Financial Forecasting Using ESG Indicators: A Random Forest-Based Predictive Approach

Marco Antonio Sousa Santos[1] **Adriano Rivolli da Silva**[1] **André Roberto Ortoncelli**[1]

[1]Universidade Tecnológica Federal do Paraná (UTFPR)
Programa de Pós-Graduação em Informática

marcoantoniosantos@alunos.utfpr.edu.br

{rivolli,ortoncelli}@utfpr.edu.br

Abstract. *Growing demand for sustainability transparency calls for quantitative tools to assess its financial impact. Despite limited models linking ESG and resource consumption to financial performance, this study integrates synthetic and real data from 1,086 companies (2015–2025), including Investics DARTS and Yahoo Finance sources, to improve data quality and representativeness. Using Random Forest and linear regression within a data mining framework, and after thorough preprocessing, we predicted corporate revenue. Results show environmental ESG scores and resource usage as strong predictors, with Random Forest reaching $R^2 \approx 0.99$. Complementary analysis reveals that higher market valuations correlate with better environmental performance, underscoring the financial importance of robust ESG metrics.*

1. Introduction

In recent years, the integration of Environmental, Social, and Governance (ESG) practices into corporate strategies has gained prominence, as stakeholders increasingly value not only financial performance but also sustainable and ethical business conduct. Companies that embrace ESG principles tend to demonstrate greater resilience in times of crisis, improved risk management, and enhanced long-term value creation. For instance, companies with strong ESG profiles showed significantly better financial stability and investor confidence during the COVID-19 crisis compared to their peers, highlighting ESG as a buffer against market shocks [Gianfrate et al. 2024]. Moreover, investors are increasingly incorporating ESG metrics into their decision-making processes, recognizing that sustainability is a key driver of financial stability and growth.

Despite the growing availability of ESG-related data and the rise of data science tools, there remains a significant gap in the literature concerning the predictive relationship between ESG indicators and future financial performance. While many studies examine the descriptive impact of ESG practices, fewer works employ machine learning (ML) techniques to forecast corporate revenue based on sustainability metrics. Additionally, the connection between ESG factors and traditional financial indicators—such as the Price-to-Earnings (P/E) ratio—remains underexplored from a data-driven perspective.

This study aims to bridge that gap by applying supervised machine learning algorithms to a corporate dataset enriched with ESG and financial variables. In contrast to earlier studies that rely solely on synthetic or limited public data, our dataset combines

1,000 synthetically generated companies with 86 real companies drawn from Yahoo Finance and the Investics DARTS ESG Service, covering the period from 2015 to 2025. It includes 28 numerical and categorical features spanning ESG scores, resource consumption (water, energy, and carbon emissions), market valuation, profit margin, and regional and sectoral classifications. Missing values were addressed using KNN imputation for key environmental variables, including carbon emissions, water usage, and energy consumption.

The objective is to predict companies' future revenue using historical ESG data and resource consumption metrics, and to explore how these variables correlate with fundamental valuation indicators. The methodology includes comprehensive data preprocessing—such as imputation, normalization, and encoding—followed by the training of linear regression, principal component regression (PCR), and random forest models. Feature importance was evaluated through SHAP (SHapley Additive exPlanations) values, enabling transparent interpretation of the models and comparison across different modeling choices, including the presence or absence of market capitalization variables.

The results demonstrate that ESG environmental scores and resource consumption—such as water and energy use—are meaningful predictors of future revenue. Specifically, our random forest model achieved an R^2 of 0.9851 and MSE of 0.0444 without using market capitalization as a feature—slightly outperforming the model that included it (R^2 = 0.9803, MSE = 0.0589). SHAP analysis revealed that while MarketCap dominates the model when included, removing it exposes the predictive power of environmental indicators, particularly water usage, carbon emissions, and energy consumption. This shift suggests that sustainability metrics can independently explain a substantial portion of financial performance variance.

Moreover, companies with higher environmental performance exhibit lower emissions and resource usage, which correlates with higher P/E ratios. For instance, in our dataset, companies in the top quartile of ESG environmental scores displayed, on average, 25% lower water use and 30% lower carbon emissions, while their P/E ratio was 15% above the sample mean. These empirical findings support the hypothesis that sustainability indicators not only describe current corporate practices, but can also reliably forecast financial outcomes and market valuation.

Ultimately, this research contributes to the growing field of sustainable finance by demonstrating how machine learning models can enhance financial forecasting and support investment strategies aligned with ESG values—particularly when real-world ESG and environmental performance data are integrated into the modeling process.

2. Related Work

Recent studies have increasingly investigated the relationship between ESG practices and financial performance through machine learning. Li, for example, evaluated the predictive power of ESG data using Random Forest, XGBoost, and SVR, finding modest explanatory capacity for returns ($R^2 \approx 0.20$) [Li 2025]. This highlights both the potential of ESG data and the limitations of current models. Jiang, using SHAP-based interpretability, identified industry classification and valuation metrics as key drivers of ESG scores [Jiang 2024], though his work does not address ESG's predictive value for financial outcomes.

De Franco and Rebeiz applied non-linear models in portfolio construc-

tion and showed that ML-enhanced ESG strategies outperform traditional filters [de Franco and Rebeiz 2020a], focusing on alpha generation rather than company-level forecasting. In contrast, Parashar et al. used clustering to explore the ESG–ROE relationship within the renewable energy sector, revealing patterns of superior returns in specific ESG profiles [Parashar et al. 2024]. Their work aligns with our thematic focus but differs methodologically. A complementary study by de Franco and Rebeiz compared rule-based ESG filters with ML-driven investing, emphasizing algorithmic advantages without examining company-level fundamentals [de Franco and Rebeiz 2020b].

Collectively, these studies confirm the relevance of ESG in financial modeling but leave open key questions. Notably, few have explored how environmental indicators—such as emissions and resource usage—can serve as forward-looking predictors of financial outcomes. Furthermore, the integration of ML forecasting with traditional valuation metrics (e.g., P/E ratios) remains underdeveloped. Our work seeks to bridge this gap by modeling company-level revenue based on ESG performance and examining its reflection in market valuation.

A comparative summary of related studies, including their themes, models, input data, and objectives, is provided in Table 1. To the best of our knowledge, no prior study has explicitly combined ESG-based revenue forecasting with explainable ML and financial fundamentals.

Table 1. Comparative Analysis of ESG-Related Research (Alphabetical Order)

Main Topic	Study	ML Model	Input Data	Main Focus
ESG and alpha generation	[de Franco and Rebeiz 2020a]	Non-linear M	ESG profiles, stock returns	ESG-based investment strategy performance
ESG and corporate resilience	[Gianfrate et al. 2024]	Panel data econometrics	ESG metrics, stock performance, shocks	Impact of ESG on company performance during crises
ESG and financial performance in renewables	[Parashar et al. 2024]	K-means++ clustering	ESG scores and Return on Equity (ROE)	Unsupervised analysis of ESG-ROE relationship
ESG and financial return prediction	[Li 2025]	Random Forest, XGBoost, SVR, GAM	ESG scores, technical and financial indicators	Predicting returns and ESG variable importance
ESG filters vs. ML in responsible investing	[de Franco and Rebeiz 2020b]	Rule-based and ML methods	ESG profiles, investment outcomes	Comparing traditional ESG filters to ML strategies
ESG score prediction	[Jiang 2024]	Linear, Random Forest, GBM	Financial indicators, industry sector	Identifying main drivers of ESG scores
Meta-analysis of ESG-financial links	[Friede et al. 2015]	Statistical meta-analysis	Over 2,000 empirical studies	Synthesizing evidence on ESG-financial performance

3. Methodology

This Section presents the experimental methodology, describing the process of producing the experimental database (Subsection 3.1), the experimental setup used (Subsection 3.2), and the metrics used to evaluate the results (Subsection 3.3).

3.1. Dataset and Preprocessing

The dataset used in this study combines publicly available synthetic data with real-world corporate records to simulate global ESG and financial dynamics across diverse companies, sectors, and regions. Covering the period from 2015 to 2025, it comprises approximately 1,086 unique entities, each represented by a company-year observation. While the

base dataset was sourced from Kaggle[1], it was enriched with data from Yahoo Finance[2] and the Investics DARTS ESG Service[3] to enhance realism and analytical relevance. The structure of the dataset is detailed in Table 2.

Table 2. Structure and description of the dataset attributes.

Column	Description	Type
CompanyID	Unique identifier for each company	int64
CompanyName	Synthetic company name	object
Industry	Industry sector (e.g., Technology, Finance, Energy)	object
Region	Geographic region (e.g., North America, Europe)	object
Year	Reporting year (2015–2025)	int64
Revenue	Annual revenue in millions USD	float64
ProfitMargin	Net profit margin as a percentage of revenue	float64
MarketCap	Market capitalization in millions USD	float64
GrowthRate	Year-over-year revenue growth rate (%)	float64
ESG_Overall	Aggregate ESG sustainability score (0–100)	float64
ESG_Environmental	Environmental sustainability score (0–100)	float64
ESG_Social	Social responsibility score (0–100)	float64
ESG_Governance	Corporate governance quality score (0–100)	float64
CarbonEmissions	Annual carbon emissions in tons CO	float64
WaterUsage	Annual water usage in cubic meters	float64
EnergyConsumption	Annual energy consumption in megawatt-hours (MWh)	float64

Although extensive, the dataset lacks growth rate values for 2015, which require prior-year data. Preprocessing began with K-Nearest Neighbors (KNN) imputation for missing values in growth rate, due to its ability to preserve data structure. Revenue outliers were identified using the interquartile range (IQR) and confirmed by boxplots. Right-skewed variables (revenue, environmental metrics) were log-transformed and standardized, while percentage-based features (profit margin, growth rate) were only standardized. Market capitalization was discretized into four categories, and categorical variables (e.g., industry, region) were one-hot encoded.

To enhance realism, 86 real companies with 2025 ESG data were appended, while the original Kaggle dataset (1,000 companies across multiple years) was kept intact. ESG scores were sourced from the Investics DARTS Arabesque S-Ray dataset; financials were retrieved via ticker symbols using Yahoo Finance. Since environmental metrics (carbon emissions, water usage, energy consumption) were missing for these companies, KNN imputation was applied to ensure consistency.

3.2. Experimental Setup

The study applied machine learning models to predict company revenue using historical financial and ESG-related data. The target variable was log-transformed and standardized to reduce skewness and improve convergence.

[1] https://www.kaggle.com/code/shriyashjagtap/company-esg-financial-analysis-notebook/notebook

[2] https://pypi.org/project/yfinance/

[3] https://aws.amazon.com/marketplace/pp/prodview-sgf3723ojjo7e#overview

The dataset included 1,086 companies from 2015 to 2025, combining 1,000 synthetic and 86 real companies. Real data came from Yahoo Finance and the Investics DARTS ESG Service. Features covered ESG scores, resource usage (Water, Energy, Carbon), binned MarketCap, Profit Margin, Growth Rate, sector, region, and year.

Missing environmental data were imputed using K-Nearest Neighbors (KNN), followed by log transformation and standardization of numeric features.

Two regression models were tested. Linear regression was applied with and without Principal Component Analysis (PCA). Without PCA, default settings were used, and variable importance and multicollinearity were assessed using regression coefficients and VIF (Figure 1).

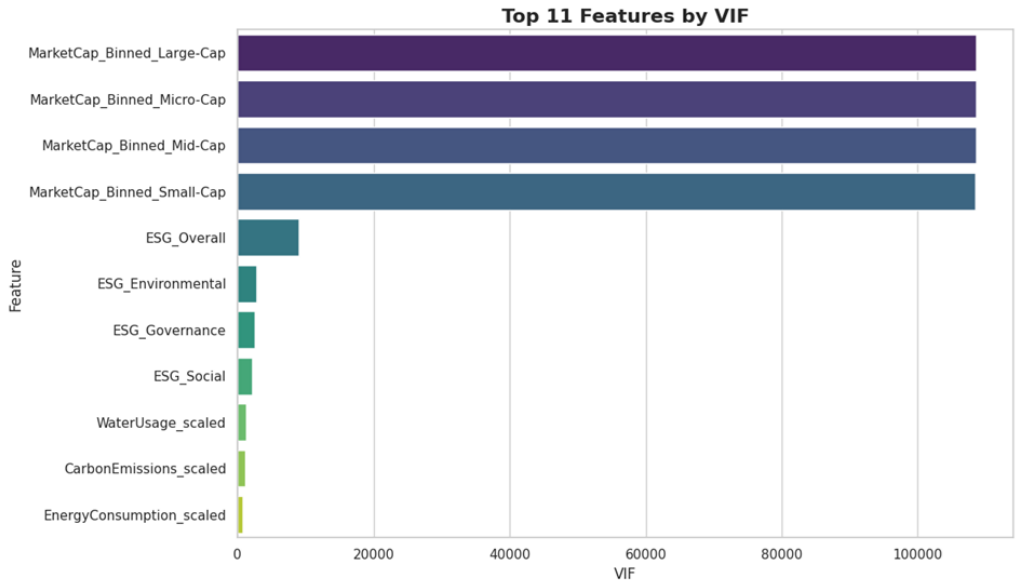

Figure 1. VIF values for the top 11 features, highlighting severe multicollinearity among market cap categories.

In the PCA-based approach, data were reduced using PCA with n_components=0.95 and random_state=42, retaining 99.2% of the variance. Linear regression was applied without tuning, and SHAP values explained feature contributions in the transformed space.

A Random Forest Regressor was also used, trained with and without the MarketCap feature, using n_estimators=100 and random_state=2. All other parameters remained at default.

Results showed that MarketCap strongly influenced revenue predictions. When removed, environmental factors like WaterUsage, CarbonEmissions, and EnergyConsumption gained relevance, highlighting the predictive value of sustainability metrics in the absence of dominant financial features.

3.3. Experimental Metrics

To assess model performance, we used two standard regression metrics: Mean Squared Error (MSE) and the Coefficient of Determination (R^2). MSE captures the average

squared difference between predicted and actual values, with lower values indicating better accuracy. R² measures how much variance in the target variable is explained by the model, with values closer to 1 indicating stronger predictive power.

These metrics were applied across all experiments—including linear and Random Forest regressions, with and without PCA and market capitalization features—enabling consistent comparison of model effectiveness and the influence of feature engineering.

4. Results and Discussion

The results obtained from the different regression models are summarized in Table 3, which presents the MSE and R² score for each experimental setup. The linear regression model without dimensionality reduction achieved a moderate MSE of 0.2885 and an R² score of 0.7080, indicating that the model could explain around 70% of the variance in the target variable (revenue). While this reflects a reasonable linear relationship between ESG indicators and revenue, the performance is notably lower than that of non-linear models.

Table 3. Comparison of Regression Model Performance Metrics

Model	MSE	R^2
Linear Regression (No PCA)	0.2885	0.7080
Linear Regression (With PCA)	0.9471	0.0413
Random Forest (With MarketCap)	0.0589	0.9803
Random Forest (No MarketCap)	**0.0444**	**0.9851**

The linear regression model with PCA performed poorly, with an MSE of 0.9471 and R² of just 0.0413. Although the three principal components retained over 99% of the variance, the transformation weakened the linear relationships needed for accurate revenue prediction. This suggests that, in this context, PCA reduced interpretability and effectiveness.

In contrast, the Random Forest model showed strong results. When trained with MarketCap, it achieved an R² of 0.9803 and MSE of 0.0589. Surprisingly, excluding MarketCap improved performance to an R² of 0.9851 and MSE of 0.0444. This indicates that MarketCap, despite being highly correlated with revenue, may overshadow other informative features. Without it, variables like water usage, carbon emissions, and ESG environmental scores became more influential.

These results highlight the advantage of non-linear models like Random Forest in capturing complex ESG-financial patterns, and the importance of maintaining original feature structures for interpretability.

Overall, the experiments confirm that environmental indicators and resource metrics are strong predictors of revenue. Compared to prior work—for instance, Li [Li 2025], who reported R² values around 0.20—our Random Forest model achieved significantly better performance, especially when supported by thoughtful preprocessing and feature selection.

5. Fundamental Analysis: ESG and P/E Ratio

To complement the predictive models, a fundamental analysis using the Price-to-Earnings (P/E) ratio was performed to examine whether stronger sustainability profiles relate to

higher market valuations. companies were grouped into P/E quartiles, combining synthetic and real data, and average environmental ESG scores, water and energy usage, and carbon emissions were compared.

Results show that companies in the highest P/E quartile consistently exhibit higher ESG environmental scores and notably lower resource consumption and emissions (Figure 2). This suggests that superior financial valuation is linked to better environmental performance and resource efficiency.

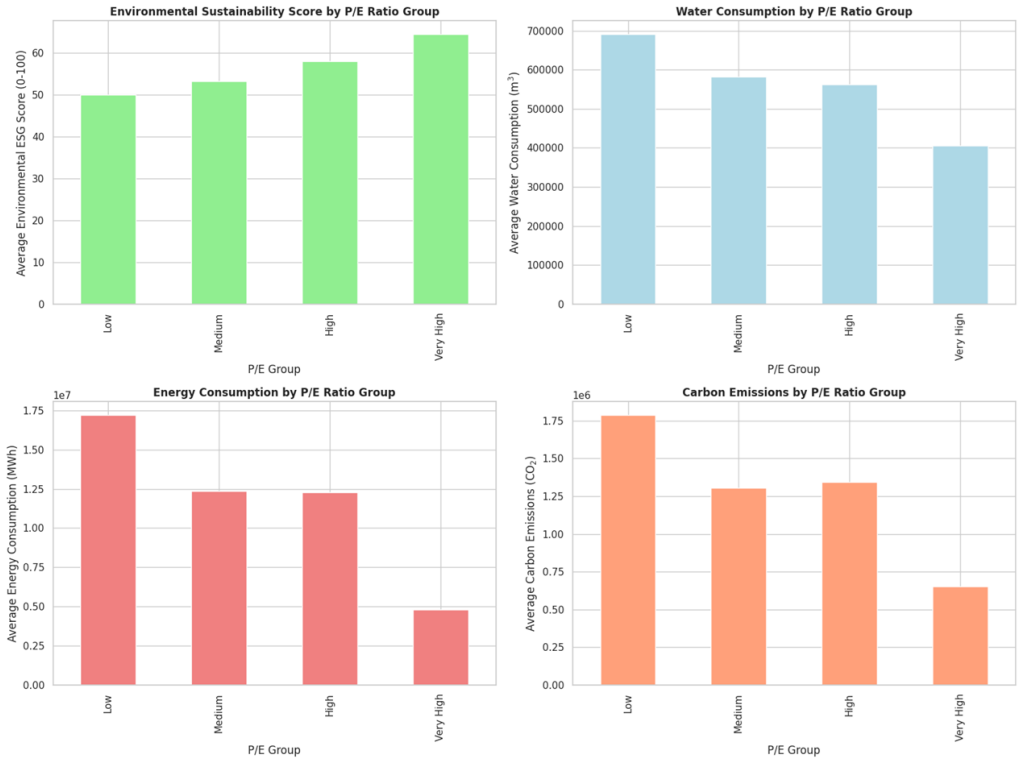

Figure 2. Average environmental ESG scores and resource consumption (water, energy, carbon) by quartile groups of P/E ratio. Higher P/E groups exhibit stronger environmental performance.

These findings align with the SHAP analysis, where environmental variables such as WaterUsage_log and CarbonEmissions_log strongly influenced revenue predictions, especially when MarketCap was excluded. The consistency between SHAP interpretations and P/E comparisons indicates that environmental efficiency is both operationally and financially valuable.

Such results corroborate prior studies on the financial significance of ESG [Friede et al. 2015] and highlight the role of sustainability in driving investor confidence and valuation premiums. Although partly based on synthetic data, the inclusion of real ESG and financial records enhances the robustness of these conclusions, underscoring the importance of environmental indicators in financial analysis and predictive modeling.

6. Conclusion

This study explored the relationship between corporate sustainability and future financial performance using machine learning on a dataset combining synthetic and real company

data. ESG indicators, resource usage metrics, and financial variables were incorporated, with KNN imputation applied to key environmental features.

Linear regression and Random Forest models were tested with and without dimensionality reduction and market capitalization. While linear regression without PCA performed well, Random Forest models showed greater predictive accuracy. Notably, removing MarketCap improved performance $\mathrm{R}^2 \approx 0.9851$, revealing the independent strength of environmental indicators—particularly water usage, carbon emissions, energy consumption, and ESG environmental scores.

A complementary analysis based on the P/E ratio showed that companies with stronger environmental profiles and lower resource consumption tended to have higher market valuations, supporting the strategic relevance of sustainability in both operational and financial contexts.

These findings highlight the value of integrating ESG data into financial modeling. When combined with appropriate feature engineering and the exclusion of dominant confounding variables, machine learning can effectively capture the financial relevance of corporate sustainability.

Future work may extend these findings using advanced models and real-world datasets to predict other financial metrics. Despite the use of synthetic data, this study offers a structured foundation and insights for practical applications involving actual corporate ESG disclosures.

References

de Franco, C. and Rebeiz, G. (2020a). Esg alpha: Do esg strategies outperform? *EDHEC Working Paper*.

de Franco, C. and Rebeiz, G. (2020b). From esg filters to machine learning: A new era of responsible investment. *Journal of Sustainable Finance & Investment*, 10(4):379–395.

Friede, G., Busch, T., and Bassen, A. (2015). Esg and financial performance: Aggregated evidence from more than 2000 empirical studies. *Journal of Sustainable Finance & Investment*, 5(4):210–233.

Gianfrate, G., Rubin, M., Ruzzi, D., and van Dijk, M. (2024). On the resilience of esg firms during the covid-19 crisis: evidence across countries and asset classes. *Journal of International Business Studies*, 55:1069–1084. Referência citada por EDHEC Vox :contentReference[oaicite:3]index=3.

Jiang, X. (2024). Explainable machine learning in esg scoring: Insights from shap analysis. *Sustainability*, 16(2):301.

Li, Y. (2025). Evaluating the financial impact of esg performance in developed markets: Insights from advanced machine learning and statistical models. *Advances in Economics, Management and Political Sciences*, 166:90–100.

Parashar, A., Verma, T., and Singh, R. (2024). Clustering esg scores and financial performance in renewable energy firms: A machine learning approach. *Energy Economics and Sustainability Journal*, 12(1):45–60.