Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/29/20230894

Machine learning for sustainable investing: Current applications and overcoming obstacles in ESG analysis

Xiaqi Liang

The Department of Mathematics, Imperial College London, London, SW7 2BX, United Kingdom

xiaqi.liang22@imperial.ac.uk

Abstract. The intersection of Environmental, Social, and Governance (ESG) issues and Machine Learning (ML) has garnered significant attention in recent years as companies and investors increasingly recognize the paramount importance of sustainable and responsible business practices. ML techniques have been actively explored to tackle various ESG-related challenges, including enhancing ESG data quality and availability, developing comprehensive and dynamic ESG risk models, and optimizing ESG portfolios. The overall process of applying ML models in ESG analysis involves data collection, preprocessing, model training and evaluation, and model interpretation. Commonly used ML models in ESG analysis include logistic regression, decision trees, random forests, and support vector machines. However, there are notable obstacles to overcome, such as the lack of standardization and transparency in ESG data, as well as the potential for bias and ethical concerns in ML-based approaches. Further research and collaborative efforts among researchers and practitioners are crucial to fully realize the potential of ML in enhancing ESG analysis while ensuring transparency, ethical use, and alignment with sustainable and responsible investing principles.

Keywords: ESG, machine learning, portfolio optimization.

1. Introduction

Environmental, Social, and Governance (ESG) issues have gained significant attention in recent years as companies and investors recognize the importance of sustainable and responsible business practices. Simultaneously, the advancements in Machine Learning (ML) and Artificial Intelligence (AI) have ushered in fresh prospects for data-driven ESG analysis and decision-making. As such, the intersection of ESG and ML has become an increasingly important topic of discussion and research in the business and finance communities [1].

Over the past few years, researchers and practitioners have been exploring the potential of machine learning to address various ESG-related issues [2]. Some of the most notable work in this area has focused on ESG data quality and availability, ESG risk modeling, and ESG portfolio optimization. However, ESG data quality and availability have posed a significant challenge due to the limitations and biases inherent in traditional ESG data sources, such as company reports and ratings agencies. To address this issue, researchers have been exploring the use of alternative data sources and ML techniques to improve ESG data quality and availability [3]. For example, Natural Language Processing (NLP) techniques can be used to analyze textual data from sources such as news articles

DOI: 10.54254/2755-2721/29/20230894

and social media to identify ESG-related events and sentiment. Researchers have also explored the use of satellite imagery and other remote sensing data to monitor ESG-related activities such as deforestation and pollution [4]. Another area of research has focused on using ML to model ESG risk, which typically relies on static data and limited variables, leading to incomplete capturing of the complexity of ESG-related risks. ML techniques can be used to analyze a broader range of data sources and variables, including unstructured data and real-time data feeds, to develop more comprehensive and dynamic ESG risk models. For example, researchers have used deep learning techniques to analyze news articles and social media feeds to identify emerging ESG risks [5]. Finally, ML techniques have also been applied to ESG portfolio optimization [3]. Traditional approaches to ESG investing often involve the use of exclusionary screens or the selection of companies based on pre-defined ESG criteria. ML techniques can be used to develop more sophisticated and data-driven approaches to ESG portfolio optimization, taking into account a broader range of factors such as market trends, company performance, and risk.

Although utilizing ML for ESG analysis has potential advantages, various obstacles must be overcome. The absence of uniformity and clarity in ESG data is one of the main obstacles. ML models rely on high-quality, consistent, and reliable data, and the lack of standardization in ESG data can make it difficult to develop accurate and reliable models [1]. In addition, there is a risk of bias and ethical issues in using ML techniques for ESG analysis. The effectiveness of ML models relies entirely on the quality of the data they are trained on [6]. Consequently, if the data is biased or incomplete, the resulting models may reflect partiality or inadequacy.

This review paper will explore the ways in which ML can be used to enhance ESG analysis, the potential benefits and drawbacks of such approaches, and the challenges that need to be addressed to fully realize the potential of this emerging field.

2. Method

2.1. Models of ESG analysis

2.1.1. Logistic regression

Logistic regression is a statistical model utilized for binary or multi-class classification tasks. In ESG analysis, it can be leveraged to forecast binary outcomes, such as the likelihood of a company displaying high or low ESG performance [4]. Logistic regression produces a probability score for each observation and subsequently uses a threshold to classify observations into different categories. One of the advantages of logistic regression as its simplicity and interpretability, as it provides coefficients that indicate the magnitude and direction of the influence of each input feature on the predicted outcome.

For example, a study by De Lucia C et al. investigated whether the adoption of ESG practices by public companies in Europe had an impact on their financial performance. By utilizing machine learning and logistic regression models, the study analyzed the fiscal year 2018-2019 data of 1038 public companies. The results indicated a strong correlation between ESG factors and financial performance, with machine learning models outperforming baseline predictions. The study identified key ESG issues, including environmental innovation, employment productivity, and diversity and equal opportunity, as important for policy implications in implementing sustainable development policies in public enterprises, aligned with the European New Green Deal and circular economy policies [7].

2.1.2. Decision tree

Decision trees are tree-based models that are adept at classification or regression tasks. In ESG analysis, the decision tree can be leveraged to predict ESG outcomes based on decision rules generated from the input features [8]. The decision tree recursively splits the data into different branches based on the values of input features, and each leaf node represents a predicted outcome. This model can

DOI: 10.54254/2755-2721/29/20230894

identify intricate non-linear patterns within the data and interactions among features, making it useful for analyzing ESG data.

As an instance, Hong X et al.'s research employed AdaBoost and SVM techniques to establish a forecasting model for cross-border mergers and acquisitions (M&A) with a focus on sustainable development and ecosystem perspectives. The research emphasized various indicators such as macroeconomic, geographic, climate, cultural, legal, deal and payment, ESG, and financial elements. The outcomes revealed that ESG-driven cross-border M&A correlated with sustainable development features and corporate governance factors, and that M&A success rates were higher for firms with high ESG ratings [9].

2.1.3. Random forest

The random forest is ensemble learning technique that build multiple trees and combine their predictions to improve accuracy and robustness [10]. In ESG analysis, the random forest can be used for classification tasks to predict ESG outcomes based on a combination of decision rules generated from multiple trees. It can handle noise in the data and is less prone to overfitting compared to individual decision trees. In addition, they can also provide feature important measures that indicate the relative importance of each input feature in predicting the outcome.

For example, a study by D'Amato et al. employed a Random Forest algorithm to examine the impact of structural data on the ESG scores of companies within the STOXX 600 Index, as determined by Thomson Reuters Refinitiv. The analysis reveals that balance sheet data plays a pivotal role in elucidating the factors that contribute to ESG scores [11].

2.1.4. Support vector machine (SVM)

SVM is a popular model for binary or multi-class classification tasks. In ESG analysis, SVM can be used to predict ESG outcomes by finding a hyperplane that best separates the data into different categories. SVM can capture non-linear relationships through the use of kernel functions, making it suitable for identifying complex patterns in the data. SVM aims to maximize the margin between the different classes, making it robust to outliers and noise in the data [12].

For example, a study by Konstantinos Petridis et al. used a SVM model to adjust the data and applied it with various distance measures and representations. The findings suggest that a higher representation of women in top management positions has a positive effect on M&A efficiency. Regression analyses are also used to examine the influence of qualitative factors on M&A performance, revealing the significance of women's representation on boards [13].

2.2. Overview of the framework

The four commonly used machine learning models in ESG analysis including logistic regression, decision trees, random forests, and support vector machines mentioned above, can be applied in a similar overall process [4, 8, 10, 12]. The process typically starts with data collection, where relevant and high-quality data on ESG-related factors is gathered. This data is then preprocessed, which may involve tasks such as data cleaning, feature selection, and engineering. Next, the preprocessed data is split into training and testing datasets for model development and evaluation. The selected machine learning models are then trained on the training dataset using appropriate algorithms and techniques, such as logistic regression or tree-building algorithms. The trained models are then evaluated using the testing dataset, and their performance is assessed using relevant performance metrics. Model performance can be further validated using techniques such as cross-validation or out-of-sample testing. Once the model is validated and deemed satisfactory, it can be used for prediction or assessment of ESG outcomes in real-world data. The interpretability of the models can provide insights into the factors influencing the predicted ESG outcomes and aid in decision-making. It's important to note that the overall process may vary depending on the specific requirements and context of the ESG analysis task at hand, and careful consideration of data quality, model selection, validation, and interpretation is crucial for reliable and meaningful results [7, 9, 11, 13].

DOI: 10.54254/2755-2721/29/20230894

3. Application and discussion

ESG factors have become a key focus for companies and investors, as sustainability and responsible investing gain prominence. As technology continues to advance, machine learning is increasingly being applied to analyze and incorporate ESG data into investment decisions, offering valuable insights and driving positive change [1, 2].

3.1. ESG data analysis and scoring

Machine learning algorithms allow for the processing and analysis of vast amounts of data from diverse sources, such as financial reports, news articles, and social media, to assess a company's ESG performance. By processing and analyzing this data, machine learning can provide a quantitative evaluation of a company's ESG performance, identifying trends and patterns over time [4, 14]. This enables investors to make informed decisions based on comprehensive and data-driven assessments of ESG performance. Additionally, machine learning algorithms can effectively highlight emergent ESG risks and opportunities, thereby enabling investors to proactively manage their portfolios.

3.2. ESG data integration and validation

ESG data can be intricate and fragmented, making it challenging to analyze and compare. Machine learning can automate the process of data collection, validation, and integration from multiple sources, improving data accuracy and consistency [6]. This enables investors to have a more reliable understanding of a company's ESG performance and make more informed investment decisions.

3.3. ESG portfolio optimization

Machine learning can optimize investment portfolios by incorporating ESG factors into the decision-making process. By leveraging machine learning algorithms, investors can identify companies that align with specific ESG goals or risk profiles. This enables them to construct portfolios that are aligned with their values while aiming for better risk-adjusted returns [3]. Machine learning can also help investors assess the potential impact of ESG factors on portfolio performance, allowing for more effective portfolio management.

3.4. ESG impact measurement and reporting

Machine learning can facilitate the measurement and reporting of the actual environmental and social impact of companies. By analyzing data from diverse sources, machine learning algorithms can provide insights into a company's true impact, beyond just self-reported data [1,5,14]. This can help investors and companies measure and report their ESG performance more accurately and transparently, promoting accountability and sustainability.

3.5. Considerations and challenges

It is pertinent to acknowledge that machine learning in ESG also faces challenges. One key consideration is the potential for bias in data and algorithms. The efficacy of machine learning models is contingent upon the quality of data used in their training, and the usage of biased or incomplete data in ESG analysis can engender biased outcomes [14]. Ensuring that the data used in machine learning models is diverse, comprehensive, and representative is crucial to avoid reinforcing existing biases. Another consideration is the need for interpretability and explainability of machine learning models in the ESG context [2, 4]. Investors and stakeholders need to understand how machine learning models arrive at their conclusions to ensure transparency and trust.

4. Conclusion

The application of machine learning in the analysis of ESG data offers significant opportunities for investors and companies to enhance their understanding of ESG performance and make more informed decisions. Through the automation of data collection, validation, integration, and analysis, machine

DOI: 10.54254/2755-2721/29/20230894

learning can provide valuable insights, optimize portfolios, and measure impact, ultimately contributing to a more sustainable and responsible approach to investing. Notwithstanding the benefits, it is important to acknowledge and address the potential challenges and considerations associated with the use of machine learning in ESG analysis. Ensuring that the data used is diverse, comprehensive, and representative, and mitigating biases in algorithms is crucial to avoid perpetuating existing biases and ensure responsible decision-making. Additionally, transparency and explainability of machine learning models in the ESG context are necessary to build trust and foster accountability. As technology continues to evolve, machine learning has the potential to revolutionize ESG analysis and contribute to a more sustainable and responsible investment landscape. By leveraging the power of machine learning in conjunction with robust ESG frameworks, investors and companies can unlock new insights, drive positive change, and create a more sustainable future for generations to come.

References

[1] Twinamatsiko E and Kumar D 2022 Incorporating ESG in Decision Making for Responsible and Sustainable Investments Using Machine Learning. In 2022 International Conference on Electronics and Renewable Systems (ICEARS) pp 1328-1334 IEEE

[2] Kumar S et al 2022 Past, present, and future of sustainable finance: Insights from big data analytics through machine learning of scholarly research. Annals of Operations Researchpp 1-44

[3] Vo N N 2019 Deep learning for decision making and the optimization of socially responsible investments and portfolio Decision Support Systems 124 113097

[4] Lopez C 2020 ESG Ratings: The Road Ahead (October 6, 2020) Available at SSRN: https://ssrn.com/abstract=3706440 or http://dx.doi.org/10.2139/ssrn.3706440

[5] Lee O et al 2022 Proposing an integrated approach to analyzing ESG data via machine learning and deep learning algorithms. Sustainability 14(14) 8745

[6] De F C 2020 Esg investments: Filtering versus machine learning approaches arXiv preprint arXiv:2002.07477

[7] Tian L 2021 Unraveling the Relationship between ESG and Corporate Financial Performance - Logistic Regression Model with Evidence from China (August 1, 2021). Available at SSRN: https://ssrn.com/abstract=3897207 or http://dx.doi.org/10.2139/ssrn.3897207

[8] De L C et al 2020 Does Good ESG Lead to Better Financial Performances by Firms? Machine Learning and Logistic Regression Models of Public Enterprises in Europe. Sustainability. 2020; 12(13):5317.

[9] Hong X 2022 Application of Machine Learning Models for Predictions on Cross-Border Merger and Acquisition Decisions with ESG Characteristics from an Ecosystem and Sustainable Development Perspective. Sustainability 14(5) 2838

[10] Abdalmuttaleb M A 2022 The role of artificial intelligence in sustainable finance Journal of Sustainable Finance & Investment 0:0 pp 1-6

[11] D'Amato V et al 2022 ESG score prediction through random forest algorithm Comput Manag Sci 19 347–373

[12] Raza H et al 2022 Applying artificial intelligence techniques for predicting the environment, social, and governance (ESG) pillar score based on balance sheet and income statement data: A case of non-financial companies of USA, UK, and Germany, Frontiers in Environmental Science, Volume 10 ISSN 2296-665X, 10.3389/fenvs.2022.975487

[13] Konstantinos P et al 2022 A Support Vector Machine model for classification of efficiency: An application to M&A, Research in International Business and Finance, Volume 61 101633 ISSN 0275-5319

[14] Sharma U et al 2022 The pertinence of incorporating ESG ratings to make investment decisions: a quantitative analysis using machine learning. Journal of Sustainable Finance & Investment pp 1-15