.g. Crunchbase).

Simple algorithms for dealing with missing data include deleting records with incomplete data, assigning missing data a value of 0, or assigning missing data an average value. More advanced algorithms include:

• K-nearest neighbors imputation;
• Random decision forests;
• Expectation maximization algorithm;
• Deep learning, especially deep autoencoder neural networks;
• Matrix completion.

Each of these advanced cases can apply either single or multiple imputation processes. In the latter, the imputation is applied multiple times to subsets of the data and then pooled into a consensus imputation.

Selecting the best algorithm depends on the structure of the missing data. Data can be missing in three basic ways: completely at random, or MCAR, in which missing data does not depend on any independent or dependent variables; missing at random, or MAR, in which missing data depends on independent variables; and missing not at random, or MNAR, in which missing data depends on dependent variables.

ESG data is often MNAR, meaning that missing values are likely biased. Companies choose to self-report ESG metrics that are likely above average and do not report metrics that are likely below average. Assigning a value drawn from this distribution would over-estimate ESG compliance. Under MNAR assumptions, the response and the presence (or absence) of response must be modeled as a joint distribution.

**Modeling Investment Return**

As ESG data improves, both through more consistent reporting and imputation techniques, analysis will increasingly evaluate the data’s materiality and provide quantitative predictions of financial and asset performance. ESG may eventually be viewed as supplementing traditional measures of financial performance, providing an orthogonal view of company performance that is predictive of future returns.
Models can either predict future financial performance metrics (revenue, profit, net margin) or future price movements. Future price movements more closely mimic the desired investing target, but are inherently noisier. Financial performance or price movements relative to a well-chosen benchmark may improve modeling performance by removing market trends from the analysis.

Machine learning does not obviate the need to discern causation and correlation, or future conditions and matching current conditions. Given a complex, dynamic system in which competition between players is a significant component, no approach ever guarantees future effectiveness. However, the best practices of machine learning can minimize these concerns. (This advice is not unique to ESG data, but applies to forecasting in time series data generally.)

To mitigate concerns of over-fitting, researchers should split data into training, testing, and validation sets. Models should be trained on the training data, and then their accuracy iteratively improved on the validation data. Validation data is used to minimize over-fitting when selecting parameters for the ML model. One should use the test data only after completing model fitting, including parameter selection; and this should be done only to confirm the model’s accuracy. Iteratively revising the model after measuring accuracy against the test data set can reduce over-fitting.

The standard practice of randomly selecting training, testing, and validation data is difficult for models based on time-series data. In these cases, information about market trends may be implicitly embedded in both training and validation sets. Instead, best practice is to develop a model using a “moving window” approach. That is, divide the data into a series of overlapping training-validation-testing sets and then iteratively train the model on the training data. Use the validation data to avoid over-fitting. Finally, test the model on the testing set. This approach simulates the actual investment activity of repeated decision points based on model training immediately prior to that time.

Outlier detection and removal is an important pre-processing step if the input data is noisy or of variable quality. Tools to identify and remove outliers include:

- Extreme value analysis: identify outliers based on mean and standard deviation (e.g. identifying data points more than 2-3 standard deviations from the mean);
- Proximity methods: identify dependent variables that vary significantly from similar data points within the data set.

After making these design decisions, ML models of increasing complexity can be applied to a problem, though best practice suggests starting with simpler models; progress to more complex models after establishing baseline performance and gaining intuition.
**Materiality.** The first area in which ML might be applied is in determining materiality – which ESG metrics are material to future financial or asset price performance. To directly estimate materiality, researchers can use linear models to predict financial performance. Multiple linear regression accepts a variety of independent variables, including ESG-related and traditional financial metrics, to predict financial performance. Linear regression coefficients can then be evaluated for their direction of impact and statistical significance to identify those factors that are materially relevant to future financial performance.

It is also possible to estimate materiality using more sophisticated models. One could, for example, train a deep-learning algorithm on traditional metrics and assess performance on validation and test data, then carry out the same process with one ESG metric. The metric is material with some level of statistical confidence if its inclusion improves the predictive power of the model.

If many different ESG metrics are under consideration, researchers must take into account the multiple testing problem (Multiple comparisons problem) (Munroe). As an illustrative example: if 20 ESG metrics are estimated to be material at the 95-percent level without correcting for multiple testing, we should expect that, on average, one of those metrics is immaterial. There are well-established methods for correcting for this, including Bonferroni’s correction and multi-armed bandit techniques. Multiple testing requires researchers and investors to collaborate in order to understand the tradeoff between classifying too many metrics as material as opposed to too few.

**Predicting Future Performance.** Machine learning algorithms can also be used to predict future performance. Although linear models are best for determining materiality, they are frequently less accurate than ML models for predicting future performance. Researchers frequently progress through a series of models, from simple to complex, measuring accuracy at each step. A typical progression might entail:

- Multiple linear regression;
- Support vector regression;
- Random forest regression;
- Xgboost;
- Recurrent neural networks (RNN)

The RNN model is a preferred neural network approach to modeling time series data. RNNs are characterized by an internal “memory” state that enables them to consume time series data. They have been used to model financial performance based on traditional financial metrics (J. B. Heaton, 2016)
**Investment Risk.** Machine learning tools can be used to model investment risk in addition to investment return. The techniques and tools are identical except for the predicted variable. Models are retuned to predict a metric of risk rather than a metric of return. Alternatively, a backtesting-based approach can be used in which risk is estimated by implementing a strategy on historical data and obtaining, for example, a Bootstrap variance estimate.

**Conclusion**

Machine learning represents a promising tool for use in ESG investing, but the development of specific ML applications must be matched to the development of the larger ESG data and reporting ecosystem.

Automated investment decisions combining ESG data and ML are currently challenged by non-standardized and incomplete data. Standardization of metrics, as promoted by SASB, will continue to improve the data situation in the long run, but ML has a role to play now. The first role is to leverage tools being developed across industries – including natural language processing and automated image analysis – to generate salient ESG estimators. Second, un-reported ESG data may be estimated via imputation methods, allowing estimate derivation for all companies, regardless of reporting format. In combination, these tools can provide complete coverage and offer real-time monitoring, enabling the cross-company and cross-sector comparisons required for foundational financial analysis.

As the completeness and quality of reported ESG data improve, the industry will evolve towards a standardized set of metrics and reporting formats. As these standardized metrics and formats are developed, investors will deploy ML to provide evidence of materiality, estimate investment risk, predict investment return, and eventually automatically deploy capital. Ultimately, ML will generate automatic investment decisions incorporating ESG factors, just as it does today in traditional finance. The investment groups that now invest in understanding and leveraging ML resources to contextualize and generate ESG data will be best positioned to advance these same technologies of automated investment as ESG data standardizes.
References


