International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 07 Issue 06 June 2024
DOI: 10.47191/ijcsrr/V7-i6-09, Impact Factor: 7.943
IJCSRR @ 2024
www.ijcsrr.org

Advanced TRST01 ESG Scoring Model with Beta Based Financial Metrics and Machine Learning Techniques

Gurucharan Kottapalli[1], Prabir Mishra[2]
[1]MBA – IIM Ahmedabad (Batch of 25), Intern @ TRST01, M.Sc. (Agricultural Statistics)
[2]CEO @ TRSTO1

ABSTRACT: In the current corporate world, assessing a company's sustainability performance is very important for investors, stakeholders, and policymakers. The TRST01's ESG (Environmental. Social and Governance) Scoring Model introduces an innovative approach integrating beta-based financial metrics with advanced machine learning techniques to comprehensively evaluate ESG credentials. This study demonstrates the development and application of the TRST01's ESG scoring model, which leverages data from the most reputable sources such as MSCI and S&P Global to ensure its reliability and accuracy. The model's unique methodology involves calculating country-specific beta values to normalize carbon emission data, thereby providing a standardized metric for meaningful comparisons across countries. Further, ESG scores are adjusted using both country and company beta values to reflect specific risk exposures, enhancing the precision and relevance of the assessments. The model ensures robust input data quality, by taking Market capital, Scope 1, Scope 2, industry wise data and beta values as predictors through extensive data preprocessing and encoding categorical variables for top 1000 listed companies. A comparative analysis of Traditional model such as Simple Linear Regression (SLR) and multiple Machine Learning (ML) models, including Gradient Boosting (GB), Support Vector Regression (SVR), and Random Forest (RF), demonstrates that the Gradient Boosting model achieves superior performance with minimal overfitting and consistent prediction accuracy. The study employs a comprehensive evaluation framework using various metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared, supplemented by detailed visualizations of actual vs. predicted values, residuals, and error distributions. This research underscores the significance of incorporating advanced financial metrics and machine learning techniques in ESG assessments, providing a reliable, accurate, and holistic framework for understanding corporate sustainability. The TRST01 ESG Scoring Model sets a new standard in sustainability evaluation, offering valuable insights for stakeholders committed to integrating sustainability into core business strategies.

KEYWORDS: Sustainability, ESG Scoring, Country Beta, Company Beta, SLR, GB, SVR, RF

1. INTRODUCTION

In today's dynamic corporate environment, evaluating a company's sustainability performance is crucial for investors, stakeholders, and policymakers. The TRST01 ESG Scoring Model is an innovative tool that offers a thorough assessment of a business's Environmental, Social, and Governance (ESG) credentials. This model distinguishes itself from traditional assessment tools by employing a beta-based scoring methodology, which integrates financial metrics with ESG evaluations to provide a nuanced and statistically robust analysis. The demand for a sophisticated and reliable ESG assessment tool is higher than ever. Companies face increasing scrutiny over their sustainability practices, and investors are eager to incorporate ESG factors into their decision-making processes.

The TRST01 model is designed to meet various needs. It provides reliable data by sourcing information from reputable sources such as MSCI and S&P, ensuring that its ESG scores are accurate and dependable. The model also enhances investment decisions by incorporating financial metrics, offering a balanced view of a company's risk profile and market dynamics, which is essential for making informed investment decisions.With a strong statistical foundation and a significant R-squared value, the model demonstrates statistical precision, ensuring consistency and precision in ESG scoring.

The TRST01 ESG Scoring Model thus represents a critical advancement in ESG evaluation, addressing the growing need for reliable, comprehensive, and precise sustainability assessments. Environmental, social, and governance (ESG) concepts in investing strategies are increasingly important in today's financial world. Investors and financial advisors are increasingly wondering if ESG ideals help or hurt financial performance. Our research examines how ESG ratings affect financial performance using the Extreme Gradient method and cutting-edge machine learning. ESG score is one of the top five predictors of mutual fund performance, according to our research. As sustainable investing becomes more important in financial markets, this study shows how ESG aspects affect investment outcomes. Socially responsible investors and financial advisors can use our information to make informed decisions that fit their financial goals with their ESG beliefs (Momparler *et al*., 2024).

Random Forests (ML model) is to examine the correlation between betas and 13 financial and non-financial characteristics for S&P 500 stocks from 2015 to 2019. The Relative Importance criterion indicates that ESG Scores are the primary factor in beta generation, determination, and sign. They confirm a correlation between betas and industrial sectors, revealing some variation across the investigated betas. Using the Random Forests methodology ensures relevant results by preventing plausible correlations between variables(Martín-Cervantes and Valls Martínez, 2023).

For private enterprises or non-traded assets, equity betas are often estimated using comparable company research. Test if the Random Forest method may improve projections. Every year from 1992 to 2018, Random Forest forecasts have fewer average and mean absolute errors in out-of-sample tests (Alanis, 2022). (Avramov et al., 2022) examined the impact of uncertainty surrounding ESG scores for publicly traded securities on the NYSE, AMEX, and Nasdaq. High uncertainty led to increased beta risk of stocks, altering the risk-return profiles of the evaluated companies.

This study examines how the global financial crisis affected multihorizon systematic risk and market risk using daily data from eight major European equities markets from 2005 to 2018. Wavelet multiscale is used in a capital asset pricing model. Beta coefficients grow at greater scales (lower frequencies) according to empirical data. During a crisis, betas and R2s tend to grow larger than before. According to scale-dependent value at risk (VaR), market risk is concentrated at lower time scales (greater frequencies) of data. Finally, our method accurately predicts time-dependent betas and VaR (Alexandridis & Hasan, 2020).

This paper introduces Weighted Forward Search (FSW) for outlier detection in asset pricing data. It down weights the most anomalous observations and tests using simulated and empirical data. It assesses outliers' impact on asset pricing model estimation and introduces statistical tests based on this new approach. It also presents an alternative robust portfolio beta estimation approach for comparing asset pricing models.(Aronne et al., 2020). This study uses ESG data to create an automated trading strategy and assess a company's ESG premium and ESG alpha. It involves establishing an ESG investing universe, conducting feature engineering on ESG data, training machine learning models to forecast stock prices, and automatically managing portfolios using an ensemble method. Trades are executed and the portfolio adjusted weekly based on forecasted stock prices.(Chen & Liu, 2020).

ESG information affected organizations valuation and performance through their systematic (lower capital costs and higher valuations) and idiosyncratic (increased profitability and reduced tail risk) risk profiles. Research suggests that ESG reforms may benefit a company's finances. ESG ratings in policy benchmarks and financial evaluations may work (Giese et al., 2019).

The empirical evidence on the performance of responsible investing (RI) that combines environmental, social, and governance (ESG) into investment decisions is mixed. Fama and French's five-factor model and an ESG-related factor apply to 1,425 US open-end equity fund returns from April 2009 to December 2016. The data suggests that RI protects against ESG-related systematic risk that diversification cannot minimize. (Jin, 2018).

The beta anomaly, or low (high) abnormal returns of equities with high (low) beta, is one of the most persistent anomalies in empirical asset pricing research. Investor desire for lottery-like equities drives the beta anomaly, as shown in this article. The beta anomaly disappears when beta-sorted portfolios are neutralized to lottery demand, regression specifications adjust for lottery demand, or factor models include a lottery demand factor. The beta anomaly only occurs in low-institutional-owned equities when lottery demand affects high-beta stocks (Bali et al., 2017).

The use of ESG scores and a European panel dataset with 8752 company-year observations from 2002 to 2014 to evaluate how Corporate Social Performance (CSP) strategies affect corporate financial performance. Thus, they show that only the second of the three ESG (Environmental, Social, and Governance) scores corporate governance activities related to the social aspect offers firms under analysis a significant opportunity to increase their value at the expense of reduced systematic risk (Sassen et al., 2016).

This study highlights how public corporations' transparency in revealing non-financial (ESG) data can reduce investment portfolio risk by lowering return rate volatility on their securities. It also identifies a significant lack of ESG reporting in the Polish market, with minimal non-financial data reporting overall. Additionally, it shows that higher ESG-rated firms' shares tend to have an above-average return rate, lower return rate volatility, and lower forecasting error in return rates, as indicated by the standard error parameter, alpha, and beta coefficients. (Czerwińska & Kaźmierkiewicz, 2015).

This study demonstrated how business sustainability affects processes and performance. They found in a matched sample of 180 U.S. companies that high sustainability companies, which voluntarily adopted sustainability policies by 1993, had distinct organizational processes by 2009, compared to low sustainability companies, which adopted almost none. Finally, high sustainability companies beat their peers in stock market and accounting success over time (Eccles et al., 2014).

This study examines CSR and corporate risk in controversial industries. After controlling for company characteristics in a large 1991–2010 U.S. sample of problematic industries like alcohol, tobacco, gambling, and others, CSR activity inversely affects business risk. They find that CSR participation in problematic industries lowers business risk using system equations and difference regressions to address endogeneity. The comparison of non-controversial and controversial firm samples shows that CSR engagement decreases risk more economically and statistically in controversial enterprises (Jo & Na, 2012).

2. METHODOLOGY

2.1. Data Collection

The carbon emission factors are sourced from **Carbon Footprint**, which provides opensource data suitable for organizations reporting their carbon emissions. These factors are collated from numerous sources and reflect the carbon emissions per kWh of electricity for different countries. The average industry scores of MSCI and S&P Global for top 1000 listed companies have taken as a base reference for calculation of adjusted scores. In addition, to this we have collected scope 1 and scope 2 emissions data from BRSR reports, Market cap, and beta value of stock for the past years 2023 and 2022 data and segregated all into respective industries. The TRST01 Adjusted Scores are calculated in a two-step process: first, adjusting the average scores based on country specific beta values, and then further adjusting these scores based on the company's beta value.
Link: https://www.carbonfootprint.com/international_electricity_factors.html

2.2. Calculation of Trst01 Adjusted scores

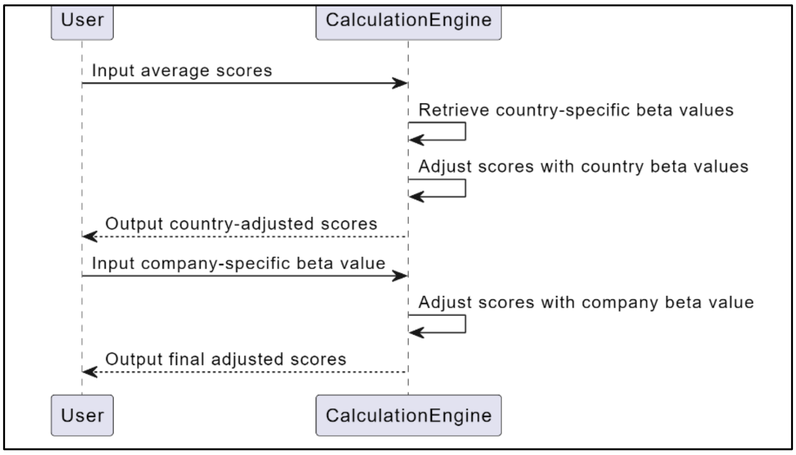
User
CalculationEngine
Input average scores
Retrieve country-specific beta values
Adjust scores with country beta values
Output country-adjusted scores
Input company-specific beta value
Adjust scores with company beta value
Output final adjusted scores
User
CalculationEngine

Fig 1. Procedure for Calculation of Scores

Fig 1 gives the overview of the calculation of Trst01 adjusted ESG scores. The calculation and methodology will be explained in detail manner.

2.3. Calculation of Country Beta

The calculation of country beta is an essential part of evaluating the environmental performance of countries based on their grid electricity carbon emissions. This process involves the use of data from Grid Electricity Conversion Factors to create a standardized metric that allows for meaningful comparisons of carbon emissions across different countries. Here's a detailed explanation of the scope and calculation of the country beta.

2.3.1. Scope

The country beta calculation involves:

- Collecting carbon emission factors for grid electricity from various countries over the past five years (2018 - 2023) and have taken two years data for future calculation in analysis to maintain uniformity across the years.
- Using these emission factors to determine a relative measure (beta) that reflects each country's carbon intensity compared to a baseline.
- Adjusting the beta values to account for variations across countries and provide a standardized metric for comparison.

2.3.2. Calculation Process

The process of calculating the country beta is as follows:

2.3.2.1. Data Collection

- Gather carbon emission factors (in kg CO2e per kWh) for grid electricity for all countries from 2018 to 2023.
- Ensure the data is aligned and consistent across the five-year period.

2.3.2.2. Baseline Emission Factor

- Set a baseline emission factor, Ebaseline, which is a reference value for comparison. In this case, Ebaseline=0.4. This average reference is taken to involve as many countries as possible.

2.3.2.3. Calculate Average Emission Factor

- Compute the average emission factor for all countries for a specific year. This is done by averaging the emission factors of all countries:

$$\frac{\sum_{i=1}^{n} E_i}{n}$$

Where Ei is the emission factor of the ith country and n is the total number of countries.

2.3.2.4. Determine Country Specific Ratios

- For each country, compare its emission factor to the baseline. If the country's emission factor is greater than the baseline, use the emission factor; otherwise, use zero.
- Calculate the ratio of this adjusted emission factor to the average emission factor

$$\text{Ri} = \frac{\text{IF (Ei>E baseline,Ei,0)}}{\text{E avg}}$$

2.3.2.5. Calculate Adjusted Beta for Each Country

- The country adjusted beta is determined by normalizing the country specific ratios by the average of these ratios across all countries:

$$\beta i \ = \frac{Ri}{\bar{R}}$$

where $\bar{R}$ is the average of all country specific ratios Ri.

The country beta provides a relative measure of a country's carbon intensity in grid electricity production compared to a baseline and an average emission factor. By normalizing these values, the TRST01 model creates a standardized metric that allows for meaningful comparisons of carbon emissions across different countries. This process ensures that stakeholders can accurately assess and report on the environmental performance of organizations in various regions.

2.3.2.6. Calculation for the Adjusted Scores as per Country Beta (India)

Adjusted E Score (Environmental Score)

$$\text{Adjusted E Score} = \frac{\text{Average E Score}}{\textit{Country Beta}}$$

Adjusted S Score (Social Score)

$$\text{Adjusted S Score} = \quad \text{Average S Score x} \left(\frac{100-\text{Average E Score}}{\text{Average S Score}+\text{Average G Score}}\right)$$

Adjusted G Score (Governance Score)

$$\text{Adjusted G Score} = \quad \text{Average G Score x} \left(\frac{100-\text{Average E Score}}{\text{Average S Score}+\text{Average G Score}}\right)$$

2.4. Calculation of Adjusted Score Considering Company Beta

Final Adjusted E Score

$$= \begin{cases} \frac{\text{Average E Country Beta Score}}{\textit{Company Beta}} & \text{if Company Beta } > 1 \\ \text{Average E Country Beta Score } \times |\text{Company Beta}| & \text{if Company Beta } < 1 \\ \text{Average E Country Beta Score} & \text{if Company Beta } = 1 \end{cases}$$

2.4.1. Intuition Behind Adjusting Scores Using Beta

The adjustment of ESG scores using the company beta is grounded in financial risk management principles. Here's a deeper look into the rationale.

2.4.1.1. Understanding Beta

Beta is a measure of a company's volatility or risk relative to the overall market.

- A beta greater than 1 implies the company is more volatile and riskier than the market. To reflect this risk, the ESG score should be reduced, indicating the company must do more to achieve the same level of ESG performance.
- A beta less than 1 implies the company is less volatile and hence less risky than the market. The ESG score should be increased to reflect the lower risk, indicating the company's ESG performance is more credible due to its lower risk profile.

Final Adjusted S Score=

$$\text{Adjusted S Score from Country Beta x} \left(\frac{100-\text{ Adjusted E Score from Country Beta}}{\text{Adjusted S Score from Country Beta + Adjusted G Score from Country Beta}}\right)$$

Final Adjusted G Score=

$$\text{Adjusted G Score from Country Beta x} \left(\frac{100-\text{ Adjusted E Score from Country Beta}}{\text{Adjusted G Score from Country Beta + Adjusted S Score from Country Beta}}\right)$$

Like the social score, the governance score is adjusted considering the adjusted E score based on the company's beta.

2.5. Model Building

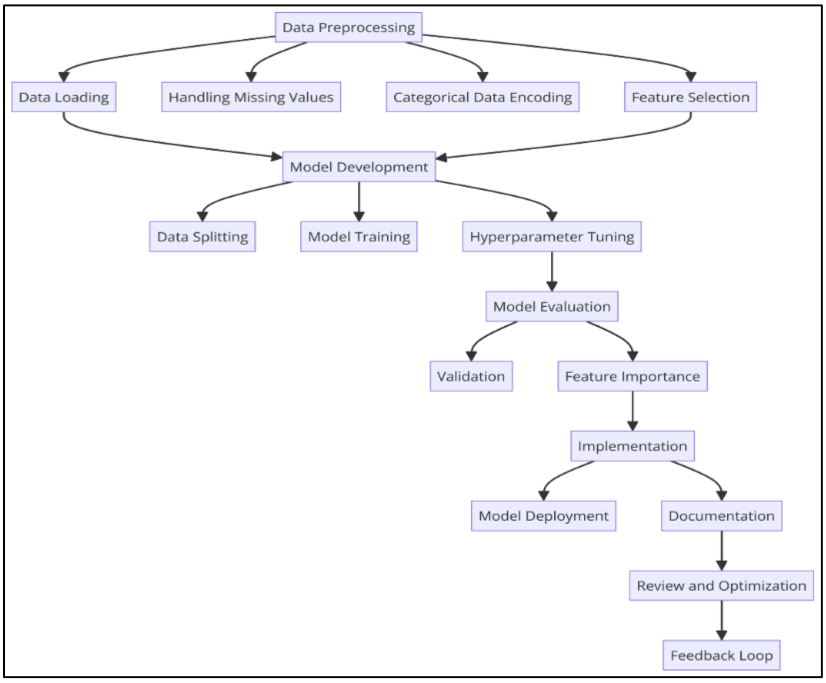

Fig 2: Schematic representation for Model Building

3. MODELS PERFORMED

3.1. Simple Linear Regression

Regression is a statistical term used for describing models that estimate the relationships among variables. Linear regression models study the relationship between a single dependent variable Y and one or more independent variables, denoted by X. If there is only one independent variable X, it is called simple linear regression; if there is more than one independent variable, it is called multiple linear regression. This Accidental Note discusses the statistical and geometric interpretation of simple linear regression models (Bangdiwala, 2018).

3.1.1. The mathematical simple linear regression model

The simple linear model is exactly that, a simple straight line that relates the one independent variable X to the dependent variable Y. It is given by the mathematical formula for a straight line.

$$Y = \beta_0 + \beta_1 X_1$$

where β_0 is called the intercept and $\beta_1$is called the slope. In the standard x–y Cartesian plane, the intercept is the point on the y-axis that is intersected by the line, and the slope is the amount of change in the y-axis for a 1 unit change in the x-axis. The intercept and slope are the two values that completely characterize a straight line, i.e. by providing these two values, one knows exactly the only straight line that they describe. The intercept is interpreted as ‘positioning’ the regression, while the slope is interpreted as ‘quantifying the association’ between X and Y. If the slope is 0, there is no association; if the slope is negative, the association is negative (for a unit increase in X, Y decreases by β_1); and if the slope is positive, the association is positive (for a unit increase in X, Y increases by β_1). The larger the absolute value of the slope, the ‘stronger’ is the association.

3.2. Random Forest (Regression)

Random forest (RF) is a flexible, easy-to-use machine learning method that, in most cases, delivers good results even without hyper-parameter tuning. Because of its simplicity and diversity, it is also one of the most often used algorithms. Random forest is a supervised machine learning algorithm. It builds a "forest" out of an ensemble of decision trees, which are generally trained using the "bagging" process. A Random Forest is made up of several trees that are built in a specific "random" manner. Each tree is made up of a distinct sample of rows, and each node is split up into a different set of features. Each tree has its prediction. After then, the average of these predictions is used to create a single result. The bagging method's general concept is that combining several learning models improves the outcome.

Form of the regression trees model

$$f(x) = \sum_{m=1}^{M} c_m . 1_{(x \in R_m)}$$

Where, R1, R2, …, RM represent a partition of feature space.

3.3. Gradient Boost

Gradient boosting regression tree algorithms employ an ensemble learning approach to create robust forecasting models by integrating multiple individual regression trees, also known as weak learners. These algorithms aim to minimize the error rate of weakly learned models, which have high bias and low variance on the training dataset. Boosting algorithms typically consist of three components: an additive model, weak learners, and a loss function. This approach can represent non-linear relationships and utilizes various differentiable loss functions. Gradient boosting machines (GBM) work by identifying the limitations of weak models through gradients and use an iterative approach to gradually improve the base learners and reduce forecast errors. This is achieved by combining decision trees through an additive model while minimizing the loss function using gradient descent.

3.4. Support Vector Regression (SVR)

Support Vector Regression (SVR), the supervised learning algorithm is used to predict discrete values. SVR's main aim is to locate the line of best fit. The best-fit line in SVR is the hyperplane with the maximum number of points. The SVR, unlike other regression models, aims to fit the best line inside a threshold value, rather than minimizing the error between the real and projected value. The distance between the hyperplane and the boundary line is the threshold value. The goal of Support vector regression is to develop a function that approximates mapping from an input domain to real numbers by using training sample data. The key goal here is to choose a decision boundary that is a distance from the original hyperplane and contains data points closest to the hyperplane or support vectors.

Fitting of Support Vector regression $f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$ can be expressed as

minimize $\{\sum_{i=1}^{n} \max[0{,}1 - y_i f(x_i)] + \lambda \sum_{j=1}^{p} \beta_j^2\}$

The main idea behind SVR is to transform the input space into a high-dimensional variable space and then build the regression or time series model in this transformed feature space.A vector of data set says $Z = \{x_i y_i\}_{i=1}^{N}$, where $x_i \in R^n$ is the input vector, y_i is the scalar output, and N is the size of the data set. The general equation SVR can be written as follows:

$$f(x) = W^T \emptyset(x) + b$$

where, W is the weight vector, b is the bias term, and superscript T denotes the transpose.

The coefficients W and b are estimated from data by minimizing the following regularized risk function:

$$R(\theta) = 1/2||w||^2 + C[1/N \sum_{i=1}^{N} L_\varepsilon(y_i, f(x_i))]$$

This regularized risk function minimizes both the empirical error and the regularized term simultaneously, helping to avoid both underfitting and overfitting of the model. The first term 1/2|| w ||2 is called the ‘regularized term’, which measures the flatness of the function. Minimizing 1/2|| w ||2 will make a function as flat as possible.

The second term $1/N \sum_{i=1}^{N} L_\varepsilon(y_i, f(x_i)$ is called the ‘empirical error’, which was estimated by Vapnik ε-insensitive loss function as follows:

$$L_\varepsilon(y_i, f(x_i)) = f(x) = \{|y_i, f(x_i) - \varepsilon|; \quad |y_i - f(x_i)| \geq \varepsilon$$

$$0, \qquad |y_i - f(x_i)| < \varepsilon$$

where,y_i is actual value and $f(x_i)$ is an estimated value. The most commonly used kernel function is the radial basis function (RBF) which is given as follows:

$$k(x_i, x_j) = \exp\{-\gamma||x - x_i||^2$$

3.5. Model comparison criteria.

3.5.1. Mean Absolute Error (MAE)

Mean Absolute error (MAE) the average of the absolute differences between the predicted values and the actual values.

MSE=$\frac{\sum_{i=1}^{N}|Y_i- \hat{Y}_i|}{N}$

Where n is number of data points, Yi is the observed values, $\hat{Y}_i$ is the predicted values

3.5.2. Mean Square Error (MSE)

Mean squared error (MSE) measures error in statistical models using the average squared difference between the observed and predicted values.

MSE = $\frac{\sum_{i=1}^{N}(Y_i- \hat{Y}_i)^2}{N}$

Where n is number of data points, Yi is the observed values, $\hat{\boldsymbol{Y}}_{\boldsymbol{i}}$ is the predicted values.

3.5.3. R Square

The coefficient of determination, R^2, is a measure to goodness of fit of a model. In the context of regression, it is a statistical measure of how well the regression line approximates the actual data.

$$R^2 = 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}}$$

4. RESULTS AND DISCUSSIONS

4.1. Simple linear regression

Table 1: Evaluation Criteria for SLR

	Mean Absolute Error	*Mean Squared Error*	*Rsquare*
E	7.175	117.595	0.023
S	8.458	132.334	0.003
G	5.207	41.57	0.132

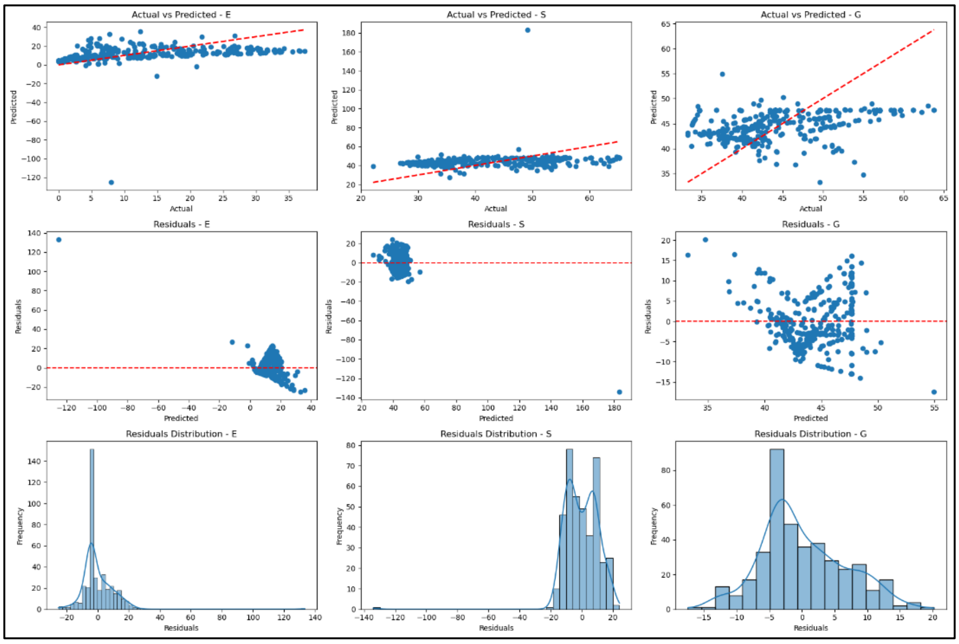

Fig 3: Graphical representation SLR (Actual vs predicted, Residuals and Error Distribution)

The Simple Linear Regression model shows poor performance across all scenarios (E, S, G) with significant prediction errors and residuals. Scenario E and S have more pronounced issues with skewed and bimodal residual distributions, respectively, while scenario G shows a more balanced error distribution but still with significant prediction inaccuracies. Table 1 clearly explains SLR is poor fit due to least r square and high errors.

4.2. Combined Actual vs. Predicted Plots

Let's analyze these combined actual vs. predicted plots for the three models (Random Forest, Gradient Boosting, and Support Vector Regressor) across the three targets (E, S, G).

Figure 4 clearly shows Gradient Boosting and Support Vector Regression models generally show better performance, with points more closely aligned with the diagonal line across different scenarios (S and G in particular). Random Forest shows slightly more spread in its predictions, indicating more variance in its accuracy, especially in the E scenario. Scenarios S and G seem to yield more consistent predictions across all models compared to scenario E, which shows more deviation in the plots.

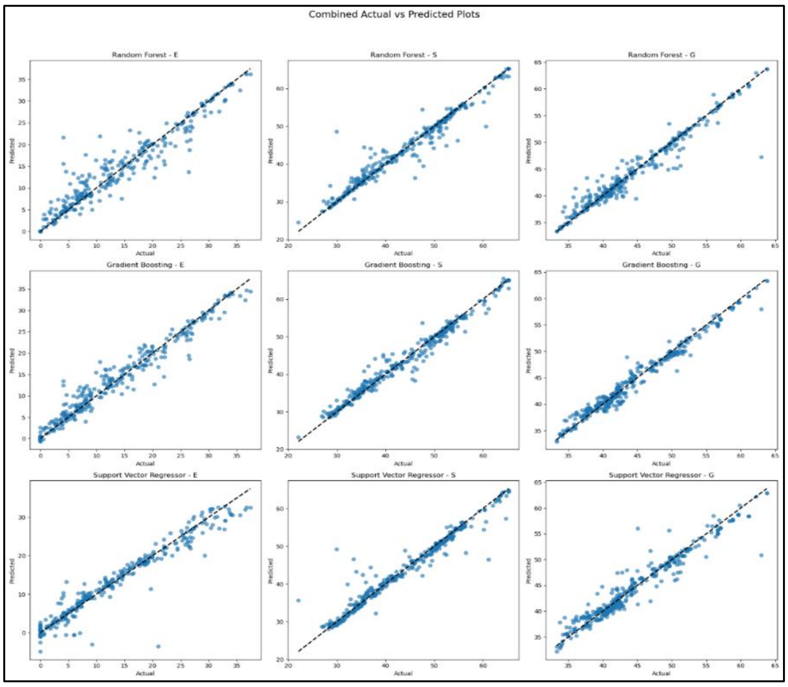

Fig 4: Actual vs predicted across models

4.3. Error Distribution Histograms

These histograms show the distribution of residuals (errors) for each model and target.Ideally, residuals should be normally distributed around zero.

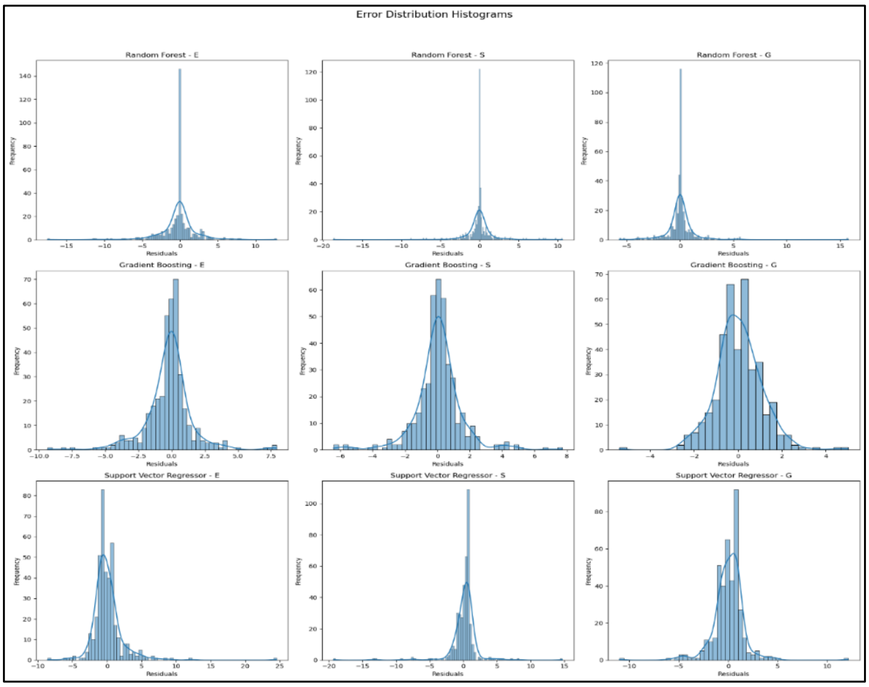

Fig 5: Residual error graphs across models

Figure 5 clearly explains that Random Forest and Support Vector Regression models generally show tightly centered error distributions around zero, indicating high accuracy. Gradient Boosting shows a more normally distributed error pattern, which may indicate a more consistent but slightly less accurate prediction performance compared to the other models.

4.4. Residual Plots

These plots show the residuals (errors) vs. the predicted values. Ideally, residuals should be randomly scattered around zero without any discernible pattern.

Figure 6 stated that Gradient Boosting and Support Vector Regression models generally show tighter clustering of residuals around the zero line, indicating higher accuracy and less variance in predictions. Random Forest shows more spread in residuals, particularly in scenario E, indicating more variability in its predictions.

4.5. Learning Curves

The Gradient Boosting and Support Vector Regression models generally show better performance with training and cross-validation scores closely aligned, indicating good generalization. Figure 7 clearly shows that the cross-validation scores improve significantly and get closer to the training scores as more data is added, reducing overfitting. This model shows higher variance in its accuracy, with consistently high training scores indicating overfitting. The cross-validation scores, while improving with more data, remain consistently lower than the training scores, especially in the E scenario.

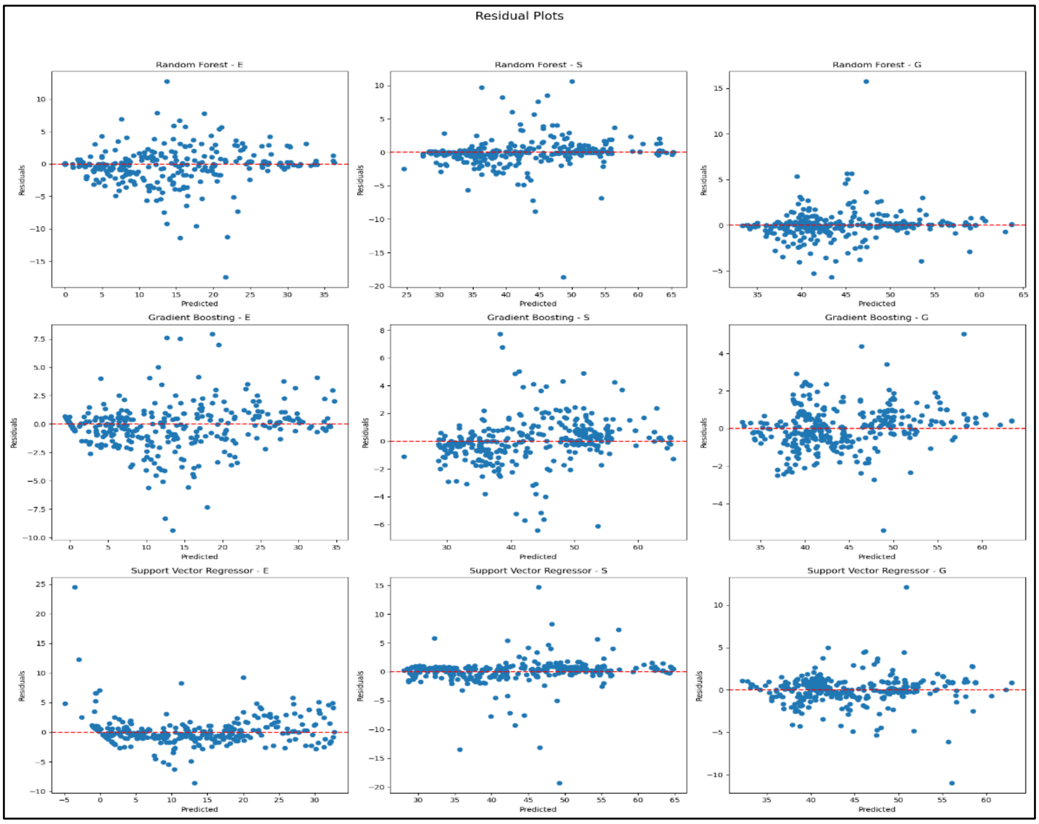

Fig 6: Residual error graphs across models

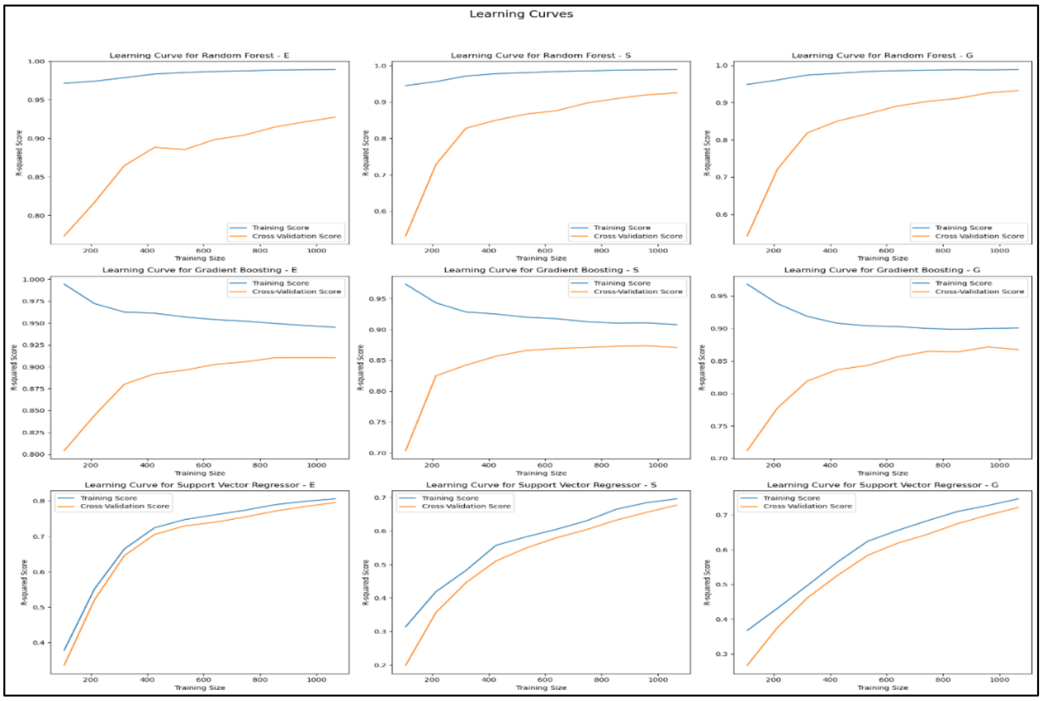

Fig 7: Learning Curves across models

4.6. Comparison of models

Table 2: Evaluation Criteria across models

ML Models		*Mean Absolute Error*	*Mean Squared Error*	*Rsquared*
Random Forest	***E***	1.334	6.114	0.942
	S	0.881	3.832	0.961
	G	0.677	2.061	0.957
Gradient Boosting	***E***	**1.144**	**3.336**	**0.968**
	S	**0.951**	**2.265**	**0.976**
	G	**0.803**	**1.157**	**0.975**
Support Vector Regressor	***E***	1.361	5.327	0.950
	S	1.070	4.596	0.953
	G	1.039	2.561	0.946

5. MODEL PRIORITIZATION

5.1. Gradient Boosting

- **Learning Curves**: Shows a good balance between training and cross-validation scores, indicating it generalizes well with more data. The cross-validation scores improve significantly and approach the training scores, reducing overfitting.
- **Actual vs Predicted Plots**: Points closely align with the diagonal line, indicating strong prediction accuracy.
- **Residuals Plots**: Residuals are more evenly distributed around zero with minimal spread, indicating low prediction errors.
- **Residuals Distribution Histograms**: Residuals are normally distributed, suggesting the model does not have significant biases and performs consistently well across different scenarios.

5.2. Support Vector Regressor

- **Learning Curves**: Demonstrates a steady improvement in both training and cross-validation scores with more data. The narrow gap between these scores indicates good generalization capabilities.
- **Actual vs Predicted Plots**: Points align well with the diagonal line, showing good prediction accuracy.
- **Residuals Plots:** Residuals are fairly evenly distributed around zero, with some minor spread indicating slight prediction errors.
- **Residuals Distribution Histograms**: Residuals are fairly normally distributed, showing the model's consistent performance and balanced error distribution.

5.3. Random Forest

- **Learning Curves:** Training scores remain high, but cross-validation scores are lower, indicating overfitting. Despite some improvement in cross-validation scores with more data, the consistent high training scores suggest overfitting issues.
- **Actual vs Predicted Plots:** Points show more spread from the diagonal line, indicating higher variance in prediction accuracy, especially in scenario E.
- **Residuals Plots:** Residuals show larger spreads and patterns, indicating inconsistent predictions and overfitting.

- **Residuals Distribution Histograms:** Residuals show more spread and skewness, indicating biases in predictions and less consistent performance compared to Gradient Boosting and SVR.

5.4. Simple Linear Regression

- **Learning Curves:** Shows poor overall prediction accuracy with significant residual spread and underprediction issues. The simplicity of the model makes it less suitable for complex prediction tasks.
- **Actual vs Predicted Plots:** Points show significant deviation from the diagonal line, indicating poor prediction accuracy.
- **Residuals Plots:** Residuals show significant spread and patterns, indicating poor model performance and underprediction issues.
- **Residuals Distribution Histograms:** Residuals are skewed and show large spread, confirming the model's tendency to underpredict and its lower performance metrics.

6. RECOMMENDATIONS

Based on the analysis of learning curves, actual vs predicted plots, residuals plots, and residuals distribution histograms, and from Table 2, the least MAE and MSE errors and high r-squared values explain that, **Gradient Boosting** is the top priority model due to its strong generalization capabilities, minimal overfitting, and consistent performance. **Support Vector Regressor (SVR)** is the second choice, offering steady performance improvements and good generalization capabilities. **Random Forest** is ranked third due to its tendency to overfit despite high training scores. Finally, **Simple Linear Regression (SLR)** is the least preferred model due to its simplicity and poor prediction accuracy. For robust and reliable predictions, prioritizing Gradient Boosting and SVR models is recommended.

7. CONCLUSION

The TRST01 ESG Scoring Model is a groundbreaking tool that sets a new standard in sustainability assessments. By integrating financial metrics with ESG evaluations, it offers a reliable, accurate, and comprehensive framework for understanding corporate sustainability. The model's ongoing evolution and expansion promise to drive significant advancements in sustainable business practices. It will be an indispensable resource for those committed to integrating sustainability into the core of their strategies.

8. ACKNOWLEDGEMENT

The author would like to thank IIM Ahmedabad for the support provided during the MBA program and TRST01 company for the opportunity to undertake this research during the internship. This research was conducted as part of my MBA program and internship at TRST01, where I focused on ESG scoring model. The insights gained during the internship provided a practical perspective that enriched the academic study.

REFERENCES

1. Alanis, E. (2022). Forecasting betas with random forests. *Applied Economics Letters*, *29*(12), 1134–1138. https://doi.org/10.1080/13504851.2021.1912278
2. Alexandridis, A. K., & Hasan, M. S. (2020). Global financial crisis and multiscale systematic risk: Evidence from selected European stock markets. *International Journal of Finance and Economics*, *25*(4), 518–546. https://doi.org/10.1002/ijfe.1764
3. Aronne, A., Grossi, L., & Bressan, A. A. (2020). Identifying outliers in asset pricing data with a new weighted forward search estimator. *Revista Contabilidade e Financas*, *31*(84), 458–472. https://doi.org/10.1590/1808-057X201909620
4. Avramov, D., Cheng, S., Lioui, A., & Tarelli, A. (2022). Sustainable investing with ESG rating uncertainty. *Journal of Financial Economics*, *145*(2), 642–664. https://doi.org/10.1016/j.jfineco.2021.09.009
5. Bali, T. G., Brown, S. J., Murray, S., & Tang, Y. (2017). A lottery-demand-based explanation of the beta anomaly. *Journal of Financial and Quantitative Analysis*, *52*(6), 2369–2397. https://doi.org/10.1017/S0022109017000928
6. Bangdiwala, S. I. (2018). Regression: simple linear. International journal of injury control and safety promotion, 25(1), 113-115.

7. Chen, Q., & Liu, X. Y. (2020, October 15). Quantifying ESG alpha using scholar big data: An automated machine learning approach. *ICAIF 2020 - 1st ACM International Conference on AI in Finance*. https://doi.org/10.1145/3383455.3422529
8. Czerwińska, T., & Kaźmierkiewicz, P. (2015). ESG Rating in Investment Risk Analysis of Companies Listed on the Public Market in Poland. *Economic Notes*, *44*(2), 211–248. https://doi.org/10.1111/ecno.12031
9. 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
10. Giese, G., Lee, L.-E., Melas, D., Nagy, Z., & Nishikawa, L. (2019). *The Journal of Portfolio Management Foundations of ESG Investing: How ESG Affects Equity Valuation, Risk, and Performance*. http://www.msci.com/prod-
11. Jin, I. (2018). Is ESG a systematic risk factor for US equity mutual funds? *Journal of Sustainable Finance and Investment*, *8*(1), 72–93. https://doi.org/10.1080/20430795.2017.1395251
12. Jo, H., & Na, H. (2012). Does CSR Reduce Firm Risk? Evidence from Controversial Industry Sectors. *Journal of Business Ethics*, *110*(4), 441–456. https://doi.org/10.1007/s10551-012-1492-2
13. Martín-Cervantes, P. A., & Valls Martínez, M. del C. (2023). Unraveling the relationship between betas and ESG scores through the Random Forests methodology. *Risk Management*, *25*(3). https://doi.org/10.1057/s41283-023-00121-5
14. Momparler, A., Carmona, P., & Climent, F. (2024). Catalyzing Sustainable Investment: Revealing ESG Power in Predicting Fund Performance with Machine Learning. *Computational Economics*. https://doi.org/10.1007/s10614-024-10618-0
15. Sassen, R., Hinze, A. K., & Hardeck, I. (2016). Impact of ESG factors on firm risk in Europe. *Journal of Business Economics*, *86*(8), 867–904. https://doi.org/10.1007/s11573-016-0819-3

Cite this Article: Gurucharan Kottapalli, Prabir Mishra (2024). Advanced TRST01 ESG Scoring Model with Beta Based Financial Metrics and Machine Learning Techniques. International Journal of Current Science Research and Review, 7(6), 3598-3611

3611 ***Corresponding Author: Gurucharan Kottapalli**

Volume 07 Issue 06 June 2024

Available at: www.ijcsrr.org

Page No. 3598-3611