Research Proposal: Do High-ESG Firms Sacrifice Profitability? Machine Learning Evidence from 2021–2024 Firm-Year Panels

Aims and Objectives, Including Research Questions

The primary aim of this project is to investigate whether firms with higher Environmental, Social, and Governance (ESG) ratings exhibit lower profit margins, using a panel dataset of listed firms from 2021 to 2024. This exploration seeks to provide empirical evidence on the potential trade-offs between ESG performance and financial profitability, contributing to the ongoing debate in sustainable finance. The objectives are structured as follows:

- **Primary Objective**: To examine the link between higher ESG ratings and profit margins using basic statistical models on firm data from 2021 to 2024.
- **Secondary Objectives**:
    1. Conduct descriptive analyses, including crosstabs and plots of ESG groups by year and industry, average ESG scores by year (with COVID annotations), correlations, and average profit margins by ESG group/score, along with distribution plots (scatter with regression line, box, bar, violin).
    2. Compare profit margins between high and low ESG groups overall and year-by-year using independent t-tests (with unequal variances) and one-way ANOVA.
    3. Develop machine learning classifiers to predict high versus low ESG groupings based on financial features (profit margin, revenue, market cap, industry dummies) in 2024, using logistic regression, random forest, gradient boosting, SVM, and KNN, with 70/30 train-test split, standardization, 5-fold CV, and metrics including accuracy, precision, recall, F1, and ROC-AUC.
    4. Visualize regression predictions (time trends by ESG group, bar by ESG quartile) and ML results (ROC curves, performance bars, confusion matrices) and explain the best ML model using SHAP values (summary dot and bar plots).

These objectives address the following research questions:

- **RQ1**: After controlling for log revenue, lagged profit margin, and firm/year effects (or industry/year dummies in pooled OLS), is a higher ESG score associated with lower profit margins? (Hypothesis: The ESG coefficient is non-positive.)
- **RQ2**: Does the profit margin differ between high and low ESG groups overall and across 2021–2024? (Hypothesis: Group means differ, as tested via t-tests and ANOVA.)
- **RQ3**: Can financial features (profit margin, revenue, market cap, industry dummies) classify high versus low ESG above chance in 2024? (Hypothesis: ROC-AUC exceeds 0.5 for at least one model.)

This project emphasizes transparency and feasibility, drawing on established econometric and machine learning methods to inform managerial and investor decisions (Angrist & Pischke, 2009; Wooldridge, 2010).

Problem Statement

The integration of ESG factors into corporate strategies has surged in recent years, driven by regulatory pressures, investor demands, and societal expectations for sustainable practices (Friede et al., 2015). However, a key challenge persists: do high-ESG firms incur short-term costs that erode profitability? Activities such as enhanced environmental reporting, supplier audits, or stakeholder engagement may increase operational expenses without immediate financial returns, potentially leading to lower profit margins (Krüger, 2015; Capelle-Blancard & Petit, 2019). This tension is particularly relevant in the post-pandemic era (2021–2024), where economic volatility, inflation, and supply chain disruptions have amplified cost pressures.

The background context includes the rapid growth of ESG investing, with global assets under management exceeding $30 trillion by 2022 (Eccles et al., 2014). Yet, inconsistencies in ESG ratings across providers complicate analysis, often leading to divergent conclusions on financial impacts (Berg et al., 2022; Gibson et al., 2021). This study addresses this by focusing on a numeric ESG score (1–7) derived from ratings like AAA to CCC/B, mapped to high/low binary groups for simplicity, excluding mid-tier firms to sharpen comparisons.

The scope is limited to 100 publicly listed firms across industries, using a curated CSV dataset yielding 400 firm-years after panel restructuring. It begins with data preparation and descriptive statistics, extends to t-tests/ANOVA, panel regressions, and machine learning, and includes visualizations like SHAP plots. Broader causal claims or long-term effects are excluded, emphasizing short-run associations. This research is relevant to managers balancing ESG goals with earnings targets, investors assessing sustainability premiums, and policymakers evaluating ESG mandates (Freeman, 1984; Chatterji et al., 2016).

Literature Overview

This project draws on academic and business sources to contextualize ESG-profitability links. Key academic works include Friede et al. (2015), a meta-analysis of over 2,000 studies showing mixed but often positive long-term ESG-financial performance associations, though short-term costs are underexplored. Eccles et al. (2014) provide evidence from firm-level data that high-sustainability firms outperform in processes and performance, but Krüger (2015) highlights negative stock reactions to ESG news, suggesting near-term penalties.

On methodology, Angrist and Pischke (2009) and Wooldridge (2010) guide fixed-effects panel regressions for controlling unobserved heterogeneity. Cameron and Miller (2015) inform clustered standard errors for robust inference (though not explicitly applied in analysis). For group comparisons, Welch (1947) underpins independent t-tests with unequal variances. Machine learning references include Breiman (2001) for random forests, Friedman (2001) for gradient boosting, Cortes and Vapnik (1995) for SVM, and Lundberg and Lee (2017) for SHAP explanations. Performance metrics follow Hanley and McNeil (1982) for ROC-AUC.

Business sources include reports from Sustainalytics and MSCI on ESG rating methodologies, highlighting divergence issues (Berg et al., 2022). Industry insights from McKinsey and Deloitte emphasize ESG's reputational benefits but warn of implementation costs (Capelle-Blancard & Petit, 2019). These sources will be critically evaluated for biases, such as selection in meta-analyses or provider-specific rating inconsistencies.

Also, ChatGPT-4 and similar generative AI tools were used for writing support and analytical guidance in the preparation of this work.

Proposed Methodology

This observational study employs a firm-year panel design (2021–2024), leveraging secondary data from a curated CSV file ('Dissertation Dataset - Final.csv') sourced from Sustainalytics and MSCI ratings. The sample comprises 100 firms (400 firm-years post-restructuring), focusing on listed entities with available ESG ratings (AAA to B), revenue, net income (NI), profit margin (NI/revenue), market cap, revenue growth, industry, and HQ country.

Variables include:

- **Outcome**: Profit margin.
- **Predictor**: ESG score (1–7, where 7=AAA/high, 3=B/low); binary high/low group (high: AAA/AA/A; low: BB/B/CCC/CC/C; excluding mid if present).
- **Controls**: Log (revenue +1) for size; lagged profit margin; industry dummies (in pooled OLS); firm and year effects (in fixed effects).

Data preparation involves loading via Pandas, checking shape/head/info/describe/isnull, mapping ESG to scores/groups, restructuring to panel format (melting years), filtering to high/low groups, sorting by firm/year, computing log revenue and log market cap (+1), and creating lagged profit margin. Rows with missing values in key variables are dropped. No winsorization or imputation is applied. Descriptive checks include summaries, correlations, crosstabs (ESG group by year/industry), and plots via Matplotlib/Seaborn (bar for crosstabs, line for average ESG by year with COVID annotations, scatter with regression line, box/bar/violin for profit margins by group/score).

Group comparisons use SciPy's stats.ttest_ind (independent t-tests with equal_var=False for unequal variances) for profit margin differences between high/low ESG overall and by year, and stats.f_oneway for one-way ANOVA by year (on ESG groups).

Main models (via Statsmodels):

1. **Pooled OLS**: ProfitMargin = ESG_Score + Log_Revenue + Lagged_Profit_Margin + industry dummies + year dummies + ε (with constant).
2. **Firm + Year Fixed Effects**: ProfitMargin ~ ESG_Score + Log_Revenue + Lagged_Profit_Margin + C(Firm) + C(Year) (via smf.ols formula).

For 2024 cross-section ML classification (high=1, low=0 ESG group):

- Features: Profit margin, revenue, market cap, industry dummies.
- Split: 70/30 train-test (stratified, random_state=42).
- Scaling: StandardScaler on train/test.
- Models: LogisticRegression (max_iter=1000), RandomForestClassifier (n_estimators=100), GradientBoostingClassifier (n_estimators=100), SVC (probability=True), KNeighborsClassifier.

- Validation: 5-fold CV on train (scoring='roc_auc').
- Metrics (on test): Accuracy, precision, recall, F1, ROC-AUC; confusion matrices; classification reports.
- Explanations: SHAP LinearExplainer on logistic model for test set, with summary dot and bar plots.

All analyses use Python (Pandas/NumPy for data, Statsmodels for regressions, Scikit-learn for ML/SHAP/Metrics, Matplotlib/Seaborn for plots). No primary data collection is involved. Regarding generative AI, tools like ChatGPT-4 may be used for writing or analysis support; all code and interpretations are manually developed.

Approach to Analyzing Findings/Outcomes

Findings will be analyzed through coefficient estimates/p-values from regressions (e.g., ESG_Score coefficient for RQ1). For group comparisons (RQ2), t-statistics/p-values from t-tests/ANOVA and group means will be reported. ML outcomes (RQ3) involve test set ROC-AUC, accuracy, precision, recall, and F1, with CV means/std for AUC. Feature importance via SHAP values (mean absolute and directional impacts) will highlight predictors of high ESG.

Results will be tabulated (e.g., regression summaries, ML metrics, classification reports) and visualized (e.g., time trend plots of actual vs. pooled predictions by ESG group, bar of actual vs. fixed effects predictions by ESG quartile, ROC curves, performance bars across metrics/models, confusion matrix heatmaps, SHAP dot/bar plots). Interpretation distinguishes correlations from causality, with no additional sensitivity checks beyond coded elements.

Limitations

Observational data limits causal inference: fixed effects mitigate time-invariant confounders but not time-varying ones (Angrist & Pischke, 2009). The short 2021–2024 window captures near-term effects but misses long-run dynamics. ESG measurement variability across providers may bias scores (Berg et al., 2022). Sample size (100 firms) risks low power and excluding mid-ESG reduces generalizability. ML relies on limited features and a small 2024 sample (~100 observations post-split). Access to proprietary data could be a barrier if CSV updates are needed.

Ethical Challenges and Possible Risks

Using secondary firm-level data avoids human subjects' issues, complying with dataset terms (no personal data). Ethical risks include misinterpretation of results leading to misguided ESG policies; caveats will emphasize non-causality. Bias in ESG ratings (e.g., favouring large firms) will be noted. Risks: data inaccuracies from CSV curation; mitigated by cleaning logs. No high risks per university tool (e.g., low financial/reputational impact). All work adheres to academic integrity.

Executive Summary Provision

The Executive Summary will be provided as a written summary (1-page) alongside the Final Report, supplemented by an infographic visualizing key findings (e.g., regression coefficients, ML ROC-AUCs, SHAP plots).

References

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46*(3), 399–424. https://doi.org/10.1080/00273171.2011.568786

Berg, F., Kölbel, J. F., & Rigobon, R. (2022). Aggregate confusion: The divergence of ESG ratings. *Review of Finance, 26*(6), 1315–1344. https://doi.org/10.1093/rof/rfac033

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources, 50*(2), 317–372. https://doi.org/10.3368/jhr.50.2.317

Capelle-Blancard, G., & Petit, A. (2019). Every little helps? ESG news and stock market reaction. *Journal of Business Ethics, 157*(2), 543–565. https://doi.org/10.1007/s10551-017-3667-3

Chatterji, A. K., Durand, R., Levine, D. I., & Touboul, S. (2016). Do ratings of firms converge? Implications for managers, investors and strategy researchers. *Strategic Management Journal, 37*(8), 1597–1614. https://doi.org/10.1002/smj.2407

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018

Eccles, R. G., Ioannou, I., & Serafeim, G. (2014). The impact of corporate sustainability on organizational processes and performance. *Management Science, 60*(11), 2835–2857. https://doi.org/10.1287/mnsc.2014.1984

Freeman, R. E. (1984). *Strategic management: A stakeholder approach*. Pitman.

Friede, G., Busch, T., & Bassen, A. (2015). ESG and financial performance: Aggregated evidence from more than 2000 empirical studies. *Journal of Sustainable Finance & Investment, 5*(4), 210–233. https://doi.org/10.1080/20430795.2015.1118917

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gibson Brandon, R., Krueger, P., & Schmidt, P. S. (2021). ESG rating disagreement and stock returns. *Financial Analysts Journal, 77*(4), 104–127. https://doi.org/10.1080/0015198X.2021.1963186

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29–36. https://doi.org/10.1148/radiology.143.1.7063747

Koenker, R. (2005). *Quantile regression*. Cambridge University Press. https://doi.org/10.1017/CBO9780511754098

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica, 46*(1), 33–50. https://doi.org/10.2307/1913643

Krüger, P. (2015). Corporate goodness and shareholder wealth. *Journal of Financial Economics, 115*(2), 304–329. https://doi.org/10.1016/j.jfineco.2014.09.008

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. In *Proceedings of the 9th Python in Science Conference* (pp. 92–96). https://doi.org/10.25080/Majora-92bf1922-011

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*(1), 1–21. https://doi.org/10.1214/09-STS313

Welch, B. L. (1947). The generalization of 'Student's' problem when several different population variances are involved. *Biometrika, 34*(1–2), 23–35. https://doi.org/10.1093/biomet/34.1-2.23

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). MIT Press.