ESG Risk Analysis and Prediction Project Report

1. Introduction and Objectives

This project focuses on a comprehensive analysis and prediction of ESG (Environmental, Social, and Governance) risk scores using the "SP 500 ESG Risk Ratings" dataset. The primary objectives were to perform a thorough exploratory data analysis, clean and preprocess the data, and build a predictive model to forecast the total_esg_risk_score. A key goal was to demonstrate a complete data science workflow, from raw data to actionable insights and a robust predictive model.

2. Data Loading and Initial Setup

The project begins with the installation of necessary Python libraries, including seaborn, pandas, numpy, and matplotlib, as well as sqlalchemy for database connectivity. The dataset was initially loaded from a CSV file into a pandas DataFrame df_kaggle. The initial dataset contained 503 entries and 15 columns. A cleaned version of the data was subsequently pushed to a MySQL database named esg_analysis into a table named kaggle_esg_risk and then read back into a DataFrame df for further analysis.

3. Exploratory Data Analysis (EDA) and Data Cleaning

Initial Data Inspection

The df_kaggle.info() and df_kaggle.describe() commands were used to inspect the data types, non-null counts, and key statistical measures of the columns. The dataset was found to have 5 float64 and 10 object data types.

Missing Value Handling

Missing values were a significant issue, particularly in the risk score columns. Controversy Score had 100 missing values, while Total ESG Risk score, Environment Risk Score, Governance Risk Score, Social Risk Score, Controversy Level, ESG Risk Percentile, and ESG Risk Level each had 73 missing values.

- The Total ESG Risk score, Environment Risk Score, Governance Risk Score, and Social Risk Score columns were imputed using the **mean** of their respective columns.
- The controversy_score column was imputed using the **median** value.

- Rows with remaining missing values in controversy_level, esg_risk_level, and esg_risk_percentile were dropped, resulting in a final dataset of 430 rows.

Feature Distributions

Histograms with KDE plots were generated to analyze the distribution of numerical risk scores. The mean and median values were overlaid on the plots.

- total_esg_risk_score had a mean of 21.53 and a median of 21.05 with a skew of 0.44.
- environment_risk_score had a mean of 5.74 and a median of 4.05 with a skew of 1.08.
- governance_risk_score had a mean of 6.73 and a median of 6.10 with a skew of 1.35.
- social_risk_score had a mean of 9.07 and a median of 8.90 with a skew of 0.37.
- controversy_score had a mean of 2.01 and a median of 2.00 with a skew of 0.58.

Outlier Analysis

The Interquartile Range (IQR) method was used to detect outliers in the risk score columns. Box plots were created to visualize these outliers.

- total_esg_risk_score: 3 outliers.
- environment_risk_score: 8 outliers.
- governance_risk_score: 20 outliers.
- social_risk_score: 4 outliers.

Correlation Analysis

A correlation matrix and heatmap were generated to visualize the relationships between numerical risk scores.

- total_esg_risk_score showed strong positive correlations with environment_risk_score (0.70) and social_risk_score (0.69).
- A weaker positive correlation was observed between total_esg_risk_score and governance_risk_score (0.35) and controversy_score (0.35).

Component Risk Heatmap by Industry

A heatmap of average ESG component scores per industry was generated to reveal variations in risk profiles.

4. Feature Engineering and Preprocessing

Several new features were engineered to provide more context and predictive power to the model. These include:

- **Outlier Flags:** Binary flags (env_outlier, gov_outlier, soc_outlier) were created to mark companies with risk scores identified as outliers, indicating unusual risk profiles.
- **High-Risk Flags:** controversy_high was created for companies with a controversy_score >= 4.
- **Anomaly Case Flag:** anomaly_case was created to identify companies with a low total_esg_risk_score (< 20) but a high controversy_score (>= 4).
- **Risk Ratios:** Features like env_to_soc_ratio and gov_to_total_ratio were created to capture the relationship between different risk scores.
- **Risk Component Products:** controversy_social_product and controversy_env_product were created by multiplying controversy_score with social_risk_score and environment_risk_score, respectively. These features are designed to capture the interaction between controversy and a specific risk dimension.
- **Controversy Anomaly Score:** controversy_anomaly_score was created to flag cases where a company has a high controversy score (> 3.5) but a low total ESG risk score (< 20).
- **High-Risk Sector Flag:** high_risk_sector_flag was created as a binary flag for sectors identified as having high risk, such as 'Energy', 'Utilities', and 'Basic Materials'.
- **Categorical Encoding:** controversy_level was ordinally encoded based on a custom order of severity. sector was one-hot encoded using pd.get_dummies with drop_first=True. industry was frequency-encoded into industry_encoded.
- **Dimensionality Reduction: Principal Component Analysis (PCA)** was performed after scaling numerical features with StandardScaler. A scree plot showed that the first two components explained the most variance.
- **Clustering: K-Means** clustering was performed on the first two PCA components, and the optimal number of clusters was determined to be 3 using the Silhouette Score. The cluster assignments were added as a new feature esg_cluster.

5. Modeling and Evaluation

Model Selection

XGBoost Regressor (XGBRegressor) was chosen as the primary model for its robustness and performance. A Random Forest Regressor was also trained for comparison.

Data Splitting

The data was split into a training set and a test set (80/20) with a random_state of 42 to ensure reproducibility. The total_esg_risk_score was designated as the target variable y, and other features were used as the input X.

Final Model Performance

The XGBoost model's performance was evaluated on the test set. It achieved an **R^2 Score of 0.8685** and a **Mean Absolute Error (MAE) of 1.8064**. This demonstrates that the model is highly accurate in predicting the total ESG risk score.

Feature Importance

A feature importance plot from XGBoost was generated to identify the most influential features. The plot showed that controversy_social_product, governance_deficit_ratio, and controversy_env_product were among the most important features for prediction.

6. Business and ESG Metrics Observations

Top Risky Companies and Sectors

The analysis identified and visualized the top 10 companies with the highest ESG risk scores. The **Energy sector** was found to have the highest average ESG risk score (32.34), followed by Basic Materials (26.72) and Utilities (26.71). This suggests that sectors involved in resource extraction and energy production generally face greater ESG-related challenges.

Most Controversial Sectors

Financial Services and Industrials had the highest count of companies with a high controversy score (>=4). This indicates that, while not necessarily having the highest average total ESG risk, these sectors are prone to frequent, significant controversies.

Correlation and Inter-Component Relationships

controversy_score was most strongly correlated with social_risk_score (0.40). This observation is a key finding, suggesting that company controversies are most often tied to social-related issues. The negative correlation of -0.22 between environment_risk_score and governance_risk_score suggests a potential trade-off or differing focus within companies.

Anomaly Detection

The project explicitly identified companies that could be considered anomalies. A scatter plot was used to visualize companies with a low total_esg_risk_score (< 20) but a high controversy_score (>=(>=4). These companies, predominantly in the Financial Services sector, represent potential red flags for analysts, as their overall risk score may not capture their high-controversy profile

Component Risk Analysis

A visualization of average social versus governance scores by sector highlighted differences in risk profiles. Healthcare and Financial Services had high average governance scores, while Consumer Defensive had a particularly high average social score.

Risk Buckets

The analysis grouped companies into risk buckets (Low, Medium, High) based on their total_esg_risk_score. The average ESG component scores within these buckets showed that High risk companies have a notably higher average environment_risk_score (13.6) and social_risk_score (12.4), while their governance_risk_score (7.6) is also high but to a lesser extent. This suggests that high-risk classifications are primarily driven by environmental and social factors.

Engineered Features' Importance

The final XGBoost model's feature importance analysis revealed that several engineered features were highly predictive. The high importance of features like controversy_social_product and governance_deficit_ratio confirms the strong connections between social issues, controversies, and total risk.

7.Conclusions and Future Work

The project successfully demonstrated a complete data science workflow, from data acquisition and cleaning to sophisticated feature engineering and predictive modeling. The XGBoost model performed exceptionally well in predicting the total ESG risk score based on a rich set of engineered features. The analysis also provided valuable insights into industry-level ESG performance, risk correlations, and specific anomaly cases.