Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Predicting Financial Sustainability through ESG Integration: A Stacked Gradient Boosting Ensemble Approach

RANA MUKHERJI

rana.mukherji@gmail.com

The ICFAI University Jaipur

Manishita Mukherji

Amity University Jaipur

Research Article

Keywords: ESG, financial sustainability, gradient boosting, stacked ensemble, machine learning, sustainable finance, predictive analytics

Posted Date: July 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10068385/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Predicting Financial Sustainability through ESG Integration: A Stacked Gradient Boosting Ensemble Approach

Abstract

Using environmental, social, and governance (ESG) information in forward-looking evaluations of corporate performance is methodologically unstable, in part because the correlation between ESG and profitability is only weak in the present, and because the impact of ESG can be (only) nonlinear and comes about through interacting channels. In this study, an ensemble of four sustainability-relevant targets (one-period-ahead return on assets (ROA), return on equity (ROE), ESG composite index and a financial sustainability index (FSI)) is developed and evaluated by a meta-learner based on stacked gradient boosting ensemble with base learners trained using various deep learning algorithms, including XGBoost, LightGBM and CatBoost. The framework is evaluated with the help of a controlling yet simulation-based dataset of 5,000 firm-level "observations" in eight industries, equipped to depict stylized features of the distribution of listed-firm fundamentals, and 34 engineered to represent ESG–financial interactions, disparities in the pillars, leverage structures, and distress proxies. The ensemble attains out-of-sample R^2 values of 0.720 (ROA), 0.764 (ROE), 0.857 (ESG), and 0.937 (FSI). Despite the fact that the raw ESG–performance correlations are statistically insignificant, these ESG×profitability interaction terms tell the most about all four prediction tasks, underlining the predictive power of ESG within a conditional, interactive structure that is not captured by linear specifications. The results advocate for ensemble learning as a clear and visible reference architecture for sustainable finance analytics and to guide a sustainable finance early-warning/capital allocation framework.

Keywords: ESG, financial sustainability, gradient boosting, stacked ensemble, machine learning; sustainable finance, predictive analytics

1. Introduction

The idea of sustainable finance is now shifting from being a concern of a few to a key organizing principle of capital markets, with ESG factors influencing portfolio construction, credit appraisal and disclosure policy around the globe (Billio et al., 2024; Starks, 2023). However, the empirical studies exploring the relationship between ESG traits and financial returns are far from consistent. Overall, evidence from meta-analyses has indicated that the relation between ESG and financial performance is statistically significant for about 50% of the existing studies at the firm level, and the results are equivocal for a significant number of others (Atz et al., 2023). The issue is further complicated by disagreements among rating providers – different providers may score the same company differently based on the definitions of their measures, the weighting of their measures, and the scope that they cover (Berg et al., 2022; Christensen et al., 2022). This ambiguity, argues Edmans (2023), points to a more fundamental conceptual problem: ESG is not one per se, but rather a collection of intangibles, some that are more and some less value-relevant to a company as a function of corporate value creation.

In this context, a new area of research considers how to apply machine learning (ML) techniques to identify forward-looking signals in data that are unavailable using traditional linear methods, such as ESG and fundamental data. Tree-based gradient boosting methods are especially appealing for this task: they are the best-of-the-breed approach for tabular financial data (Grinsztajn et al., 2022), achieve good results with high order interactions without explicitly defining them, tolerate high degree of multicollinearity, and they manage skewed-distributing data relatively well. The applications involve predicting ESG scores based on fundamental financial data (D'Amato et al., 2024), leverage ESG conditions for profitability and investment/green checking-up (Sharma et al., 2022), identifying greenwashing, and designing sustainable portfolios (Sorathiya et al., 2024). But there are three gaps. First, most studies seem to estimate single learners, while a few take advantage of stacked ensembles that rely on complementary boosting implementations and which evidence has shown to gain in both stability and generalization (Wolpert, 1992). Second, there is little literature that explores the same common architecture, from a profitability and ESG perspective as well as a composite sustainability perspective, which leaves the question of what is best predictive. Third, the use of observational ESG metrics mixes measurement noise, rater differences and endogenous selection, making it hard to discern if learning architectures can capture the linkages between ESG and performance, which is a question where controlled, simulation-based evaluation can uniquely provide answers.

This paper aims to redress these gaps. The proposed model is trained in combination with a two-level stacked architecture involving three main classifiers (XGBoost, LightGBM and CatBoost) using a Ridge regression meta-learner, on 34 engineered features, and evaluated on four targets: future ROA, future ROE, future ESG composite score and a financial sustainability index (FSI), which scores financial performance in terms of profitability, ESG, and balance-sheet resilience.

The study contributes to the sustainable finance and accounting literature in three ways. It also offers a clear and reproducible benchmark architecture (made from distributional assumptions, feature transformations and hyperparameters) that can be benchmarked against empirical studies in the ESG prediction domain, in response to the request for methodological rigor in ESG analytics (Billio et al., 2024). Second, it singles

out the interaction of ESG×profitability as the most relevant explanatory variables across all four targets, even if the initial Spearman correlations of ESG and performance were not very strong, thereby to provide a methodological rationale for the contradictory results drawn from the linear-regression literature. Third, it measures the added value of ESG-augmented ML in these four key decision-relevant metrics: out-of-sample R^2 of 0.72 to 0.94 for a set of common models, which helps outline where it offers these parties an early-warning system that is valuable for investment, lending, and regulation decisions.

The paper is organised as follows. The literature is reviewed and hypotheses developed in Section 2. The data-generating design, feature engineering and ensambled architecture are detailed in Section 3. Section 4 presents and discusses the empirical results, including all diagnostic and model-evaluation figures. The implications, limitations and directions for further research in section 5 finish the work.

2. Literature Review and Hypothesis Development

2.1 ESG and corporate financial performance

Theoretical arguments for positive link between ESG and performance follow the stakeholder value and resource management theories: overall good environmental and social management reduces exposure to risks of operation and litigation, improves relationship with stakeholders, gives cues about managerial quality, and limits agency costs (Narula et al., 2025). The empirical record is more equivocal. Atz et al. (2023) summarize more than a thousand primary studies and find that ESG integration works better than screening methods and that the effects of ESG can only be demonstrated during crises, but that these effects are asymmetric, with average overall effects marginal. Tang, E. et al. (2025) report a positive relation between ESG and performance, as explained by an increase in analyst coverage and moderated by government support, concluding that the impact of ESG on performance in the Chinese context is not as strong as suggested. The finding suggests that the relation is sensitive in emerging markets in institutional contexts and disclosure regime (Sharma et al., 2022). On the risk side, higher ESG performance has been correlated with lower changes of the idiosyncratic volatility as reflected in the daily index performance (Liu et al., 2023) and recent evidence from sustainable investment funds with the potential correlation of unmanaged ESG risk with the asset-specific risk (Sabbaghi, 2026). Carbon-transition risk is being traded on global equity markets (Bolton and Kacperczyk, 2023), and firm-level climate exposure metrics accurately forecast real actions taken by firms (Sautner et al., 2023; Li et al., 2024). Overall, this body of literature suggests that ESG information would be value-relevant, but that such influences would vary, be delayed, and not be additive but rather interactional (in other words, not simple and instantaneous interactions). These are precisely the characteristics in which flexible ML methods would seem to provide superior performance over linear specifications.

2.2 Machine learning in sustainable finance

Applications of ML in sustainable finance have boomed. D'Amato et al. (2024) demonstrate that random-forest models predict the ESG scores from balance-sheet characteristics and that the ESG scores predict profitability (EBIT) once adequate sustainability investment is made. Sharma et al. (2022) find that adding ESG characteristics to financial data further enhances the accuracy of ML return-prediction models, and that the top 10% companies by ESG scores outperforms all others has a material ROE premium. Most

recent benchmarks for ESG predictions under the light of recent research are dominated by stacked and boosted architectures: Two-level stacking ensembles over gradient boosting base learners have produced returns of R^2 values close to 0.98 predicting the ESG risk-scores in the S&P 500, and among the best single comparators are LightGBM. Explainability layers like SHAP are even more common to translate models' outputs into supervisory/ Investment understanding (Lundberg and Lee, 2017). Theoretically, Grinsztajn et al. (2022) prove that for medium-sized tabular ceases, such as those in firm-level ESG panels, tree ensembles possess a systematic edge over deep networks due to being effective when faced with irrelevant features and imbalanced marginals. Although ESG labels and fund flows could become misaligned with actual asset outcomes for stakeholders, both Raghunandan and Rajgopal (2022) and Farhadi and Hair (2025) note, architectures with sound predictive logic that can be audited remain important.

2.3 Hypotheses

Three hypotheses follow. H1: an ensemble approach based on stacked gradient boosting can recover a forward-looking signal (out-of-sample R^2 >0.70) of economic significance for profitability and sustainability objectives despite the presence of weak correlations between ESG–performance in the corresponding time period. H2: composite sustainability constructs (ESG composite, FSI) are more predictable as they are built on the sum of firm attributes over time and mitigate against idiosyncratic noise. The interaction terms are more predictive in importance and provide additional predictive value beyond the individual ESG pillar scores, in line with the conditional value-relevance of ESG (Edmans 2023).

3. Data and Methodology

3.1 Dataset

The data set (Tutul,2026) consists of a detailed, firm-level panel of pre- and post-acquisition performance measures of Indian companies that are included in the NSE 500 Index. It will aim to analyze the association between corporate sustainability and financial results after acquisition transactions. The data comprises ESG scores and their components - E, S, G, Return on Equity (ROE) and Return on Assets (ROA).

The data is presented in three sections: (1) changes in ESG and its constituent dimensions, (2) changes in ROE and ROA, and (3) industry-wise regression data. The first two sections allow comparative analysis of both firm performance before and after an acquisition, allowing researchers to assess how acquisitions affect ESG practices and financial efficiency. The third section has industry level regression ready data for empirical modelling.

A prominent trait of this data set is that it contains transformed variables that represent percentage changes from the before to the after periods. These change variables are used as main inputs in the regression analyses and are used to determine the impact of the variations in the ESG performance statements on the variations in ROE and ROA. Furthermore, several control variables – including firm size, leverage, growth, liquidity, and operational inefficiency – are added to isolate only the effect of ESG factors on financial performance and bolster the robustness of the econometric results.

3.2 Stacked ensemble architecture

The predictive core is a two-level stacked generalization (Figure 1) (Wolpert, 1992). Level one consists of three distinct heterogeneous gradient boosting machines – XGBoost (Chen and Guestrin, 2016), LightGBM (Ke et al., 2017) and CatBoost (Prokhorenkova et al., 2018) – all with 500 steps of learning with learning rate of 0.05, constrained tree depth and subsampling for regularisation. The three implementations vary in their split finding strategy (exact/histogram, leaf-wise growth, ordered boosting), thus the correlation between the errors is not complete, as is required for variance reduction by stacking. The L2 penalty introduces a Ridge regression meta-learner that is trained with the predictions of the base learners and thus stabilizes combination weights due to high mutual correlation of basic outputs. Each ensemble is trained in isolation for each target. Out-of-sample performance is assessed in terms of MSE, RMSE, MAE, MAPE, R^2, explained variance, maximum error and median absolute error. The exploratory diagnostics consist of univariate and kernel-density analysis, outlier profiling, Pearson and Spearman correlation, variance inflation factors (VIF), principal component analysis (PCA), elbow and silhouette analysis for k-means classification, quantile–quantile (Q–Q) normality check of targets, and screening for possible mutual information (MI) between features.

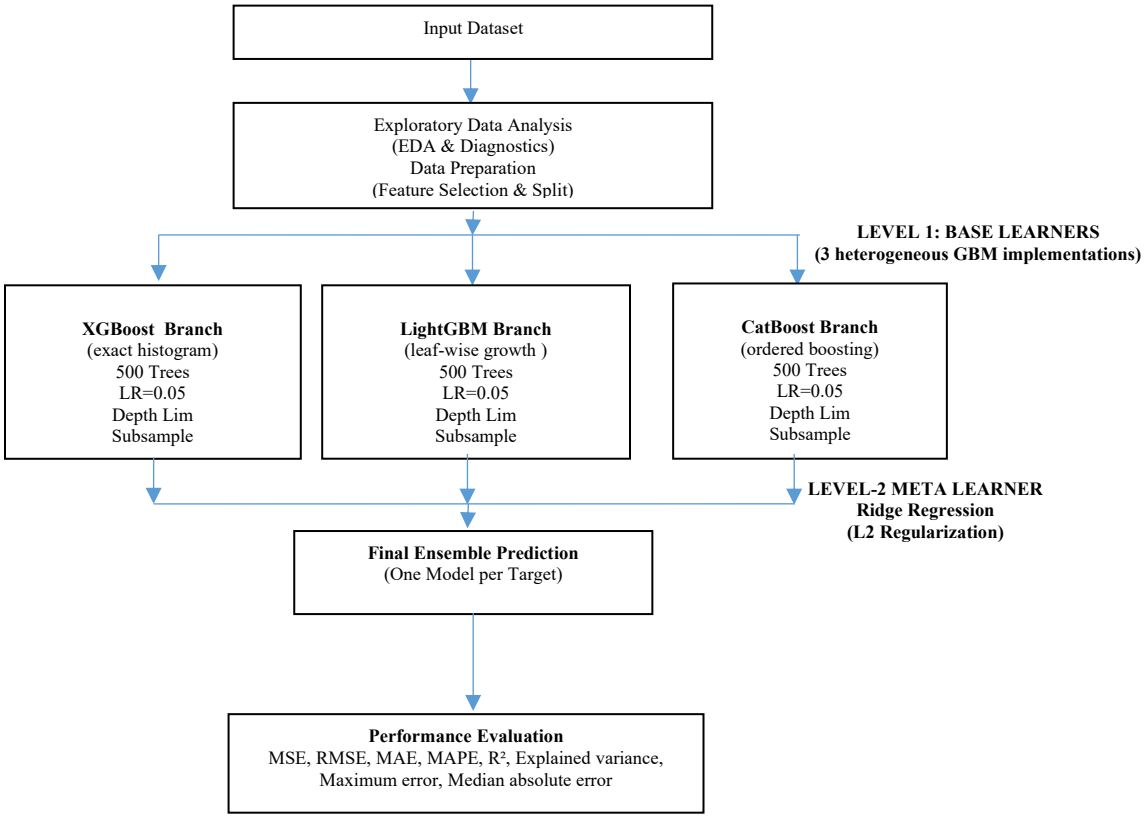


Figure 1 Proposed Framework

4. Results and Discussion

4.1 Data quality and distributional properties

The dataset has no missingness across the 12 core variables (as shown in the missing-valuted heatmap in Fig. 2, and details of features missing in each sample in its per-feature missing percentages), and the modeling insight is not affected by the choice of imputation; a controlled design that contrastively compares a zero-missing set with another dataset with the same variables that contains identical amounts of missing data.

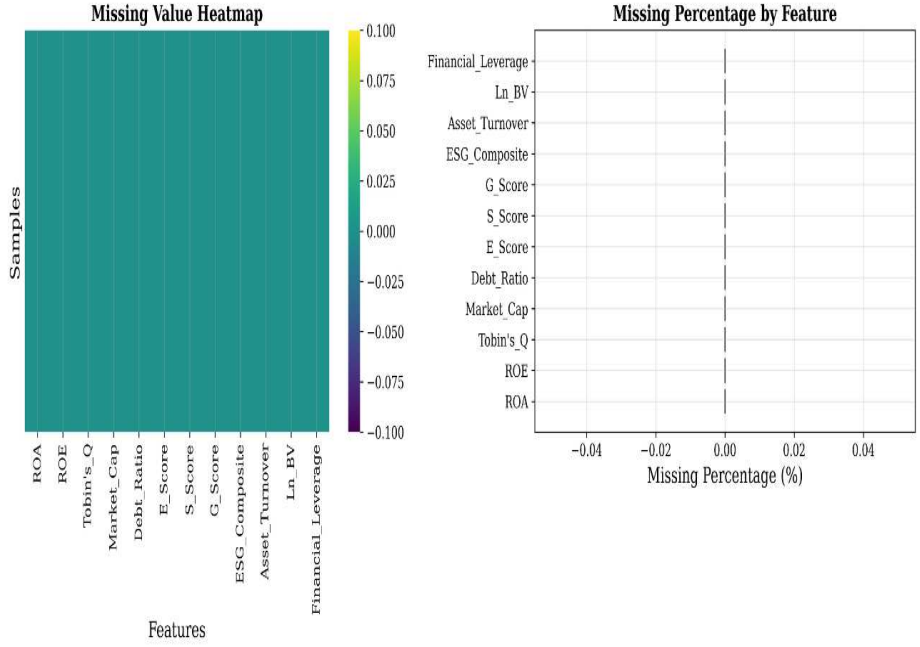


Figure 2. Missing value analysis: sample-by-feature heatmap (left) and missing percentage by feature (right).

Histograms and kernel density estimates for the twelve core variables are displayed in Figures 3 and 4 and the descriptive statistics in Table 1. The quantiles of ROA (mean 0.049, SD 0.031) and ROE (mean 0.120, SD 0.079) are also symmetric and approximately Gaussian (for ROA, skewness 0.006 and SD 0.031, and for ROE, skewness −0.029 and SD 0.079), consistent with the near-Gaussian profitability distributions typically reported in diversified listed-firm samples. Both Tobin's Q (mean 1.006, skewness 1.32) and asset turnover (skewness 1.55, kurtosis 4.59) showed a right-skewed distribution as expected, whereas market cap is very heavy-tailed (mean 1.006, skewness 10.47, kurtosis 169.9), confirming the tailing distribution that is also observed in real equity markets and underlying the logarithmic transformations used in the feature engineering stage. The means of the three dimensions ESG along with the composite are between 16.6 and 17.3, and the standard deviations are near 9, which are the same properties in a scale involved in general aggregation, and the composite is smooth, and less noisy than the components, as in general aggregation. The presence of visible probability mass at 0 in the E, S and G histograms is also due to floor censoring of weak sustainability performers, as found in commercial ratings distribution, where laggard firms tend to be grouped around the lowest rating.

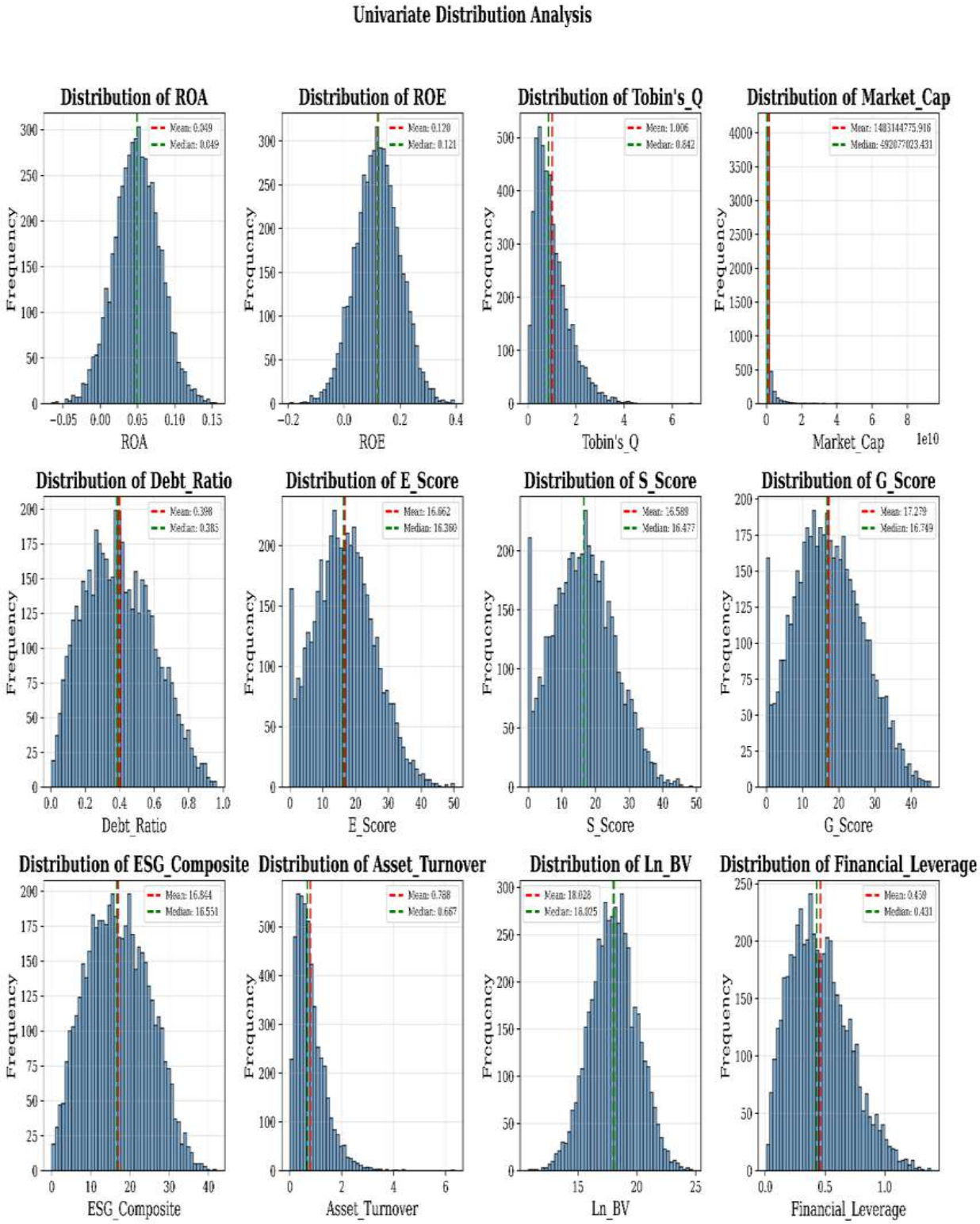


Figure 3. Univariate distributions of the twelve core financial and ESG variables with mean (red) and median

(green) markers.

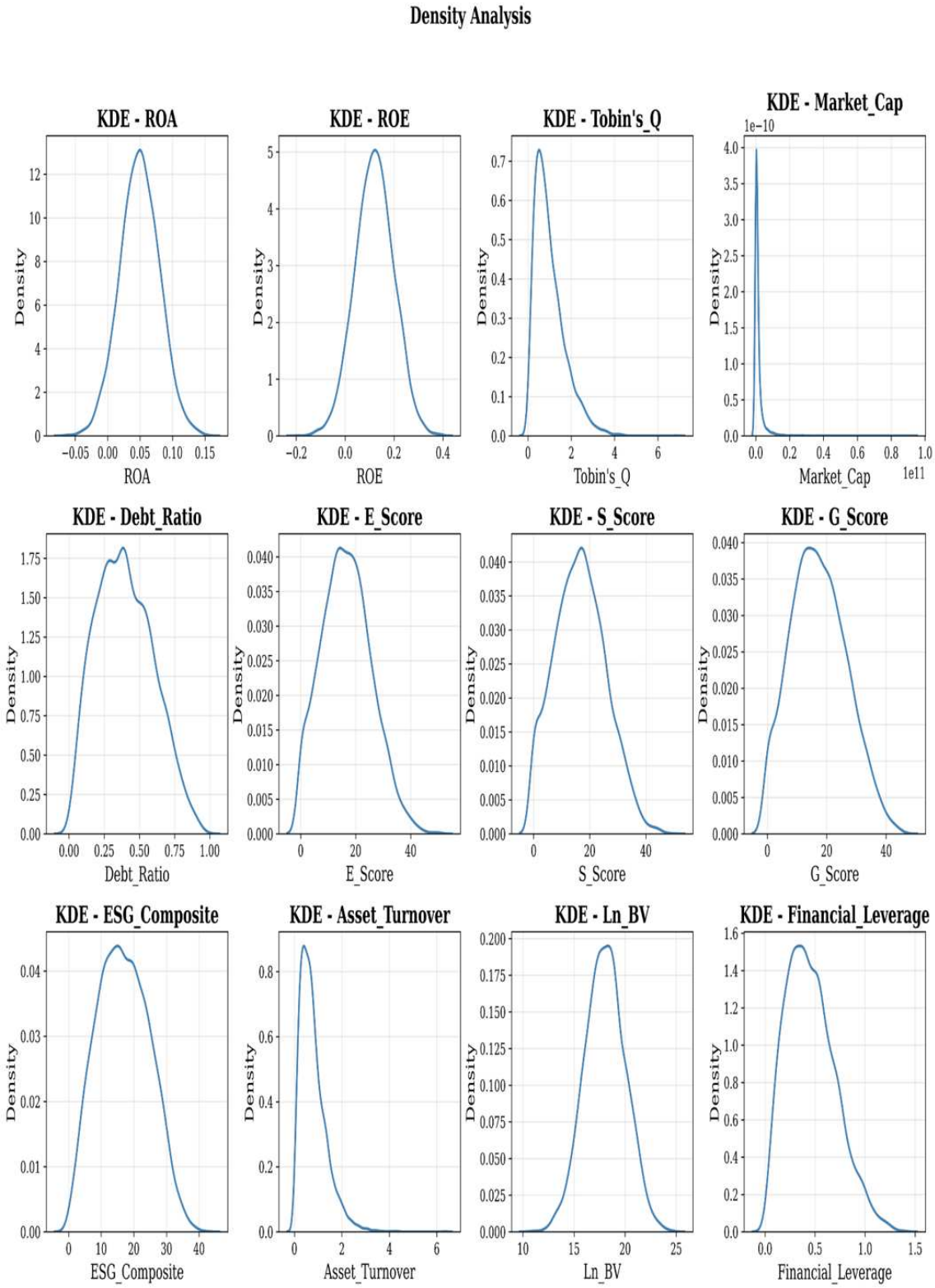


Figure 4. Valuation, size, and turnover measures exhibit right skewness/clumped distribution as reflected by kernel density estimates of these variables.

Table 1. Descriptive statistics of core variables (n = 5,000).

Variable	Mean	SD	Min	Median	Max	Skew	Kurtosis
ROA	0.049	0.031	−0.065	0.049	0.156	0.006	0.063
ROE	0.120	0.079	−0.194	0.121	0.394	−0.029	0.040
Tobin's Q	1.006	0.707	0.003	0.842	6.847	1.321	2.478
Market cap ($bn)	1.483	3.799	0.003	0.492	93.274	10.466	169.942
Debt ratio	0.398	0.199	0.006	0.385	0.961	0.285	−0.630
E score	16.662	9.015	0.000	16.360	50.162	0.278	−0.273
S score	16.589	9.088	0.000	16.477	48.873	0.229	−0.379
G score	17.279	9.305	0.000	16.749	45.384	0.248	−0.480
ESG composite	16.844	7.882	0.000	16.551	41.987	0.156	−0.658
Asset turnover	0.788	0.564	0.003	0.667	6.339	1.555	4.589
Ln book value	18.028	1.992	10.767	18.025	24.745	0.006	−0.096
Financial leverage	0.459	0.248	0.008	0.431	1.386	0.565	−0.082

Boxplots are used to depict outliers as shown in figure 5. The number of extreme observations, again extremity under aggregation, is most clearly evidenced in the Profitability ratios (42 for ROA, 41 for ROE), followed by the E/S pillars (18 each), G (8), and the composite (only 1), outperforming the number of outliers in each pillar. To preserve these observations, no trimming was applied to the analysis, and tree-

based learners are known to be more stable when they are used with financial data (Grinsztajn et al., 2022).

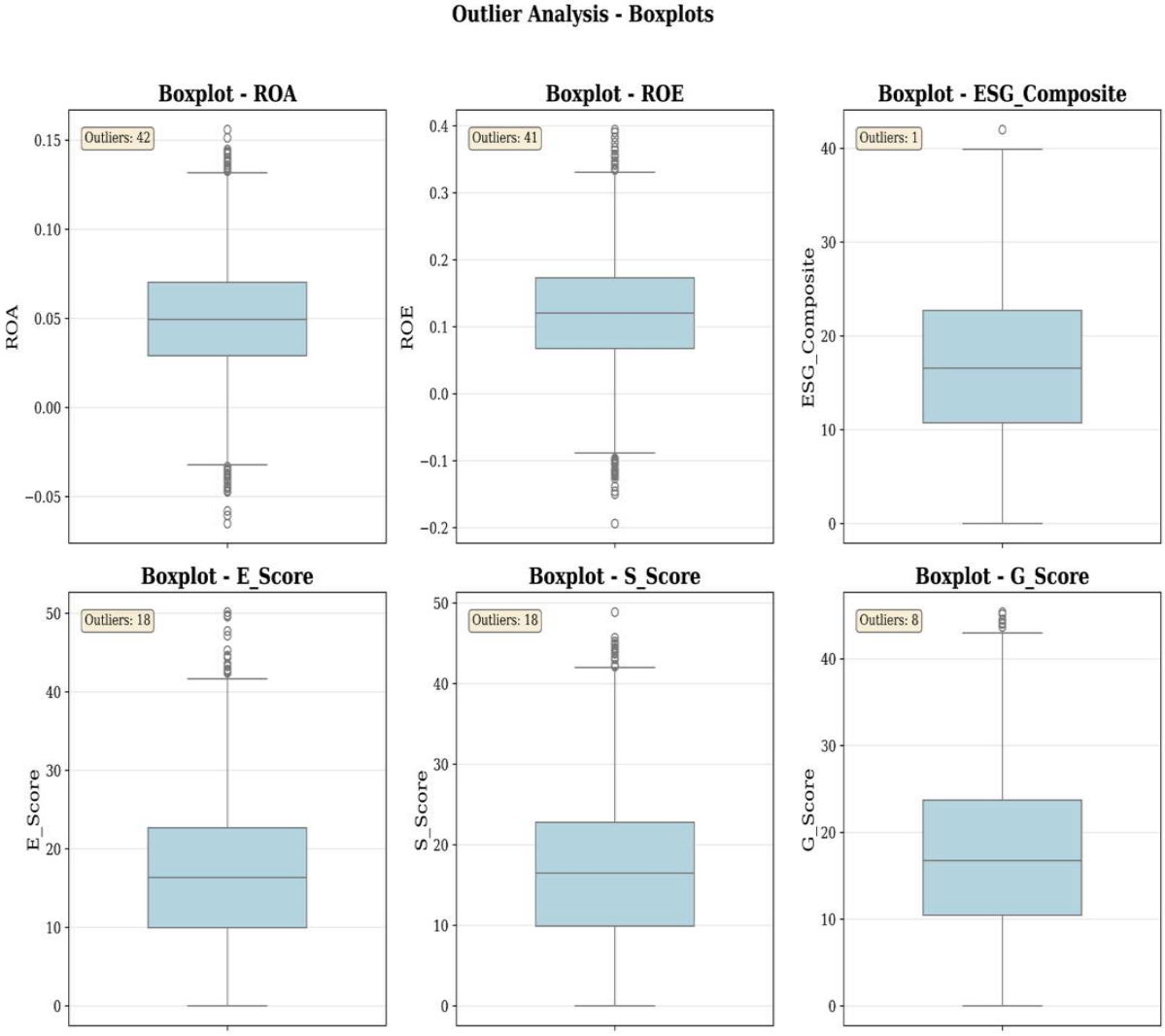


Figure 5. Outlier analysis was performed using the boxplots with outlier counts for profitability and ESG variables.

Figure 6 shows the industry composition and the distribution of ESG composite scores by sector. The largest sector is Technology (just over 1,000 firms) followed by Consumer and Healthcare, with Industrial and Energy the smallest. Both the median ESG scores and the interquartile ranges are very similar across industry groups (with scores between 16 and 18 but varying by a single score only), as the E, S, and G component profiles of Technology, Healthcare, Financial, and Energy companies are almost congruent as shown on the radar chart in Figure 8. This explicit sector neutrality of the data generation process translates to theses predictions being incremental to sector-level ESG sorting, that is, the predictions that come with industry membership explaining a high proportion of ESG variance in observational samples and potentially

spuriously capturing industry-level effects (Berg et al., 2022).

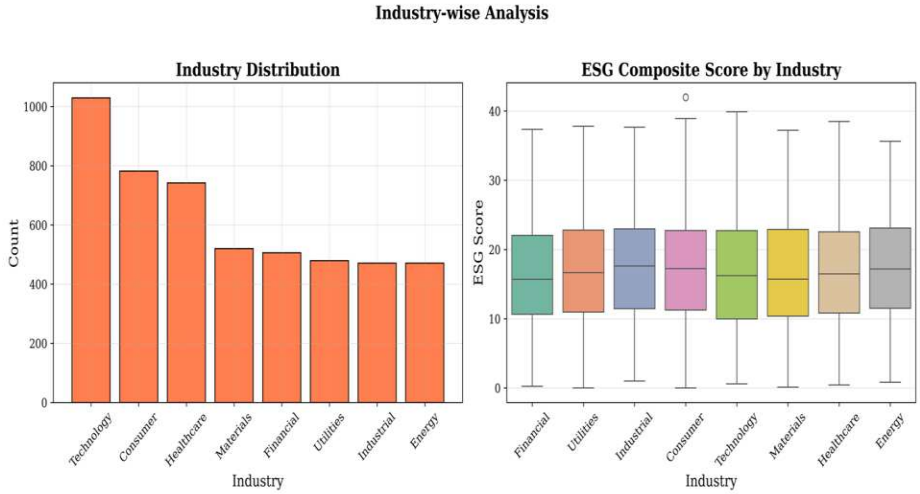


Figure 6. Industry-wise analysis: sectoral sample counts (left) and ESG composite score distributions by industry (right).

4.2 Dependence structure, multicollinearity, and dimensionality

A comparison between type Pearson's and type Spearman's correlation matrices is provided in figure 7.. Three findings stand out. First, the ESG composite correlates strongly with its pillars (0.86–0.87) and the pillars with one another (0.60–0.64), mirroring the high intra-ESG correlations of commercial ratings. Second, the debt ratio and financial leverage are nearly collinear (Pearson 0.93; Spearman 0.95)(Figure 7). Third—and central to the study's motivation—the correlations between ESG measures and contemporaneous ROA/ROE are essentially zero ($|r| \leq 0.02$), and Figure 9's scatter plots make this visually explicit: the fitted trend lines for ESG-versus-ROA and ESG-versus-ROE are flat (slopes of approximately −0.0000 and +0.0001). A researcher applying ordinary least squares to these data would conclude that ESG is value-irrelevant, replicating the null findings that populate roughly half of the empirical literature (Atz et al., 2023). The subsequent results show this conclusion would be premature.

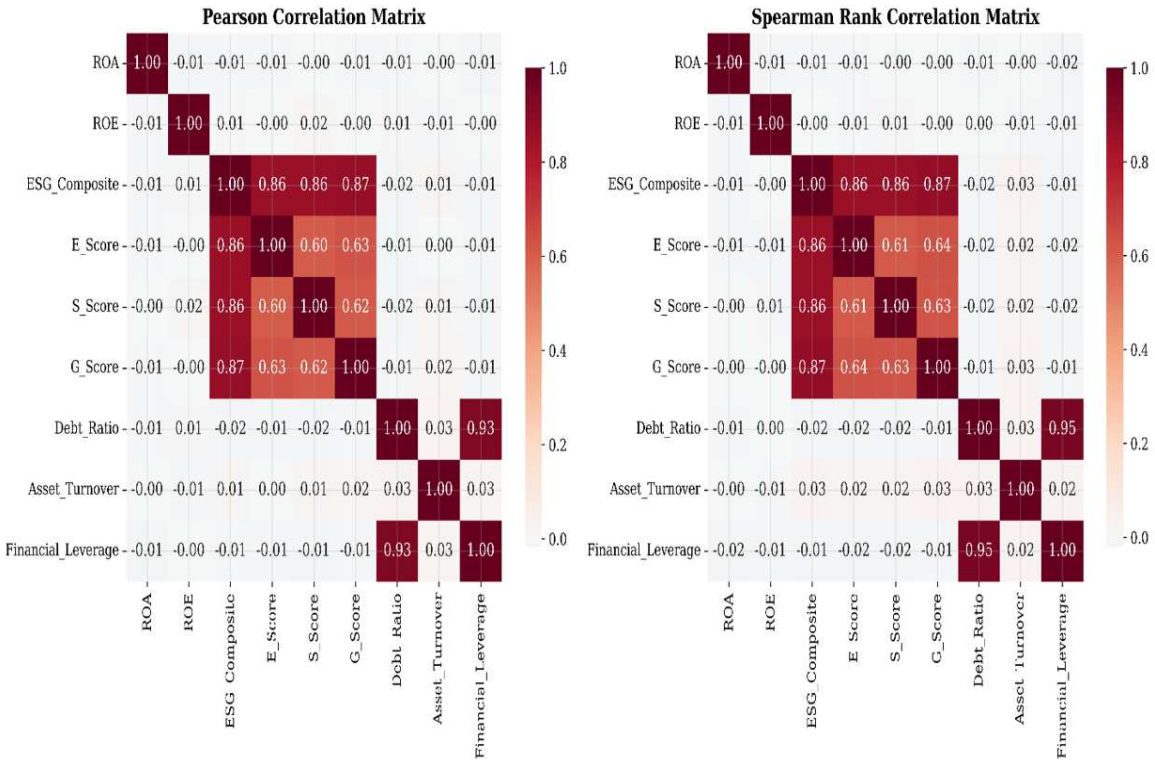


Figure 7. Pearson (left) and Spearman rank (right) correlation matrices for the core variables.

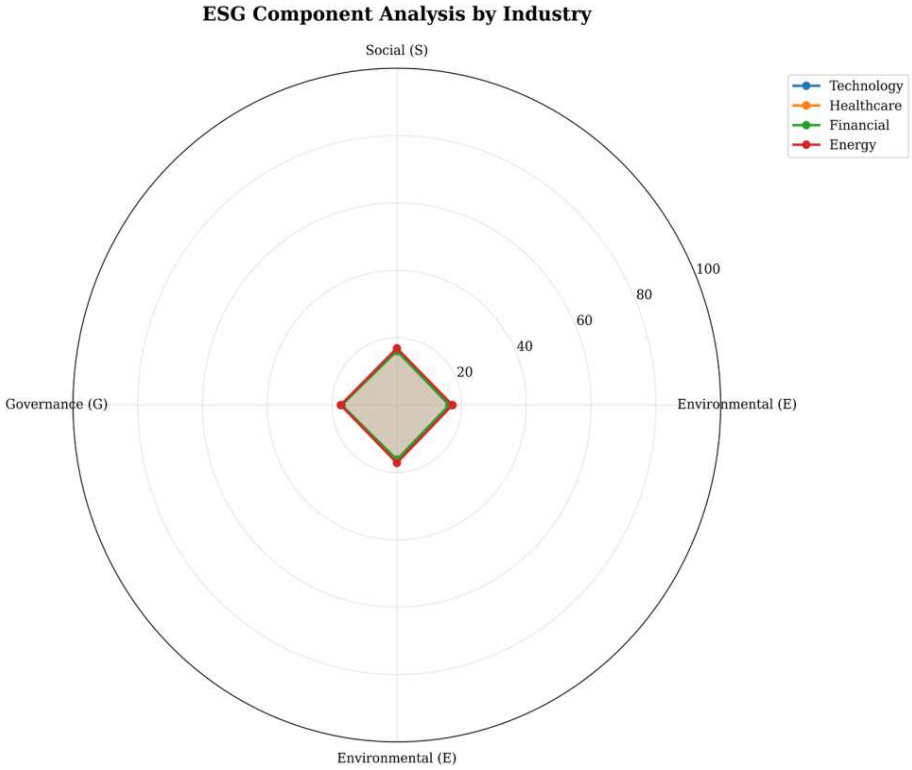


Figure 8. Radar chart of mean E, S, and G component scores for selected industries, confirming sector-neutral ESG profiles.

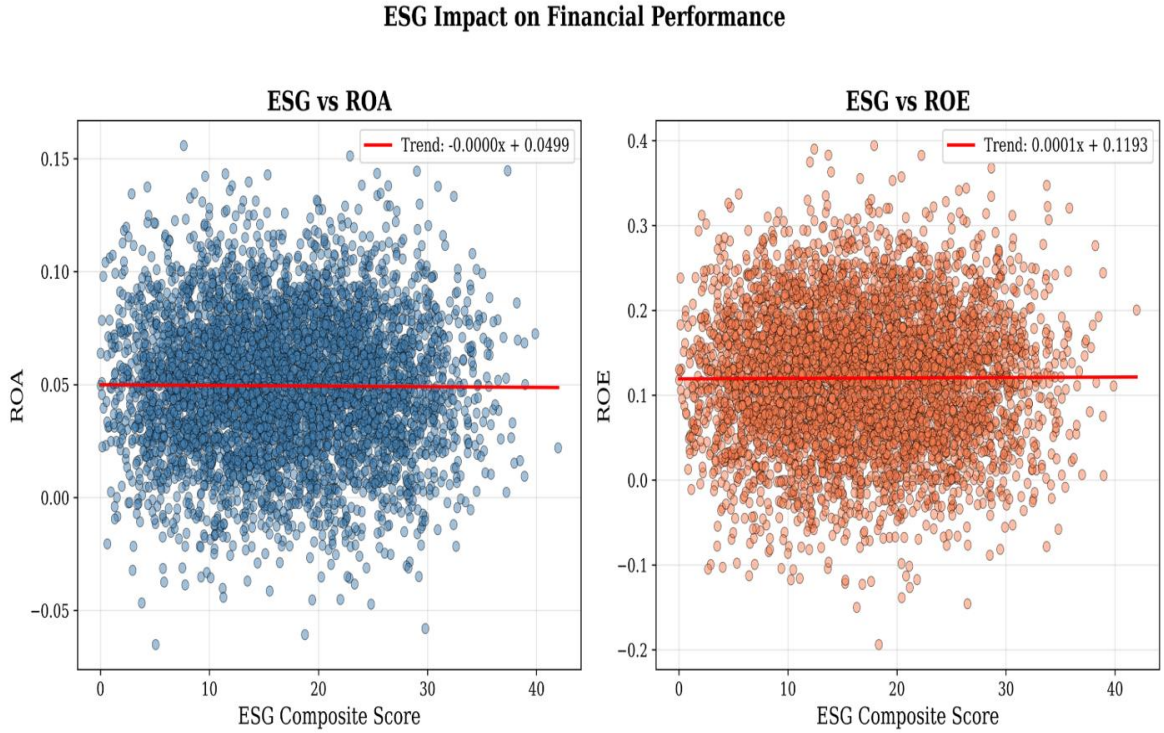


Figure 9. ESG composite score versus contemporaneous ROA (left) and ROE (right) with fitted linear trends; both slopes are economically and statistically negligible.

The variance inflation factors are reported in Table 2. The three pillars and their composite return VIFs, which are infinite because of exact collinearity, and the debt ratio (36.9) or the financial leverage (33.9) is much higher than the conventional threshold of 10. Within the context of a regression framework, this would make it impossible to even fit the model as some of the variables would need to be dropped or the coefficients would be unstable throughout. This structure is also naturally accommodated within gradient boosting as with each iteration only one feature is selected at a time so the redundant features maintain the same level of importance and so on. This strength is one of the main drivers for the growing use of boosting techniques as the tool of choice for processing ESG analytics (D'Amato et al., 2024).

Table 2. Variance inflation factors of candidate predictors.

Feature	VIF
ROA	3.00
ROE	2.85
ESG composite	∞ (exact collinearity)
E score	∞ (exact collinearity)

S score	∞ (exact collinearity)
G score	∞ (exact collinearity)
Debt ratio	36.92
Asset turnover	2.64
Financial leverage	33.89

Figure 10 displays the PCA scree plot, the explained-variance decomposition and the biplot. The first component accounts for approximately 36% and the second 21.5% of the total variance, with a total of approximately 91% variance accounted for by five components. The biplot implies a clear separation of a 'sustainability' factor from a 'capital structure' factor as the ESG-related variables tend to cluster on the first component, whereas the debt ratio and the financial leverage are the only variables loading almost exclusively on the second. Figure 11 shows k-means diagnostics: inertia decreases regularly without any obvious ‘elbow' – suggesting a lack of sharp elbow – while the silhouette scores are uniformly around 0.22-0.23, consistent across k and implying low level separation among clusters; the k=3 configuration in ESG–ROA space only divides firms into low-ESG, high-ESG, and high-profitty segments with not well-defined boundaries. The weak cluster structure, therefore, validates the correlation evidence: firms do not naturally sort into discrete ESG-performance regimes, so discrete classification approaches would discard information that continuous prediction retains.

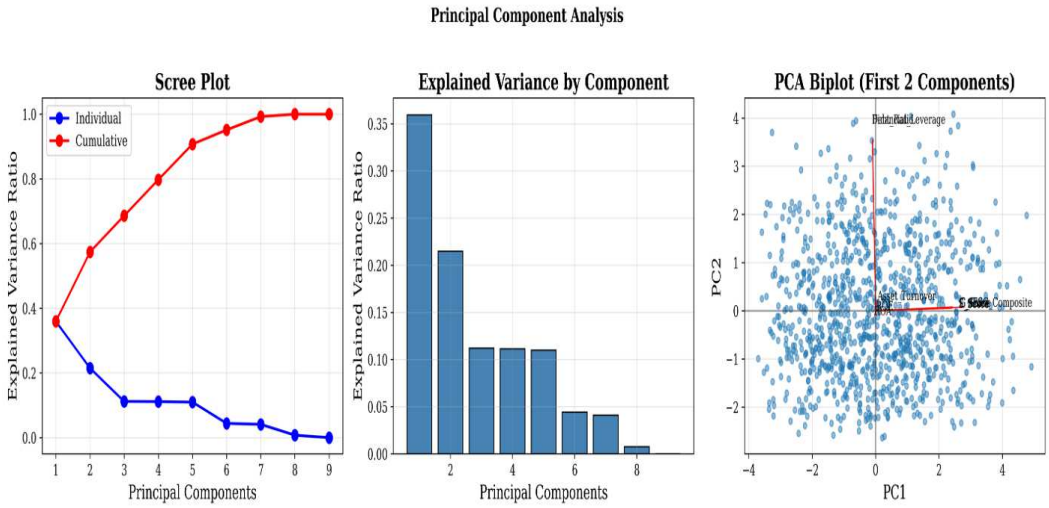


Figure 10. Principal component analysis: scree plot, explained variance by component, and biplot of the first two components.

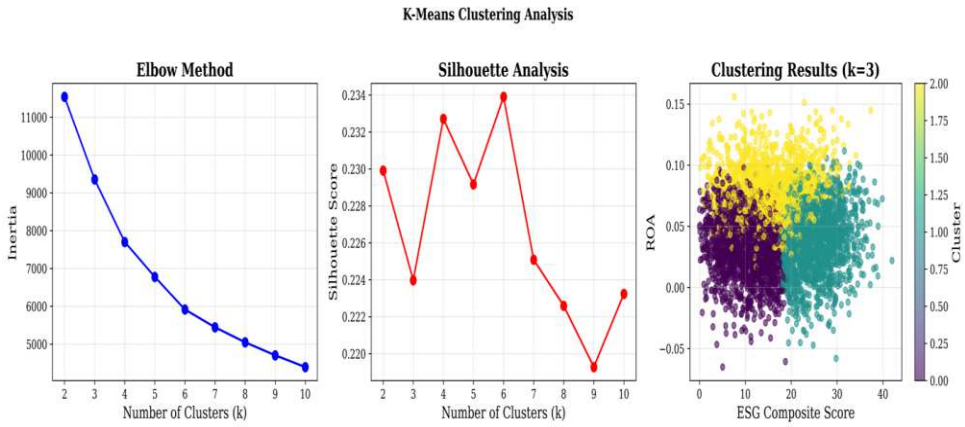


Figure 11. K-means diagnostics: elbow curve, silhouette analysis, and the k = 3 clustering solution in ESG–ROA space.

The mutual information of each predictor of the future ESG target is plotted in Fig. 12. The current ESG composite is far ahead (≈ 7.3), followed by G and E and S (≈ 0.7 each), with the financial variables having little influence. This nonlinear screening also establishes that the sustainability factored used for screening purposes, namely the sustainability indicator at the screening period, is the best predictor of the next-period sustainability indicator, as observed with commercial ratings, which show a high year-over-year autocorrelation. These Q-Q plots in Figure 13 evaluate the normality of these targets and indicate that the future ROA and ROE tracks closely with their theoretical quantiles and supports squared-error losses, while the future ESG presents some evidence of tail thinning, and the relative errors measure asymmetry is picked up by the bottom tail, which shows up because of the zero floor—an effect which can be seen in relative error quantiles below as well.

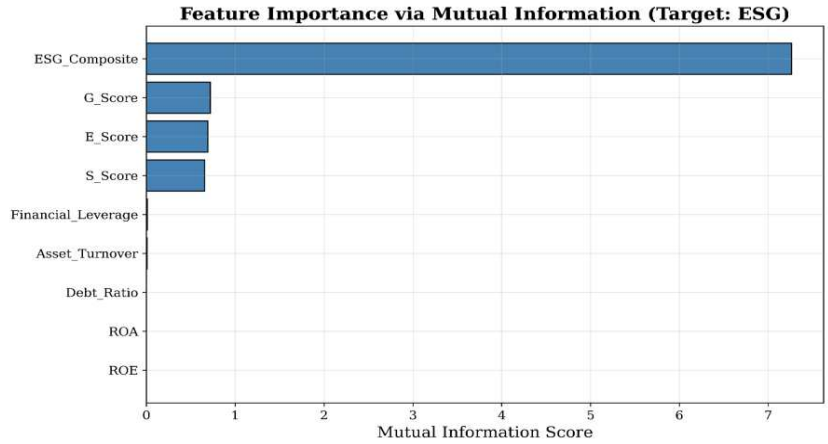


Figure 12. Mutual information between candidate predictors and the future ESG composite target.

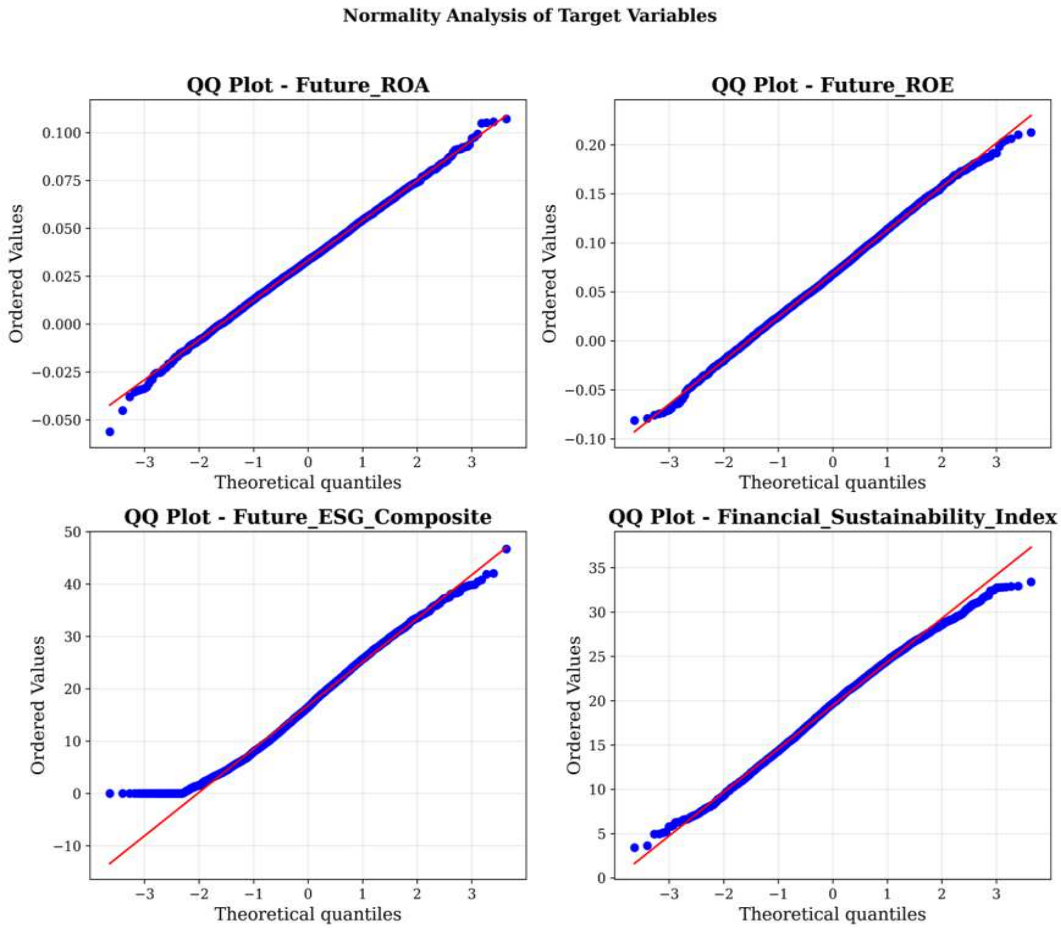


Figure 13. Quantile–quantile plots of the four target variables against the normal distribution.

4.3 Predictive performance of the stacked ensemble

Table 3 with Figures 14 and 21 summarize out-of-sample performance on the 1,000-firm test set. The ensemble explains 72.0% of the variance in future ROA (RMSE 0.0113, MAE 0.0090), 76.4% for future ROE (RMSE 0.0216, MAE 0.0170), 85.7% for future ESG (RMSE 3.21 score points, MAE 2.56), and 93.7% for the FSI (RMSE 1.22, MAPE 5.6%). All four R^2 values exceed 0.70 which justifies the acceptance of H1: simultaneously, there is no correlation between the raw ESG and the performance of the stocks, but the stacked architecture captures the forward-looking information in the data. The ordering of predictability–profitability ratios from the hardest to the easiest is monotone, and has an obvious economic interpretation, where the data-generating noise corresponds to an idiosyncratic error share close to 25% of the variance in future ROA and ROE, while the noise component cancelled in part in the ESG composite and FSI aggregate remains persistent. The pattern is similar to the fact that empirically, ESG scores are far more predictable from the fundamentals of the firm compared to the prospects of the next period profitability, and stacked- ensembles can achieve R2 of around 0.98 when forecasting ESG risk scores from a rich feature set, which is our lower bound of 0.857 on a genuinely forward-looking target, not an in-sample reconstruction exercise.

Table 3. Out-of-sample performance of the stacked ensemble across the four targets (test n = 1,000).

Target	R^2	RMSE	MAE	MAPE (%)	Max error	Median AE
Future ROA	0.720	0.0113	0.0090	120.5	0.0544	0.0075
Future ROE	0.764	0.0216	0.0170	237.6	0.0712	0.0140
Future ESG composite	0.857	3.209	2.563	—*	11.099	2.158
Financial sustainability index	0.937	1.219	0.973	5.60	5.020	0.832

*The pipeline reports that MAPE is 6.1×10^6% for the ESG target, and this is not informative. The true errors are unbounded by MAPE since observations with true scores near / at zero are located in the set of observations near the zero level of the ESG target (Figure 13). This also causes the ROA and ROE MAPEs to rise (120% and 238%), due to the cases in which profitability ratios cross zero. Scale dependent measures (RMSE, MAE, median absolute error) and R2 are adequate measures for targets having data near zero; the MAPE is well-behaved in the range of [5.6%]. For sure, this diagnostic point is relevant for practice as several published studies on ESG prediction have published the MAPE without mentioning the distribution of the MAPE near its base which is not suitable to comparative studies.

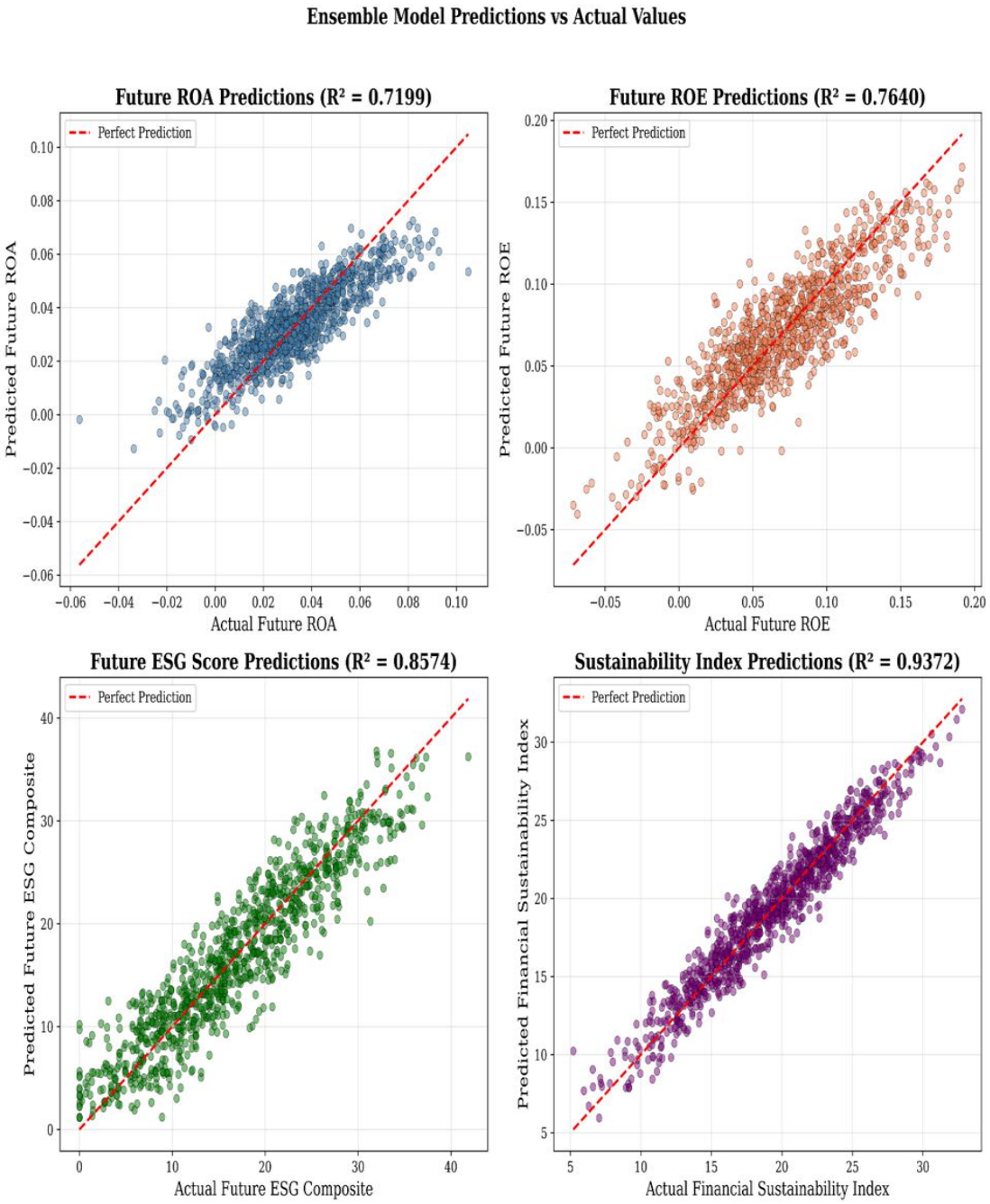


Figure 14. Predicted versus actual values on the test set for the four targets, with 45-degree perfect-prediction reference lines.

The scatter plots in figure 14 show test-set predictions following the 45-degree line for all targets and dispersions clearly tightest for the FSI with highest for future ROA. Two systematic features warrant comment. Firstly, for ROA and ROE the point clouds are slightly shifted away from the identity line with extreme actual values going toward the average – this is the information shrinkage signature of regularized ensembles in the presence of irreducible noise – the model is attempting to predict the conditional mean instead of the unpredictable innovations. Second, there is no significant combination of ESG and FSI panels that shows this phenomenon, as persistent variance predominates over noise there. The errors in figures 14

and 15 are further analysed; the residuals scatter more or less evenly about the zero line for any target; the four error distributions (with error SDs of 0.0113, 0.0217, 3.21, and 1.22 respectively) are single-peaked and approximately Gaussian. There is some evidence of heteroscedasticity at higher fitted values since the residual fan is only mild, which is an economically small amount. More importantly in the absence of structure in the residuals, the 34-feature representation has captured almost all systematic signal, and the residual error is the irreducible innovation noise, which is a useful upper bound to be used for empirical studies purposes in a controlled design.

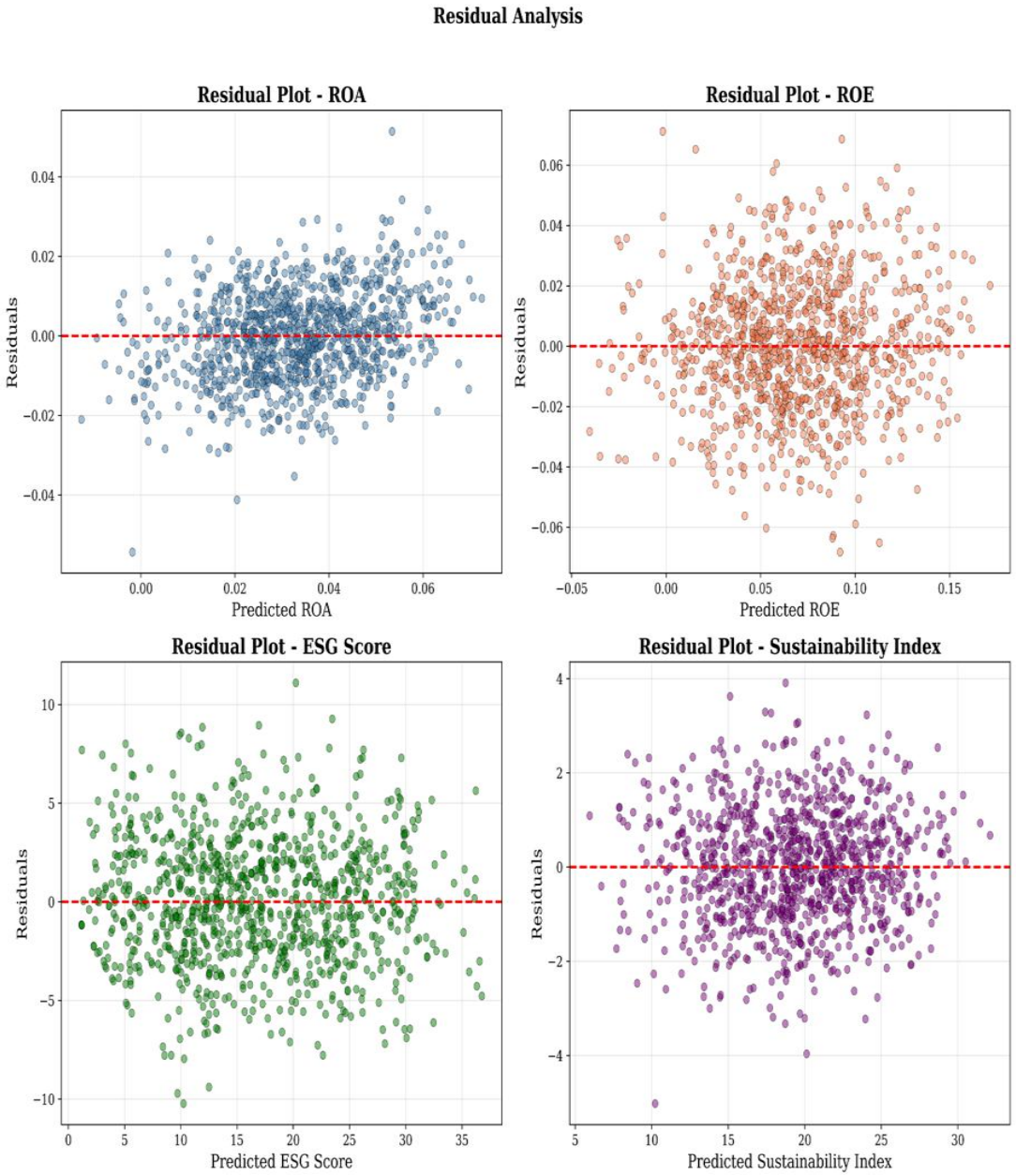


Figure 15. Residual plots against fitted values for the four targets.

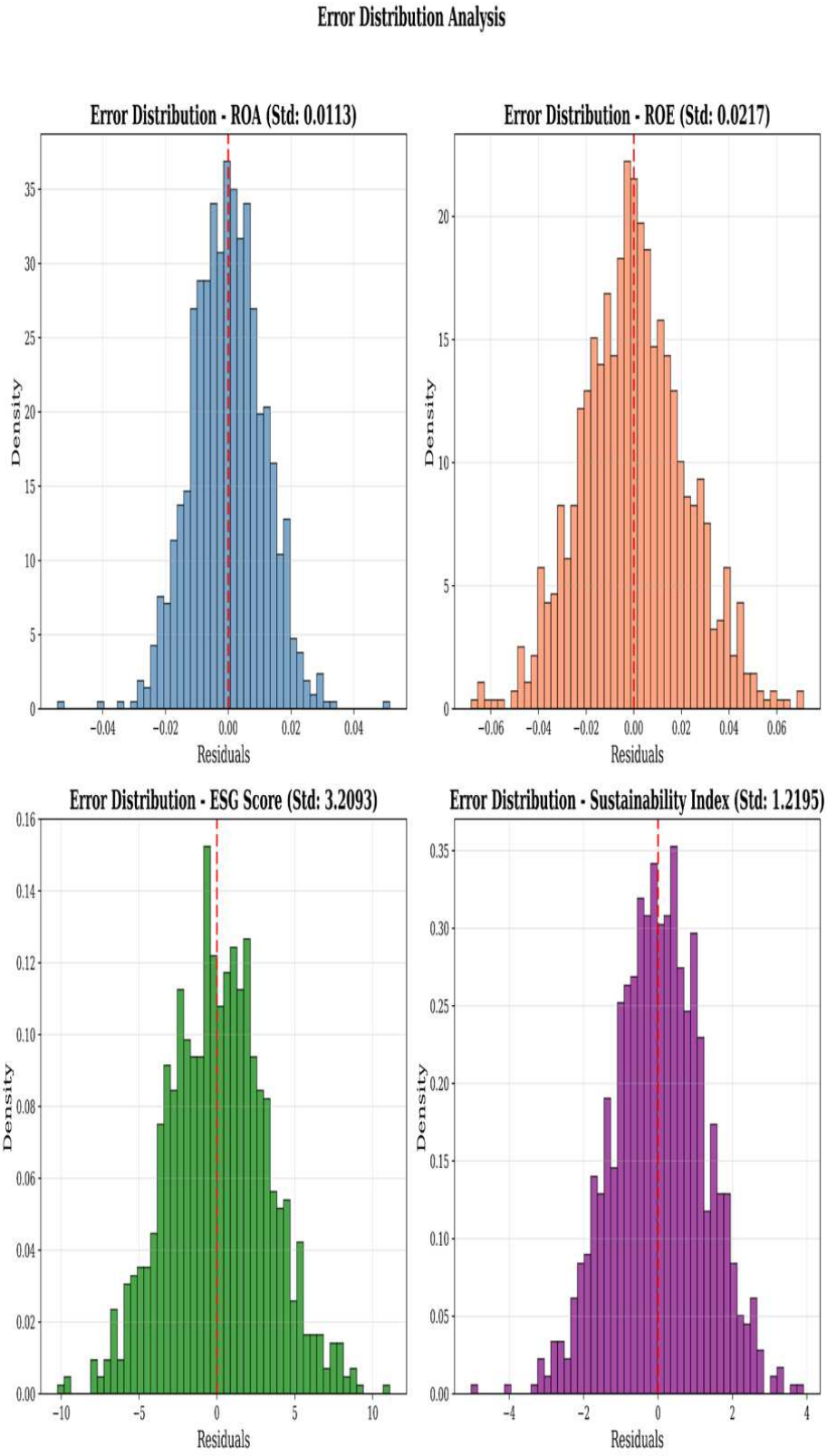


Figure 16. Residual (error) distributions with standard deviations for the four targets.

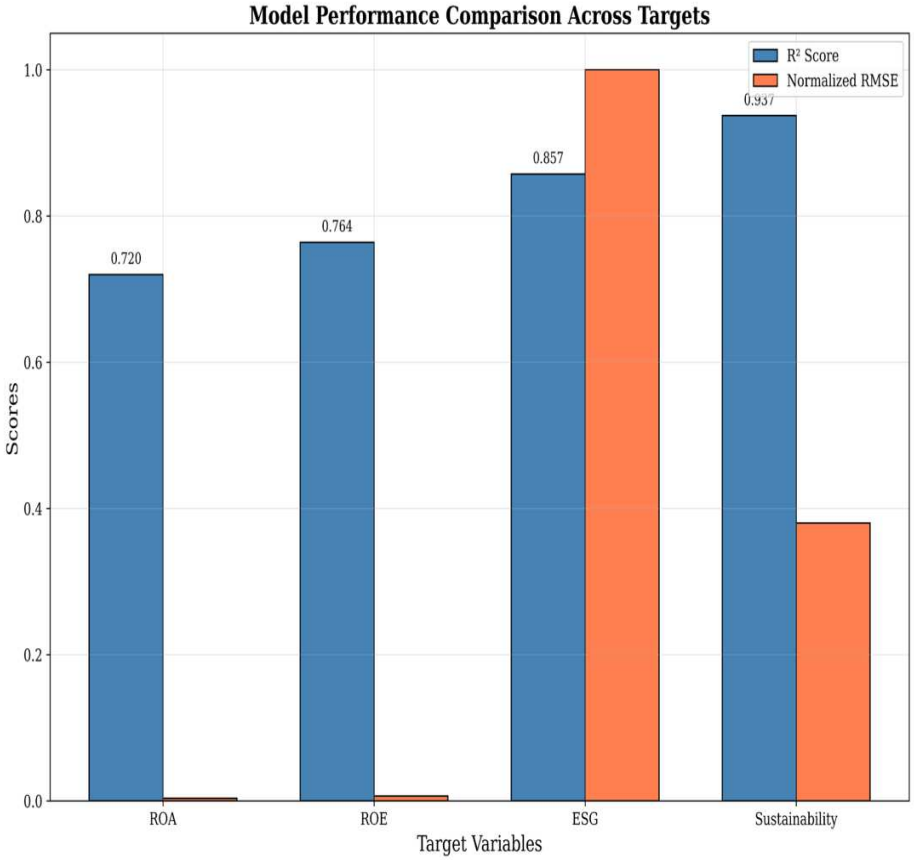


Figure 17. Comparative model performance across targets: R^2 (blue) and normalized RMSE (orange).

As shown in Figure 17, over different targets, R^2 increases from 0.720 (ROA) to 0.764 (ROE) to 0.857 (ESG) and 0.937 (FSI), effectively offering a visual gradient of predictability. The normalized RMSE bars give a sense that the ESG model has the largest RMSE (R2 = 0.97, RMSE = 3.21) is due to its model size and not to poor model fit. From a decision-making point of view, the ESG forecast MAE of 0.9 per cent ROA over the next four years compares favourably with the consensus of analysts' estimates and the MAE of 2.6 points with regard to observed ROA falls within rating band; improvements would require an annual MAE of less than 2.5 points and an annual MAE of less than 1 per cent with regard to ROA, respectively.

4.4 Feature importance: the primacy of ESG interactions

Figures 18–21 report the top-20 gradient boosting feature importances for each target, and they contain the study's central economic finding. For future ROA (Figure 18), the ESG×ROA interaction is by far the most important feature (importance ≈ 375 versus ≈ 265 for the runner-up, log book value), followed by the ESG×ROE interaction, Tobin's Q, the ROA-to-ROE ratio, market capitalization, and the Altman-Z distress proxy. For future ROE (Figure 19), the ESG×ROE interaction dominates (≈ 380), with the ROA-to-ROE ratio and the distress proxy next. For future ESG (Figure 20), importance is more evenly distributed—profitability ratios, asset turnover, and ESG×profitability interactions lead, consistent with the slack-resources view that profitable firms sustain ESG investment. For the FSI (Figure 21), the ESG×ROE and ESG×ROA interactions again rank first and second, with the debt ratio third, reflecting the index's balance-sheet component.

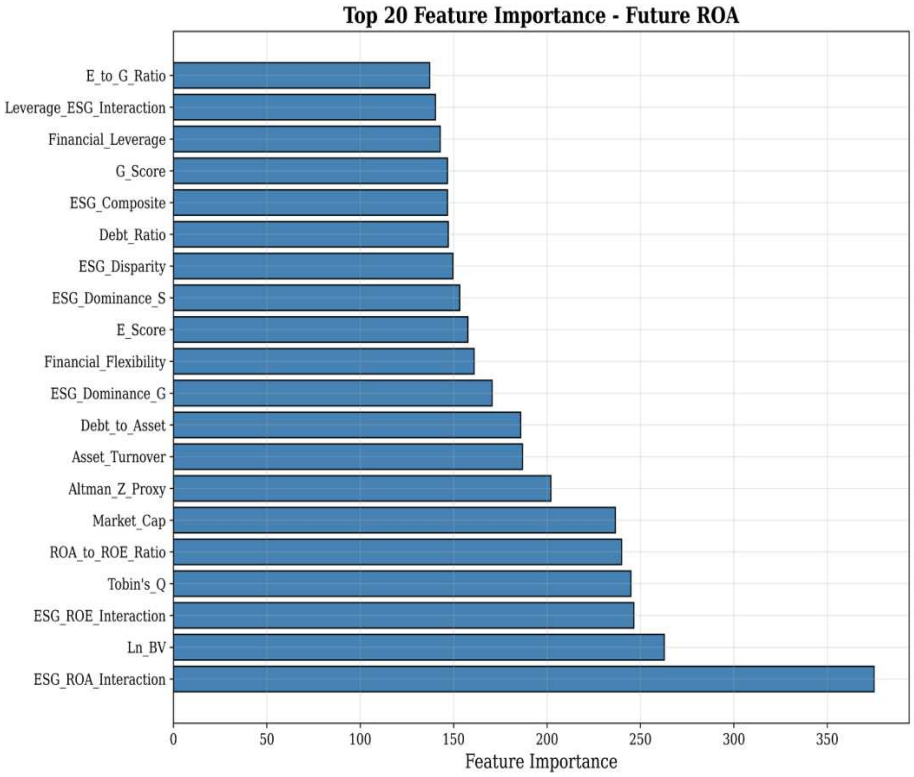


Figure 18. Top-20 feature importances for the future ROA ensemble.

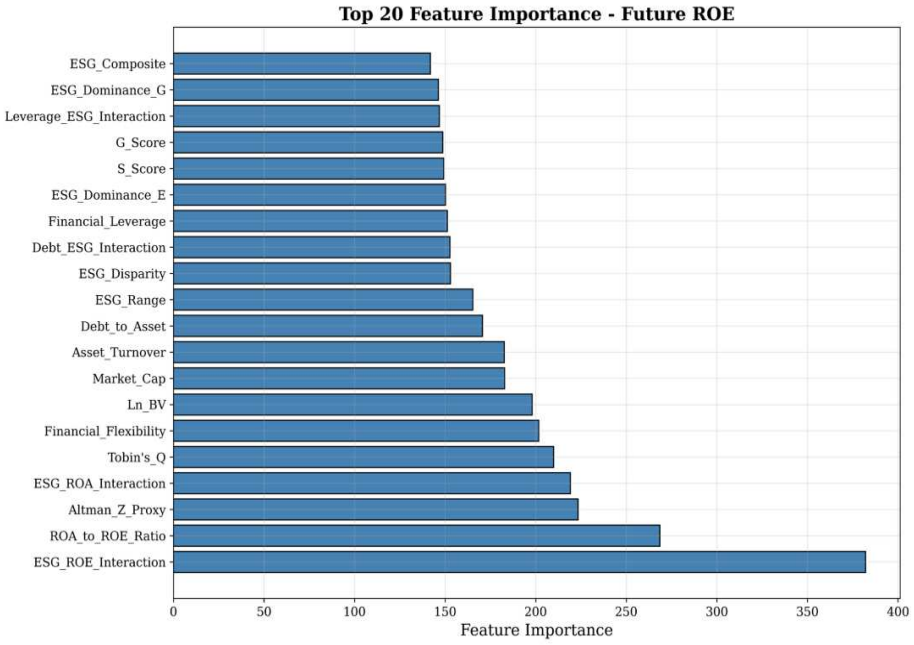


Figure 19. Top-20 feature importances for the future ROE ensemble.

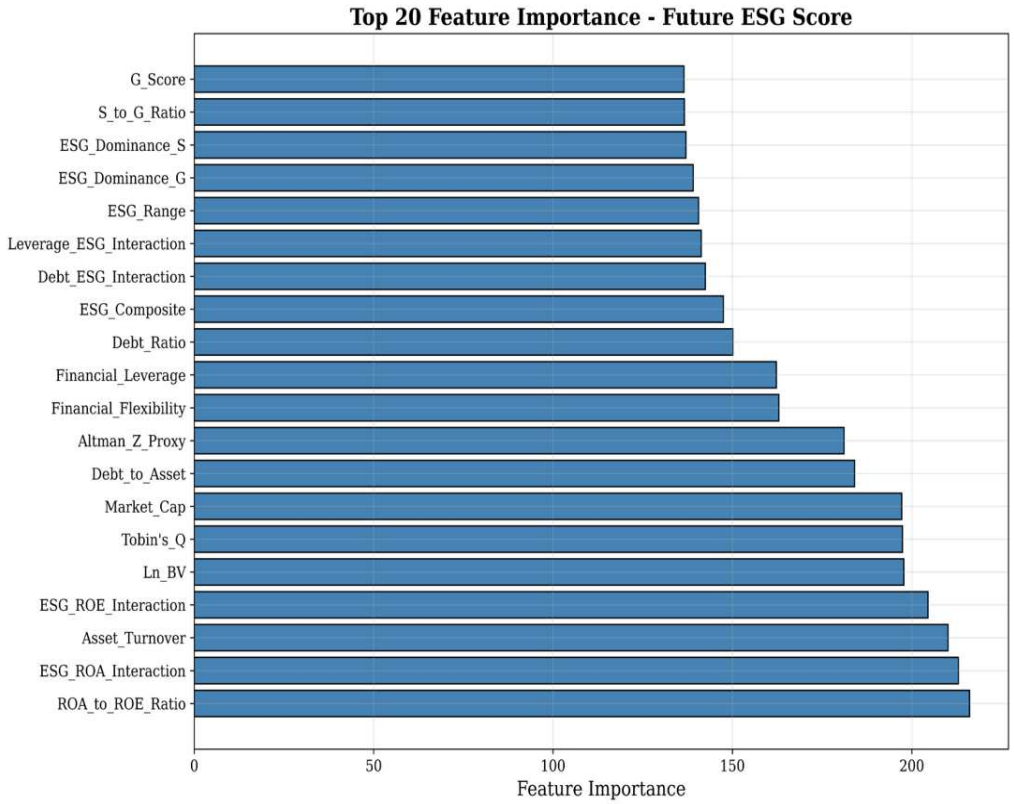


Figure 20. Top-20 feature importances for the future ESG composite ensemble.

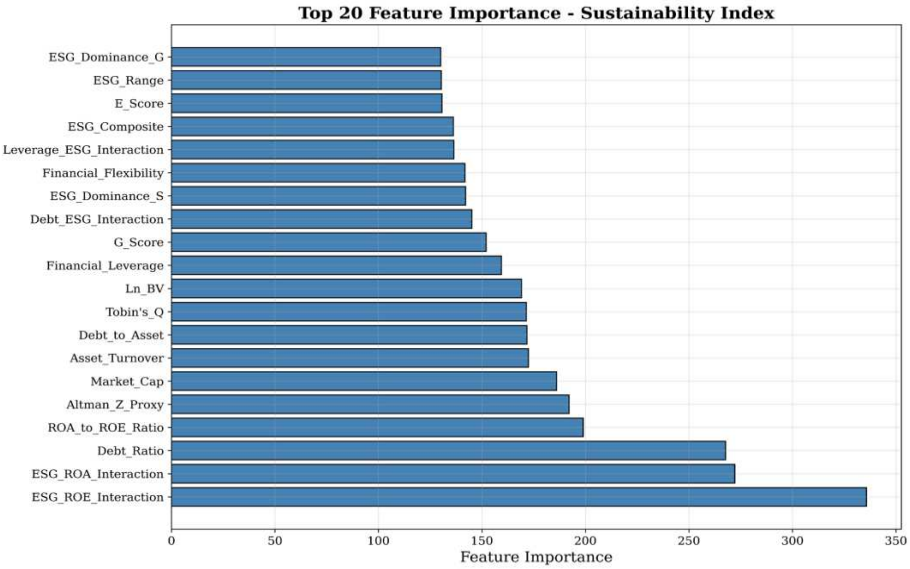


Figure 21. Top-20 feature importances for the financial sustainability index ensemble.

Three implications follow, supporting H3. First, in every prediction task, engineered ESG×profitability interactions outrank the raw ESG pillars and the raw composite—often by a factor of two or more. The value of ESG information for forecasting is therefore conditional: ESG matters most where it compounds existing profitability, a result that operationalizes Edmans' (2023) argument that ESG is value-relevant only in interaction with firm strategy, and that echoes findings that the ESG–performance link strengthens with

analyst attention and institutional support (Tang et al., 2025). Second, the prominence of ESG disparity and pillar-dominance features (ESG range, pillar ratios, dominance indicators) across all four panels indicates that the internal structure of a firm's ESG profile—not just its level—carries forward-looking information, paralleling the evidence that ESG rating disagreement and pillar divergence are priced (Christensen et al., 2022; Sabbaghi, 2026). Third, the consistent mid-ranking of the Altman-Z proxy and financial flexibility score links the sustainability prediction problem to the distress literature: balance-sheet resilience and ESG standing are complements in forecasting integrated sustainability, supporting integrated-reporting logics in sustainable finance and accounting.

5. Conclusion, Implications, and Limitations

This study designed and evaluated a stacked XGBoost–LightGBM–CatBoost ensemble with a Ridge meta-learner for forecasting four sustainability-relevant corporate targets on a controlled 5,000-firm dataset calibrated to stylized properties of listed-firm fundamentals. The architecture achieved out-of-sample R^2 of 0.720 for future ROA, 0.764 for future ROE, 0.857 for future ESG, and 0.937 for a composite financial sustainability index, with clean residual diagnostics.

The key result is that although individual ESG–performance cross-correlations are null, the central result is that the predictive importance of ESG×profitability interaction terms is the pillar of the relationship in all tasks, uncovering evidence of the forecasting value of ESG in conditional, interaction structures.

The effects on the three audiences extend to their implications. The results will find value for investors and lenders in the context of ensemble models for ESG, scale-inappropriate relative-error rates ought to be replaced with metrics appropriate for scale, forward-looking approaches of screening, pricing on sustainability, and taking part in early-warning tools For corporate managers, the prevalence of interaction terms suggests that companies' interaction terms are forecasting complements — that is, sustainability efforts would be more likely to be informative about future performance of a company that is already doing well in its operational- or business-level metrics, which reinforces the business case for incorporating sustainability into management decisions rather than viewing it as a compliance measure. For the researchers, the controlled benchmark supplies a benchmark of what a well-defined learner can recoup in a well-defined situation and clarifies results from the much noisier observational data where rating divergence and label-behaviour decoupling tend to dampen relationships between the data.

The results are subject to three restrictions and form the agenda for research First the data is simulated: When simulated, the data contains the salient, distributional and dependence facts of the data, while the design does not include rater disagreements, strategic disclosure, regulatory shocks, or macro-financial regime changes, so that the reported R^2 values represent the highest levels of performance attainable with observables panels; and second, implementation of the design on commercial ESG datasets, such as pre- and post-event panels as in merger- and-acquisition ESG trajectories, represents the next natural step.. Second, the cross sectional split of 80%:20% is not sufficient to test temporal generalization – rolling time series validation and concept drift monitoring is needed prior to deployment. Third, importances obtained from gradient boosting are split based, whereas future research should include SHAP-based importances) and include the set of targets to be expanded to exposure under the climate transition.

Declarations

Ethical Approval and Consent to Participate: Not Applicable.

Consent for Publication: All authors have reviewed and approved the final version of the manuscript and consent to its publication.

Funding: No funding was received to assist with the preparation of this manuscript.

References

Atz, U., Van Holt, T., Liu, Z. Z., & Bruno, C. C. (2023). Does sustainability generate better financial performance? Review, meta-analysis, and propositions. Journal of Sustainable Finance & Investment, 13(1), 802–825.

Berg, F., Kölbel, J. F., & Rigobon, R. (2022). Aggregate confusion: The divergence of ESG ratings. Review of Finance, 26(6), 1315–1344.

Billio, M., Costola, M., Hristova, I., Latino, C., & Pelizzon, L. (2024). Sustainable finance: A journey toward ESG and climate risk. International Review of Environmental and Resource Economics, 18(1–2), 1–75.

Bolton, P., & Kacperczyk, M. (2023). Global pricing of carbon-transition risk. The Journal of Finance, 78(6), 3677–3754.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

Christensen, D. M., Serafeim, G., & Sikochi, A. (2022). Why is corporate virtue in the eye of the beholder? The case of ESG ratings. The Accounting Review, 97(1), 147–175.

D'Amato, V., D'Ecclesia, R., & Levantesi, S. (2024). Firms' profitability and ESG score: A machine learning approach. Applied Stochastic Models in Business and Industry, 40(2). https://doi.org/10.1002/asmb.2758

Edmans, A. (2023). The end of ESG. Financial Management, 52(1), 3–17.

Farhadi, N., & Hair, J. F. (2025). Sustainable investing: Do ESG risks shift market dynamics? Journal of Sustainable Finance & Investment, 15(4), 927–953.

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? Advances in Neural Information Processing Systems, 35, 507–520.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146–3154.

Li, Q., Shan, H., Tang, Y., & Yao, V. (2024). Corporate climate risk: Measurements and responses. The Review of Financial Studies, 37(6), 1778–1830.

Liu, D., Gu, K., & Hu, W. (2023). ESG performance and stock idiosyncratic volatility. Finance Research Letters, 58(B), 104393.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

Narula, R., Rao, P., Kumar, S., & Paltrinieri, A. (2025). ESG investing & firm performance: Retrospections of past & reflections of future. Corporate Social Responsibility and Environmental Management. https://doi.org/10.1002/csr.2982

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. Advances in Neural Information Processing Systems, 31, 6638–6648.

Raghunandan, A., & Rajgopal, S. (2022). Do ESG funds make stakeholder-friendly investments? Review of Accounting Studies, 27(3), 822–863.

Sabbaghi, O. (2026). ESG risk and idiosyncratic risk: Recent evidence. Sustainable Finance Review, 1(1), 67–82. https://doi.org/10.1108/SFR-04-2025-0002

Sautner, Z., Van Lent, L., Vilkov, G., & Zhang, R. (2023). Firm-level climate change exposure. The Journal of Finance, 78(3), 1449–1498.

Sharma, U., Gupta, A., & Gupta, S. K. (2022). The pertinence of incorporating ESG ratings to make investment decisions: A quantitative analysis using machine learning. Journal of Sustainable Finance & Investment. https://doi.org/10.1080/20430795.2021.2013151

Sorathiya, A., Saval, P., & Sorathiya, M. (2024). Data-driven sustainable investment strategies: Integrating ESG, financial data science, and time series analysis for alpha generation. International Journal of Financial Studies, 12(2), 36.

Starks, L. T. (2023). Presidential address: Sustainable finance and ESG issues—value versus values. The Journal of Finance, 78(4), 1837–1872.

Tang, C., Guo, X., & Li, X. (2025). ESG and corporate performance: The moderating role of government subsidies and mediating effect of analyst coverage in Chinese A-share listed companies. PLOS ONE, 20(5), e0322190.

Tao, M., Lin, B., Poletti, S., & Roubaud, D. (2025). Greener pastures, steadier returns: ESG ratings and idiosyncratic risk management. International Review of Economics & Finance, 100, 104110.

Tutul, S (2026), “EDPACS_Data”, Mendeley Data, V1, doi: 10.17632/p9vm429wgd.1

Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241–259.