Hochschule Fulda
University of Applied Sciences

Profitability Prediction through ESG & Sustainability Metrics

Course

Application of Data Science in Finance(AI5247)

Department

Data Science

Submitted by

Pavithra Lakshmi Venugopal

1565315

Fulda, September 2025

Table of Contents

Abstract

This project investigates whether Environmental, Social, and Governance (ESG) and sustainability factors can predict company profitability without relying on direct financial indicators such as Profit Margin, Revenue, and Market Capitalization. The analysis followed the CRISP-DM framework, progressing through business understanding, data exploration, preparation, modeling, evaluation, and application.

The dataset contained ESG scores, financial attributes, and categorical information such as industry, region, and year. Profitability was categorized into Low, Medium, and High classes based on profit margin thresholds. Growth Rate, which contained approximately nine percent missing values, was imputed using industry-specific medians. Exploratory analysis showed moderate correlations between ESG scores and profitability, with Environmental and Governance dimensions emerging as the most influential predictors.

Three machine learning models: Decision Tree, Random Forest, and Logistic Regression, were applied. Initial experiments yielded near-perfect accuracy due to leakage from direct financial variables. Once these were removed, Decision Tree and Logistic Regression achieved modest predictive power, while Random Forest produced stronger accuracy. After probability calibration, Random Forest emerged as the best-performing model, achieving 86 percent accuracy, a ROC AUC of 0.96, a KS statistic of 0.80, and a Log Loss of 0.39.

The findings confirm that ESG and sustainability metrics, when modeled correctly, can provide reliable insights into profitability. The project highlights the importance of preventing data leakage, applying calibration, and using advanced evaluation metrics, demonstrating that ESG is not only an ethical consideration but also a financially significant factor.

1. Introduction

In recent years, the integration of Environmental, Social, and Governance (ESG) practices into financial analysis has gained significant attention from academics, practitioners, and policymakers. ESG metrics are designed to measure how sustainable and responsible a company is in its operations, providing indicators that extend beyond traditional financial reporting. They capture aspects such as environmental efficiency, corporate governance structures, and social responsibility, all of which are increasingly associated with long-term corporate value. Investors and regulators now routinely monitor ESG disclosures, regarding them as potential signals of risk management, operational performance, and profitability.

Despite this growing emphasis, the financial relevance of ESG metrics remains debated. Many firms disclose ESG data alongside conventional financial statements, yet it is unclear whether these nonfinancial variablesindicators can reliably predict profitability. Traditional financial metrics such as Profit Margin, Revenue, and Market Capitalization dominate predictive models, but their inclusion introduces data leakage, where information that directly determines profitability is inadvertently used as a predictor. This produces artificially inflated accuracy, undermining the validity of the models and obscuring the true contribution of ESG variables.

This project examines whether ESG and sustainability factors can independently predict firm profitability when direct financial indicators are excluded. The analysis also evaluates the importance of calibration, since in financial applications, reliable probability estimates are as critical as accurate classifications.

1.1 Problem Statement:

The central problem is to determine whether ESG and sustainability factors can predict profitability without relying on direct financial indicators such as Profit Margin, Revenue, and Market Capitalization, which would otherwise cause data leakage and inflated accuracy.

1.2 Subproblems:

To address this central question, the study focuses on three sub-problems. The first is to examine how data leakage affects model validity and why controlling it is essential for building trustworthy predictive models. The second is to identify which ESG and sustainability factors are most relevant for predicting profitability, clarifying not only whether ESG matters but also which of its dimensions drive outcomes. The third is to investigate how probability calibration can improve the reliability of predictive models, ensuring that outputs are both accurate in classification and meaningful in terms of probability estimates.

1.3 Objectives:

This project investigates whether Environmental, Social, and Governance (ESG) and sustainability factors can be used to predict the profitability of investments. The study examines whether ESG features alone are sufficient for profitability prediction and evaluates the performance of three models: Decision Tree, Random Forest, and Logistic Regression.

The analysis assessed not only classification accuracy but also advanced evaluation metrics such as Log Loss and McFadden's R^2 to measure both predictive performance and the

reliability of probability estimates. Calibration techniques were applied to ensure that probability outputs were suitable for real-world financial decision making. The overarching objective was to evaluate the predictive value of ESG information, validate the robustness of the evaluation process, and demonstrate how data science methods can provide financial insights beyond traditional numerical indicators.

2. Business understanding

The business relevance of this project lies in understanding whether ESG and sustainability factors can be used as predictors of profitability, independent of financial indicators. In the field of finance, profitability is commonly explained through metrics such as Profit Margin, Revenue, or Market Capitalization, which are linked to financial outcomes. While these indicators have strong predictive power, their inclusion in data science models leads to data leakage, where the target variable is indirectly revealed to the model through its predictors. Such leakage leads to artificially high accuracy, giving the false impression of model performance while hiding the true contribution of ESG features.

From a business perspective, this problem is critical because firms, investors, and policymakers need tools that can assess the financial relevance of sustainability practices without bias. A model that relies only on ESG and related contextual factors could provide insights into how responsible corporate practices contribute to long-term profitability. Furthermore, in financial decision-making, calibrated probability estimates are essential. Whereas a classification output can predict whether a company is likely to be in a high or low profitability category, decision makers rely on probability estimates to quantify risk and allocate resources under uncertainty. Thus, beyond the accuracy of prediction, the project focuses on probability calibration as a crucial step in creating a suitable model for business use.

3. Data understanding

The dataset used for this study consisted of both financial and nonfinancial variables collected across 1000 firms, 9 industries, and 7 regions. The financial variables included Profit Margin, Revenue, Market Capitalization, and Growth Rate, while the nonfinancial variables captured ESG scores in four dimensions: Environmental, Social, Governance, and an Overall ESG score. Additional categorical columns such as Industry, Region, and Year provided contextual information. The central aim of this phase was to develop a clear understanding of the overall structure and quality of the dataset and to identify the relationships between ESG metrics and profitability.

The target variable, Profitability, was created by categorizing Profit Margin into three distinct classes: Low (Lower than 5 percent), Medium (5 - 10 percent), and High (Greater than 10 percent). An initial inspection revealed that the distribution of classes was imbalanced, with the High-profit category significantly outnumbering the Medium- and Low-profit categories. This imbalance suggested the need for techniques such as SMOTE or class weighting to avoid bias in the training process.

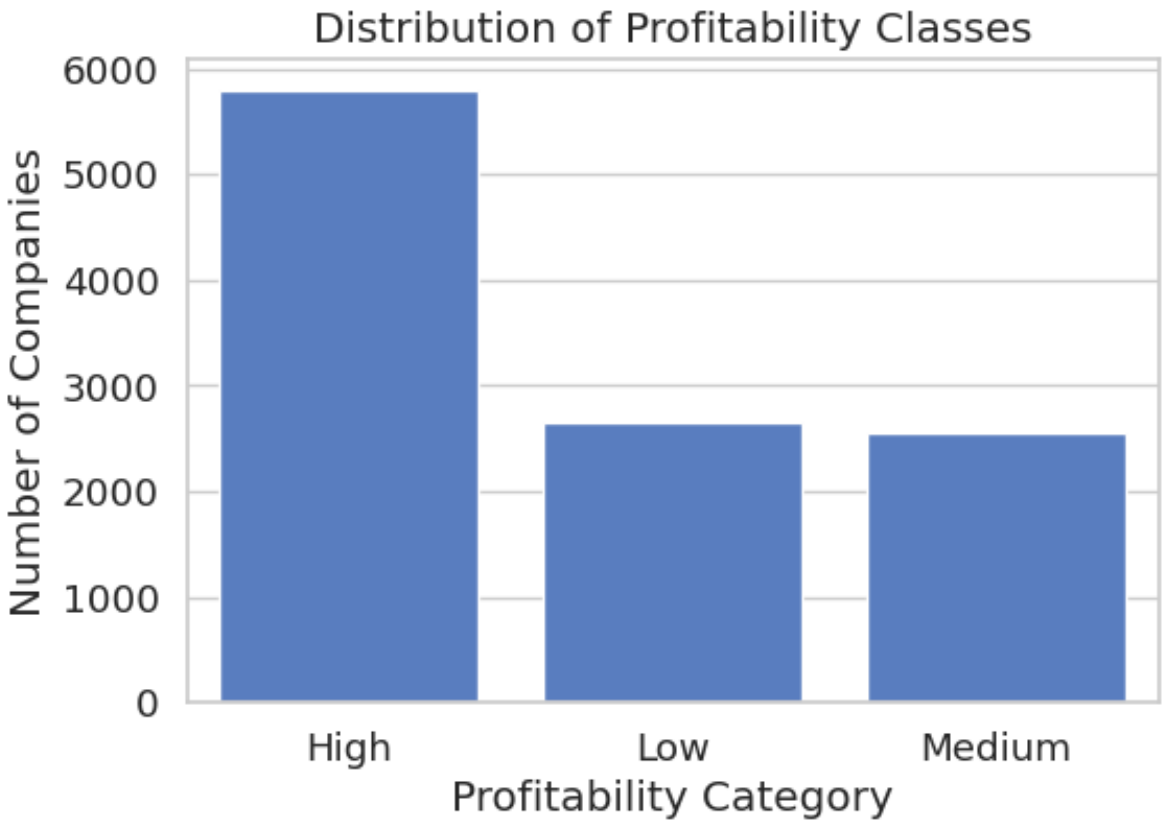

Figure 1: Distribution of profitability classes.

This figure shows the class imbalance across Low, Medium, and High categories.

A correlation analysis was conducted to examine associations between numerical variables. The results demonstrated moderate correlations between ESG Overall and Profit Margin, providing initial evidence that ESG factors may influence profitability, though not overwhelmingly. Interestingly, the Environmental dimension showed stronger correlation than the Social or Governance dimensions.

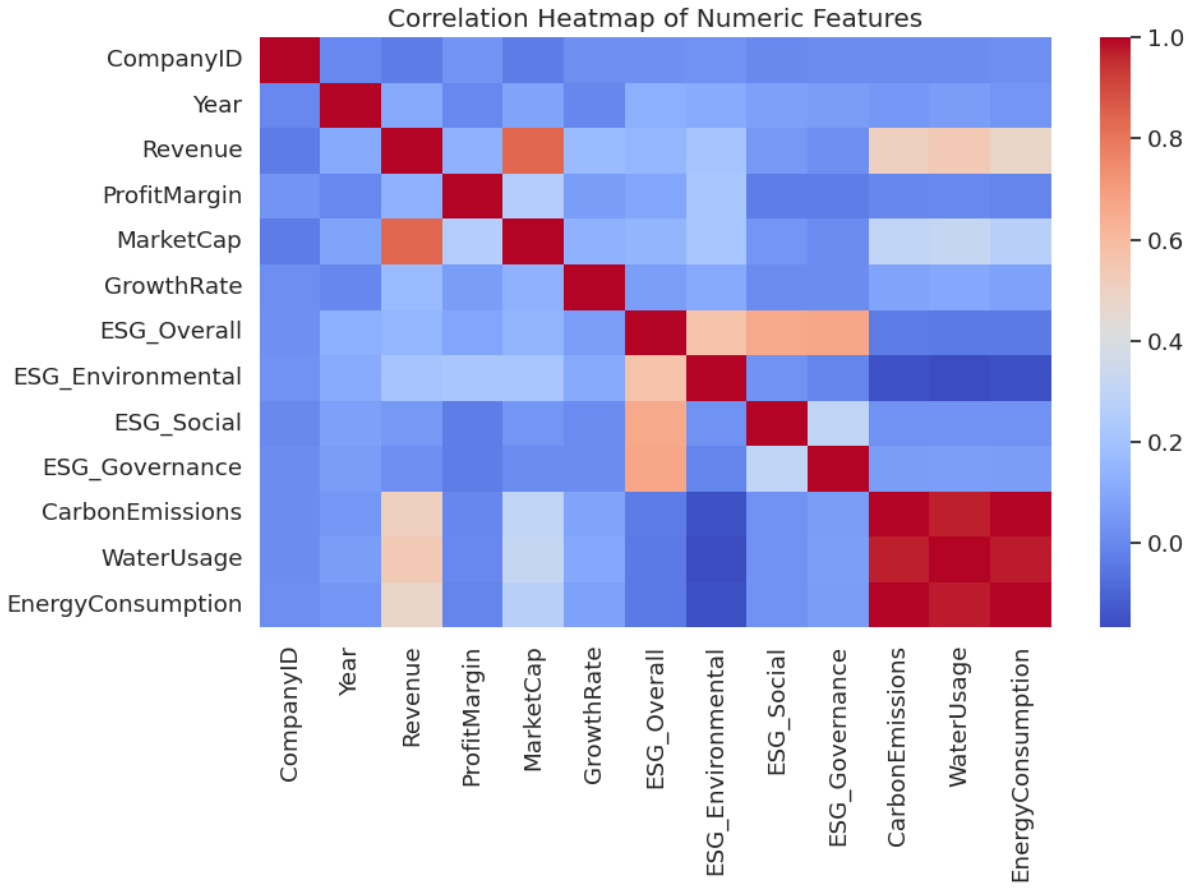

Figure 2: Correlation heatmap of financial and ESG variables.

This figure highlights the correlations among Profit Margin, Growth Rate, ESG scores, and other financial metrics.

Correlation heatmap of numeric variables showing relationships among ESG scores, financial metrics, and sustainability features. Red shades indicate positive correlations, while blue shades indicate negative correlations. Numeric coefficients are not annotated in this visualization.

Further exploration of ESG variables revealed unique distributional characteristics. The Environmental scores displayed clustering near their maximum values, which may indicate either reporting bias or the effect of regulatory incentives. Social and Governance scores were more evenly distributed, while the Overall ESG score combined these dimensions into a moderately skewed distribution.

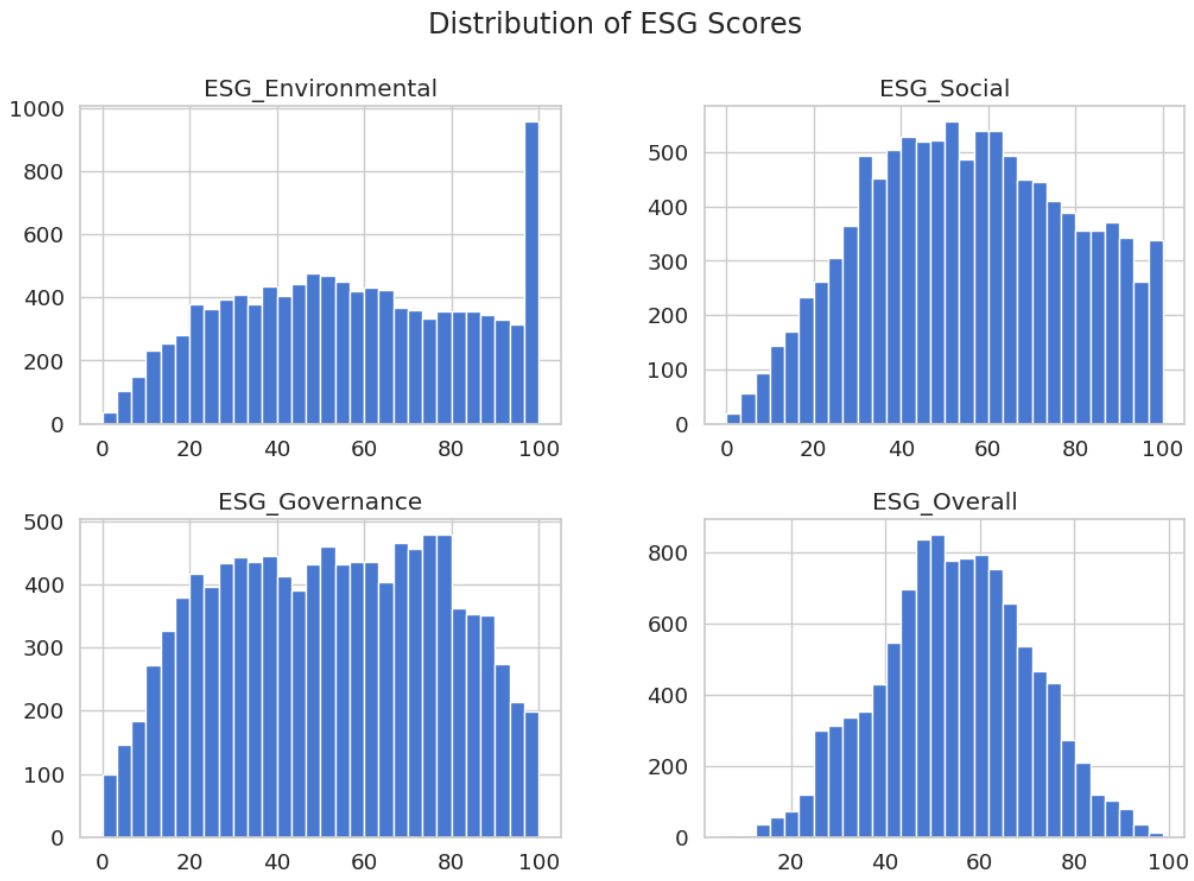

Figure 3: Distribution of ESG scores (Environmental, Social, Governance, Overall)

This figure presents the histograms of each ESG component and the composite ESG score.

To examine the relationship between ESG metrics and profitability categories, boxplots were generated for ESG Overall across Low, Medium, and High classes. The plots indicated that firms in the High-profit category tended to report higher ESG Overall scores. However, significant overlaps across categories also showed that ESG is not the sole determinant of profitability.

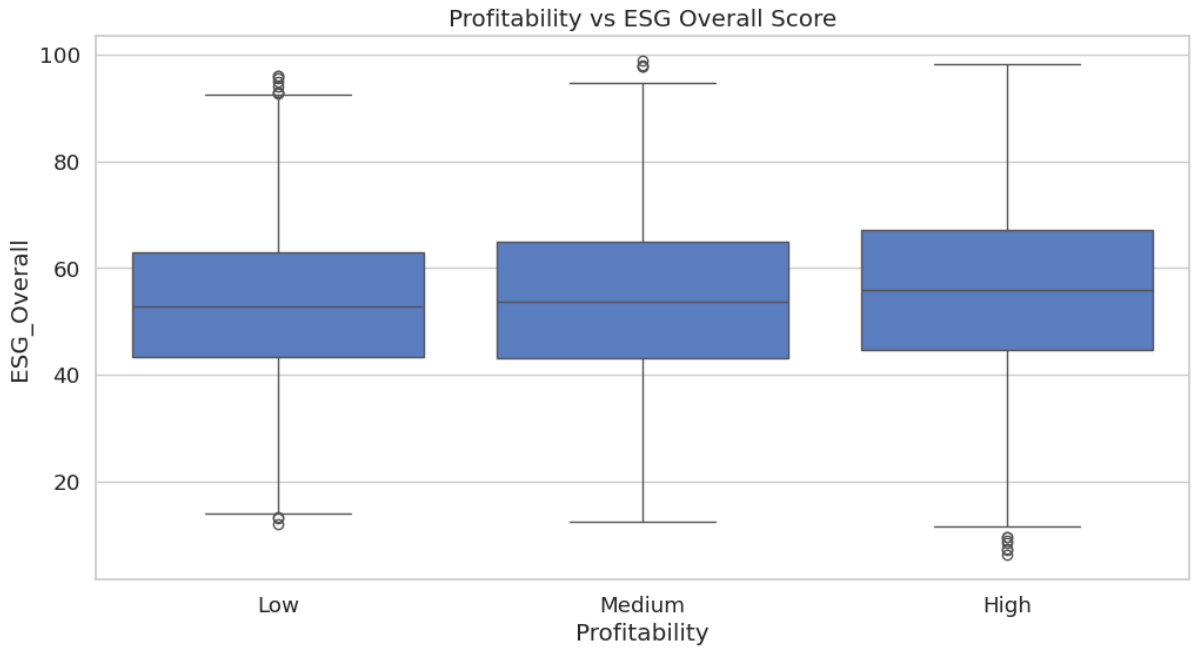

Figure 4: Boxplot of ESG Overall across profitability categories.

This figure illustrates differences in ESG performance across profitability classes, with overlaps highlighting the complexity of the relationship.

In summary, the Data Understanding phase revealed that ESG variables, particularly Environmental and Governance scores, had visible but not dominant relationships with profitability. The imbalance in profitability classes and the unusual clustering in Environmental scores suggested the need for careful preprocessing before modeling.

4. Data Preparation

The Data Preparation phase involved transforming the raw dataset into a more suitable form for machine learning modeling while ensuring that issues such as data leakage, missing values, and class imbalance were properly addressed. This step was crucial because decisions taken during preprocessing directly affect both the fairness and reliability of the final predictive models.

The first major task was the handling of missing values. The Growth Rate variable contained approximately nine percent missing entries, which, if left untreated, could have distorted model training and reduced predictive performance. Rather than applying a simplistic mean imputation across the entire dataset, industry-specific medians were used. This approach preserved sector-level differences and ensured that the imputed values were realistic, reflecting the business conditions of firms operating within the same industry.

The second task concerned the prevention of data leakage. Variables such as Profit Margin, Revenue, and Market Capitalization are inherently determinative of profitability. Their inclusion in predictive modeling would have allowed the models to indirectly access the target information, leading to artificially inflated accuracy. Initial experiments confirmed this risk: models achieved nearly perfect performance when these variables were included, but this

performance was unreliable and unsuitable for drawing meaningful conclusions about ESG. Consequently, these direct financial variables were deliberately excluded from the feature set before training, leaving ESG scores, Growth Rate, Industry, Region, and Year as the main predictors.

A new categorical variable, Profitability, was engineered from Profit Margin to serve as the target outcome. Firms were classified as Low, Medium, or High profitability based on predefined thresholds. This transformation converted the problem into a supervised multiclass classification task, aligning the predictive objective with the project's research question.

The third preprocessing step was the treatment of categorical attributes. Industry, Region, and Year were encoded into dummy variables so that they could be processed by machine learning algorithms without introducing bias from categorical labels. These variables were important to retain because sectoral, geographical, and temporal differences may shape the relationship between ESG practices and profitability.

Finally, the imbalance in the target classes was addressed. The High-profit category contained disproportionately more observations than the Low- and Medium-profit categories, creating the risk that models might be biased toward the majority class. To mitigate this issue, a combination of Synthetic Minority Oversampling Technique (SMOTE) and class weighting was employed. SMOTE generated synthetic examples of the minority classes to balance the training set, while class weighting ensured that model optimization penalized misclassifications of minority classes more strongly. Together, these measures improved the fairness of the models and strengthened their ability to generalize across all three categories of profitability.

Through these preparation steps, including missing value imputation, leakage prevention, target engineering, categorical encoding, and class balancing, the dataset was transformed into a robust foundation for predictive modeling. This ensured that the subsequent modeling stage would fairly assess the predictive contribution of ESG and sustainability features.

5. Modeling

The modeling phase applied three machine learning algorithms, namely Decision Tree, Logistic Regression, and Random Forest, to predict firm profitability using ESG and sustainability features. Each algorithm was chosen for its distinct strengths: the Decision Tree provided transparency and straightforward interpretability, Logistic Regression offered statistical insights into the direction and magnitude of feature effects, and Random Forest combined the predictive strength of ensemble methods with the capacity to capture complex non-linear interactions.

Early experiments highlighted the critical importance of leakage prevention. When Profit Margin and Revenue were included among the predictors, all models achieved near-perfect accuracy. However, this performance was misleading because these variables directly defined profitability. Once the financial indicators were excluded, the true predictive power of ESG and contextual variables became evident.

The Decision Tree achieved an accuracy of approximately 57 percent. Its structure was easy to interpret and provided insight into the sequential decision rules that classified firms into

profitability categories. However, the model was prone to overfitting and lacked robustness when applied to unseen data, which limited its reliability in practical contexts.

The Random Forest improved upon this performance by aggregating multiple decision trees and reducing variance. Without calibration, it achieved an accuracy of around 65 percent, outperforming both the Decision Tree and Logistic Regression. Feature importance analysis showed that Environmental and Governance scores were strong positive contributors to profitability, while high energy consumption and carbon emissions reduced the likelihood of high profitability. Industry categories also added explanatory value by capturing structural differences across sectors.

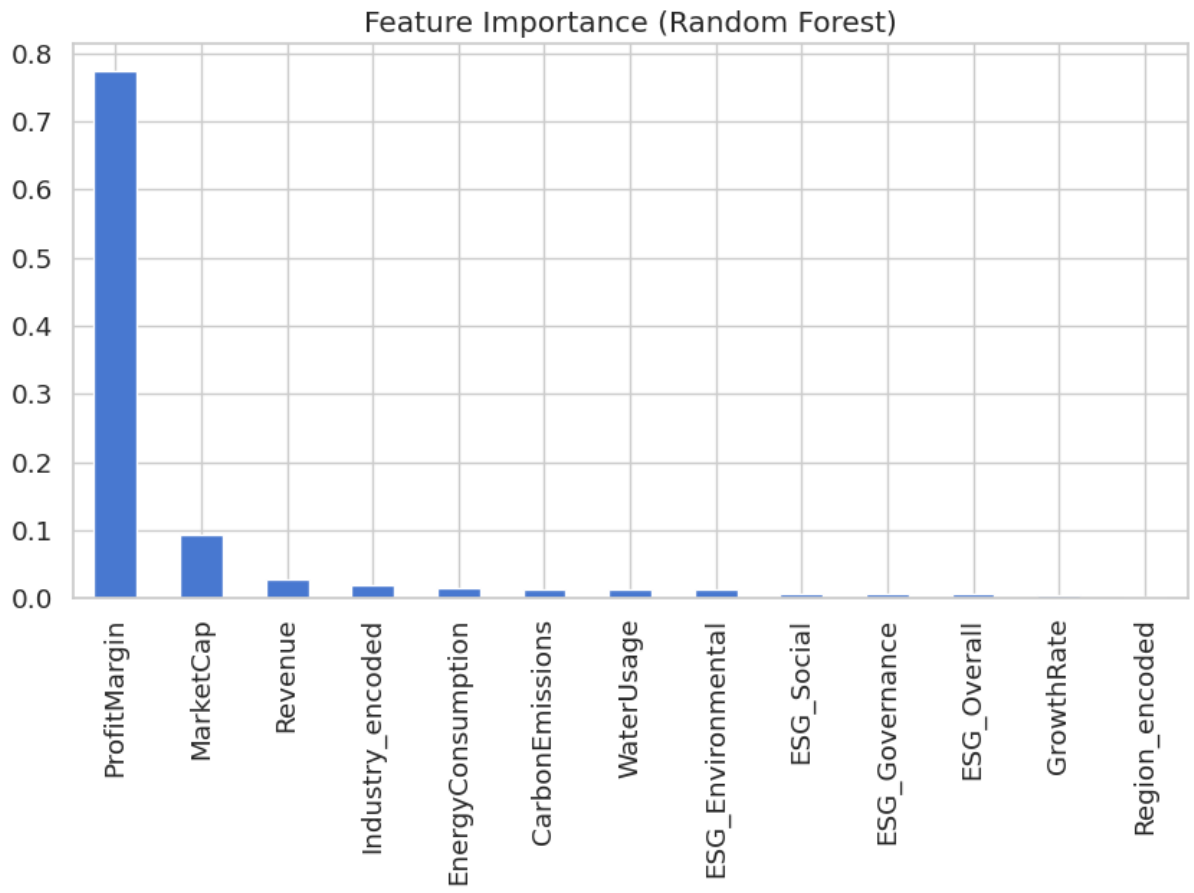

Figure 5: Feature importance before removal of leakage variables.

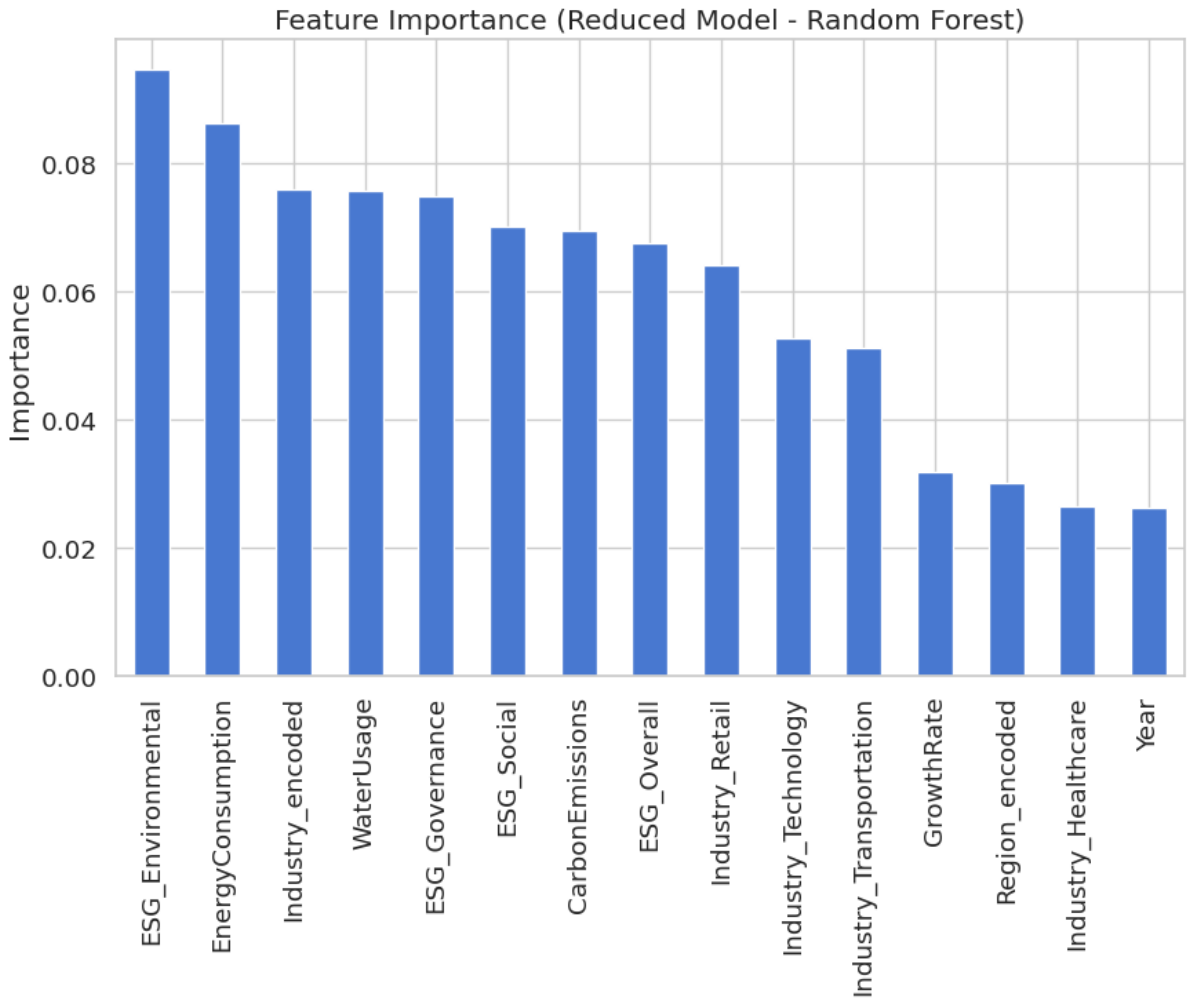

Figure 6: Feature importance after removal of leakage variables.

Logistic Regression, applied as a multinomial classifier, achieved an accuracy of approximately 55 percent. Although it was weaker than Random Forest in terms of predictive strength, it provided valuable interpretability through its coefficients. The model revealed that Environmental scores consistently increased the odds of being in higher profitability categories, whereas the influence of Social and Governance scores was less consistent and sometimes negative. High levels of carbon and energy consumption reduced profitability, which aligned with the economic burden of inefficiency.

The estimated Logistic Regression equation was:

logit(P(Profitability)) = 0.0000 + (0.0126 * ESG_Environmental) + (0.0037 * ESG_Overall) + (0.0017 * GrowthRate) + (0.0006 * Year) - (0.0012 * ESG_Governance) - (0.0004 * ESG_Social) + (Industry/Region Effects)

These coefficients indicate that Environmental and Overall ESG scores positively influence profitability, while Governance and Social dimensions showed slight negative effects. Growth Rate and Year contributed positively, and industry/region effects captured contextual variation.

Although Random Forest demonstrated stronger classification accuracy, its probability outputs were poorly calibrated. The uncalibrated model produced a ROC AUC of only 0.27, indicating that its predicted probabilities did not correspond to actual event frequencies. This lack of calibration limited its applicability in finance, where decision makers depend on reliable probability estimates for investment and risk management.

To address this issue, probability calibration was performed. After calibration, Random Forest achieved remarkable improvements: the ROC AUC increased to 0.96, the Kolmogorov-Smirnov statistic reached 0.80, and Log Loss fell to 0.39. These results demonstrated that calibration transformed Random Forest from a strong but misaligned classifier into a suitable classifier for business use.

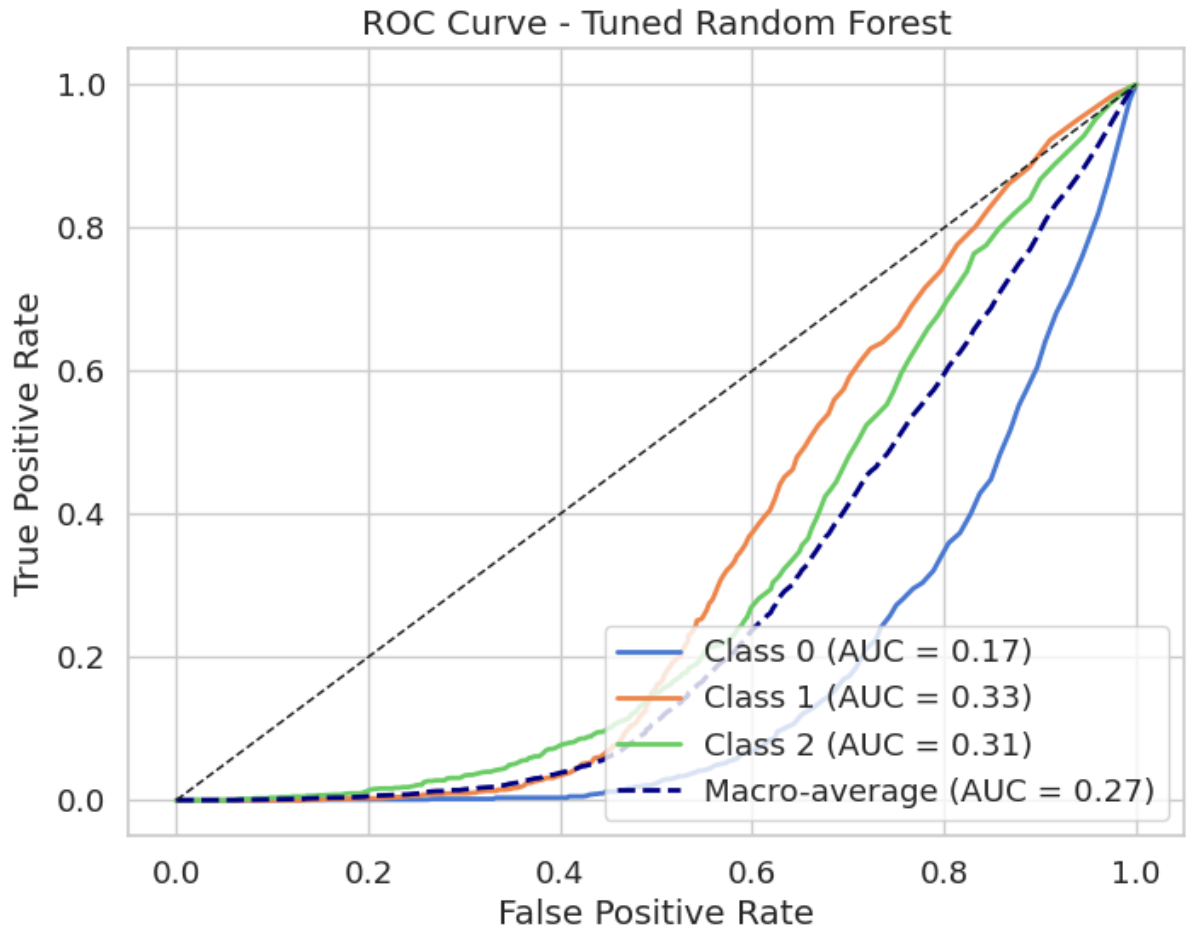

Figure 7: ROC curve of Random Forest before calibration.

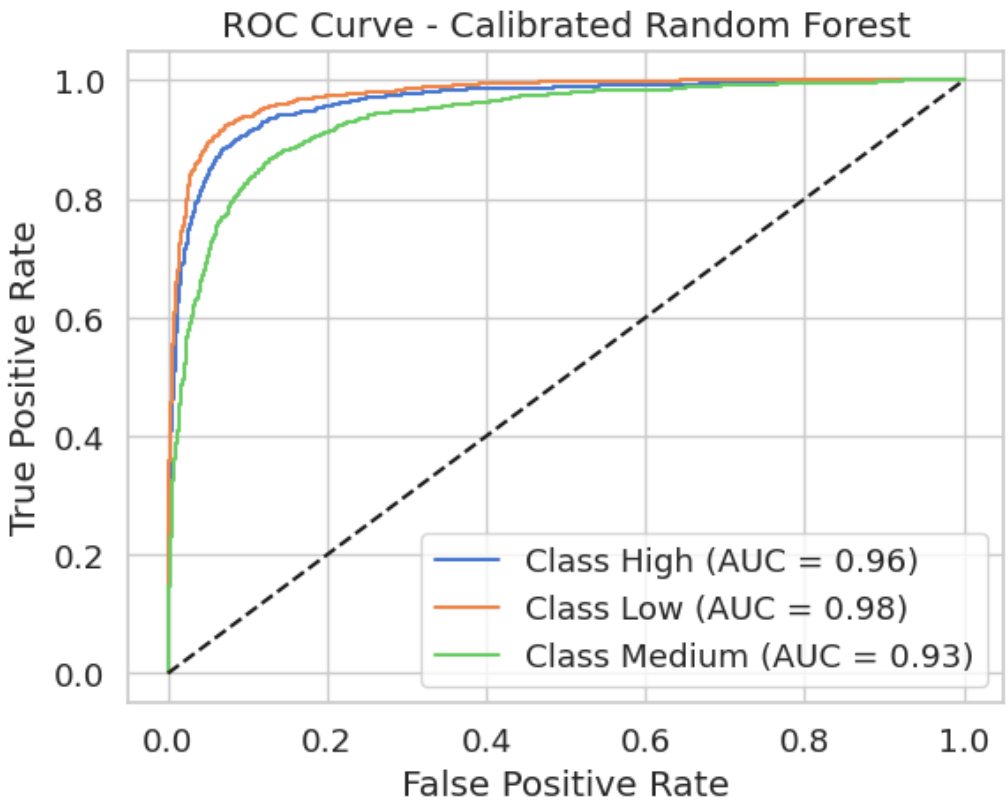

Figure 8: ROC curve of Random Forest after calibration.

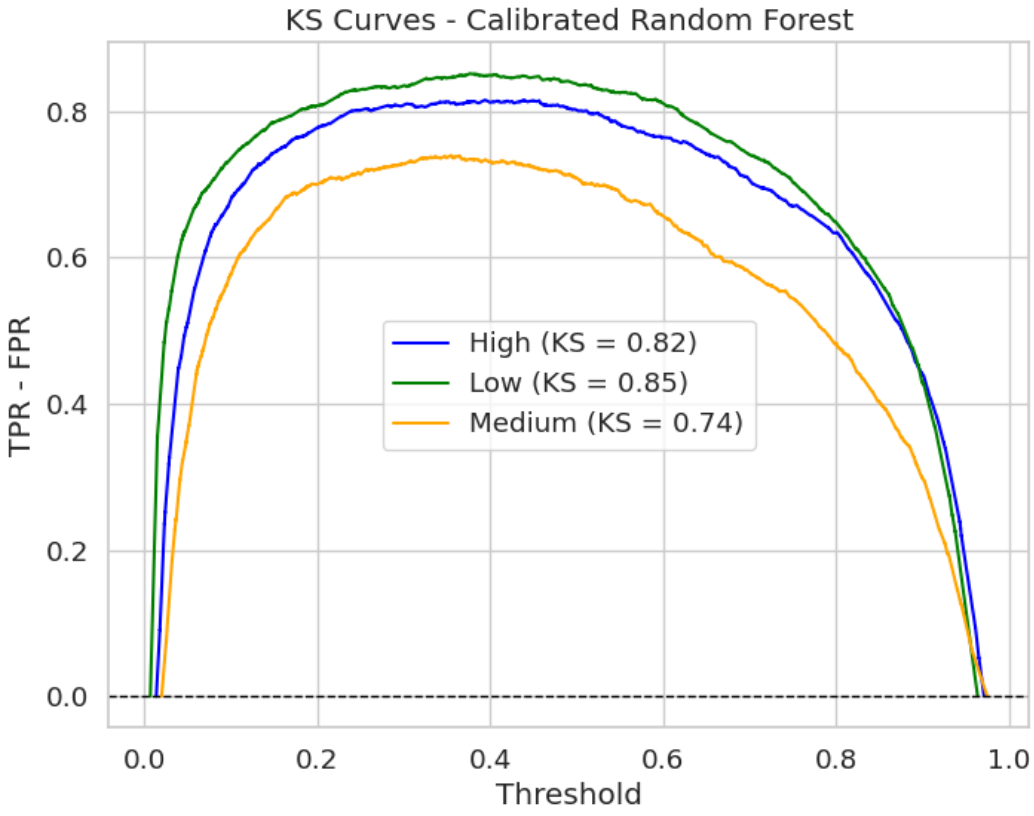

Figure 9: Kolmogorov–Smirnov curve showing class separation after calibration.

In summary, the Decision Tree offered interpretability but limited predictive power, Logistic Regression provided statistical transparency with moderate accuracy, and Random Forest emerged as the most effective approach. After calibration, Random Forest combined strong accuracy with reliable probability estimates, making it the most suitable model for financial applications where both classification and reliably calibrated probabilities are required.

6. Evaluation

The evaluation phase compared the performance of the three models using both standard accuracy and advanced metrics. Relying solely on accuracy was insufficient, as the dataset contained class imbalance and the purpose of the project required well-calibrated probability estimates. For this reason, additional measures such as Log Loss, McFadden's R^2, Brier Score, and the Kolmogorov-Smirnov statistic were employed to provide a more complete assessment of model quality.

The Decision Tree achieved an accuracy of 57 percent. This confirmed that ESG and sustainability features carried predictive information, but the model was limited by its sensitivity to small variations and its tendency to over-fit. Logistic Regression performed similarly with an accuracy of 55 percent. Although weaker in predictive power, it provided interpretability through its coefficients, showing that Environmental scores were positively related to profitability, while Social and Governance scores displayed weaker or sometimes negative associations. Both of these models demonstrated the relevance of ESG but were constrained in their ability to deliver reliable predictions.

The Random Forest achieved a higher baseline performance with an accuracy of 65 percent. Its ensemble design enabled it to capture complex relationships between ESG features, contextual factors, and profitability. Feature importance analysis confirmed that Environmental and Governance scores were consistently influential, while excessive energy consumption and carbon emissions reduced the probability of high profitability. Industry differences also contributed explanatory power by reflecting sectoral conditions.

Despite its superior classification performance, the Random Forest initially suffered from poorly calibrated probability estimates. The model achieved an ROC AUC of only 0.27 before calibration, indicating that its probability predictions did not correspond to actual observed outcomes. This limitation reduced its usefulness in finance, where decision-making depends heavily on reliable probability values rather than only class labels.

After calibration, Random Forest performance improved dramatically. The ROC AUC rose to 0.96, reflecting excellent discrimination between profitability classes. The Kolmogorov-Smirnov statistic reached 0.80, confirming strong separation of distributions. Log Loss decreased to 0.39, indicating that the model's confidence was aligned with the correctness of its predictions. These improvements transformed the Random Forest into a model that was both accurate and business-ready.

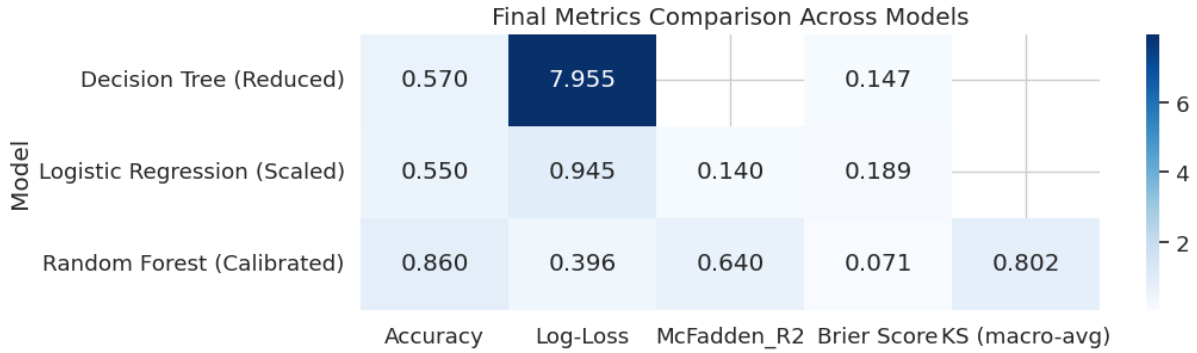

Figure 10: Heatmap of evaluation metrics for Decision Tree, Logistic Regression, and Random Forest.

Model	Accuracy	Log Loss	McFadden R^2	Brier Score	KS
Decision Tree (Reduced)	0.57	7.9545	N/A	0.1471	N/A
Logistic Regression (Scaled)	0.55	0.9445	0.1403	0.1892	N/A
Random Forest (Calibrated)	0.86	0.3955	0.6400	0.0714	0.8019

Table 1: Comparison of Model Performance Metrics.

This table summarizes the performance of Decision Tree, Logistic Regression, and Random Forest across accuracy, Log Loss, McFadden's R^2, Brier Score, and KS statistic. It shows that while Decision Tree and Logistic Regression offered interpretability with modest predictive ability, Random Forest after calibration, delivered the strongest performance across all evaluation criteria.

Taken together, the evaluation results demonstrated that Decision Tree and Logistic Regression offered interpretability but modest predictive ability. Random Forest, once calibrated, combined high accuracy with reliable probability estimates, making it the most effective model for assessing the financial relevance of ESG and sustainability practices.

7. Conclusion

This study set out to examine whether Environmental, Social, and Governance (ESG) and sustainability factors can be used to predict firm profitability when direct financial indicators are excluded. Following the CRISP-DM framework ensured that the process was systematic, beginning with business understanding and data exploration, continuing through careful preparation and modeling, and culminating in robust evaluation.

The results confirmed that ESG metrics, particularly Environmental and Governance dimensions, carry predictive value for profitability. High energy consumption and carbon emissions were consistently associated with reduced profitability, while industry-level attributes provided additional explanatory power. Although ESG factors are not the sole determinants of profitability, their influence was statistically significant and quantifiable.

From a methodological perspective, the project highlighted two critical lessons. First, controlling for data leakage is essential for ensuring model validity. Early experiments that included Profit Margin and Revenue produced near-perfect results, but these were misleading

and unsuitable for meaningful analysis. Once leakage was removed, the models reflected true predictive power. Second, probability calibration was shown to be vital in financial applications. Raw classification accuracy is insufficient in contexts where probability estimates inform investment and risk management decisions. The calibrated Random Forest, with a ROC AUC of 0.96, a Kolmogorov–Smirnov statistic of 0.80, and a Log Loss of 0.39, demonstrated that machine learning models can be transformed into tools that produce not only accurate predictions but also reliable probability estimates.

From a business perspective, the findings demonstrate that ESG metrics should not be regarded merely as compliance requirements or ethical considerations. They have measurable financial relevance and can be incorporated into investment scoring models, credit risk assessments, and portfolio optimization strategies. By capturing the link between sustainability practices and profitability, the study shows that data science can generate insights that extend beyond traditional financial indicators.

Future work could extend this analysis by applying advanced ensemble methods such as gradient boosting or XGBoost, integrating time series data to examine the dynamic effects of ESG practices, and testing models across multiple regions and sectors. Such extensions would reinforce the role of ESG as both a sustainability measure and a financial signal.

In conclusion, this project demonstrated that ESG and sustainability factors, when properly modeled and calibrated, are not only relevant for responsible business practices but also for financial performance prediction. The results provide a strong case for integrating ESG into financial analytics, supporting decision-making that aligns profitability with sustainability.

Dataset Link: 🌍 ESG & Financial Performance Dataset