Journal of International Education and Development
2025, VOL. 9, NO. 1, 144-147
DOI: 10.47297/wspiedWSP2516-250031.20250901

Risk and Return Prediction of ESG Investment based on Machine Learning Models

Qingyu Du[1,2], Dennis Wong[3]

[1]Macao Polytechnic University, Faculty of Applied Sciences, Macau, China;

[2]School of Insurance, Guangdong University of Finance, Guangzhou, 510521, Guangdong, China;

[3]Macao Polytechnic University,Faculty of Applied Sciences ,Macau, China

ABSTRACT

Environmental, social and governance (ESG) investment focuses not only on financial returns, but also on the sustainable development and social responsibility of enterprises, which aims to promote social and environmental well-being and achieve economic benefits. However, the risk and return prediction of ESG investment is complex and changeable, and traditional prediction methods often fail to accurately respond to rapid changes in the market. In recent years, machine learning technology has shown great potential in the financial market analysis with its excellent data processing capacity and predictive accuracy. Therefore, this paper discusses the risk and return prediction of ESG investment based on machine learning models, in order to analyze and predict the potential risks and returns of ESG investment by establishing and training various machine learning models.

KEYWORDS

machine learning model; ESG investment; risk and return

As investors and regulators continue to focus on sustainable investment, ESG investment has moved from a peripheral topic to an important part of mainstream investment decisions. This investment strategy focuses not only on traditional financial returns, but also on enterprise performance in terms of environmental protection, social responsibility and governance structures. Enterprises with high ESG scores often have better financial performance and lower levels of risks[1]. Good social responsibility practice can help enterprises build a stronger reputation in the market and reduce legal and operational risks, thereby protecting investors from potential negative events. However, the complexity of market dynamics and the incompleteness of available data often make it difficult for investors to make decisions based on comprehensive information[2]. In this context, machine learning technology show great potential due to its ability to process large and complex data sets and extract valuable insights from them. In summary, the introduction of machine learning provides new perspectives and tools for ESG investment, and enables investors to make more informed decisions in a complex and ever-changing market environment. The purpose of this paper is to explore the effectiveness of machine learning in predicting the risks and returns of ESG investment, provide investors with practical tools to help them achieve financial goals while taking social responsibility into account, and drive the entire financial industry towards a more sustainable and responsible direction.

1 Research Method

1.1 Selection of Machine Learning Models

In this study, three mainstream machine learning models are used to handle the prediction task, each of which is selected for its data processing capacity and applicability.

Linear Regression: Linear regression is a statistical method to predict a numerical response variable based on the linear relationship between the independent and dependent variables. In the return prediction of ESG investment, it is assumed that there is a linear relationship between the rate of return RRR and multiple ESG factors, and the model can be expressed as follows:

$$R=\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n+\epsilon$$

In this model, X1, X2,... Xn is the explanatory variable (various dimensions of ESG scores), β0, β1,... , βn is the model parameter, and ϵ is the error term.

Random Forest: Random forest is an ensemble learning technique that builds multiple decision trees and outputs a pattern of classification or regression. It performs well when dealing with a large number of input variables and complex interaction relationships between variables, and is suitable for dealing with the non-linear relationship of ESG data.

Support Vector Machine (SVM) : SVM is a powerful classification and regression technique that finds the best linear segmentation plane or the hyperplane in a high-dimensional space by mapping data into that space. Support vector

regression (SVR) is used in regression problems, which aims to find a function that is as close as possible to all training data points within a specific tolerance:

$$\min_{\omega,b,\xi,\xi^*} \frac{1}{2} /\!/ \omega /\!/^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*)$$

In this model, C is the regularization parameter, ω is the normal vector of the hyperplane, b is the error term, and ξ,ξ∗ is the slack variable.

1.2 Model Training

Model training aims to use historical data to train selected machine learning models, so that these models can learn the patterns and relationships present in the data and make accurate predictions based on them. Data preparation requires collecting historical data on ESG investment from multiple sources, including the ESG score, financial performance, market environment variables of enterprises. The collected data must be carefully cleaned and pretreated so that it can be effectively processed by the model. During the model configuration phase, it is crucial to set the appropriate parameters for each selected model[3]. For example, in the random forest, parameters such as the number of trees, the depth of trees and the minimum sample size required to split a node need to be determined. For SVM, the important parameters include the choice of kernel functions, the value of regularization parameter C and the parameters of kernel functions.

1.3 Data Sources

This study uses data sets from a variety of sources, including public financial statements, scoring data published by ESG rating agencies, and market transaction data. Specifically, the annual and quarterly financial reports provide the necessary financial indexes, and the ESG rating data is sourced from globally renowned ESG rating agencies, which provide a comprehensive evaluation of the enterprise's performance in environmental protection, social responsibility and enterprise governance.

2 Result Analysis

2.1 Evaluation of Model Performance

In this study, various statistical indicators are used to evaluate the predictive efficiency of the models, including accuracy, precision, recall, F1 score and area under curve, as shown in Figure 1.

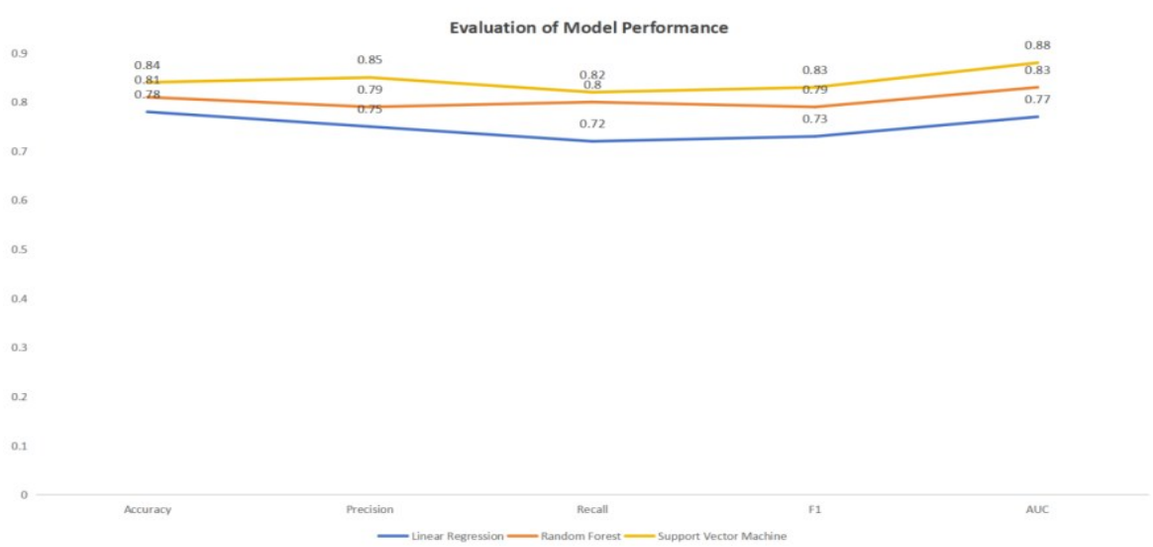

Figure 1 Evaluation of Model Performance

The random forest has the best performance on almost all evaluation indicators, and shows high accuracy and recall, which means that it can identify positive samples in the data set better, and has the low error rate. Support vector machine also shows good performance in data processing, which indicates that it is excellent in balancing precision and recall ratios. In contrast, although linear regression performs relatively robustly, it has some shortcomings in various indicators, and especially in recall and area under curve. The reason may be that the linear assumption limits the ability to capture the complex relationship of ESG data.

2.2 Risk and Return Prediction Results

The predicted rate of return, risk assessment (such as predicted loss and volatility), and confidence level of investment advice are shown in Table 1.

Table 1 Risk and Return Prediction Results

Condition	Model	Predicted Rate of Return	Predicted Volatility	Predicted Loss	Confidence Level
Market Advances	Linear Regression	12.00%	8.00%	3.00%	70.00%
	Random Forest	15.00%	7.00%	2.00%	85.00%
	Support Vector Machine	13.00%	8.00%	2.50%	80.00%
Market Declines	Linear Regression	-3.00%	10.00%	5.00%	65.00%
	Random Forest	-1.00%	9.00%	4.00%	75.00%
	Support Vector Machine	-2.00%	10.00%	4.50%	70.00%
Market Stable	Linear Regression	5.00%	5.00%	1.00%	80.00%
	Random Forest	7.00%	4.00%	0.50%	90.00%
	Support Vector Machine	6.00%	5.00%	1.00%	85.00%

The random forest shows a higher predicted rate of return and a lower predicted volatility under various market conditions, which shows its strong ability in processing complex nonlinear data. In contrast, linear regression is more conservative in the conditions of market advancing and stable, and the predicted loss is larger in the market declining, due to the model's inability to adequately capture complex dynamics in the data. Support vector machine shows the good volatility and loss prediction in various conditions, and provides a relatively balanced perspective on risk control. These results highlight the importance of selecting the right model in practical applications to effectively predict risks and returns. The performance of random forest is particularly outstanding, which not only predicts high returns, but also shows superiority in risk control.

2.3 Practical Applications of Prediction Results

In the area of risk and return prediction of ESG investment, practical applications of machine learning models demonstrate their ability to translate complex data analysis into concrete investment decisions. In portfolio management, the output of prediction models can directly affect the decision process of asset allocation. By analyzing the risk and return data predicted by the models, investment managers can identify which ESG investments have potentially high return and low risk characteristics, and adjust their portfolios accordingly to maximize the return-to-risk ratio. It can help investors achieve their financial goals while maintaining social responsibility, and provide data support for continuous monitoring and response to market changes.

For example, the performance advantage of random forest in predicting highly volatile market conditions makes it particularly valuable in the dynamic market environment and it can provide timely and reliable data support for investment decisions[4]. In addition, by showing the prediction process and results of the models, investors are able to see the data-driven logic of investment decisions, thereby gaining a deeper understanding of the potential value and risks of ESG investment.

3 Discussion

3.1 Research Limitations

Although this study has shown the certain potential and effectiveness of machine learning models in predicting the risk and return of ESG investment, there are some limitations in the research process and results that cannot be ignored. On the one hand, the availability and quality of data have a decisive impact on the accuracy and reliability of the models. ESG data often comes from different sources with different degrees of standardization, and the inconsistency and incompleteness of data may lead to deviations during model training, thus affecting the accuracy of prediction results. On the other hand, although machine learning models are excellent at handling large amounts of data and recognizing complex patterns, these models typically require a lot of parameter tuning and optimization to achieve optimal performance, which often relies on a lot of experimentation and expertise. It increases the complexity of model training and leads to overfitting of the models.

At the same time, the "black box" nature of machine learning models, which is the opacity of the model's decision-making process, can somewhat limit the use of models in certain application scenarios that require a high degree of transparency and interpretability.

3.2 Prospect

With the development of block chain and other distributed ledger technologies, it is believed that the transparency and reliability of ESG data recording and verification will be significantly improved in the future. At the same time, the application of natural language processing (NLP) technology can extract key information about ESG performance from unstructured data sources, thereby enriching the data base of the analysis model. In addition, the interpretability of models is an important aspect of improving the application of machine learning in financial investment decisions. Explicable machine learning models such as decision tree, principle learning, and model visualization techniques can be explored to provide investors with more intuitive understanding and trust. Finally, real-time updates and dynamic adjustments of models will be a necessary direction to respond to rapidly changing market conditions. As computing power increases and algorithms are optimized, the real-time data processing and online learning algorithm will enable ESG investment analysis models to respond to market changes in real time and provide more timely investment advice.

4 Conclusion

Through the performance of linear regression, random forests, and support vector machine under different market conditions, the study reveals the advantages and limitations of each model, and highlights the effectiveness of machine learning methods in processing volume and multi-dimensional ESG data. The results show that the proper selection and adjustment of machine learning models can significantly improve the accuracy of investment risk assessment and return prediction. In conclusion, the application of machine learning to risk and return prediction of ESG investment can not only enhance the scientific nature and effectiveness of investment strategies, but also promote the concept of responsible investment at a broader level, and have a positive impact on promoting social and environmental well-being.

Funding

Her research is supported by the Macao Polytechnic University research grant (Project code: RP/FCA-05/2023 and fca.445f.c2c3.4)

References

[1] Lu S F, Guo J R, Huang K. Local Government Debt Management System Reform and Corporate ESG Performance: Causal Inference based on Double Machine Learning Model[J]. Research on Financial and Economic Issues, 2024(11): 76-88.

[2] Lin J X, Sun Y, GAO Y L. Research on Investment Portfolios based on Machine Learning[J]. Science Technology and Industry, 2024, 24(20): 238-245.

[3] Yu D F, Li X, Zhang Y. ESG Performance and Supply Chain Risks of Enterprises[J/OL]. Contemporary Finance and Economics,1-13[2024-11-26].

[4] Shen J X, Li Y H. PPP Social Capital Investment Forecasting Model based on Machine Learning[J]. Jiangsu Commercial Forum, 2024(4): 81-84.