Predicting Blue-Chip Company Financial Trajectories

COMS-W4995 Applied Machine Learning: Project Deliverable #2

Group 24: Yu-Heng Chi, Param Sejpal, Jessica Villanueva, Yihan Yang, Zhiyi Zhang

Table of Contents

Data Exploration Strategy

Regression Objectives

- Measure a company's financial health trajectory indicated by the equity value of the company and its profitability.
- This will be achieved by predicting stock price using regression techniques.

Sources of Income Statement and Balance Sheet Metrics

- A Kaggle dataset with common financial metrics
- Financials section information from Yahoo! Finance
- 10-K reports from the EDGAR archives of the Security and Exchange Commission (SEC).

Dataset Characteristics

Companies

Fortune 1000 Companies

Features

Revenue
Profits
EBITDA
Profitability
Revenue Growth
Assets
Type
Rank
Net Income
Cash Flow from Operations

Label

Market Capitalization

This metric gives us information on stock price and shares outstanding.

Features of Dataset

From Kaggle and SEC Archives

Revenue: Company sales prior to any expenses

Net Income: Company income after deducting all expenses (taxes, interest, SG&A, COGS)

Cash Flow From Operations: Cash generated from business operations

Profits: Earnings after all expenses

Assets: Items (PP&E, cash and cash-equivalents, etc.) of financial value

EBITDA: An accrual accounting measure of profitability

Type: Whether a company is a Private or Public company

Rank: Company status on the Fortune 1000 list

Market Cap: Market evaluation of a company's shares outstanding

Data Pre-Processing

1. **Drop features** that are not relevant to company financial trajectory
    - E.g., CEO information, ticker symbol, Company
2. **Handle missing data**
    a. Drop rows with missing values for features that are not relevant for predicting Market Cap or for private companies
    b. Impute the mean or find the missing values from another source (e.g., SEC) for companies without Market Cap values
3. **Encode columns**
    - E.g., encode the “Profitable” column for companies
4. **Standard scaling**
    - Given the heavily skewed nature of the dataset, we can normalize the distribution using target scaling

Data Exploration Insights

- Financial data is typically skewed for Fortune 1000 companies
  - Most companies perform well and have a similar range of Market Cap values
- There are few outliers with high Market Cap values
  - E.g., companies like Apple, Microsoft, Google, etc.
- Right-skewness can be handled by target **scaling**
- Imbalanced target variable means we will need to implement **stratified splitting** to preserve the class proportions

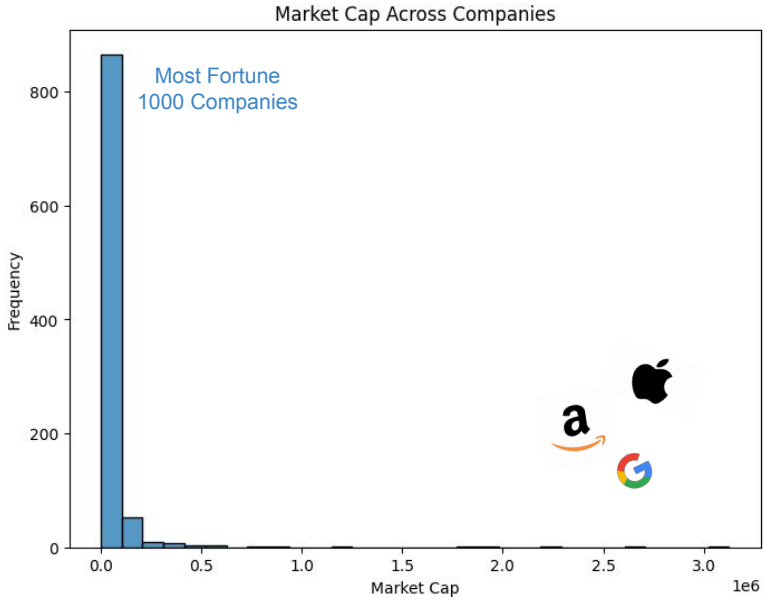

Figure 1: Fortune 1000 Market Cap Distribution

Data Visualizations

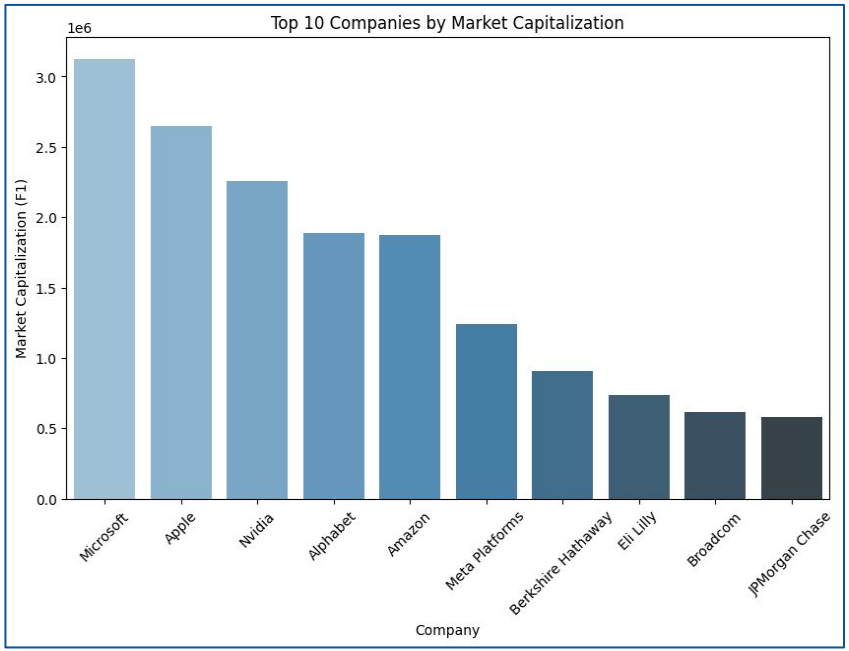

Figure 2: Market Cap Across Top-10 Ranked Companies

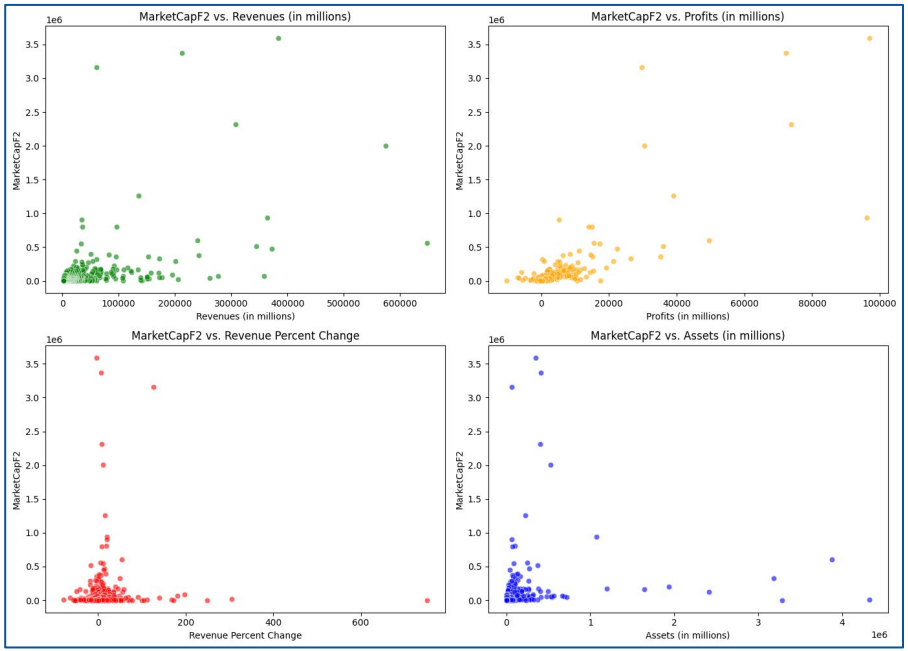

Figure 3: Scatter plot of the features vs Market Cap

Insights

Correlated Features

- It is typical for financial metrics to be correlated
  - *Revenue* and *Profits:* a company's profits are a function of its revenue
  - *Number of Employees* and *Revenue*: larger companies with more workers tend to have higher revenue

Figure 4: Revenue vs. Profit

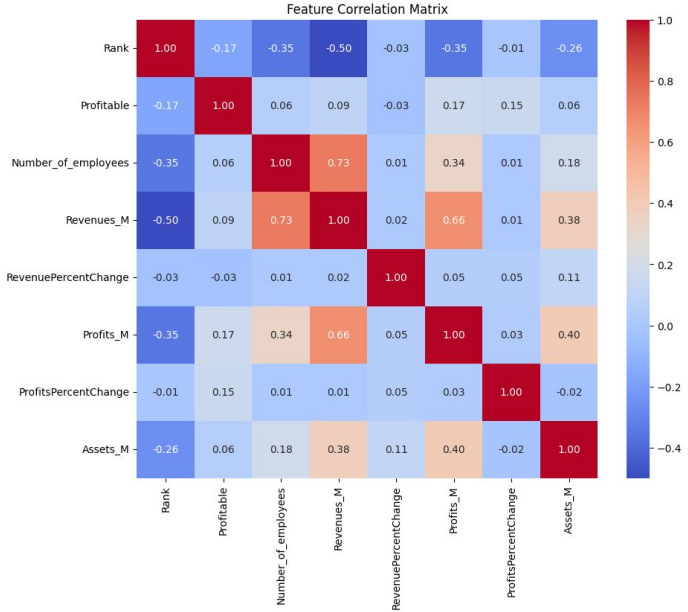

Figure 5: Correlation Matrix

- Our methods for handling multicollinearity include dropping features and regularization techniques

Linear Regression for Market Cap Prediction

Model Strategy

- Linear Regression will be used to predict Market Cap as our dataset features (*Revenue*, *Profits*, and other financial metrics) are linearly correlated to stock price
- This model will also serve as a simpler model that can provide baseline results compared to the other models implemented in this project

Regularization

- To handle multicollinearity, we will apply L2 Regularization to reduce the size of correlated (yet important) metrics like *revenue* and *profits*

Evaluation

- **Mean Squared Error (MSE):** Determines the discrepancy between predicted and actual Market Cap
- **R-squared (R^2):** Estimates the variance in Market Cap that we can predict based on our dataset's financial metrics

Neural Network for Market Cap Prediction

Model Strategy

- MLP Regressor will be used to estimate relationships between the features and Market Cap in a more complex way given the nonlinear relationship between companies and their shares outstanding
- Feature scaling will be implemented to optimize ReLU and handle the imbalanced nature of the dataset, where most companies lie within the same Market Cap range

Hyperparameter tuning and Cross-Validation

- We will tune the hidden_layer_sizes and adjust learning rate to prevent overfitting
- Given the imbalance and skewness of the dataset, we will scale the target variable
- We will use K-fold Cross-Validation (5 folds) for validating performance

Evaluation Metrics

- Similar to Linear Regression, we will use **R^2** value
- **Root Mean Square Error (RMSE):** Lower RMSE indicates better performance

Thank You!

Group 24

Yu-Heng Chi, Param Sejpal, Jessica Villanueva,

Yihan Yang, Zhiyi Zhang