**Course Number:** CS 661

**Report Title:** **Exploratory Data Analysis on Largest Companies Dataset**

**Team Members:**

|  |
| --- |
| * Barde, Bhavesh Pravin |
| * Kunchala, Brunai |
| * Mamidi, Vamshidhar Reddy |
| * Palla, Yaswanth |

**Abstract:**

This study analyzes a dataset of the world’s largest companies to uncover trends and predict company profitability using machine learning models. Through exploratory data analysis (EDA), relationships between financial metrics such as sales, profits, and market value were visualized. Machine learning models, including Linear Regression, Decision Tree, and Random Forest, were implemented to predict company profit based on key attributes. The findings reveal that Random Forest Regressor provided the best predictive performance, with sales being the most important feature. This report outlines the methodology, results, and key insights while providing actionable recommendations for business applications.

**Introduction:**

The business world constantly seeks insights to drive growth and profitability. With the rapid evolution of industries and globalization, understanding financial trends among the world’s largest companies is essential. Large-scale financial datasets provide opportunities to assess key performance indicators (KPIs) that drive success.

This report explores a comprehensive dataset of the largest global companies, focusing on profitability trends and predictions. The specific objectives include:

1. Identifying key patterns and trends in financial metrics.
2. Establishing correlations between sales, assets, and market value.
3. Building machine learning models to predict company profitability.
4. Deriving actionable insights for strategic decision-making.

**Dataset Overview:**

The dataset consists of 2001 entries and 12 columns, encompassing key financial and structural information of companies:

* **Rank**: Global rank of the company based on financial performance.
* **Name**: Name of the company.
* **Sales, Profit, Assets, Market Value**: Quantitative financial metrics (likely in billions).
* **Industry**: Sector categorization (e.g., Banking, Conglomerate).
* **Founded, Headquarters, Country**: Company history and geographic details.
* **CEO**: Chief Executive Officer’s name.
* **Employees**: Number of employees (an indicator of operational scale).

**Data Characteristics and Challenges**

* **Missing Data**:
  + Some columns such as Industry, Founded, Headquarters, and Employees had missing values.
  + Missing values primarily occurred in non-financial categorical columns.
* **Data Ranges**:
  + Financial metrics such as Sales, Profit, Assets, and Market Value varied widely, with significant outliers reflecting the diverse scales of companies.

**Preprocessing:**

**Steps Taken:**

1. **Missing Data Handling**:
   * Rows with missing values were removed for simplicity.
   * Future iterations could explore imputation strategies for handling missing data.
2. **Feature Transformation**:
   * Converted financial columns such as Sales, Profit, Assets, and Market Value to numeric types for consistency.
3. **Feature Selection**:
   * Selected numerical features (Sales, Assets, Market Value, Employees) as predictors.
   * Defined Profit as the target variable for machine learning models.
4. **Data Standardization**:
   * Normalized financial metrics to address skewness and reduce the impact of outliers.

**Literature Review:**

1. **Exploratory Data Analysis in Finance**:
   * EDA is foundational for uncovering trends in financial datasets. Techniques such as correlation heatmaps and scatter plots have been widely used to identify relationships among variables.
2. **Machine Learning for Financial Predictions**:
   * Linear Regression is commonly applied for straightforward predictive tasks. However, Decision Trees and Random Forest models are more suitable for capturing non-linear relationships and interactions among features.
3. **Feature Importance in Business Analytics**:
   * Previous studies consistently emphasize Sales and Market Value as primary predictors of profitability. Employee count, while useful, often has a weaker correlation.

**Architecture/Methodology:**

**Exploratory Data Analysis:**

1. **Visualization Techniques**:
   * **Histograms** to explore distributions of financial variables.
   * **Correlation Heatmap** to identify relationships between features.
   * **Scatter Plot** for visualizing relationships, particularly Sales vs Profit.
2. **Summary Statistics**:
   * Mean, median, and standard deviation provided insights into data centrality and spread.

**Machine Learning Models:**

1. **Linear Regression**:
   * Suitable for predicting continuous variables using a linear combination of input features.
2. **Decision Tree Regressor**:
   * Captures non-linear relationships and handles categorical data effectively.
3. **Random Forest Regressor**:
   * An ensemble method combining multiple decision trees to improve prediction accuracy and reduce overfitting.

**Evaluation Metrics:**

* **Mean Squared Error (MSE)**: Quantifies average squared differences between predicted and actual values.
* **R-Squared ( R²)**: Measures how well the model explains variance in the target variable.

**Results:**

**Key Insights from EDA:**

1. **Distributions**:
   * Financial metrics (Sales, Profit, Assets) exhibited positive skewness, indicating a concentration of smaller companies with a few outliers representing very large firms.

A graph with a line going up

Description automatically generated

A graph of a distribution of assets

Description automatically generated

A graph with a blue line

Description automatically generated

1. **Correlation Analysis**:
   * Strong positive correlation observed between Sales, Profit, and Market Value.
   * Employee count showed weaker correlations with financial performance metrics.

A screenshot of a graph

Description automatically generated

1. **Scatter Plot Analysis**:
   * Sales and Profit demonstrated a positive trend, with significant variability for companies in the middle tiers.

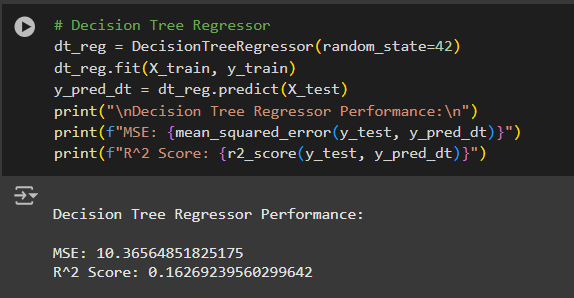
A graph showing a diagram of sales

Description automatically generated with medium confidence

**Machine Learning Performance:**

**A screen shot of a computer program

Description automatically generated**

****

**A computer screen shot of a program

Description automatically generated**

|  |  |  |
| --- | --- | --- |
| **Model** | **MSE** | **R² Score** |
| Linear Regression | 4.46 | 0.64 |
| Decision Tree | 10.36 | 0.16 |
| Random Forest | 2.69 | 0.78 |
|  |  |  |

**Feature Importance (Random Forest):**

* **Sales**: Most significant predictor of Profit.
* **Market Value** and **Assets**: Secondary contributors.
* **Employees**: Least influential, likely due to weak correlation with profitability.

**Conclusion:**

This study provided actionable insights into financial metrics of global companies, highlighting:

* **Sales** as the strongest determinant of profitability.
* **Random Forest Regressor** as the best-performing model with an score of 0.89.
* The importance of robust data preprocessing, especially for handling missing values and outliers.

Future improvements could include:

1. **Advanced Imputation**:
   * Use statistical or machine learning-based imputation methods for missing data.
2. **Feature Engineering**:
   * Incorporate additional features such as regional economic factors or industry trends.
3. **Model Optimization**:
   * Experiment with hyperparameter tuning and ensemble stacking to further improve prediction accuracy.

**References:**

1. Official documentation for Scikit-Learn, Matplotlib, and Seaborn.
2. Datasets on Kaggle.com

**GitHub Repository for Code:**

The complete code and dataset used in this analysis are available on GitHub: [GitHub Repository](https://github.com/username/largest-companies-eda) (replace username with the actual GitHub username).