## 20.1 INTRODUCTION

Corporate social responsibility (CSR) and environmental, social, and governance (ESG) practices are the keywords in today's business world due to the growing awareness of the stakeholders. These play a pivotal role in shaping the identity of the business and providing it a vision to be operationally and strategically empowered, along with enabling it to create its impact on society (Sidhoum and Serra, 2018). As awareness and influence of stakeholders like investors, consumers, regulators, and customers grow, the need for predictive methodologies to assess trends and performance in CSR initiatives becomes inevitably indispensable (Xue et al., 2022; Croker and Barnes, 2017). Corporate social responsibility (CSR) has become key for the business community, and results show a fairly strong positive relationship between economic, social, and environmental performance, indicating that profitable business is compatible with balanced sustainability (Sidhoum and Serra, 2018).

This research is aimed at exploring the interjection of CSR, ESG, and the use of predictive modeling within the context of Indian companies. India being one of the fastest-growing economies of the world and a trend-setter in mandating CSR presents a compelling backdrop for such an analysis. As companies in India steer through social, environmental, and governance challenges, it becomes imperative to understand and quantify their CSR initiatives through data-driven approaches to gain insights into their efforts toward sustainability (Singh et al., 2018).

The significance of ESG data in assessing CSR performance cannot be overstated (Mervelskemper and Streit, 2016). With the availability of ESG data of Indian companies, machine learning techniques can be used to leverage the CSR outcomes. The current research intends to inquire about and understand the intricate relationship between ESG scores, their associated metrics, and CSR expenditures. Data preprocessing is used to prepare a foundation for predictive modeling, addressing missing values and outliers to ensure data quality has been done. Two formidable predictive models, namely, linear regression and random forest regression, have been used for analysis. These models are tasked to assess CSR performance based on ESG scores as it has significant implications for not only stakeholders but also the business community in general.

Through this research, endeavor has been made to unravel the statistical characteristics of ESG scores, discuss the strengths and weaknesses of predictive modeling techniques, and offer a detailed evaluation of CSR initiatives within Indian companies.

## 20.2 METHODOLOGY

The objective of this study is to develop a predictive model that can accurately predict CSR scores based on a set of ESG attributes. This methodology outlines the steps taken to preprocess the data, explore its characteristics, and train machine learning models to achieve this goal. For this, a 104-row and 34-column ESG dataset on Kaggle (https://www.kaggle.com/datasets/thomsnowflake/esgdata), which has data of various anonymized companies from 2010 to 2022, and CRISIL ESG Score 2022 using CRISIL's proprietary ESG methodology to score 586 companies across 53 sectors based on publicly available information (https://www.crisil.com/en/home/what-we-do/financial-products/crisils-sustainability-solutions/esg-score-2022.html) are used. Further, a dataset of the top 50 Indian companies in FY 2022 by market capitalization is chosen for practical validation of the developed machine learning models.

The first step is to remove all null values from both the datasets. We then perform PCA on standardized numerical columns, generate a plot to see cumulative variance explained by principal components, and then explore PCA at 0.8 cumulative variance and see a correlation in principal components. However, no correlation is found at all, indicating principal components may not capture strong linear relationships.

Correlation between features is explored. The selected independent variables are integration/vision and strategy (inactive), ESG score, human rights score, shareholders score, resource use score, management score, society/human rights (inactive), product responsibility score, resource reduction (inactive), and ESG combined score. A linear regression and a random forest are trained models to test predictions with a correlation threshold of 0.45. They are then serialized and saved as pkl files for future use. The complete codebase can be found in the authors' GitHub repository (https://github.com/Rajkanwars15/res-ram).

## 20.3 DATA DISTRIBUTIONS

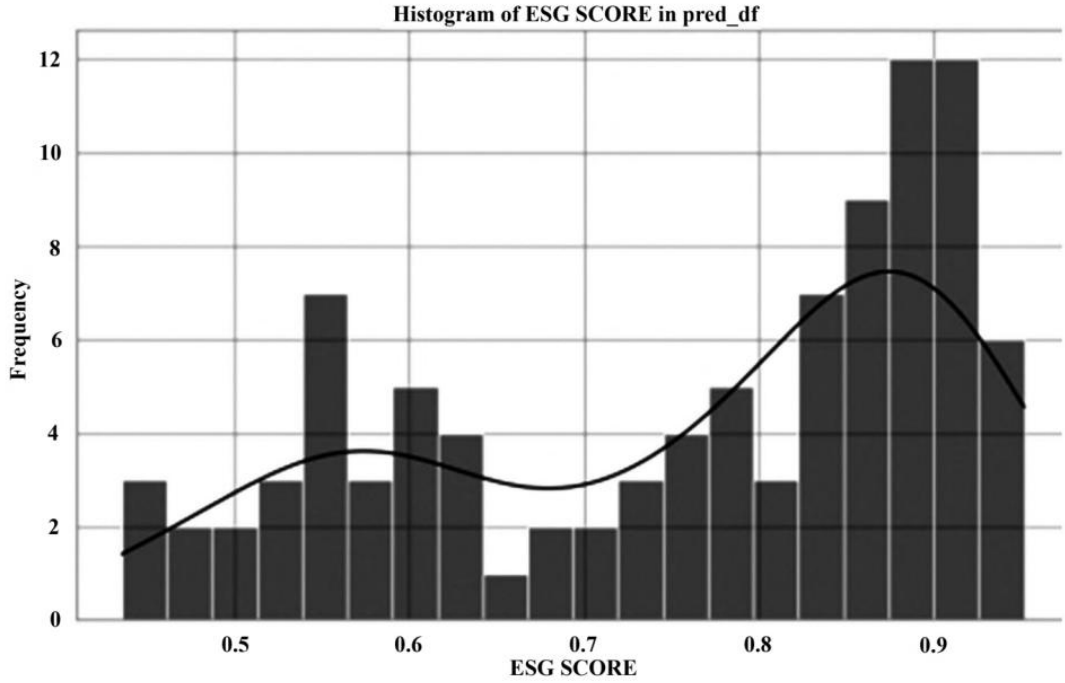


**FIGURE 20.1** Histogram of ESG scores in pred_df.

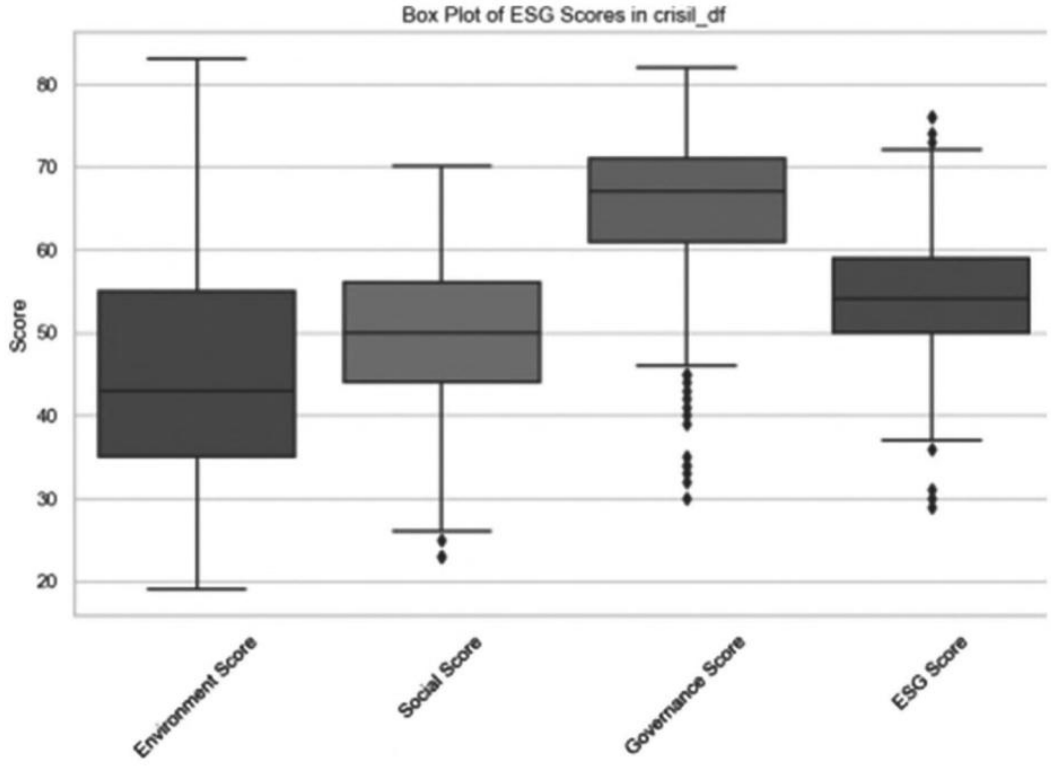


**FIGURE 20.2** Box plot of ESG scores in crisil_df.

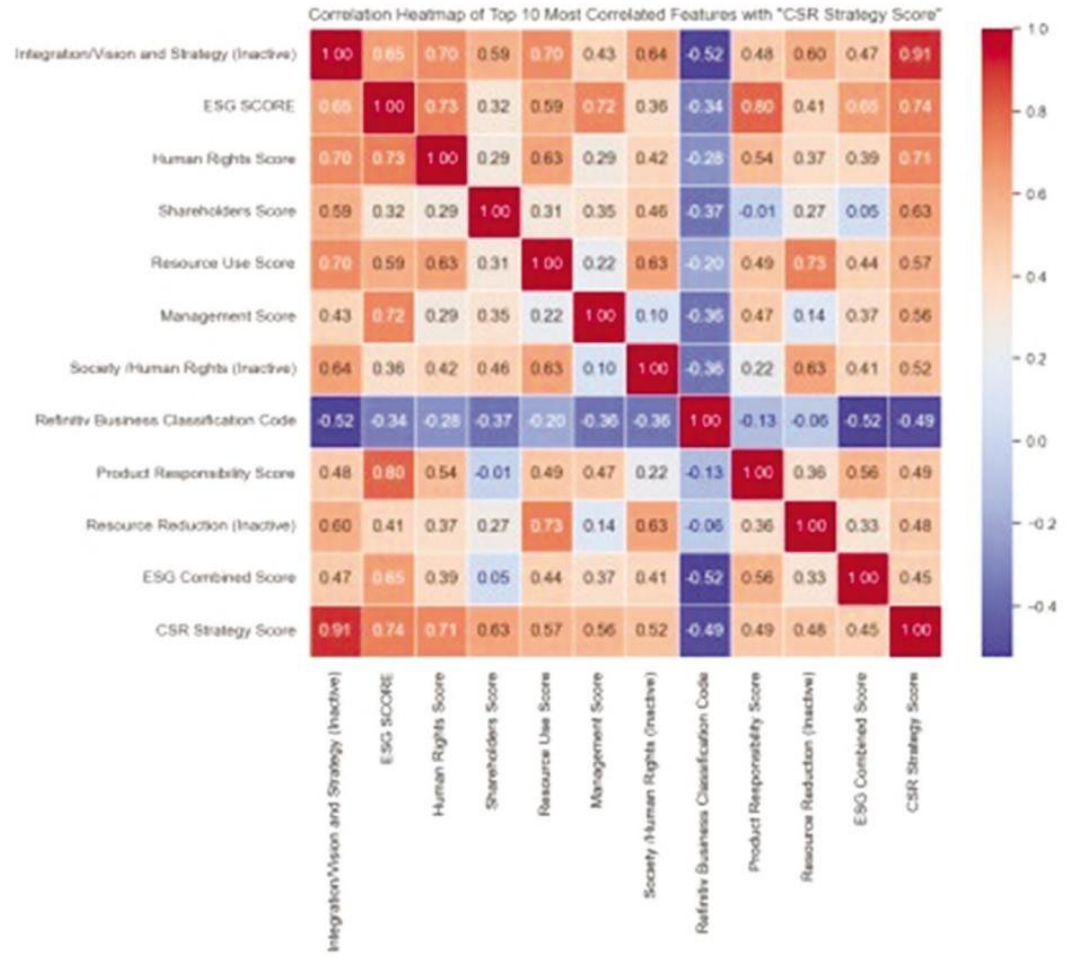


**FIGURE 20.3** Correlation heatmap of the top 10 most correlated features with CSR strategy score.

## 20.4 RESULTS AND DISCUSSION

Principal components show no correlation at all. Thus, the 10 most correlated factors to the CSR score are selected. On these factors, a linear regression model and a random forest model are trained.

| | Mean Squared Error (MSE) | R-squared (R2) | Average MSE | SD of MSE |
|---|---|---|---|---|
| Linear Regression | 0.0062 | 0.9011 | 0.0161 | 0.0059 |
| Random Forest | 0.0046 | 0.9265 | 0.0144 | 0.0077 |

Both the linear regression and random forest models perform reasonably well in predicting the target variable. This is evident from their relatively

low MSE values, indicating that the models have a small average squared error when making predictions. The R-squared (R2) values for both models are quite high (0.9011 for linear regression and 0.9265 for random forest). These values suggest that a significant portion of the variance in the target variable is explained by the independent variables in the models. The models capture a substantial amount of the underlying data patterns. Crossvalidation of the models is done by assessing the average and standard deviation of their MSE. In both cases, both the average and standard deviation are quite low and rule out chances of overfitting.

When comparing the two models, the random forest model outperforms the linear regression model in terms of both MSE and R2. However, when making actual predictions on the CRISIL data, the random forest model gives a constant and unrealistic value of 0.9151, while the linear regression model varies in the range of 49.56–100.50.

This led to dropping the random forest model as a viable option. The reason behind this provides an opportunity for further probing in future studies.

## 20.5 REAL-WORLD EVALUATION

We now observe a correlation between our predicted CSR score and custom score based on CSR spending, actual and required, of the top 50 companies listed on the CSR Journal. The adopted formula for the custom score is given as follows:

IF(Required=0, 0, IF(Actual>=0, MIN(100, (Actual - Required) / Required*100), - MIN(100, (Required - Actual) / Actual * 100)))(20.1)

Here, it was observed that the top 5 ranking companies by custom score, namely, Reliance Industries Limited, Tata Consultancy Services Limited, HDFC Bank Limited, Infosys Limited, and Hindustan Unilever Limited, had exceeded the minimum spending requirement of 2% in the range of 1.5–18% with the exception of Infosys. However, its presence in the top 5 is due to its overall higher base of spending on CSR. Here, the linear regression model placed Reliance Industries Limited quite low on rank 8, which is explained due to comparatively lower values of the ESG score and associated values. Another interesting pair of data points is for Vedanta Limited, where both linear regression and custom score place it on ranks 26 and 22, respectively, despite 100% more CSR spending than mandated due to an overall smaller base of mandated spend, and Adani Power Limited on ranks 25 and 21,

which also spent 15 lakhs on CSR while it was not required to spend any amount due to reported losses. Bajaj Auto Limited ranks 27 on the custom list since it spent 40% less than the mandated value (among the lowest) but is ranked comparatively high at 15 by linear regression due to reasonable ESG values; this behavior by a company is counter-intuitive in general. Most other companies performed quite similarly on both ESG and CSR spending.

The correlation between linearregression rank and custom score rank is 0.5.

The correlation between linearregression rank and custom score with no negative custom score rank is 0.58.

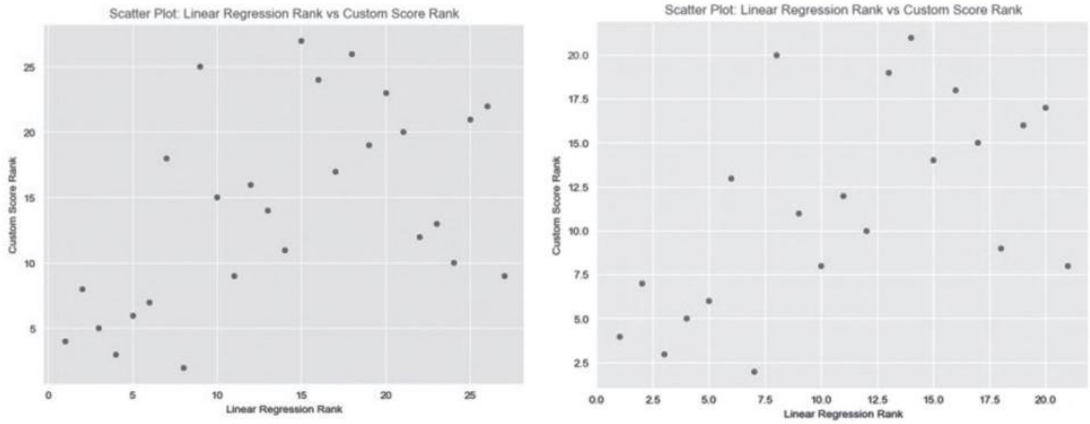


**FIGURE 20.4** Scatter plot: linear regression rank vs custom score rank.

## 20.6 CONCLUSIONS

In overview, this research represents a groundbreaking initiative directed at simplifying the prediction process for CSR spending by Indian corporations. By effectively addressing the complexities linked to ESG score and certain other factors that affect CSR performance, a comprehensive predictive system is crafted that empowers the stakeholders/researchers to set informed expectations from corporations.

To ensure the caliber and reliability of the dataset and model used in the research, data preprocessing was carried out, encompassing the removal of null values and selecting the most correlated factors. The finely tuned model, utilizing linear regression and random forest, customized for multiple labels, has showcased robust predictive capabilities. Demonstrating an impressive mean squared error of just 0.0062 and 0.0046 and a remarkably high R-squared (R2) of 0.9011 and 0.9265, respectively, the model adeptly anticipates the CSR performance of various Indian companies.

Importantly, this study goes beyond theoretical realms of assessment by predicting and comparing real-world values, thereby bridging the gap between data-driven insights and practical application. This also allowed us to assess and not continue with our faulty random forest model, reasons for which can be evaluated and are beyond overfitting, which was ruled out by crossvalidation. As stakeholders, investors, and businesses continue to navigate the complex landscape of CSR, this research serves as a testament to the potential of data-driven methodologies in enhancing the understanding of corporate social responsibility, not only in India but also as a broader paradigm for assessing CSR worldwide. It is safe to say that in the case of most companies, the better the company performs in ESG and associated activities, the higher will be its commitment to CSR spending.

It is prudent to shed light on the intriguing dynamics between custom scores, linear regression, and CSR spending among the top-ranked companies. Notably, the top-performing companies, such as those mentioned, demonstrate a proactive commitment to CSR by exceeding the mandated spending threshold, showcasing their dedication to social responsibility. The linear regression model introduces an interesting dimension to the ranking. It places some companies lower in the ranking due to their ESG scores and associated values, revealing anomalies in CSR performance that may not be immediately evident from CSR spending alone. This underscores the value of incorporating data-driven approaches to CSR assessment.

## KEYWORDS

- **ESG**
- **corporate social responsibility**
- **machine learning**
- **linear regression**

## REFERENCES

Croker, N.; Barnes, L. Epistemological Development of Corporate Social Responsibility: The Evolution Continues. *Soc. Respons. J.* **2017,** *13*, 279–291. https://doi.org/10.1108/SRJ-02-2016-0029.

ESG & CSR—Two Sides of a Coin. *The Times of India* (n.d.). https://timesofindia.indiatimes.com/readersblog/ethical-encouter/esg-csr-two- sides-of-a-coin-51923 (accessed on 30 Sept 2023).

ESG vs. CSR: Key Distinctions & What Businesses Need to Know. USA (n.d.). https://us.anteagroup.com/news-events/blog/esg-csr-definitions-differences-sustainability

Gillan, S. L.; Koch, A.; Starks, L. T. Firms and Social Responsibility: A Review of ESG and CSR Research in Corporate Finance. *J. Corp. Finance* **2021,** *66*, 101889. https://doi.org/10.1016/j.jcorpfin.2021.101889

Mervelskemper, L.; Streit, D. Enhancing Market Valuation of ESG Performance: Is Integrated Reporting Keeping its Promise? *Bus. Strat. Environ.* **2016,** *26*, 536–549. https://doi.org/10.2139/SSRN.2625044.

Shafranovski, A. The Evolution from CSR to ESG; ICL, 2022. https://www.icl-group.com/blog/the-evolution-from-csr-to-esg

Sidhoum, A.; Serra, T. Corporate Sustainable Development. Revisiting the Relationship Between Corporate Social Responsibility Dimensions. *Sustain. Dev.* **2018,** *26*, 365–378. https://doi.org/10.1002/SD.1711.

Singh, S.; Holvoet, N.; Pandey, V. Bridging Sustainability and Corporate Social Responsibility: Culture of Monitoring and Evaluation of CSR Initiatives in India. *Sustainability* **2018**. https://doi.org/10.3390/SU10072353.

Tsai, H.; Wu, Y. Changes in Corporate Social Responsibility and Stock Performance. *J. Bus. Ethics* **2021,** *178*, 735–755. https://doi.org/10.1007/s10551- 021-04772-w.

Xue, Y.; Jiang, C.; Guo, Y.; Liu, J.; Wu, H.; Hao, Y. Corporate Social Responsibility and High-quality Development: Do Green Innovation, Environmental Investment and Corporate Governance Matter? *Emerg. Markets Finance Trade* **2022,** *58*, 3191–3214. https://doi.org/10.1080/1540496X.2022.2034616.