# Prediction of Investment Revolution of Green Enterprises: A Two-dimension Approach to the Combination of Characteristics and Temporalities of Green-specific Characteristics

Shiyao Lu

{sl12748@nyu.edu}

Computer Science, New York University Shanghai, Shanghai, 200000, China

**Abstract.** The paper suggests a dual-dimension feature system combining financial indicators, ESG measures, Personalized green features and lagged return features. The paper evaluates the regression algorithms on the use of a panel dataset of 300 observations that consist of 50 green enterprises (2018-2023). It is shown through the results of the experiment that Gradient Boosting algorithm provides optimal results, with a Mean Squared Error (MSE) of 0.011 and R-squared (R2) of 0.80. According to the feature importance analysis, the core predictors are the composite ESG rating, the ratio of green R&D investment, and the lagged returns. The proposed methodology is a powerful analytical tool that will enable investors in green finance to analyze and decide by quantifying green features and providing a combination of the static and dynamic dimensions.

**Keywords:** Green enterprises; Investment return prediction; ESG metrics; Temporal dependencies; Gradient boosting.

## 1 Introduction

Green finance has become one of the most important blocks of the sustainable development of the world and the market has witnessed the rapid growth in recent years [1]. It is highly important to forecast investment returns in the case of green enterprises accurately, but the traditional prediction models do not always succeed in predicting the proprietary nature of such enterprises [2]. At the same time, national green development policies are becoming more cumulative in needing data-oriented tools to aid in their decision-making guidance [3]. Predictive models that are currently in use do not have the ability to respond to the specific characteristics of the green businesses and the changing nature of the financial time series, which results in the reduced accuracy of predictions and less practical value. As a result, the paper provides a gap in the research methods and addresses existing practical needs and, thus, it is a very important research topic in the field of financial modeling.

In order to address the above-presented problems, the paper creates a special panel data set

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365224

that will cover the time frame between 2018 and 2023, namely, in terms of the new energy and environmental protection equipment industries. The main aim of the research is to develop measurable and standardized customary green attributes to include intrinsic competitiveness related to green and external policy influences of enterprises. In addition, the paper comes up with a dual-dimension feature representation, which combines the fixed environmental features that are green and dynamic over time. Through the use of the annual relative return as a stationary target variable, the paper seeks to compare the different regression algorithms to confirm the effectiveness of this system and key predictors to make investment decisions.

The main contributions of the paper are three times. First, the paper would numerically measure the proprietary green features into similar ratio variables, a major issue of having been quantitative dependence on qualitative description, and low model scalability of the literature. Second, the paper makes use of fixed green characteristics and time-varying dependencies defeating the restrictions of single-dimensional modelling. Last, the study develops a very specific dataset of green features which is cross-validated with data on multi-source and successfully staves off the applicability problems normally faced when working with general-purpose financial datasets.

## 2 Literature review

### 2.1 The traditional asset pricing and ESG integration

Anywhere in financial forecasting, the traditional asset pricing models, including the current extensions by Zhang et al. [2], can provide some explanations to green enterprises but does not pay much attention to their green nature. The early models offer a theoretical basis, but their use cannot be applied to the modern day because of the lack of environmental parameters. As a way of overcoming this gap, Li et al. had to incorporate ESG ratings to their models, with the effect of enhancing prediction accuracy and smooth progression into more holistic assessment [4]. However, this method also is based on the generalized non-financial indicators and does not measure endogenous green technology investments of enterprises.

### 2.2 Time-related conditions of forecasting

As far as the modeling of temporal dependency is concerned, Chen and his in-group revealed the importance of the implementation of lagged return variables, but, using generic data was not flexible enough to apply to the green-specifics [5]. Correspondingly, Rebeschini pointed to the importance of incorporating stationary relative returns as opposed to absolute prices in time-series forecasting, which is an important recommendation that greatly increases the robustness of a model [6].

### 2.3 Green features quantification

Regarding the features on processing green, Wang et al. referred greatly on the qualitative descriptions, and Zhao and Liu resorted to binary policy variables[7][8]. Even with these developments, the current research explained has found it difficult to section green features into continuous, similar variables. In general, the existing literature is characterized by a number of drawbacks such as the inability to quantify proprietary green features, insufficient comparability between enterprises, and the separation of temporal dependencies and green features.

## 3 Methodology

### 3.1 Problem interpretation

The identification of the problems and the target variable is presented in 3.1. The paper responds to the question as a regression problem in order to forecast the mean annual relative returns of green enterprises. To establish data stationarity and prevent the bias that exists with absolute prices, the focus variable is the annual relative return. The target variable calculation can be written as shown in Formula (1):

$$\text{Annual Relative Return} = (\text{To the Current Year Stock Price} - \text{To the Previous Year Stock Price}) / \text{To the Previous Year Stock Price} \quad (1)$$

### 3.2 Data collection and feature engineering

A total of eight features make up the predictors, namely containing three financial features, two ESG features, two custom green features, as well as one temporal feature. The paper extracted historical database financial information, MSCI ESG rating data indicators and raw green information (green R&D expenses) in annual corporate reports using manual methods. In order to model temporal variations, the paper provided the lagged feature of annual relative returns of the past year. All the selected features have been defined and described in more detail and outlined in Table 1.

**Table 1.** Dataset field structure.

| Field type | Field name | Description |
|---|---|---|
| Identification | company_id | Unique enterprise ID (G001–G050) |
| Identification | year | 2018–2023 |
| Financial feature | debt_to_asset | Asset-liability ratio (0.2–0.7) |
| Financial feature | net_profit_growth | Net profit growth rate (-0.1–0.3) |
| Financial feature | current_ratio | Current ratio (1.0–2.5) |
| ESG feature | carbon_emission | Carbon emission intensity (0.1–0.8; lower = greener) |
| ESG feature | ESG_rating | ESG composite rating (1.0–5.0; higher = better) |
| Custom green feature | green_rd_ratio | Green R&D expense / annual revenue (0.01–0.08) |
| Custom green feature | env_policy_support | Environmental subsidy / annual revenue (0.005–0.04) |
| Temporal feature | lag_ 1_annual_return | Previous year’s annual relative return (-0.15–0.4) |
| Target variable | annual_relative_return | Annual relative return (-0.15–0.4; prediction target) |

### 3.3 Data processing and splitting of datasets

The ultimate data arrangement contains 300 data points in 50 enterprises during a span of six years and it includes 11 head-on numerical variables with none of them being void after data processing. Forward filling has been used to fill in missing financial and ESG information to ensure the continuity of time series when all the green data could not be obtained and data gaps. Interquartile Range (IQR) was used to truncate outliers of all the numerical features. The paper also used the standardization of Z-score to remove the dimensions of variables. The time-series dataset was split sequentially without shuffling and data after 2018 was taken as the training data and after 2022 was taken as the testing data to prevent the chances of future leakage of data [9].

### 3.4 Model selection and hyperparameter tuning

The paper is comparing three regression algorithms, the first one is the Linear Regression used as a control, the second one is the Random Forest used to cope with non-linear interaction, and finally is the Gradient Boosting used as the core ensemble algorithm because it uses iterative residual optimization. Model testing is mainly based on the R-Squared or the R-squared (R 2) the mean squared error or the error of prediction in models and the prediction explanatory power of the models respectively.

To avoid data leakage further as well as ensure better model performance, the paper involves grid search and five-fold cross-validation of the training set to optimize the hyperparameters. Table 2 shows the particular tuned parameters and the best values that were used to model ensemble.

**Table 2.** Hyperparameter tuning results.

| Model | Tuned parameters | Optimal values |
|---|---|---|
| Random Forest | n_estimators,max_depth,min_samples_split | 100, 8, 5 |
| Gradient Boosting | learning_rate, n_estimators, max_depth | 0. 1, 150, 6 |

## 4 Experimental Results

### 4.1 Exploratory data analysis

In addition, the target variable exhibits an approximation of a normal distribution which means that it is well suited to be used in the training of a regression model. The target variable has the distribution depicted in Figure 1.

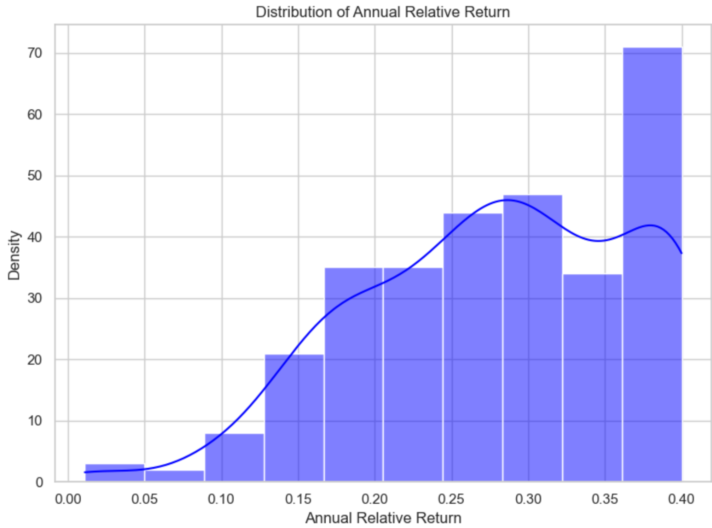


**Fig. 1.** Distribution of Annual Relative Return. (Picture credit: Original)

Correlation analysis conducted as a result of the exploratory data analysis shows that the target variable has a positive correlation with ESG ratings, the ratio of the green R&D investment, and lagged annual returns. The analysis proves there is no multicollinearity of severe nature regarding the data given because the correlation coefficients between the features are all below the reasonable values. Figure 2 presents the particular correlation table.

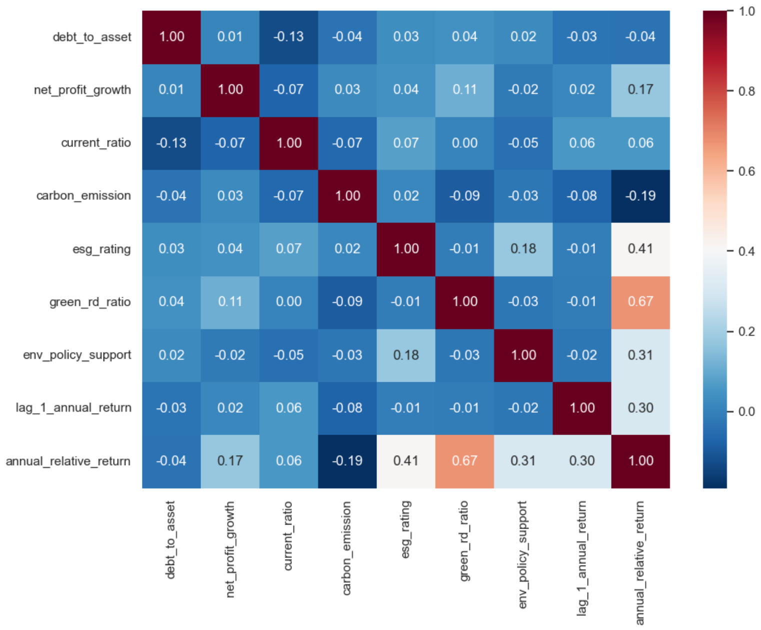


**Fig. 2.** Correlation Matrix of Features. (Picture credit: Original)

### 4.2 Model performance comparison

An evaluation of the performance of the different models in the testing set results in the finding that Gradient Boosting algorithm beats both Random Forest and Linear Regression by a long margin. To be more precise, Gradient Boosting algorithm has a Mean Squared Error = 0.011, R-squared = 0.80. Table 3 shows the total performance metrics of each of the models.

**Table 3.** Model performance on test set.

| Model | MSE | $R^2$ |
|---|---|---|
| Linear Regression | 0.027 | 0.57 |
| Random Forest | 0.017 | 0.74 |
| Gradient Boosting | 0.011 | 0.80 |

Besides, the explanatory power of all the models enhanced compared to the models without any temporal features on the introduction of the lagged annual return feature which was indeed a justification of the significance of capturing temporal dependencies [10].

### 4.3 Feature importance and robustness checks

The importance of the features using the optimal Gradient Boosting model has discovered that the three main predictors are the composite ESG rating, the proportion of green R&D investment, and lagged returns. Most of the weight in predictions is shared by custom green features and the temporal feature. Figure 3 shows the relative significance of each of the features.

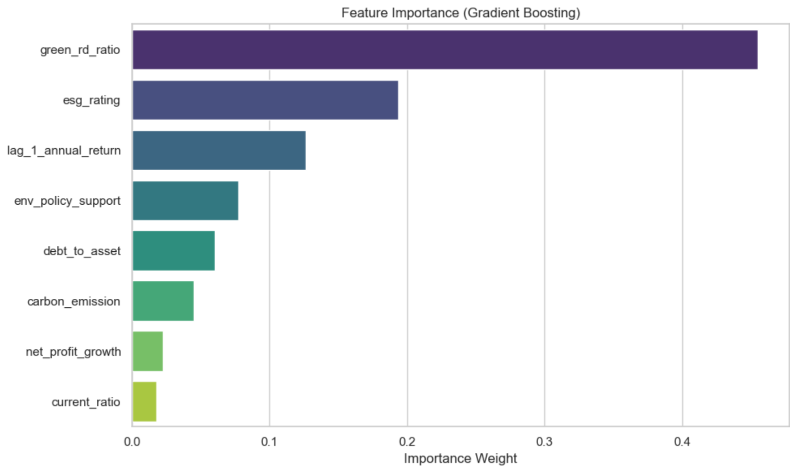


**Fig. 3.** Feature Importance Chart based on the Gradient Boosting Model. (Picture credit: Original)

In order to test the stability of such findings, the paper performs robustness checks, where it compares different lagged features. The obtained findings prove that the model with the one-year lagged annual return has the highest value of R-squared, which supports the validity of the chosen structure of features.

## 5 Conclusion

The article defines a two-dimensional feature state-system comprising of financial indicators, ESG stage, tailor-made green features and lagged return features to precisely estimate the returns of green business investments. The results of the experiment prove that the Gradient Boosting algorithm suits the worst performance. The developed proprietary green design, which focuses on the development of custom green functionality, is measured in a useful manner and contributes significantly to the predictive aspect of the model. A combination of these two traits of green solidity along with the dynamic time effects become vital in increasing the total explanatory strength of the model.

These results confirm that endogenous green competitiveness and the temporal trends are simultaneously important in determining the investment returns of green enterprises. The research therefore provides feasible investment advice and recommends that an investor should focus on companies that have high levels of ESG compliance, excellent green R&D investments and consistent returns over the last few years in order to maximize the returns in their portfolio.

The current research is limited in spite of the positive results. The sample of 300 observations may restrict the overall extrapolation of the findings and one lagged feature can be not enough to account for complex temporal dynamics. Moreover, the process of data extraction is very time-consuming when performed by human hands. Future studies will be to broaden the dataset with pertinent database robustness, incorporate multi-period lagged macroeconomic variables, and performance of deep learning models to advance automated screening instruments of the investments.

## References


[1] MSCI.: MSCI ESG ratings methodology. MSCI Inc., New York. (2022)

[2] Zhang, Y., Li, X., Wang, H.: Research on investment return prediction of listed companies based on financial indicators. Journal of Financial Research. Vol. 38(5), pp. 123-135. (2021)

[3] National Development and Reform Commission.: Green finance development guidelines. National Development and Reform Commission, Beijing. (2022)

[4] Li, M., Zhang, Q., Zhao, J.: ESG rating and corporate investment return: Empirical evidence from China's A-share market. China Industrial Economics. Vol. 40(3), pp. 89-106. (2022)

[5] Chen, L., Yang, H., Wu, X.: Application of machine learning algorithms in financial time-series prediction. Journal of Systems Engineering. Vol. 38(2), pp. 289-302. (2023)

[6] Rebeschini, P.: Review session: Final project guidelines. Unpublished manuscript. (2024)

[7] Wang, Z., Liu, Y., Chen, S.: Research on green finance development and enterprise performance. Environmental Science & Technology. Vol. 44(7), pp. 189-196. (2021)

[8] Zhao, T., Liu, J.: The impact of environmental subsidies on corporate green innovation. Journal of Cleaner Production. Vol. 380, pp. 134-145. (2023)

[9] Liu, Y., Wang, S.: Machine learning in portfolio management: An empirical study. Finance Research Letters. Vol. 46, pp. 102-110. (2022)

[10] European Investment Bank.: Green finance and sustainable development report. European Investment Bank, Luxembourg. (2023)