PREDICTING MARKET VALUATION AND CATEGORY TIERS USING ESG & FINANCIAL DATA

Presented by
Anamika Sharma RBA26
Palak Sahu RBA51
Sourav Manna RBA70

DATASET

This dataset simulates the financial and ESG (Environmental, Social, and Governance) performance of 1,000 global companies across 9 industries and 7 regions from 2015 to 2025. It contains realistic financial metrics (e.g., revenue, profit margins, market capitalization) alongside comprehensive ESG indicators, including carbon emissions, resource usage, and detailed ESG scores.

Dataset Link:- https://www.kaggle.com/datasets/shriyashjagtap/esg-and-financial-performance-dataset?resource=download

OBJECTIVES

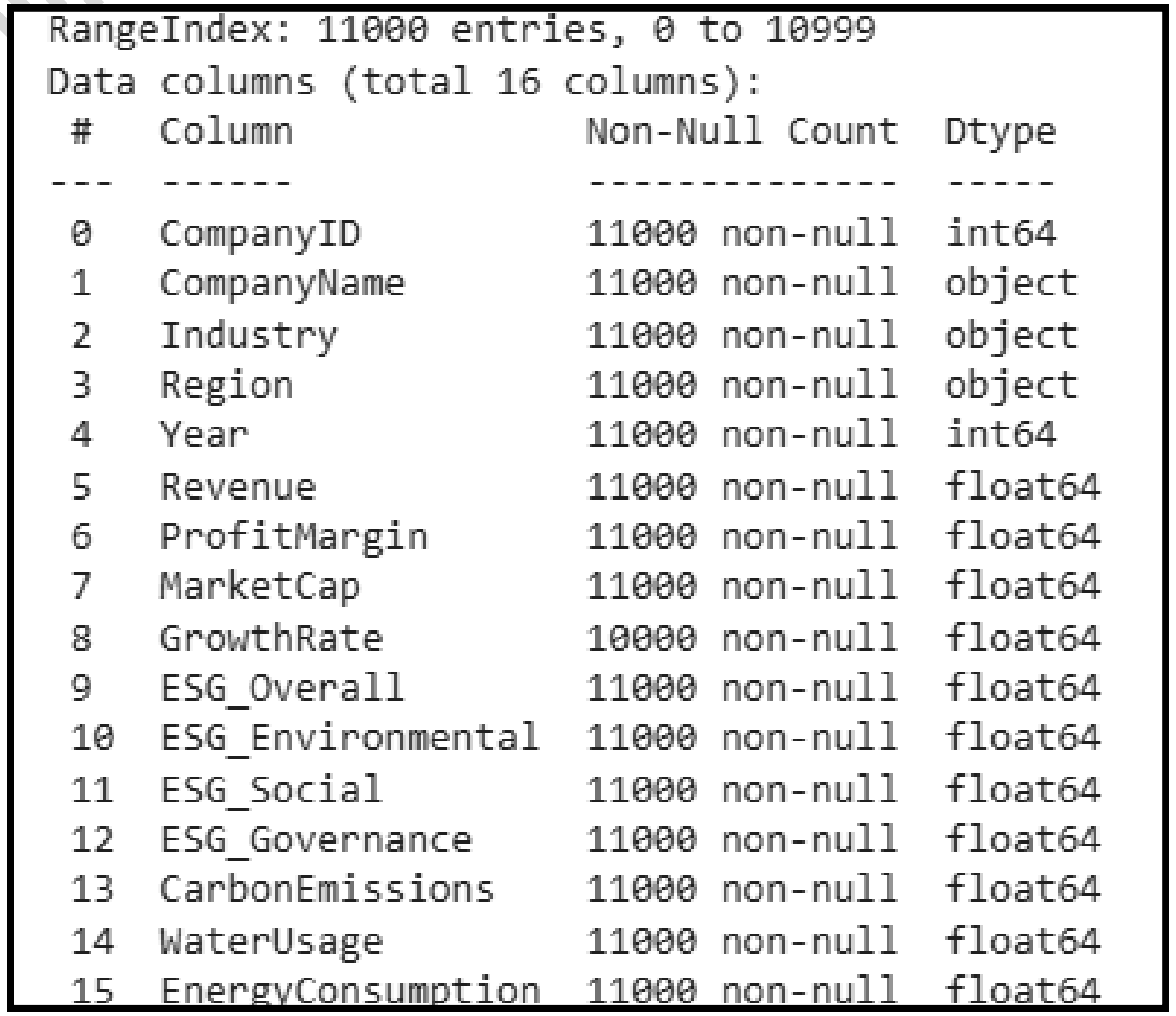
- Objective 1: Regression (Prediction)

Task: Predicting the continuous numerical value of Market Capitalization.

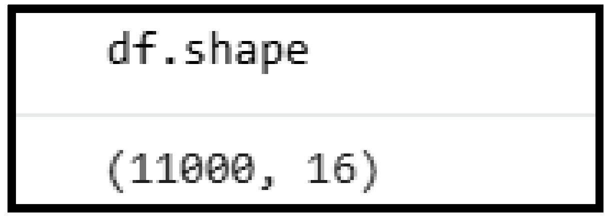
- Objective 2: Classification (Categorization)

Task: Classifying companies into Small, Mid, and Large-Cap tiers.

DATASET OVERVIEW

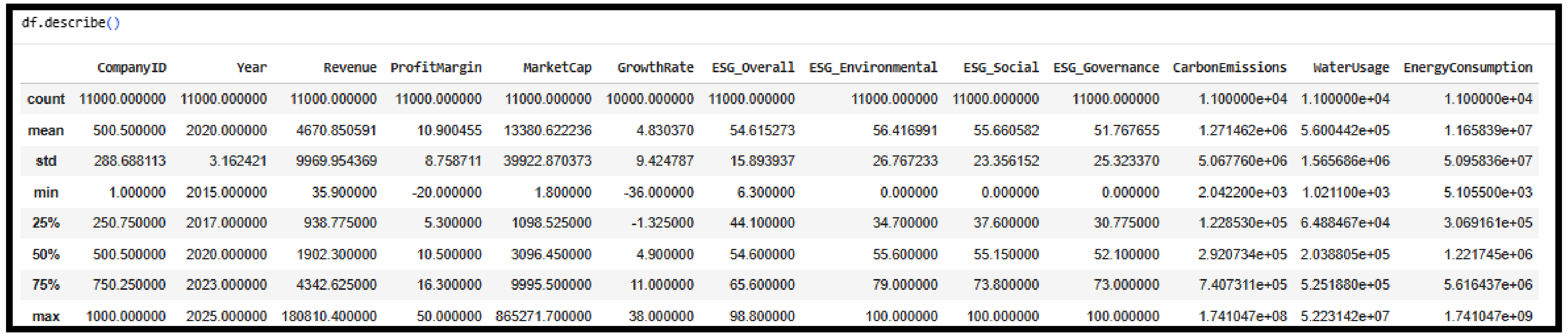


```
RangeIndex: 11000 entries, 0 to 10999
Data columns (total 16 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   CompanyID           11000 non-null  int64
 1   CompanyName         11000 non-null  object
 2   Industry            11000 non-null  object
 3   Region              11000 non-null  object
 4   Year                11000 non-null  int64
 5   Revenue             11000 non-null  float64
 6   ProfitMargin        11000 non-null  float64
 7   MarketCap           11000 non-null  float64
 8   GrowthRate          10000 non-null  float64
 9   ESG_Overall         11000 non-null  float64
 10  ESG_Environmental   11000 non-null  float64
 11  ESG_Social          11000 non-null  float64
 12  ESG_Governance      11000 non-null  float64
 13  CarbonEmissions     11000 non-null  float64
 14  WaterUsage          11000 non-null  float64
 15  EnergyConsumption   11000 non-null  float64
```



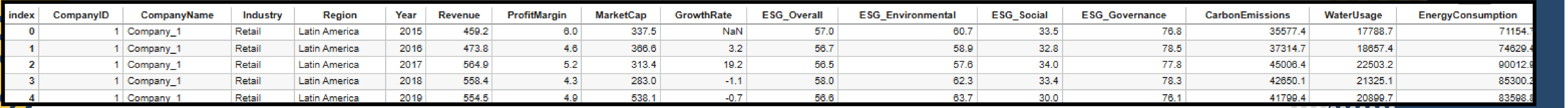
```
df.shape

(11000, 16)
```



```
df.describe()
```

	CompanyID	Year	Revenue	ProfitMargin	MarketCap	GrowthRate	ESG_Overall	ESG_Environmental	ESG_Social	ESG_Governance	CarbonEmissions	WaterUsage	EnergyConsumption
count	11000.000000	11000.000000	11000.000000	11000.000000	11000.000000	10000.000000	11000.000000	11000.000000	11000.000000	11000.000000	1.100000e+04	1.100000e+04	1.100000e+04
mean	500.500000	2020.000000	4670.850591	10.900455	13380.622236	4.830370	54.615273	56.416991	55.660582	51.767655	1.271462e+06	5.600442e+05	1.165839e+07
std	288.688113	3.162421	9969.954369	8.758711	39922.870373	9.424787	15.893937	26.767233	23.356152	25.323370	5.067760e+06	1.565686e+06	5.095836e+07
min	1.000000	2015.000000	35.900000	-20.000000	1.800000	-36.000000	6.300000	0.000000	0.000000	0.000000	2.042200e+03	1.021100e+03	5.105500e+03
25%	250.750000	2017.000000	938.775000	5.300000	1098.525000	-1.325000	44.100000	34.700000	37.600000	30.775000	1.228530e+05	6.488467e+04	3.069161e+05
50%	500.500000	2020.000000	1902.300000	10.500000	3096.450000	4.900000	54.600000	55.600000	55.150000	52.100000	2.920734e+05	2.038805e+05	1.221745e+06
75%	750.250000	2023.000000	4342.625000	16.300000	9995.500000	11.000000	65.600000	79.000000	73.800000	73.000000	7.407311e+05	5.251880e+05	5.616437e+06
max	1000.000000	2025.000000	180810.400000	50.000000	865271.700000	38.000000	98.800000	100.000000	100.000000	100.000000	1.741047e+08	5.223142e+07	1.741047e+09



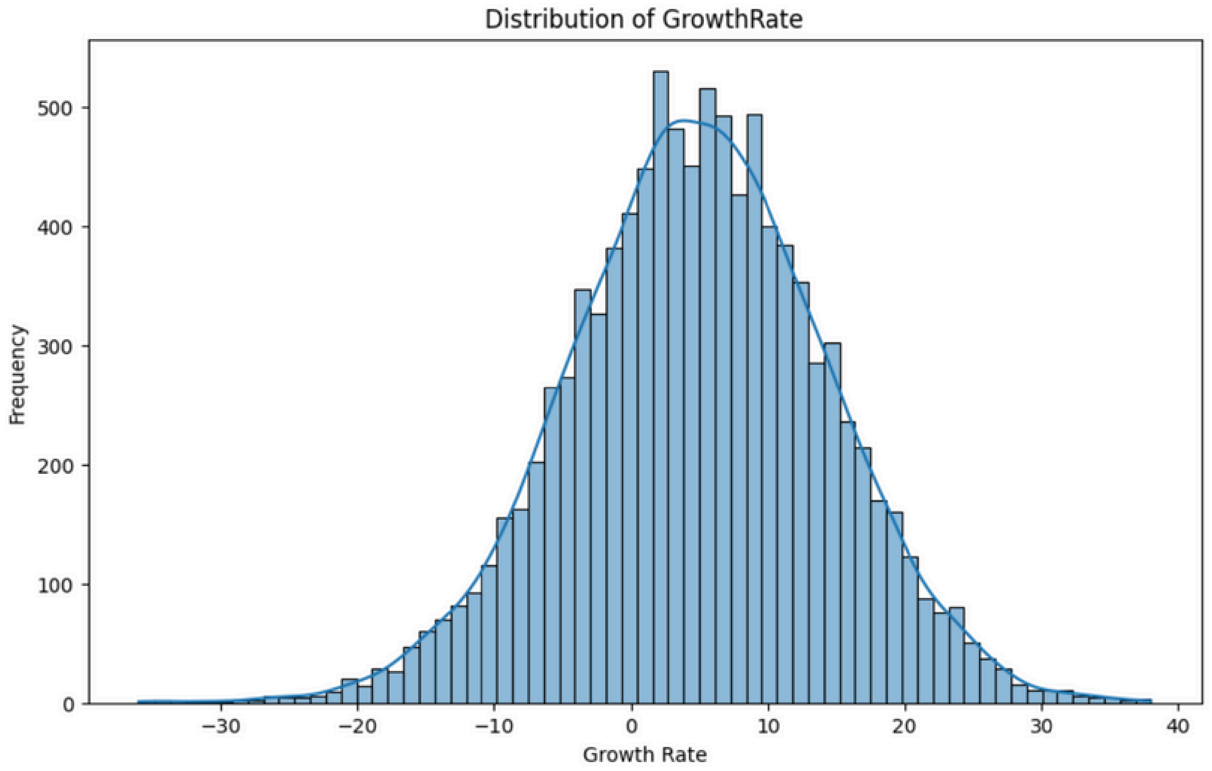
index	CompanyID	CompanyName	Industry	Region	Year	Revenue	ProfitMargin	MarketCap	GrowthRate	ESG_Overall	ESG_Environmental	ESG_Social	ESG_Governance	CarbonEmissions	WaterUsage	EnergyConsumption
0	1	Company_1	Retail	Latin America	2015	459.2	6.0	337.5	NaN	57.0	60.7	33.5	76.8	35577.4	17788.7	71154.7
1	1	Company_1	Retail	Latin America	2016	473.8	4.6	366.6	3.2	56.7	58.9	32.8	78.5	37314.7	18657.4	74629.4
2	1	Company_1	Retail	Latin America	2017	564.9	5.2	313.4	19.2	56.5	57.6	34.0	77.8	45006.4	22503.2	90012.9
3	1	Company_1	Retail	Latin America	2018	558.4	4.3	283.0	-1.1	58.0	62.3	33.4	78.3	42650.1	21325.1	85300.3
4	1	Company_1	Retail	Latin America	2019	554.5	4.9	538.1	-0.7	56.6	63.7	30.0	76.1	41799.4	20899.7	83598.8

SHAPE OF THE DATA

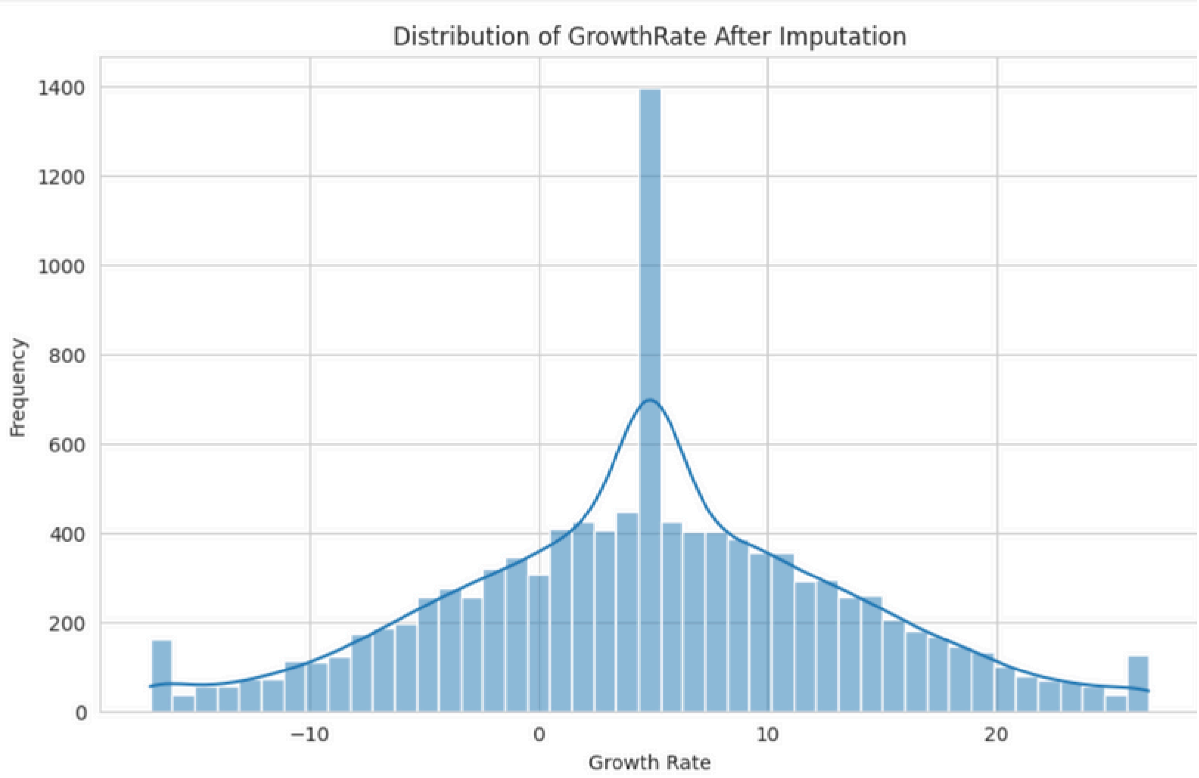
(11000, 16)

MISSING VALUES

Only found in the 'Growth' column.
Had 1000 null values. Imputed the missing values using Median method.



Before Imputation



After Imputation

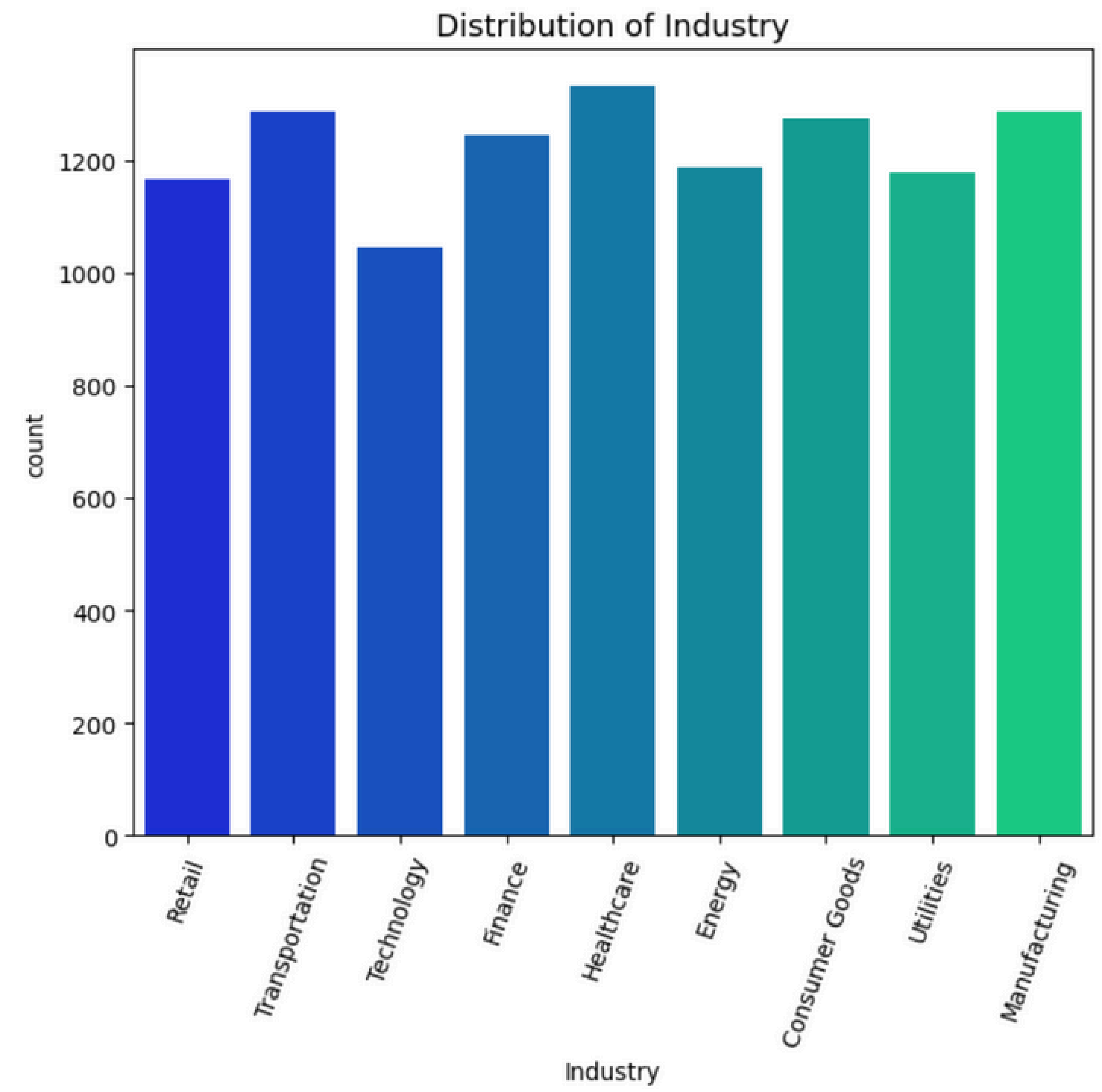
DUPLICATE VALUES

None found

DROPPING COLUMNS

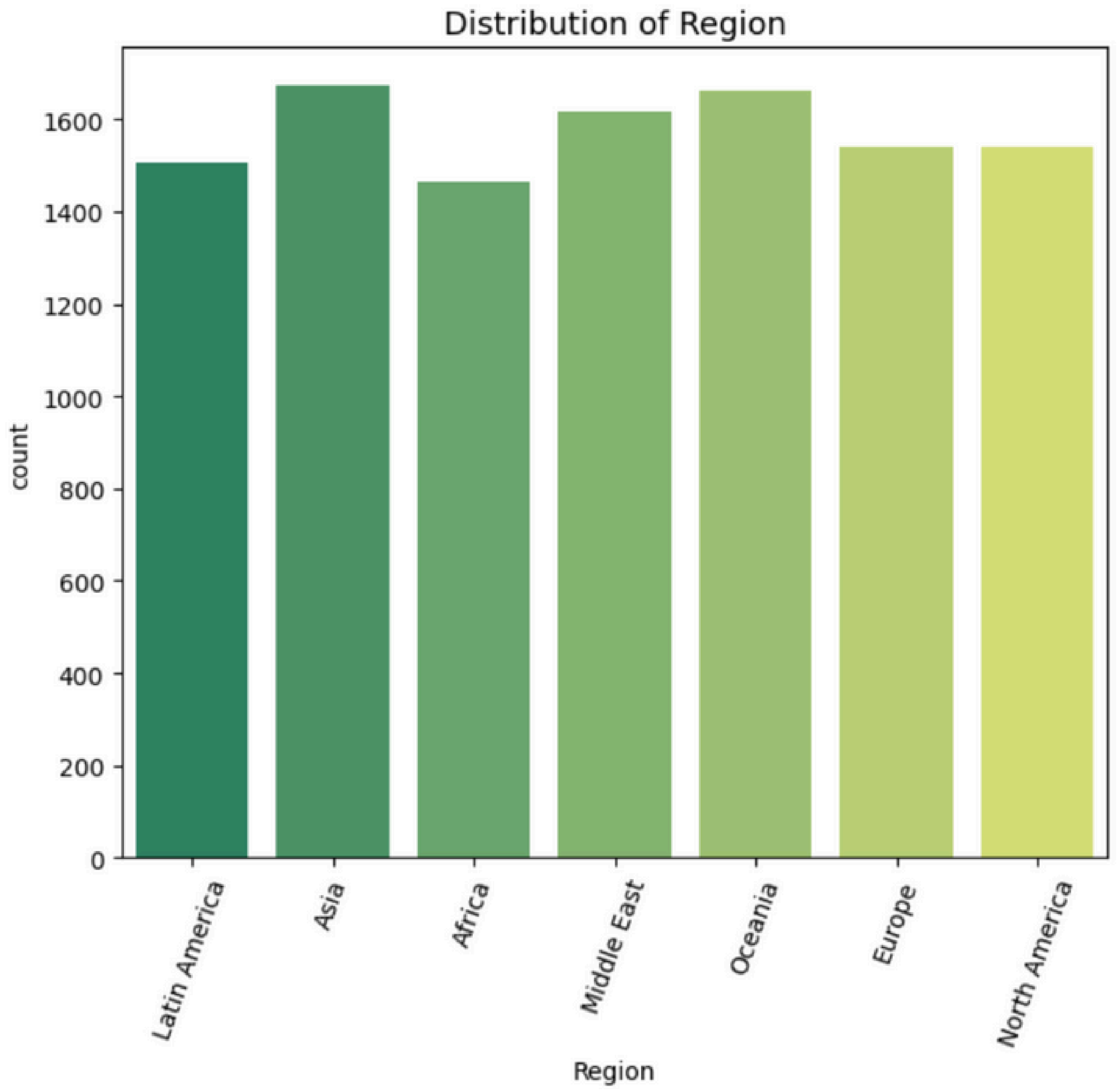
Dropped columns 'CompanyID', 'CompanyName' as they are not relevant in predicting or classifying the Market Cap.

EDA



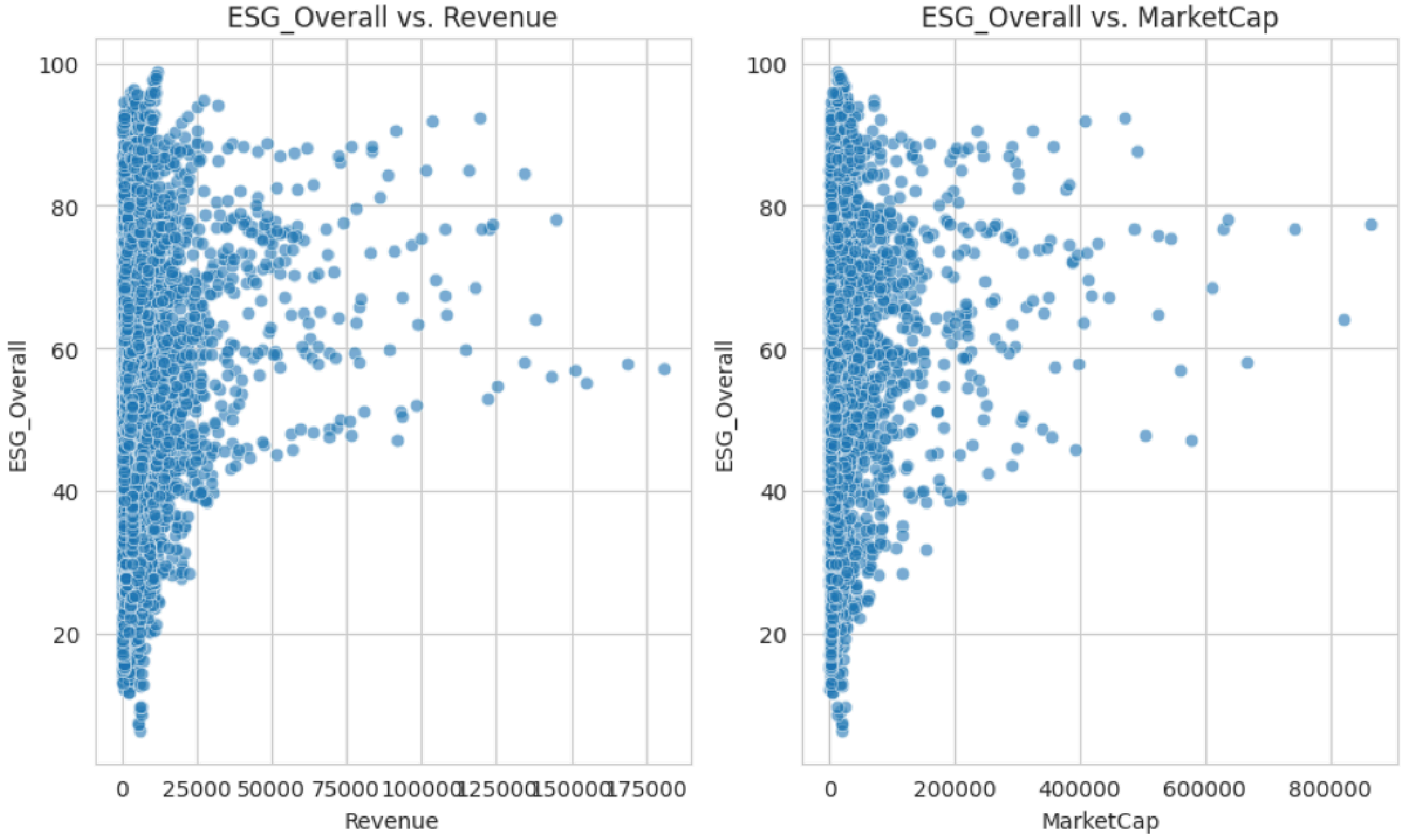
The dataset covers companies across multiple industries, including Manufacturing, Healthcare, Finance, Energy, and Technology.
No single industry dominates the dataset, ensuring balanced sectoral coverage.

EDA



- Companies are distributed across Asia, Europe, North America, Middle East, Africa, and Oceania.
- No region is underrepresented. Meaning our findings won’t be region-specific and can support international decision-making.

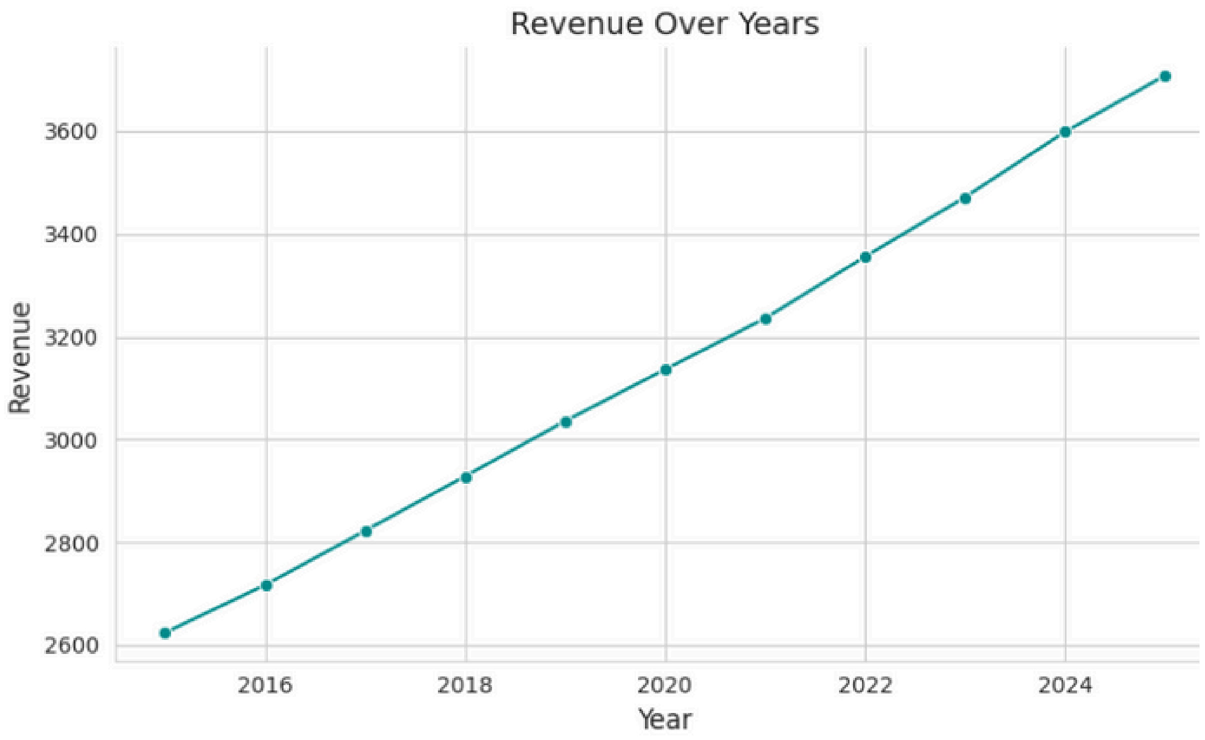
EDA

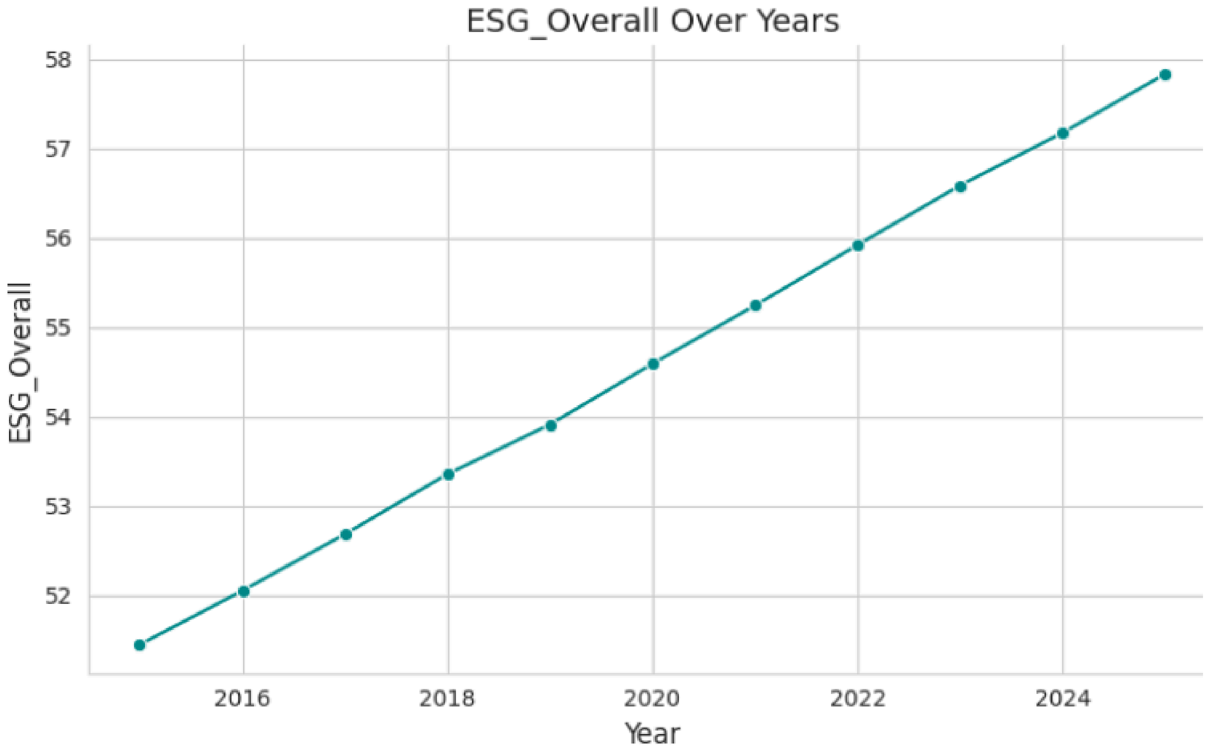


There is a "Weak Correlation" between Revenue/Market Cap and ESG_Overall scores. There is no strong linear pattern. High revenue or MarketCap does not consistently imply high ESG_Overall scores.

Insight: *High revenue does not automatically buy a high ESG score. This proves that sustainability is an independent performance metric, not just a byproduct of being a large company.*

TREND ANALYSIS





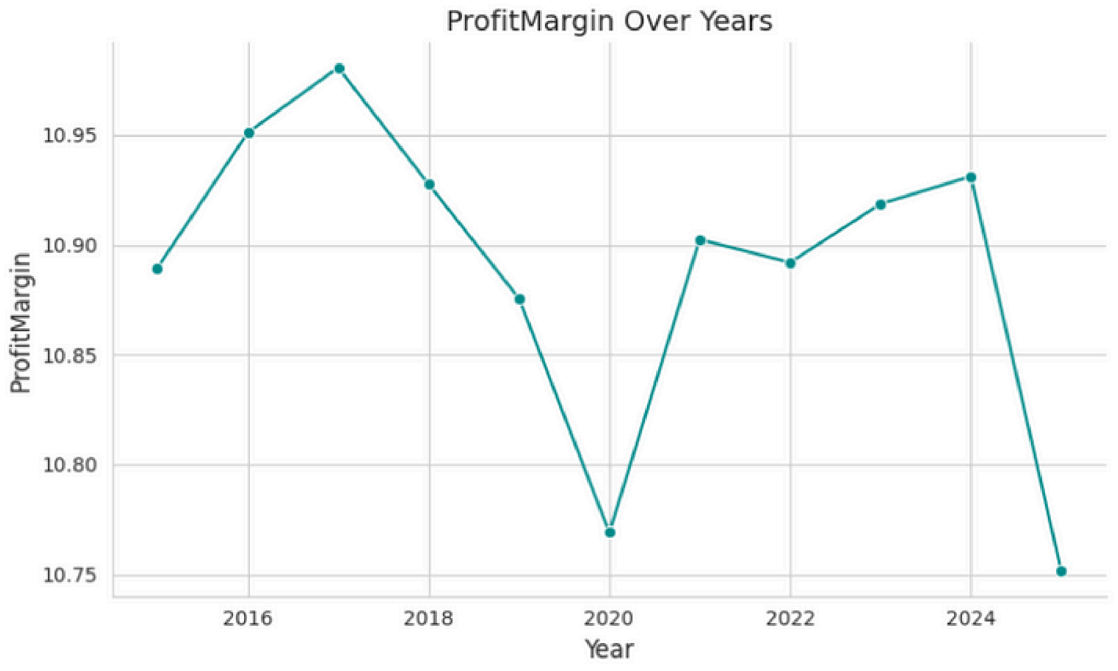
Revenue shows a consistent upward trend across the entire period.

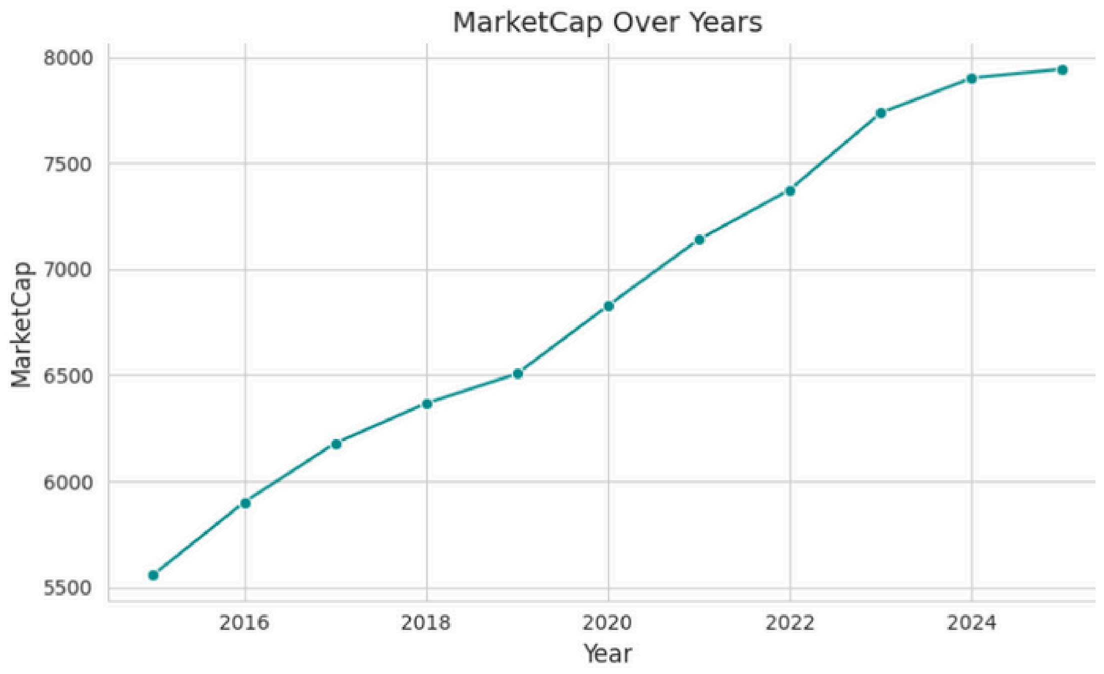
Insight: *Confirms revenue as a robust predictor in performance modeling.*

ESG Overall scores show a steady and continuous improvement across years.

Insight: *Indicates growing emphasis on sustainability, governance, and social responsibility.*

TREND ANALYSIS





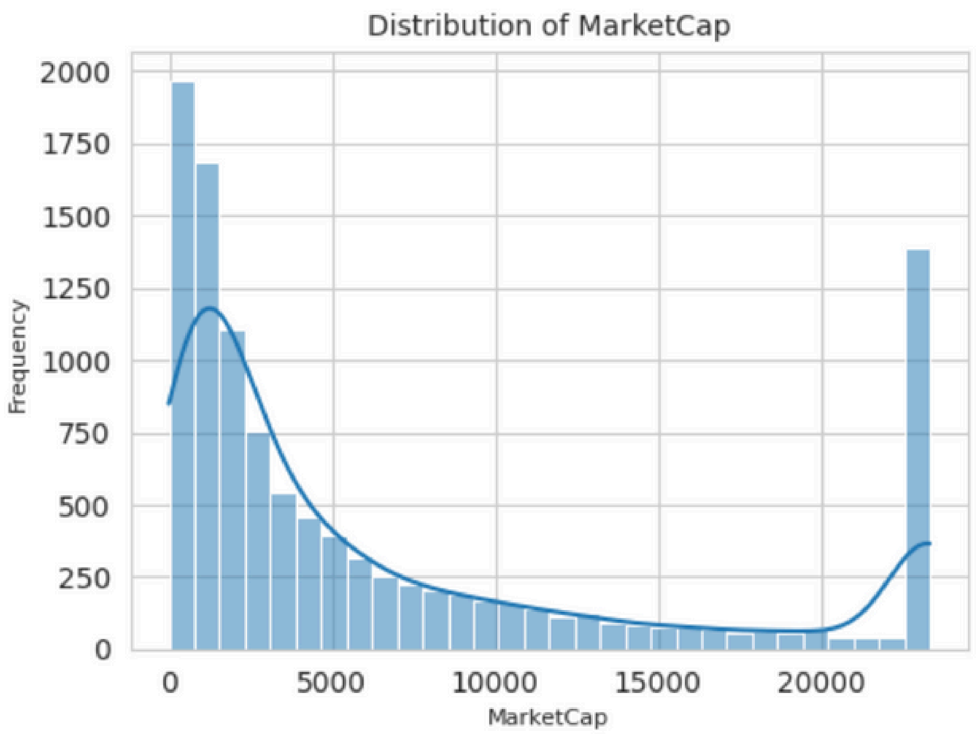
A temporary dip in 2020 is observed (likely due to covid operations disrupted), followed by recovery in subsequent years.

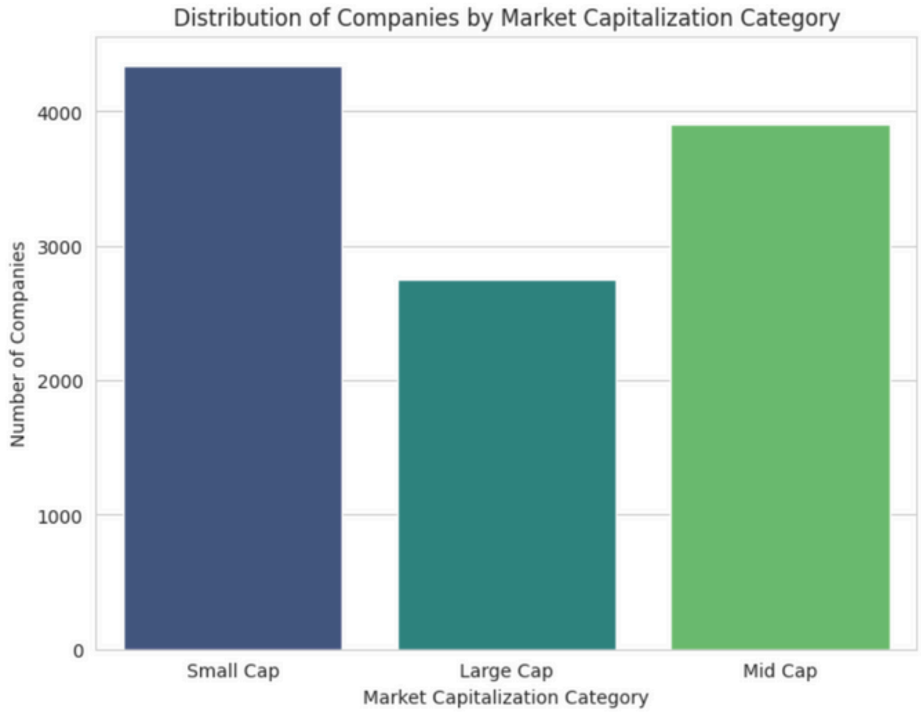
***Insight:** Quick recovery post 2020 indicates operational resilience and cost control.*

Market capitalization shows a strong upward trajectory over time.

***Insight:** Rising market capitalization reflects increasing investor confidence.*

DERIVED FEATURES FOR ANALYSIS



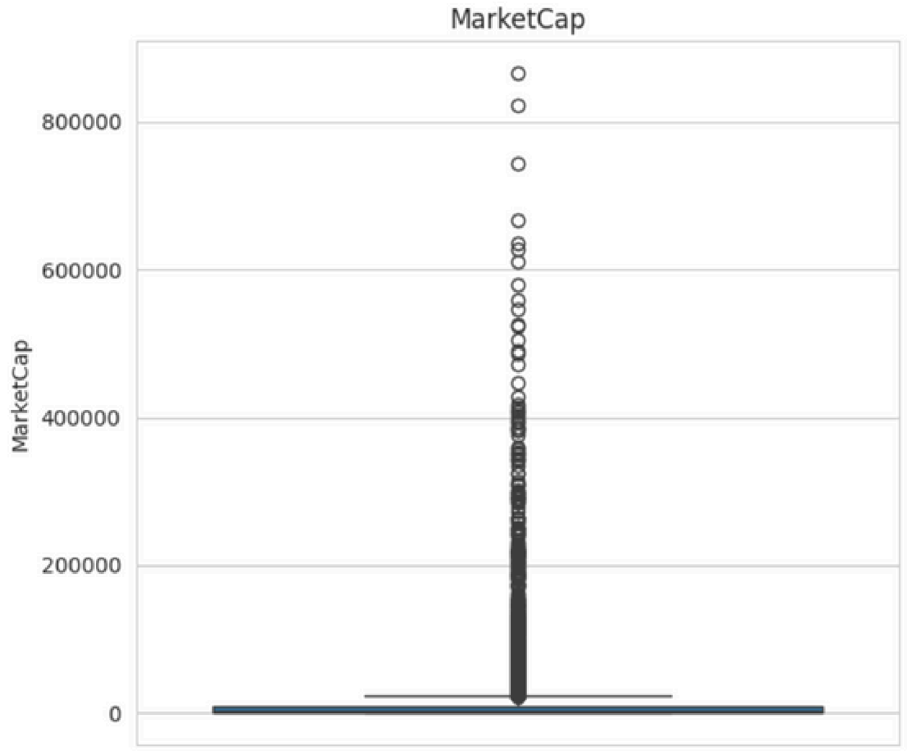


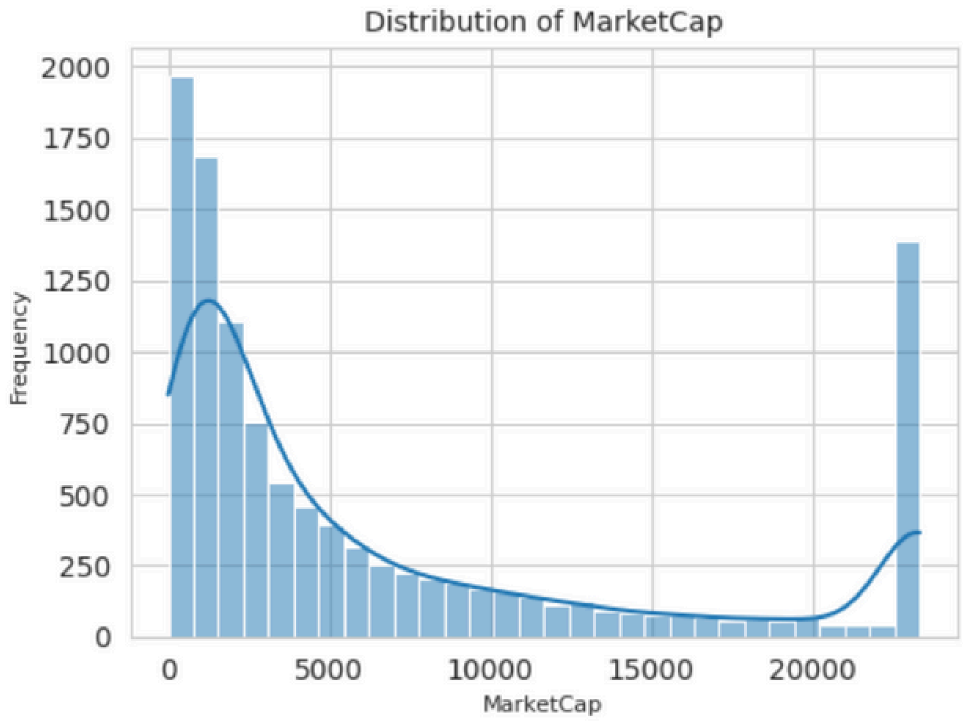
Market capitalization exhibits strong right skewness, as indicated by its long-tailed distribution and the concentration of firms at lower values. Converting it into size-based categories to improve interpretability .

Market Capitalization was categorized into Small Cap, Mid Cap, and Large Cap to improve interpretability.

Categorization allows models to capture non-linear size effects more effectively.

TARGET VARIABLE ENGINEERING





- *MarketCap was selected as the primary target variable.*

- Raw market cap values showed heavy right skew and extreme outliers.

- Applied log transformation to stabilize variance and improve model learning.

- ***Log Transformation*** *was used* because MarketCap spans several orders of magnitude. Log transformation compresses extreme values, improves model stability

MARKETCAP PREDICTION: MODEL TRAINING

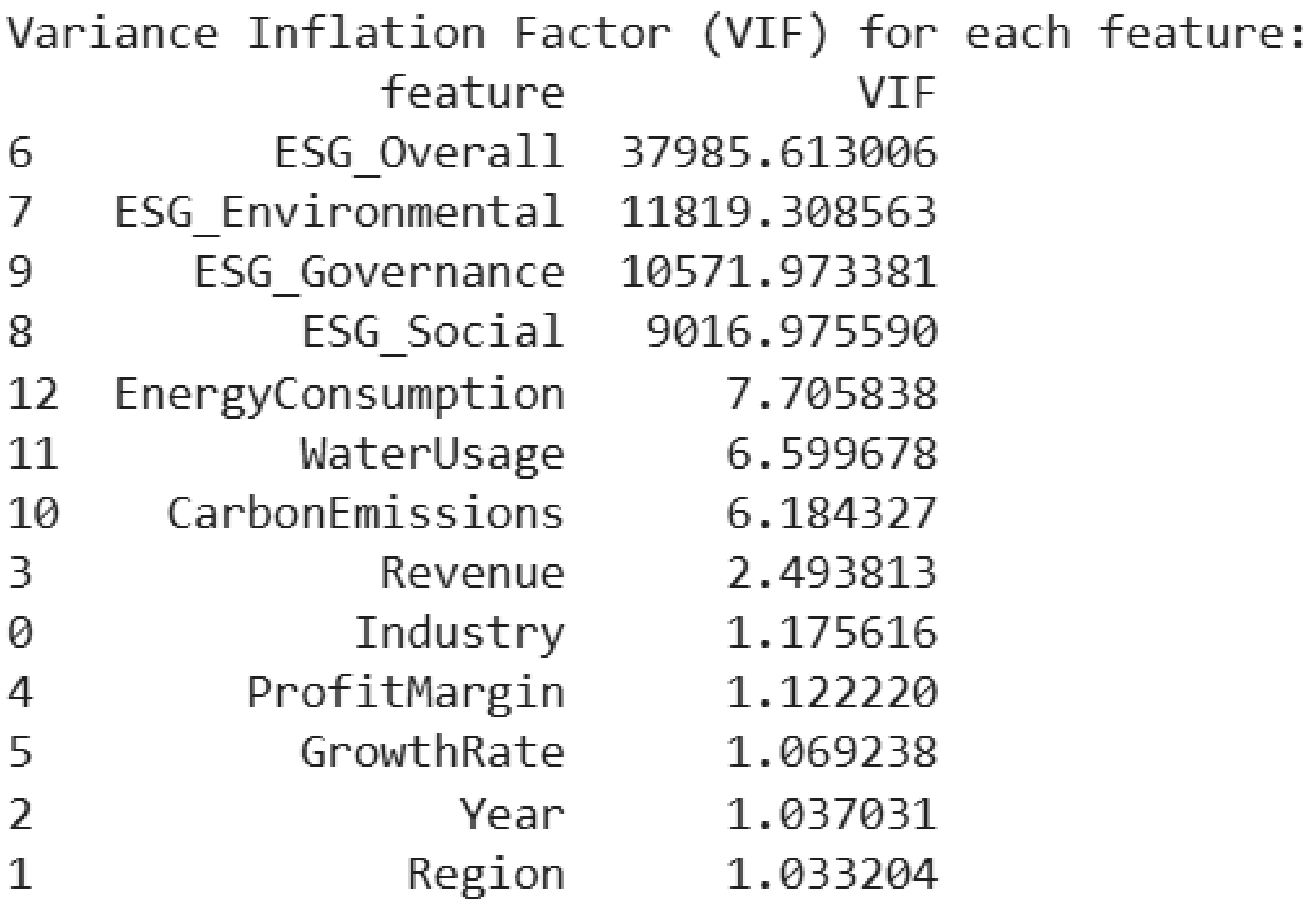
Dataset was split into 80% training and 20% testing.

Feature Scaling using StandardScaler

- Financial, ESG, and environmental variables exist on very different scales.
- Applied standardization to normalize feature influence.
- Scaling was done so that the model evaluates each driver fairly rather than favoring large-magnitude metrics.

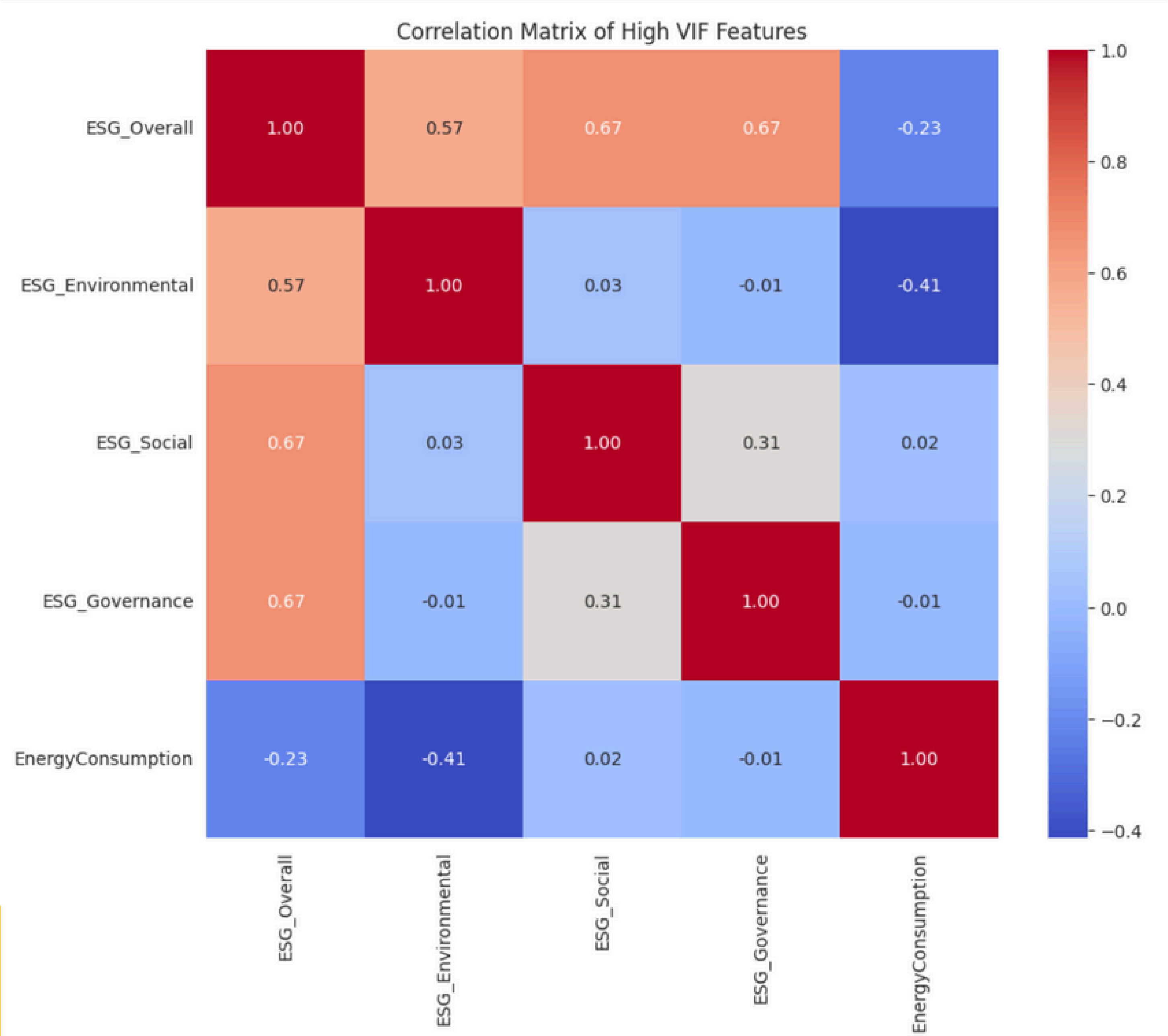
MULTICOLLINEARITY DIAGNOSIS

- Extremely high multicollinearity detected among ESG variables.
- ESG Overall, Environmental, Social, and Governance scores strongly overlap.



```
Variance Inflation Factor (VIF) for each feature:
               feature           VIF
6          ESG_Overall  37985.613006
7    ESG_Environmental  11819.308563
9       ESG_Governance  10571.973381
8           ESG_Social   9016.975590
12   EnergyConsumption      7.705838
11          WaterUsage      6.599678
10     CarbonEmissions      6.184327
3              Revenue      2.493813
0             Industry      1.175616
4         ProfitMargin      1.122220
5           GrowthRate      1.069238
2                 Year      1.037031
1               Region      1.033204
```

CORRELATION HEATMAP OF HIGH-VIF FEATURES



Correlation Matrix of High VIF Features

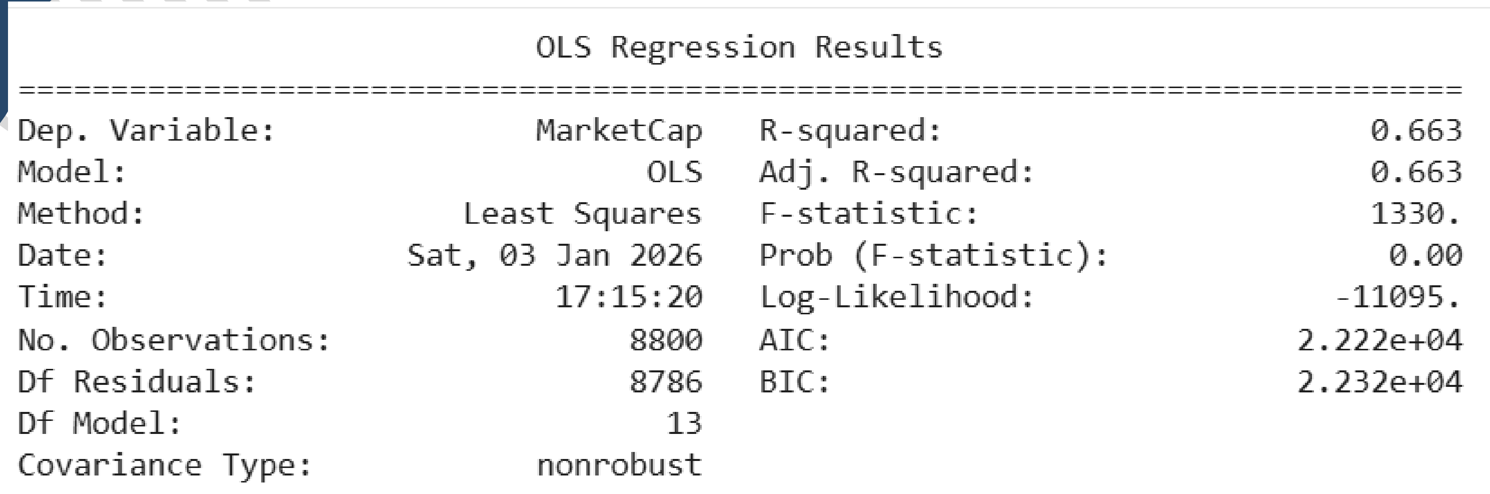
	ESG_Overall	ESG_Environmental	ESG_Social	ESG_Governance	EnergyConsumption
ESG_Overall	1.00	0.57	0.67	0.67	-0.23
ESG_Environmental	0.57	1.00	0.03	-0.01	-0.41
ESG_Social	0.67	0.03	1.00	0.31	0.02
ESG_Governance	0.67	-0.01	0.31	1.00	-0.01
EnergyConsumption	-0.23	-0.41	0.02	-0.01	1.00

Interdependence Among ESG Metrics

- Strong correlation between ESG Overall and ESG Environmental, Social and Governance.
- Confirms ESG acts as a composite signal.
- Explains inflated VIF values.

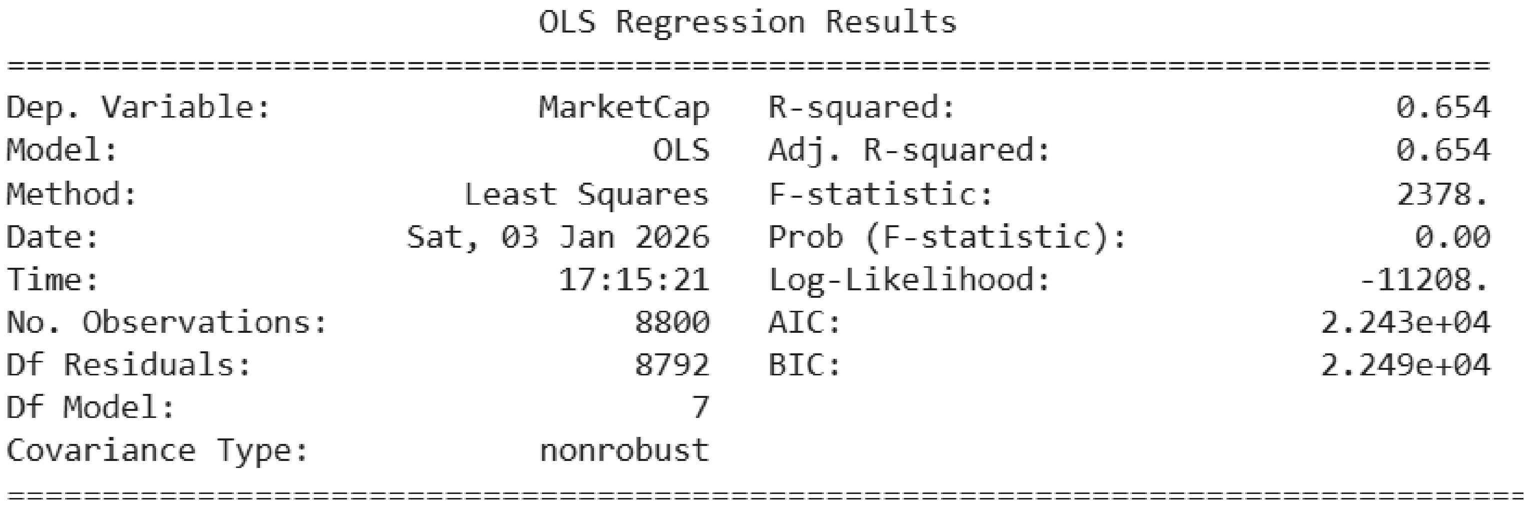
Insight- Sustainability based features should be interpreted holistically, not as isolated dimensions.

MULTIPLE LINEAR REGRESSION (MLR): KEY FINDINGS & BUSINESS INSIGHTS



```
                            OLS Regression Results
==============================================================================
Dep. Variable:              MarketCap   R-squared:                       0.663
Model:                            OLS   Adj. R-squared:                  0.663
Method:                 Least Squares   F-statistic:                     1330.
Date:                Sat, 03 Jan 2026   Prob (F-statistic):               0.00
Time:                        17:15:20   Log-Likelihood:                -11095.
No. Observations:                8800   AIC:                         2.222e+04
Df Residuals:                    8786   BIC:                         2.232e+04
Df Model:                          13
Covariance Type:            nonrobust
```

First MLR Model Statistics



```
                            OLS Regression Results
==============================================================================
Dep. Variable:              MarketCap   R-squared:                       0.654
Model:                            OLS   Adj. R-squared:                  0.654
Method:                 Least Squares   F-statistic:                     2378.
Date:                Sat, 03 Jan 2026   Prob (F-statistic):               0.00
Time:                        17:15:21   Log-Likelihood:                -11208.
No. Observations:                8800   AIC:                         2.243e+04
Df Residuals:                    8792   BIC:                         2.249e+04
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
```

MLR Model statistics after removal of columns based on VIF and Heatmap Multicollinearity

Final MLR explains ~65% of variation in MarketCap (Adjusted R^2 = 0.654)

Insights-

Revenue is the biggest factor behind market value

Companies that earn more revenue generally have higher market valuations.

Profit margins matter, not just company size

Even among similar-sized firms, those with better margins tend to be valued higher.

Growth plays an important role in valuation

Companies showing strong growth are rewarded by the market for their future potential.

Overall ESG score affects how investors value companies

Investors look at the overall sustainability performance rather than individual ESG components.

Water usage reflects operational efficiency

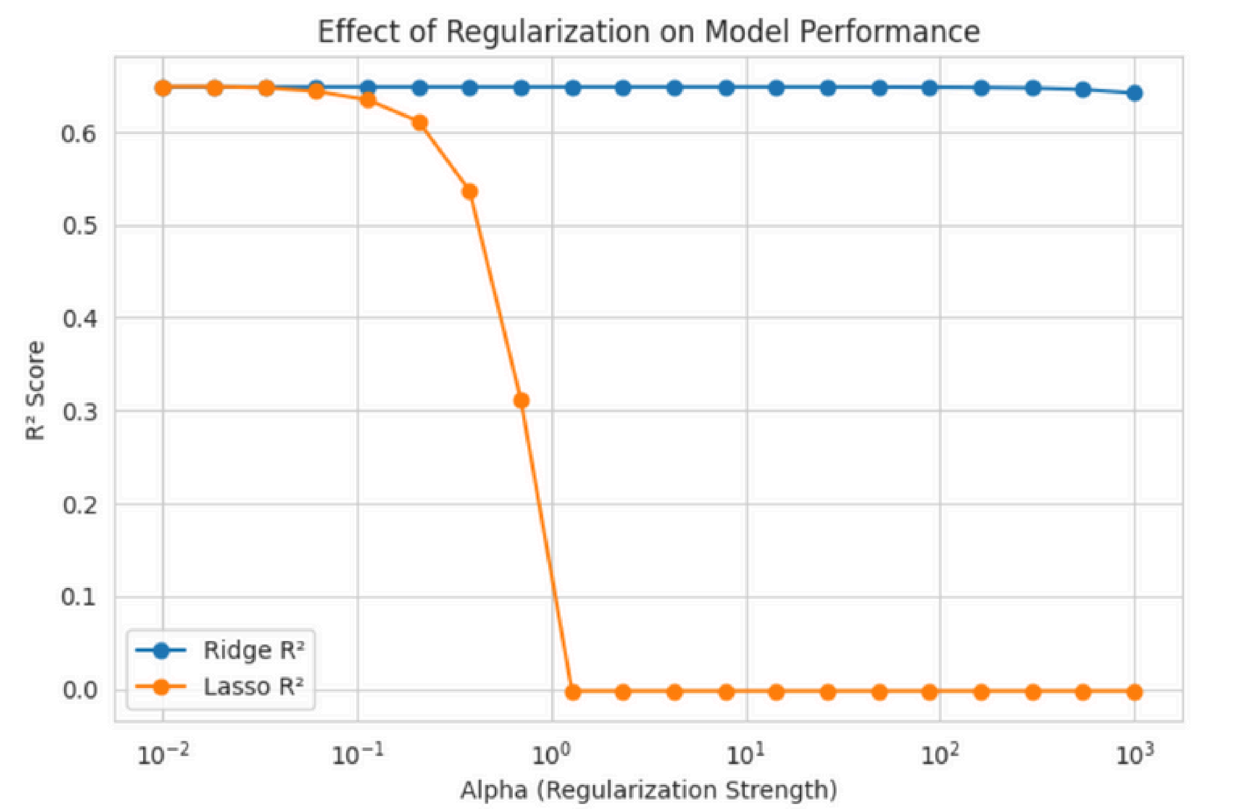
Firms that manage water consumption better are viewed more favorably by the market.

Industry type influences market valuation

A company's sector impacts its valuation beyond what financial numbers alone explain.

MODEL COMPARISON – LINEAR VS RIDGE VS LASSO

Model	*R² (Test)*	*RMSE*
Multiple Linear Regression	~0.65	~0.89
Ridge Regression	~0.65	~0.89
Lasso Regression	~0.64	~0.90



1. All three models perform similarly, indicating strong underlying relationships in the data
2. Ridge does not significantly improve accuracy over multiple regression
3. Lasso removes important drivers, leading to a drop in predictive power
4. Revenue, profitability, growth, and ESG signals are too important to be eliminated. Therefore, Lasso is not the model chosen.

Final Model Selection-

- Multiple Linear Regression was selected as the final model
- It delivers comparable accuracy while remaining easy to interpret
- Business drivers and their impact on valuation remain transparent
- Ridge adds complexity without meaningful performance gains
- Lasso over-penalizes correlated business variables in this dataset

RIDGE & LASSO:

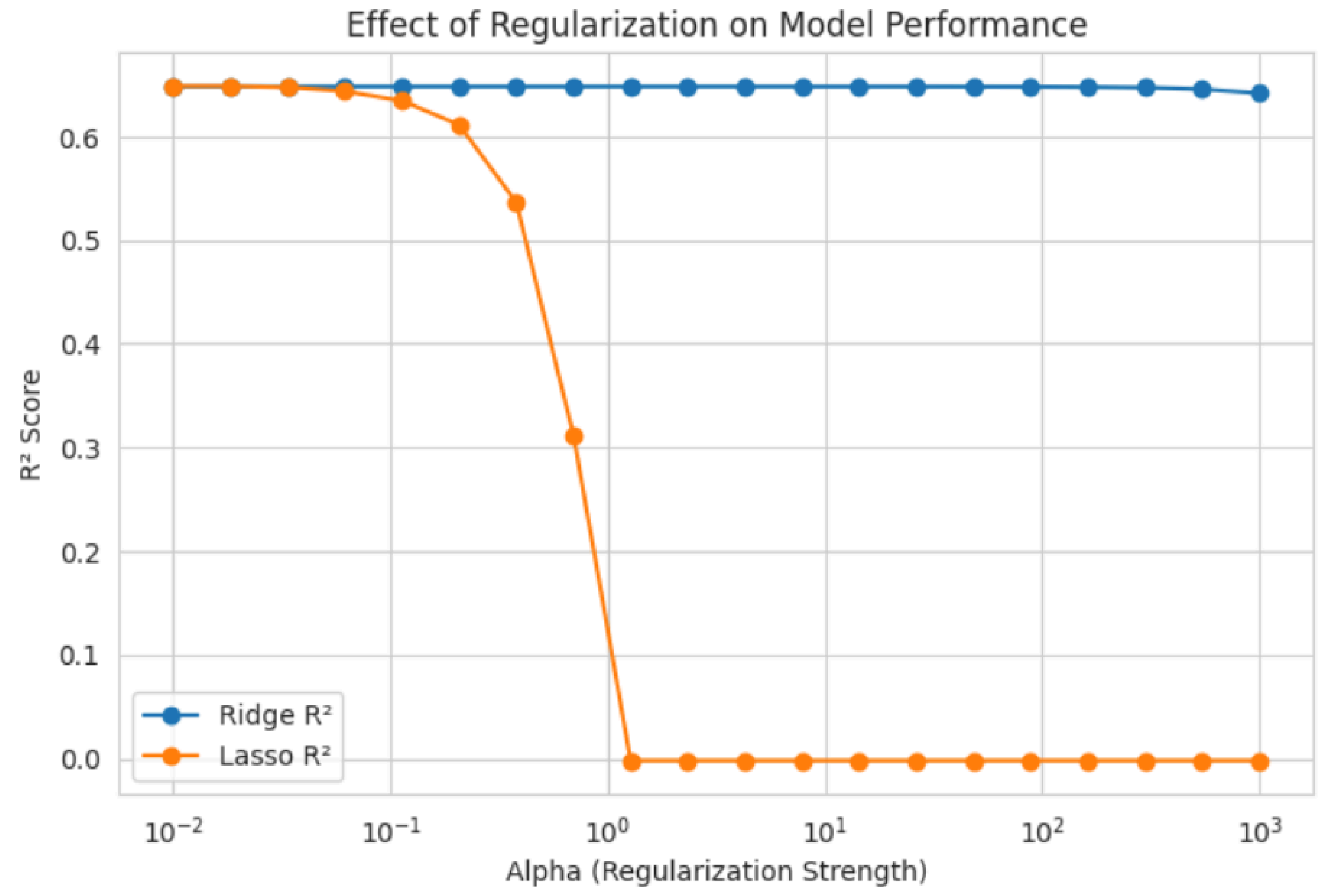
Regularization was tested to assess whether multicollinearity affected model stability.

Model	***Test R²***	***RMSE***	***Insight***
MLR	~0.65	~0.89	Stable, interpretable
Ridge	~0.65	~0.89	No meaningful improvement
Lasso	~0.64	~0.90	Information loss

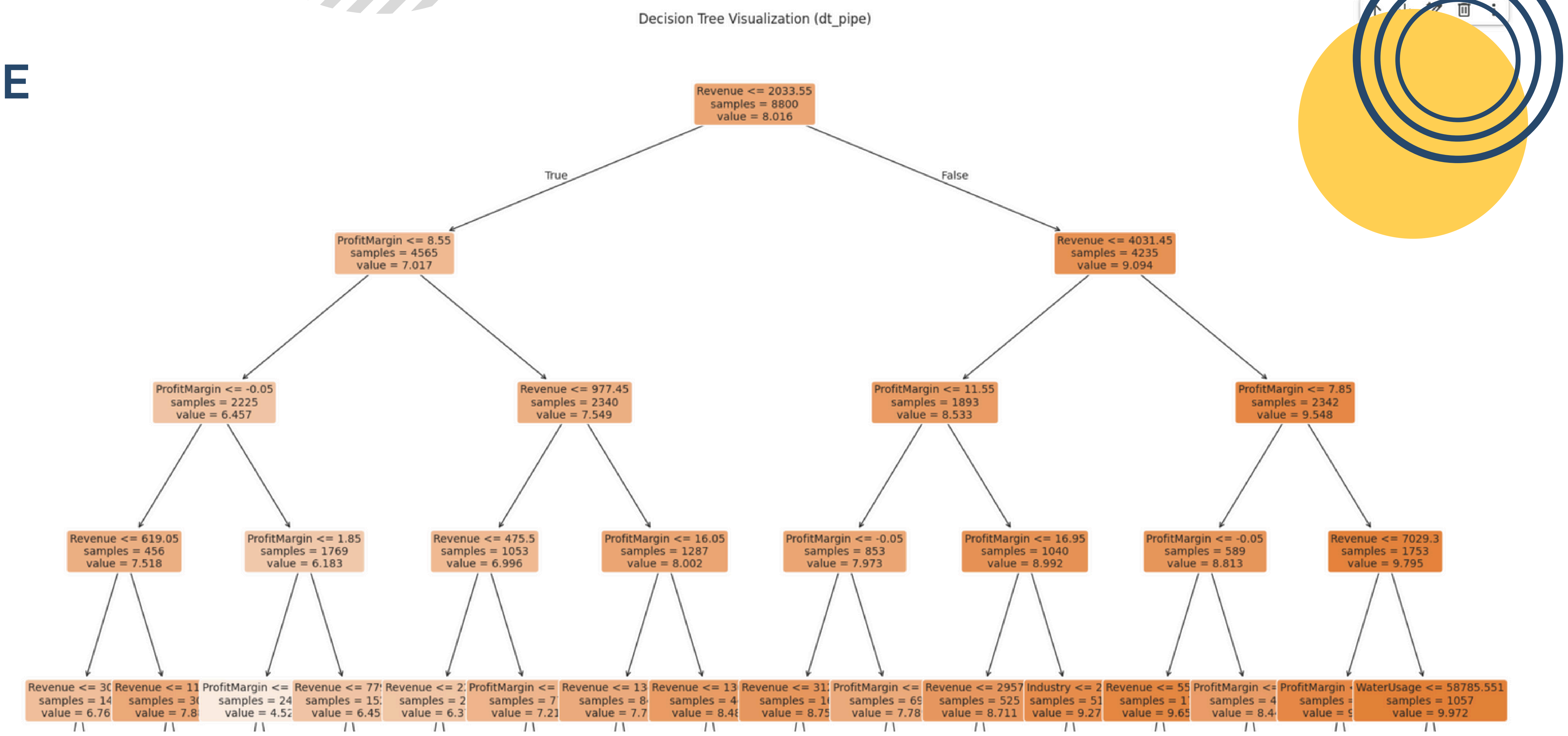
1. All three models perform similarly, indicating strong underlying relationships in the data
2. Ridge does not significantly improve accuracy over multiple regression
3. Lasso removes important drivers, leading to a drop in predictive power
4. Revenue, profitability, growth, and ESG signals are too important to be eliminated. Therefore, Lasso is not the model chosen.

Findings-

- Lasso aggressively removes correlated financial and ESG signals, leading to underfitting.
- Ridge confirms that multicollinearity is controlled and not distorting results.



DECISION TREE


Decision Tree Visualization (dt_pipe)
Revenue <= 2033.55 samples = 8800 value = 8.016
True
False
ProfitMargin <= 8.55 samples = 4565 value = 7.017
Revenue <= 4031.45 samples = 4235 value = 9.094
ProfitMargin <= -0.05 samples = 2225 value = 6.457
Revenue <= 977.45 samples = 2340 value = 7.549
ProfitMargin <= 11.55 samples = 1893 value = 8.533
ProfitMargin <= 7.85 samples = 2342 value = 9.548
Revenue <= 619.05 samples = 456 value = 7.518
ProfitMargin <= 1.85 samples = 1769 value = 6.183
Revenue <= 475.5 samples = 1053 value = 6.996
ProfitMargin <= 16.05 samples = 1287 value = 8.002
ProfitMargin <= -0.05 samples = 853 value = 7.973
ProfitMargin <= 16.95 samples = 1040 value = 8.992
ProfitMargin <= -0.05 samples = 589 value = 8.813
Revenue <= 7029.3 samples = 1753 value = 9.795
Revenue <= 30 samples = 14 value = 6.76
Revenue <= 11 samples = 30 value = 7.8
ProfitMargin <= samples = 24 value = 4.52
Revenue <= 77 samples = 15 value = 6.45
Revenue <= 2 samples = 2 value = 6.3
ProfitMargin <= samples = 7 value = 7.21
Revenue <= 13 samples = 8 value = 7.7
Revenue <= 13 samples = 4 value = 8.48
Revenue <= 31 samples = 1 value = 8.75
ProfitMargin <= samples = 69 value = 7.78
Revenue <= 2957 samples = 525 value = 8.711
Industry <= 2 samples = 51 value = 9.27
Revenue <= 55 samples = 1 value = 9.65
ProfitMargin <= samples = 4 value = 8.4
ProfitMargin samples = value = 9
WaterUsage <= 58785.551 samples = 1057 value = 9.972

1

Revenue remains the primary driver of valuation, as it consistently appears at the top of the decision tree

2

Profit margin further separates companies with similar revenues, highlighting the role of operational efficiency.

3

Industry plays a secondary but meaningful role, refining valuations once size and profitability are accounted for.

4

Water usage appears at deeper levels of the tree, suggesting that resource efficiency influences valuation for certain firm profiles rather than across the board.

ENSEMBLE MODELS:

Tree-based ensemble models were also tested to capture non-linear interactions.

Why Ensemble Models?

- MarketCap shows non-linear relationships that linear models cannot fully capture.
- Ensemble models better reflect how investors actually price firms in practice.
- These models are more suitable for valuation forecasting and scenario analysis than traditional linear approaches.

Model	***Test R^2***	***RMSE***	***MAE***
MLR	~0.65	~0.89	~0.65
Decision Tree	~0.89	~0.50	~0.37
Random Forest	**~0.96**	**~0.31**	**~0.23**
Gradient Boosting	~0.96	~0.31	~0.24
XGBoost	~0.95	~0.33	~0.24

Random Forest, Gradient Boosting, and XGBoost deliver the highest predictive accuracy, with test R^2 close to 0.95.

Based on this **Random Forest** was selected because:

- Delivered highest predictive accuracy on unseen data (strong test R^2, low RMSE).
- More stable than a single Decision Tree as it reduces overfitting.
- Performs consistently without aggressive feature elimination.

HYPERPARAMETER TUNING – RANDOM FOREST

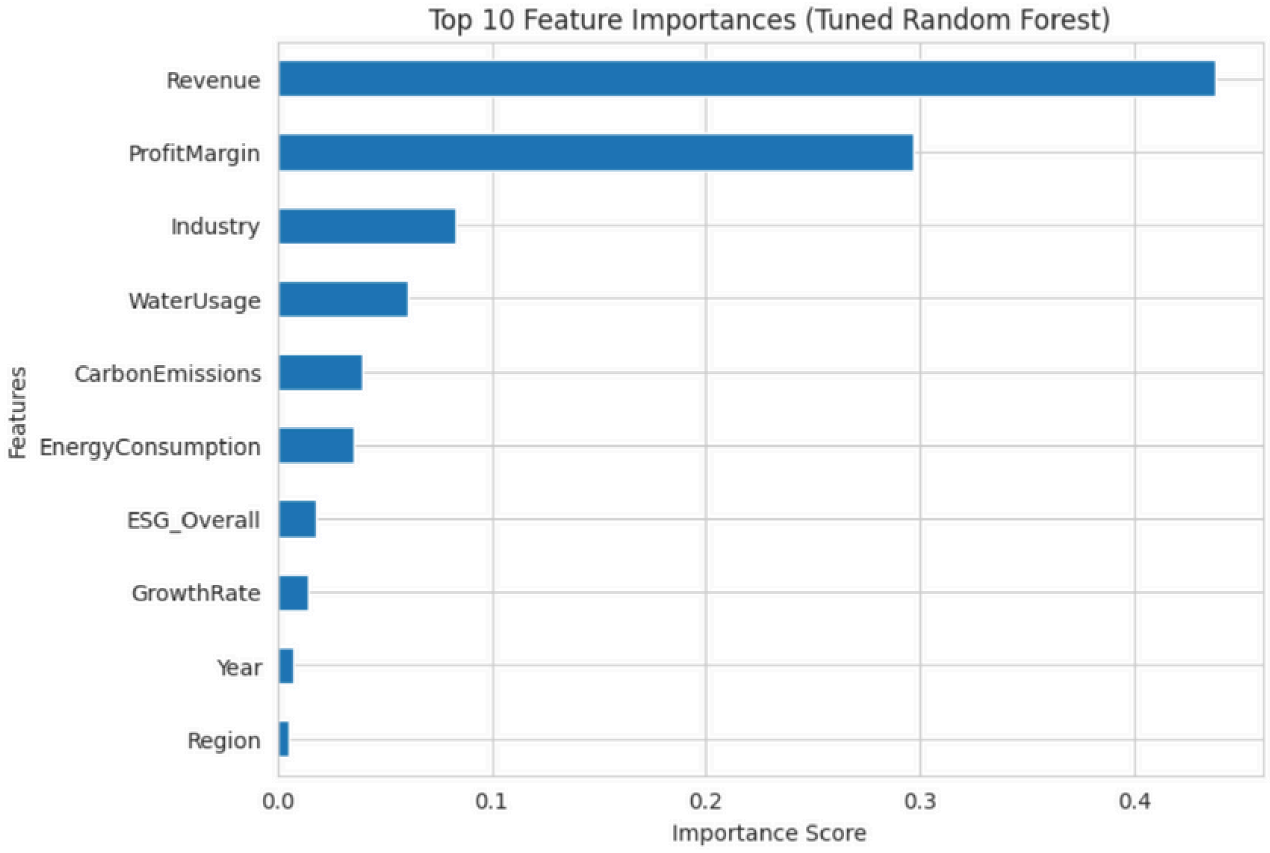
Before tuning:

1. Test R² ≈ 0.956
2. Very low RMSE and MAE, indicating strong fit but higher risk of overfitting

After tuning:

- Test R² = 0.93
- Test RMSE = 0.39, Test MAE = 0.27
- Slight drop in R²

Although there is a small drop in accuracy after tuning, but tuning reduced overfitting and produced a more reliable model for unseen companies,



Revenue clearly dominates valuation decisions
Company size remains the strongest signal driving market capitalization.

Profit Margin is the second most influential factor
Firms that convert revenue into profits more efficiently receive higher valuations.

Industry plays a meaningful role
Valuation expectations differ across sectors, even for companies with similar financials.

Water Usage stands out among environmental factors
Resource efficiency is increasingly reflected in how markets price companies.

Carbon Emissions and Energy Consumption have moderate impact
Environmental costs matter, but they influence valuation indirectly rather than as primary drivers.

BUSINESS IMPLICATIONS

1. Company valuation is driven first by scale (revenue), then by efficiency (profitability).

2. Sustainability matters, but investors look at the overall ESG picture rather than individual ESG scores.

3. How efficiently a company uses resources, particularly water, is starting to influence its market value.

4. A company's industry plays a role in how it is valued, as different sectors are judged by different standards.

STRATEGIC RECOMMENDATIONS

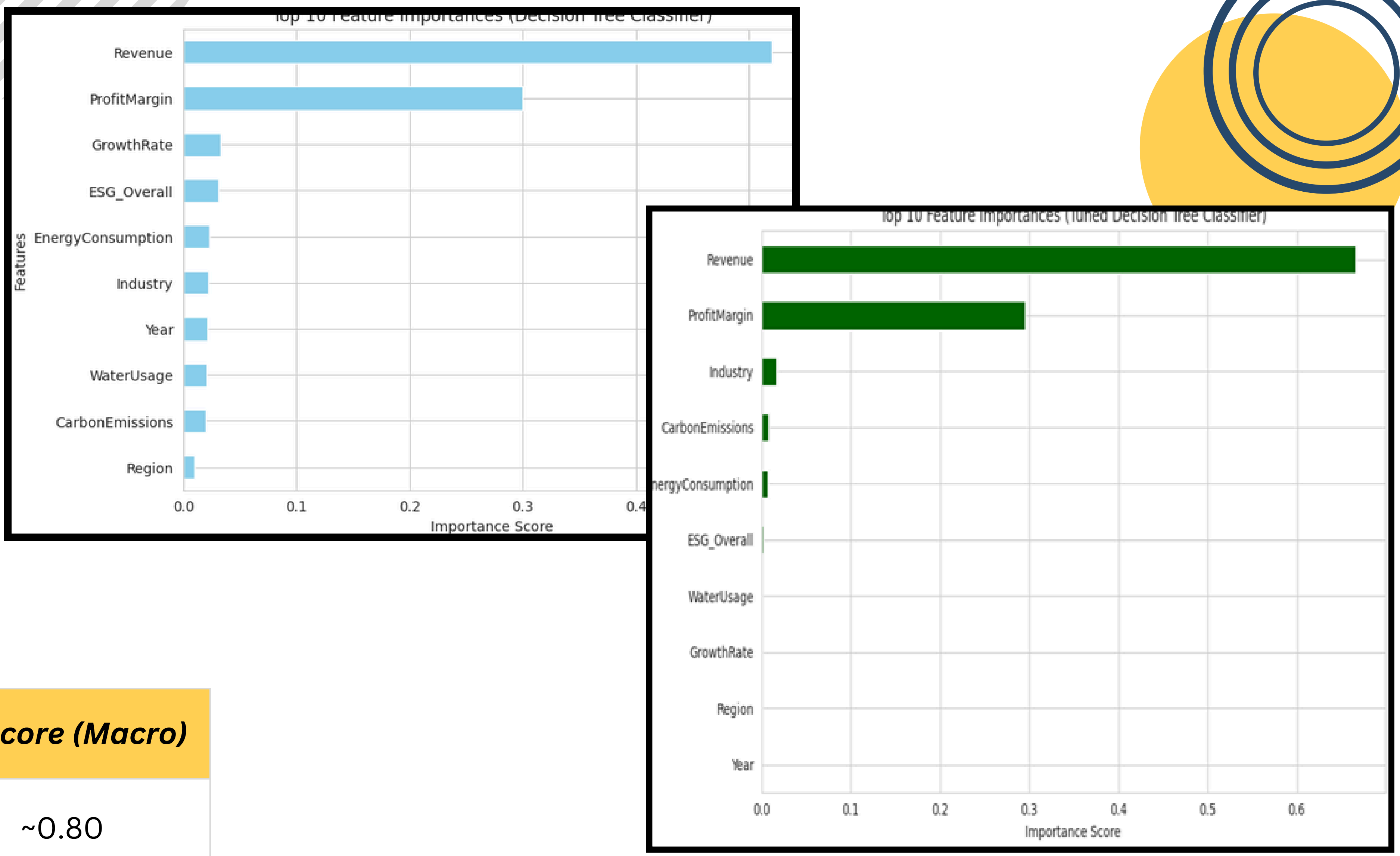
1. Companies should focus on growing revenue while keeping costs under control to improve overall valuation.

2. Sustainability efforts should target areas that improve efficiency, such as reducing water and resource usage, rather than focusing only on ESG labels.

3. ESG should be treated as a factor that strengthens valuation once financial fundamentals are strong, not as a replacement for them.

MARKETCAP CLASSIFICATION: MODEL TRAINING

- Standardization applied where required.
- All models were trained using a stratified train–test split to preserve class balance.



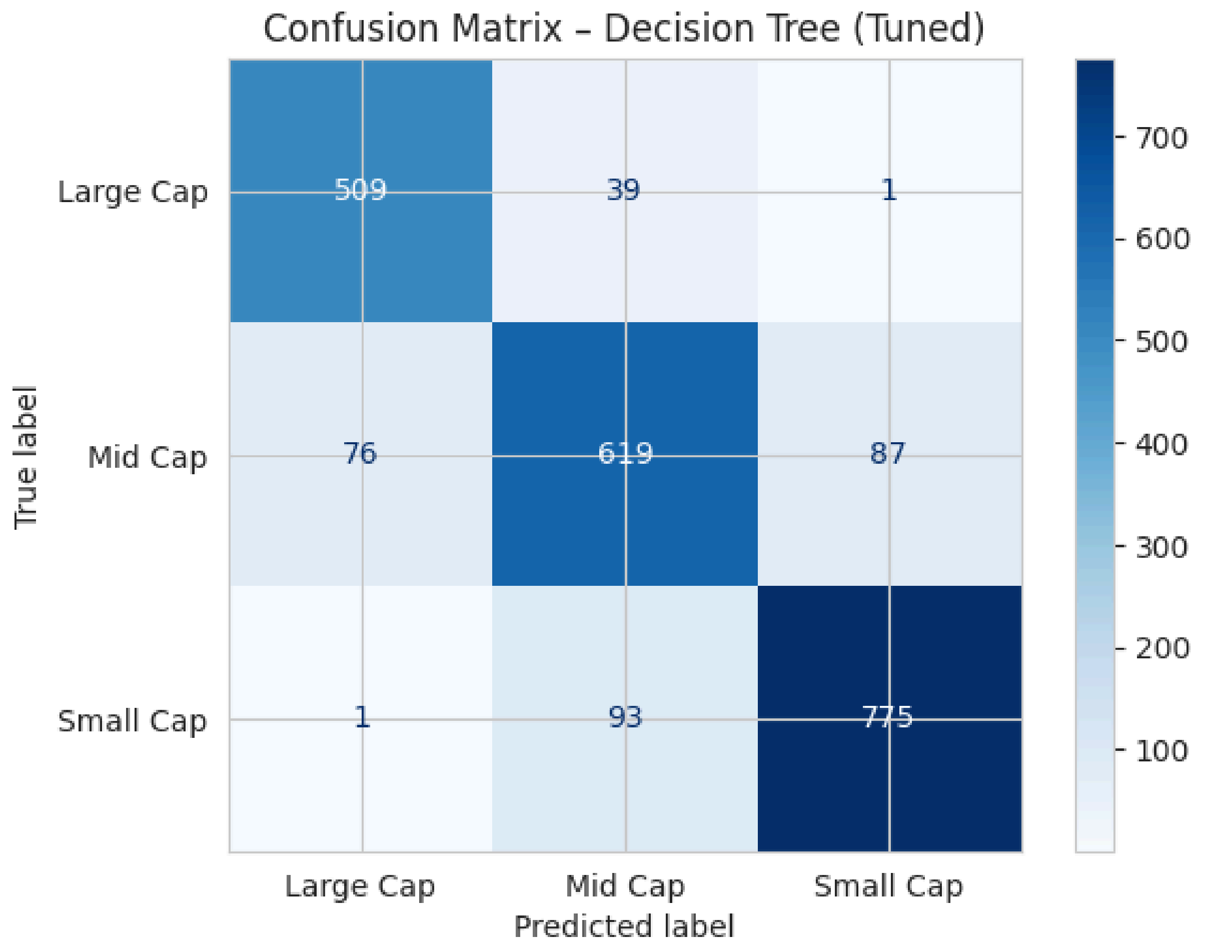
MODEL COMPARISON:

Model	*Accuracy*	*Precision (Macro)*	*Recall (Macro)*	*F1-Score (Macro)*
Logistic Regression	~0.80	~0.81	~0.80	~0.80
KNN	~0.81	~0.81	~0.81	~0.81
SVM – Linear	~0.80	~0.81	~0.80	~0.81
SVM – Polynomial	~0.81	~0.82	~0.81	~0.81
SVM – RBF	~0.83	~0.84	~0.83	~0.83
SVM – Sigmoid	~0.65	~0.65	~0.66	~0.65
Decision Tree Classifier (tuned)	**~0.865**	**~0.862**	**~0.87**	**~0.87**

Final Model Selected: Decision Tree Classifier

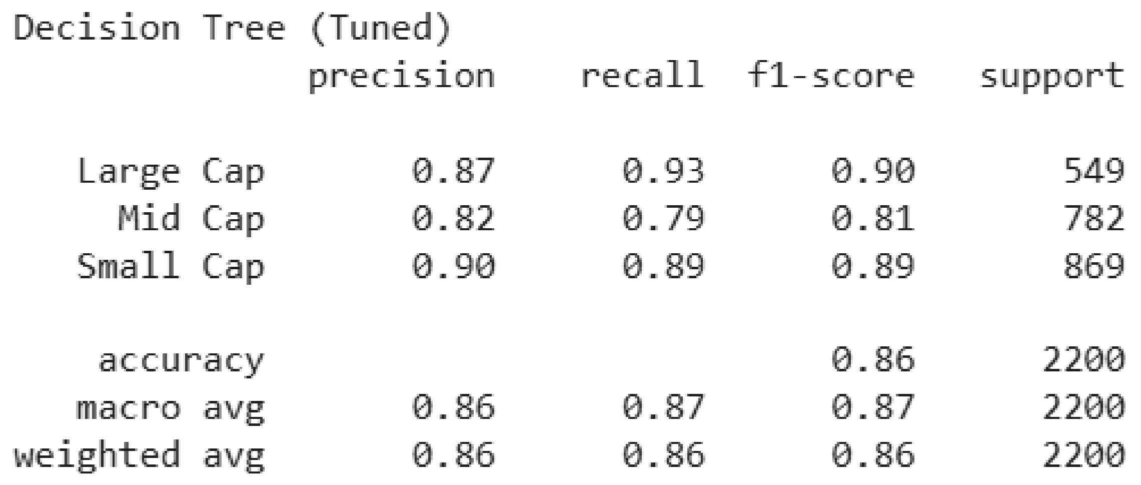
- The **tuned DTC** delivered the best performance, outperforming Logistic, KNN, and SVM models.
- Hyperparameter tuning improved generalization and reduced overfitting, ensuring stable classification across all market-cap categories.
- **Revenue and Profit Margin** are the most dominant predictor
- **Growth Rate and ESG** Overall play a supporting role, indicating ESG factors enhance

CONFUSION MATRIX & CLASSIFICATION REPORT INSIGHTS



Confusion Matrix – Decision Tree (Tuned)

True label \ Predicted label	Large Cap	Mid Cap	Small Cap
Large Cap	509	39	1
Mid Cap	76	619	87
Small Cap	1	93	775



```
Decision Tree (Tuned)
              precision    recall  f1-score   support

   Large Cap       0.87      0.93      0.90       549
     Mid Cap       0.82      0.79      0.81       782
   Small Cap       0.90      0.89      0.89       869

    accuracy                           0.86      2200
   macro avg       0.86      0.87      0.87      2200
weighted avg       0.86      0.86      0.86      2200
```

Large Cap

- Out of 549 Large-cap firms, 509 are correctly classified.
- Recall of 0.93 shows the model is very effective at identifying large firms.

Mid Cap firms

- 619 out of 782 Mid-cap firms are correctly classified.
- Mid-cap has the lowest recall (~0.79), making it the hardest class to predict.
- **Misclassifications** mainly occur into Small or Large Cap, which is expected due to **overlapping firm sizes**.

Small Cap

- **775 out of 869** Small-cap firms are correctly identified.
- High Precision (0.90) and Recall (0.89) indicate very low risk of misclassifying small firms.
- Small-cap firms are rarely predicted as Large Cap, showing strong separation

Overall Model Insight

- The tuned Decision Tree achieves **~86%** accuracy with balanced precision and recall.
- Most errors occur between adjacent categories, not extreme ones (Small & Large).

BUSINESS APPLICATIONS OF CLASSIFYING MARKETCAP

Investment Screening: Helps quickly group companies by size for portfolio selection and risk assessment.

Growth Tracking: Helps identify firms that may move from Small to Mid or Mid to Large Cap over time.

Strategic Benchmarking: Allows fair comparison of companies operating at a similar market scale.

Resource Allocation: Supports decisions on capital investment, strategy, and sustainability focus based on company size.

INTERACTIVE MARKETCAP FORECASTING TOOL

BUSINESS VALUE OF THIS TOOL

1 Enables quick valuation estimates without running complex financial models.

2 Helps decision-makers understand how changes in revenue, margins, or sustainability efforts impact market value.

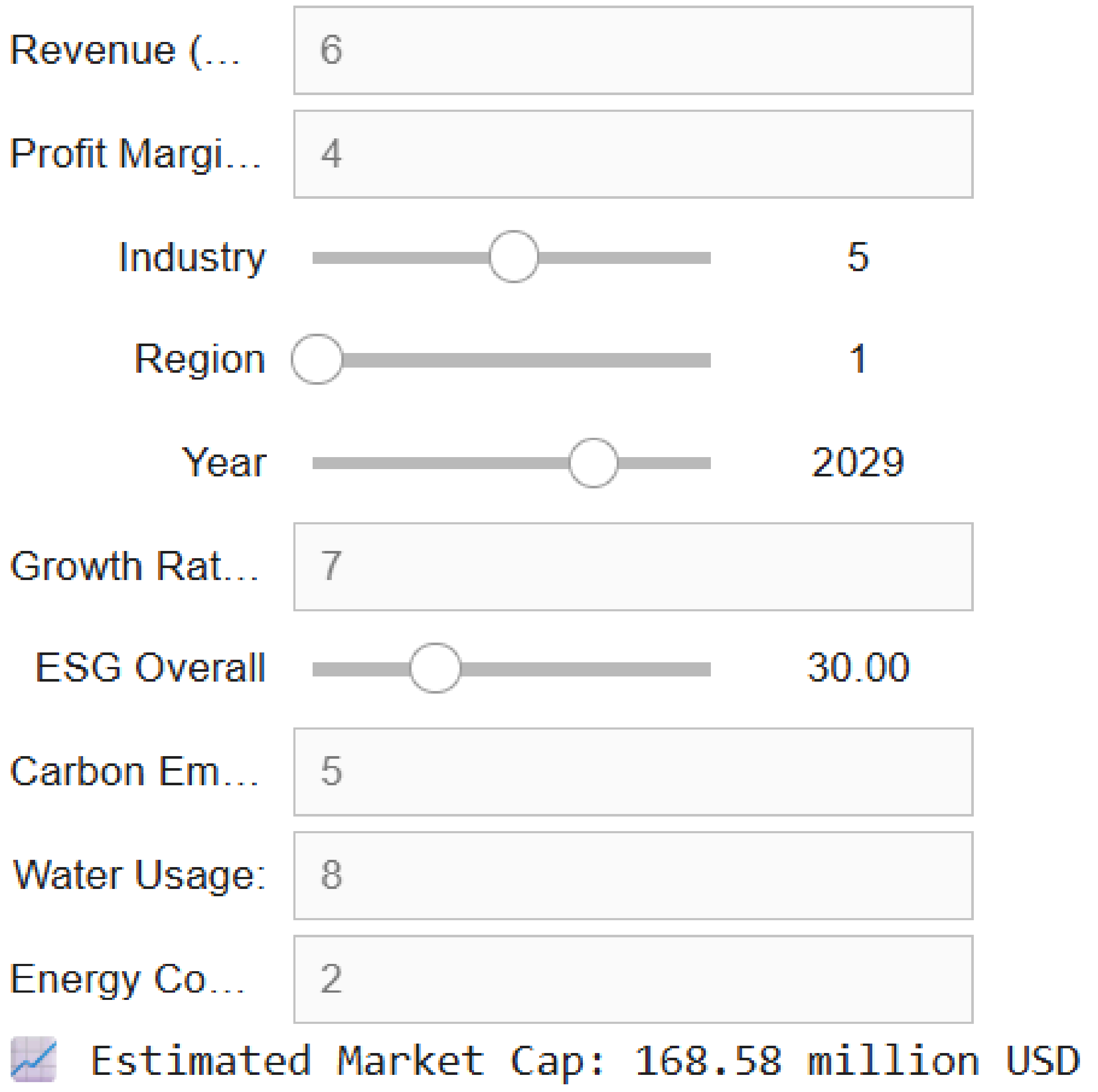
3 Makes the model actionable, not just analytical.

PRACTICAL USE CASES

1 ***Investment Analysis:*** Test valuation impact under different growth or margin scenarios.

2 ***Strategic Planning:*** Assess how operational or ESG improvements may influence future valuation.

3 ***Management Decision Support:*** Compare outcomes across industries, regions, or time horizons.



Revenue (…	6
Profit Margi…	4
Industry	5
Region	1
Year	2029
Growth Rat…	7
ESG Overall	30.00
Carbon Em…	5
Water Usage:	8
Energy Co…	2

```
📈 Estimated Market Cap: 168.58 million USD
```

THANK YOU