Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241590

Leveraging AI and machine learning for ESG data analysis and sustainable investment decision-making

Xiaolong Zeng[1], Li Zheng[2,4], Chenyang Cui[3]

[1]The University of Queensland, St Lucia QLD 4072, Australia
[2]School of Economics and Management, Kunming University, Yunnan, China
[3]Lund University, Lund, Sweden

[4]rara481846778@gmail.com

Abstract. AI and ML may be used to process large amounts of ESG data to assess the sustainability of a company as well as its ability to generate financial returns. We are exploring the disruptive approach to processing ESG data with applications of AI and ML and will focus on building predictive models using ESG factors for both sustainability and investment performance. The data used in this research will be collected from a wide range of public and private sources. Supervised and unsupervised learning based on downsampling, feature scaling and binning methods will be used to process ESG data. We will also investigate the potential to apply various types of ensemble models, which provide a significant improvement in terms of model robustness and accuracy. Additionally, the paper presents case studies illustrating how demonstrable data enable us to explore causality among financial performance and sustainability factors in the sectors where ESG is of paramount importance. The aim of this approach behind digital transformation of ESG data is to help investors extract deeper financial implications for ESG factors, particularly for building long-term financial returns as well as for making more informed and sustainable investment decisions.

Keywords: ESG Data Analysis, Artificial Intelligence, Machine Learning, Sustainable Investment, Predictive Models.

1. Introduction

The prominence of ESG factors in investment considerations such as portfolio selection and stock valuation have increased tremendously in recent years as compared to traditional investment techniques . Concerns for environmental and social issues have paved the way for organisations to extend their focus beyond the sole profitability of their operations, and to consider the practices through which they create and allocate value effectively . Consequently, the analysis of ESG data for sustainable investment has far exceeded traditional methods, mostly because of the complexity and volume of the resulting datasets . AI and ML are thus the best available tools to accomplish such analysis because they enable sophisticated data processing and predictive analytics that provide high-level outputs. ESG scores are obtained from corporate reporting, independent rating agencies such as Sustainalytics , Refinitiv , and data providers of financial information such as Bloomberg , Reuters , MSCI , and FTSE Russell. Despite the wealth of raw information within this dataset, the main challenge is connected to the inconsistencies and biases that stem from self-disclosed and reported metrics of the companies . Integration with data

DOI: 10.54254/2755-2721/87/20241590

from external sources such as Sustainalytics, Refinitiv, and other financial providers such as Bloomberg and Reuters can solve this vital issue, and increase the robustness and reliability of the analysis. However, best practices in data preprocessing are fundamental to guarantee the robustness of the ML models. Within the pre-learning stage, raw data must be cleaned to remove inconsistencies, fixed for integrity, and consistent numbers with NaN values should be replaced with zeros if possible, and otherwise with mean or median values . After this initial data cleaning phase, normalisation techniques through z-score and min-max scaling can be used to standardize the data, giving the opportunity to process large volumes of data as if they had the same starting point. The availability of low and high values makes this processing more straightforward and improves the convergence of the ML algorithms. In addition to this, feature selection and engineering, which consists in focusing on the most pertinent variables among all available, and reducing the dimensionality of the problems by constructing new features, are known techniques to improve the performances of ML models . For example, a meaningful way to accomplish this is via correlation analysis, principal component analysis (PCA), through an unsupervised technique. As for the correlation analysis, numbers above 0.5 are suggestive, although they should not be interpreted in an absolute manner . Furthermore, although feature engineering helps to capture a feature's intricate relationships through various techniques, PCA delivers a meaningful directive in ESG analysis by reducing the dimension of the problem to a few main components while still identifying the feature space . Because most ESG scenarios are multidimensional, PCA can be used to fit into the geometrical space, and assist in making sense of the results. If achieved, this approach becomes valuable in capturing hidden patterns and structures through a more intuitive data visualisation . Besides, most ML models should be trained to accomplish the stated goal. Another feature of these models is that they establish an input-output relationship, as compared with simple aggregating decision rules, which are completely imprecise. Supervised learning models, based on traditional statistical methods such as linear regression, decision trees, and random forests, enable the modellers to predict future performance based on the knowledge generated with historical data. Some indicators require evaluating the goodness of the modelling process and predictive ability, so the adoption of R-squared, accuracy, and F1-score is necessary for this purpose . It is worth noting that these ML models can be evaluated quantitatively with random selection to form a new benchmark for estimating specific outcomes (validation). Last but by no means least, unsupervised learning techniques constitute an effective complement to supervised models in the ESG analysis due to their remarkable property of extracting hidden patterns and structures from the analysed data without prior information . For example, k-means clustering and PCA can be particularly useful for such purposes. K-means clustering is an unsupervised algorithm to estimate the number of customer segments in the data through a square-error [1]. PCA is a statistical technique for data exploration and reducing the input variables to fewer uncorrelated variables that explain as much variability as possible in the underlying data while stripping off all noise . Therefore, it significantly reduces the complexity of large multidimensional data, and ideally brings the data to a two-dimensional space, a trade-off that ESG researchers must decide on.

2. Data Collection and Preprocessing

2.1. Sources of ESG Data

This ESG data can be derived from the company reports, third-party rating agencies, and financial data providers like Bloomberg Terminal, as well as MSCI ESG Ratings, which likewise provides an array of ESG datasets. However, the richness of ESG data can lead to a lack of data quality, as it varies due to differences in reporting standards, and has the potential to be biased by self-reported metrics. For instance, a company might inflate its progress in the annual report, over reporting the company's public awareness of the sustainability efforts to persuade the reader. An attempt to overcome this is the combination of data from multiple sources to provide a more robust assessment of the financial and sustainability performance. For example, Reuters and Bloomberg Terminal, integrating data from corporate self-reports with that from independent agencies such as Sustainalytics or Refinitiv, can provide more exhaustive information. On the other hand, financial data providers, such as Reuters and

DOI: 10.54254/2755-2721/87/20241590

Bloomberg, provide financial metrics related to ESG [2]. Table 1 below gives an example of how metrics from different sources can be combined to form a more comprehensive view of the sustainability performance of companies and the associated financial risks.

Table 1. Sample ESG Data Table

Company	Bloomberg ESG Score	MSCI ESG Rating	Sustainalytics Score	Refinitiv ESG Score	Revenue (in millions)	Net Income (in millions)
Company A	75	AA	65	70	500	50
Company B	65	BB	75	60	300	20
Company C	80	A	70	75	450	45
Company D	70	BBB	80	65	350	30
Company E	85	AA	60	80	600	60

2.2. Data Cleaning and Normalization

Proper preprocessing is important for building accurate ML models. For example, data cleaning involves removing inconsistencies, correcting any errors, and addressing situational issues such as missing values. Incorrect assumptions about the data may lead to different outcomes. Imputation, outlier detection and other preprocessing techniques are then applied. An example of normalisation is z-score and min-max scaling. These methods of normalisation bring ML models to a comparable baseline in order to accelerate the convergence of the ML algorithms. For example, if ESG scores come from different vendors, and some providers may use a different scale of measurement, normalising the data would be essential prior to the analysis being done based on that data. In this case, the Z score-normalisation method would be applied. For instance, if Currency A score has been transformed to 2.8 after the data analyst applied the Z score-normalisation method, that score would be the same as if it was raw data converted to 2.8. Unlike the Z score-normalisation method, the min-max scaling sets a strict range for the data as one of the inputs needed for this technique would be the largest value in the data set. The main objective of min-max scaling is to reshape the data so that it is comparable. It depends on the purpose of the ML algorithms, and some algorithms are sensitive to scale, so min-max scaling could be more appropriate [3].

2.3. Feature Selection and Engineering

Relevant ESG indicators can be chosen through a correlation analysis and PCA, for instance to build new variables through combination of subsets of the ESG scores, which would be more representative of complex relationships. Feature selection tends to improve model performance reducing dimensionality. It could be based on correlation analysis between most variables and the response variable so as to identify combinations of ESG factors with strongest associations to the financial performance and then to reduce the complexity of features by PCA, which consists in a data transformation into principal components that retain maximum variation, so as to build new features through combination of subsets of the ESG scores (such as environmental impacts and financial data like profit margins would be combined to build a new feature representing financial efficiency of the environmental practices). So, the models will be pushed to learn from the most relevant data [4]

3. Machine Learning Models for ESG Analysis

3.1. Supervised Learning Techniques

Supervised learning algorithms, such as linear regression, decision trees, random forests and boosting trees, are trained on data from financial reports and indices, recommend details, annual reports, ESG scores, etc. These algorithms are designed to get people's risk appetite by selecting historic data with better financial performance than the financial market. By feeding greater financial and ESG data into these algorithms, the models can learn the features of values, companies and sectors significantly linked

DOI: 10.54254/2755-2721/87/20241590

to outperformance. When trained an tested, these algorithms can produce results for evaluation metrics, such as R-squared, accuracy and F1-score to measure the model's performance on the testing dataset. Linear regression is a supervised learning algorithm that predicts continuous or numeric values by examining the historical performance of funds that scored highly on various ESG themes. With linear regression model, we can detect the cause-and-effect relationships between ESG scores and financial returns, specifically highlight the ESG score categories that have more predictive power to drive better financial performance. This can effectively help investors choose their ESG portfolio selection. In contrast, a decision tree can classify the trained dataset into two categories, such as high-risk and low-risk investments. [5] Table 2 shows how supervised learning models can filter ESG data to predict financial performance and evaluate model index and performance metrics. The table demonstrates how the various supervised learning models are used to predict financial performance outcome.

Table 2. Sample Supervised Learning Techniques Data Table

Company	ESG Score	Stock Return (%)	Risk Category	Predicted Stock Price	R-squared	Accuracy	F1-score
Company A	75	10	Low	105	0.85	0.92	0.91
Company B	65	5	High	90	0.75	0.85	0.83
Company C	80	12	Low	110	0.88	0.94	0.93
Company D	70	7	Medium	95	0.80	0.88	0.87
Company E	85	15	Low	120	0.90	0.95	0.94

3.2. Unsupervised Learning Techniques

Unsupervised learning methods such as k-means clustering and PCA identify hidden trends and structures within ESG data. Both k-means clustering and PCA are useful for visualising and interpreting data, which can benefit ESG analysis. They help analysts distinguish between outliers and trends. For example, through k-means clustering, the companies' ESG profiles can be grouped companies in each cluster. K-means clustering enables analysts to identify groups that share certain characteristics, and then to visualise their differences. PCA can reduce dimensionally large data into lower-dimensional information, which is easier to analyse and visualise [6]. For example, ESG scores can be reduced to two lower-dimension data components through PCA. Analysts can then identify the key elements of sustainability processes among companies in different industries.

4. Predicting Long-Term Financial Performance

4.1. Model Training and Validation

The training process involves splitting data into training and testing sets, followed by cross-validation and hyperparameter tuning. Ensuring model generalization is crucial to avoid overfitting. For instance, a GBM might be tuned using cross-validation to predict the Return on Investment (ROI) of companies based on their ESG scores [7]. Hyperparameter tuning involves adjusting parameters such as the learning rate and the number of trees in a GBM to optimize model performance. Cross-validation, where the data is repeatedly split into training and testing sets, helps ensure that the model performs well on unseen data [8]. For example, a model trained on ESG data from 2010-2019 can be validated on data from 2020-2021 to assess its predictive accuracy. The process of data splitting and model training can be mathematically formulated as follows:

Data Splitting

The dataset D is divided into training set D_{train} and testing set D_{test}

DOI: 10.54254/2755-2721/87/20241590

$$D_{train} = \{X_{train}, y_{train}\}$$
$$D_{test} = \{X_{test}, y_{test}\} \quad (1)$$

Cross-Validation

In k-fold cross-validation, the dataset is divided into k subsets.Each subset D_i is used as a validation set once, and the remaining $D \backslash D_i$ as the training set. The cross-validation score is:

$$\text{Cross-validation score} = \frac{1}{k}\sum_{i=1}^{k} \text{score}(D_i) \quad (2)$$

Hyperparameter Tuning

Hyperparameter tuning involves finding the optimal hyperparameters θ by minimizing the average loss over k-fold cross-validation:

$$\theta^* = \underset{\theta}{\text{argmin}} \frac{1}{k}\sum_{i=1}^{k} \text{Loss}(D_i, \theta) \quad (3)$$

Gradient Boosting Machine (GBM) Prediction The GBM prediction $\hat{y}$ is the weighted sum of the predictions of all M trees:

$$\hat{y} = \sum_{m=1}^{M} \alpha_m f_m(x) \quad (4)$$

These formulas give ESG data specialists a framework to train and evaluate machine learning models using unbiased data [9]. This enables more robust, reliable and specific prediction, enabling investors to incorporate more nuanced ESG considerations into their investment processes.

4.2. Evaluating Financial Returns

Predictive models are applied to a future time point – for example, the return on investment of company x next year. This is then compared with a benchmark, such as Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). An output is generated on how 'close' the predictions are – for purposes of example, if ESG scores 'correctly' predict stock prices in the future (meaning the algorithm that computes these predictions is efficient), an RMSE of 5 per cent signifies that when an ESG score is applied to the prediction of a stock price, the prediction is on average 5 per cent away from the actual stock price. Predicted vs actual financial returns will identify companies that are 'over-performing' or 'under-performing' their ESG scores (if the benchmark was RMSE of 5 per cent, for example). These insights can be used once again to inform investment decisions[10].

5. Conclusion

With these advanced methods of data processing and predictive modelling, they are likely able to develop more nuanced insights into the financial implications of ESG factors. The paper presents methodologies for data collection, preprocessing and the use of supervised and unsupervised learning techniques applied in practice to several types of sustainable investment products involving equity such as renewable energy, technology in a real-world context. To conclude, the future of ESG analysis will be driven by further developments in both AI and ML technologies. In particular, the newer sub-fields of AI and ML, such as deep learning and reinforcement learning, will improve the accuracy and precision of models in predicting various ESG factors. NLP will mean that unstructured information from corporate reports, earnings calls and social media can be analysed and reflected on all models. Real-time data analysis in the investment context will also be possible due to the influx of the IoT and social media, which reduces the reporting lags and enables dynamic and responsive investment strategies to adapt to changing market conditions and ESG risks. As technologies continue to evolve, there will be an increasing emphasis on transparency, ethics and sustainability in the investment decision process. Investors can reap the benefits by making more informed investment decisions, thereby contributing to a more sustainable and inclusive economy in the long run.

DOI: 10.54254/2755-2721/87/20241590

References

[1] Erol, Isil, Umut Unal, and Yener Coskun. "ESG investing and the financial performance: A panel data analysis of developed REIT markets." Environmental Science and Pollution Research 30.36 (2023): 85154-85169.

[2] Soori, Mohsen, Behrooz Arezoo, and Roza Dastres. "Artificial intelligence, machine learning and deep learning in advanced robotics, a review." Cognitive Robotics 3 (2023): 54-70.

[3] Seo, Hang Ju, Dong Hyuk Jo, and Zhi Pan. "ESG News Analysis Using News Big Data: Focusing on Topic Modeling Analysis." Software Engineering and Management: Theory and Application: Volume 16. Cham: Springer Nature Switzerland, 2024. 15-27.

[4] Bilyay-Erdogan, Seda, Gamze Ozturk Danisman, and Ender Demir. "ESG performance and dividend payout: A channel analysis." Finance Research Letters 55 (2023): 103827.

[5] Bhat, Mamatha, et al. "Artificial intelligence, machine learning, and deep learning in liver transplantation." Journal of hepatology 78.6 (2023): 1216-1233.

[6] Entezari, Ashkan, et al. "Artificial intelligence and machine learning in energy systems: A bibliographic perspective." Energy Strategy Reviews 45 (2023): 101017.

[7] Amsterdam, Daniel. "Perspective: limiting antimicrobial resistance with artificial intelligence/machine learning." BME frontiers 4 (2023): 0033.

[8] Sarkar, Chayna, et al. "Artificial intelligence and machine learning technology driven modern drug discovery and development." International Journal of Molecular Sciences 24.3 (2023): 2026.

[9] Higgins, Oliver, et al. "Artificial intelligence (AI) and machine learning (ML) based decision support systems in mental health: An integrative review." International Journal of Mental Health Nursing 32.4 (2023): 966-978..

[10] Mhlanga, David. "Artificial intelligence and machine learning for energy consumption and production in emerging markets: a review." Energies 16.2 (2023): 745.