Comprehensive Investment Analysis of Top 20 US Companies: 2003-2022

DSCI 510: Principles of Programming for Data Science

Final Project Report

Christopher Bou Saab & Pardi Bedirian

Table of Contents:

1) Introduction
2) Data Collection
3) Data Cleaning
4) Data Analysis
5) Data Visualization
6) Future Work

1. Introduction

This project aims to provide a comprehensive investment analysis of the top 20 US companies by market capitalization over a two-decade period, from 2003 to 2022. The analysis focuses on evaluating the financial performance, stock market trends, and the effects of overall economic indicators on these leading companies.

A pivotal aspect of this analysis is the computation of an aggregate score for each company in a specific year. This score is derived by taking into account all the aforementioned financial and economic factors, allowing for a comprehensive assessment of each company's performance. Based on the aggregate scores, companies are assigned to specific tiers for each year, representing their relative performance and stability. Building on this classification, a predictive model is developed and trained using this data. This model is designed to predict the 'Final Score' or tier of a specific company in a given year, offering valuable foresight into potential future performances. This analytical approach and predictive modeling are crucial for investors, financial analysts, and market observers.

2. Data Collection

This section details the diverse data sources and collection methods utilized in the study.

a. ***Market Capitalization Data:*** Market capitalization data, crucial for identifying the top 20 US companies, was obtained from a CSV file named 'marketcap.csv'. This csv was downloaded from companiesmarketcap.com and modified before usage in the project. This dataset in the csv was used to sort and select the top companies based on their market capitalization."
b. ***Company Financial Data:*** Financial data for the selected companies, including earnings per share (EPS), revenues, and net income loss, were gathered from the Securities and Exchange Commission (SEC) filings that were accessed by API.
c. ***Economic Indicators:*** Economic indicators such as GDP growth, inflation rate, and unemployment rate were sourced from the Federal Reserve Economic Data (FRED) database using the 'fredapi' Python library. These indicators provide a broader economic context affecting the market and companies.
d. ***CIK Numbers and SEC Data:*** CIK (Central Index Key) numbers, essential for accessing detailed financial reports from the SEC, were retrieved using a web request to the SEC's official website. This allowed for the collection of specific financial data directly related to each company.
e. ***Stock Price Data:*** Historical stock price data was downloaded using the yfinance library. This data included various price metrics such as opening, closing, high, and low prices, enabling the analysis of stock price trends over the years.

3. Data Cleaning

All the collected data was then integrated and prepared for analysis. This involved merging, formatting, and structuring the data into a consistent format suitable for the subsequent stages of cleaning, analysis, and visualization.

This section outlines the critical data cleaning and preprocessing steps undertaken to ensure the accuracy and reliability of the data used in our analysis.

- **Market Capitalization Filter:** The initial step involved sorting the market capitalization data (sourced from 'marketcap.csv') by rank. We selected the top 20 companies from the data for our analysis.
- **Initializing Dataframes:** For each of the top 20 companies, individual dataframes were initialized for the years 2003 to 2023. These dataframes included placeholders for various financial metrics such as stock price, market cap, revenue, EPS, and others, which were initially set to zero.

- **Reindexing and Sorting:** The dataframes were sorted based on the companies' market capitalizations for each given and reindexed. This reindexing facilitated easier data manipulation and integration in subsequent steps.
- **Handling Missing Values:** In the process of retrieving financial data from different sources, missing values were encountered. These were managed by setting respective fields to 'None' when data was unavailable, ensuring consistency across the dataframes.
- **Adjusting Time Series Data:** For stock market data retrieved via yfinance, adjustments were made to align data points with their corresponding years. This involved handling time zone differences and ensuring that data for each year correctly reflected the market status at that year's end.
- **Updating Financial Metrics:** The company-specific dataframes were updated with actual financial data, including 'EPS (Earnings per Share)', 'Revenues', and 'NetIncomeLoss'. This step involved systematically replacing the initial placeholder values with the real data obtained from the various sources.
- **Calculation of Key Metrics:** We computed crucial financial ratios and growth metrics, such as P/E Ratio and Revenue Growth, directly within the dataframes. This involved writing custom functions to calculate these metrics based on the available financial data.
- **Consolidation for Analysis:** Finally, the cleaned and updated company-specific dataframes were consolidated into a single, comprehensive dataframe. This unified dataframe served as the foundation for the subsequent data analysis, visualization, and modeling stages."

4. Data Analysis

This section delves into the methodologies and outcomes of our analysis, focusing on predicting the investment potential of top US companies. Predicting the investment potential of Microsoft for the year 2023 was used as an example. Our approach combines historical data analysis with future projections to offer a robust method for assessing investment potential in the US stock market.

1) Model Selection:
   a) Rationale for Model Choice
      - The RandomForestClassifier was chosen due to its proficiency in handling complex datasets and its capability to classify the target variable – the 'Final Score' – into five distinct classes. These classes represent varying levels of investment recommendation, with 5 indicating a strong recommendation to invest and 1 suggesting a more cautious approach.
2) Model Training and Evaluation:
   a) Training Process
      - The RandomForestClassifier was trained on standardized data, encompassing a wide range of financial metrics and economic indicators for each company over the past 20 years.

b) Evaluation Metrics
  - The model's performance was evaluated based on accuracy, supported by a detailed classification report. These metrics provided insights into the model's effectiveness in correctly predicting the 'Final Score' for each company.

3) Future Performance Forecasting:

a) Utilization of ARIMA Model
  - To forecast the future performance of Microsoft for the year 2023, we employed the ARIMA model. ARIMA is particularly effective in capturing historical trends and patterns, making it ideal for projecting financial and economic features of a company into the future.

4) Predictive Analysis:

a) Application of RandomForestClassifier
  - Leveraging the trained RandomForestClassifier, we predicted Microsoft's 'Final Score' for the year 2023. This prediction was based on the forecasted financial and economic data generated by the ARIMA model.

b) Outcome of the Analysis
  - The outcome of this predictive analysis provided an estimated investment potential for Microsoft in 2023. This estimation is grounded in both historical data analysis and projected future performance, offering a comprehensive view of the company's investment potential.

5) Insights and Implications:

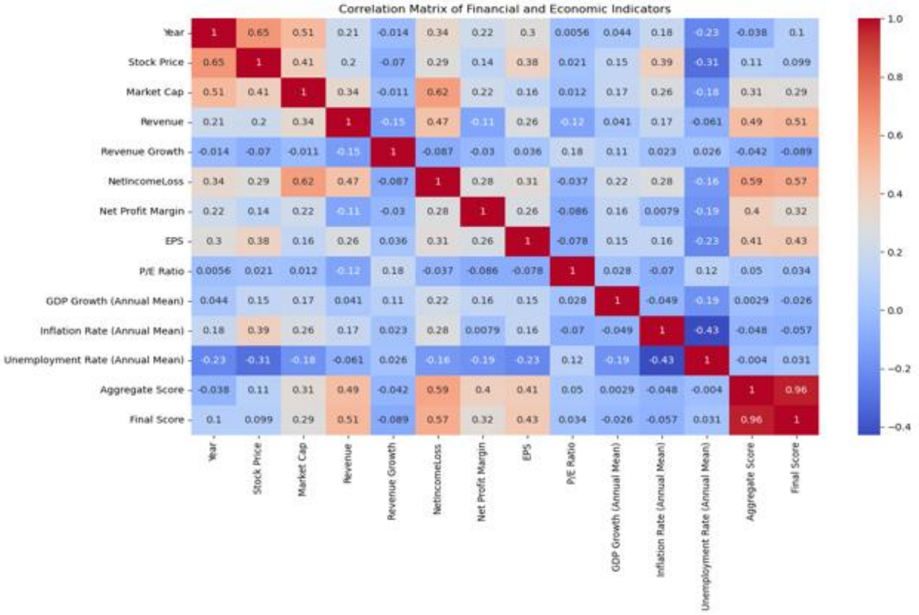
a) Interpretation of Results
  - The analysis yielded significant insights into the financial health and potential growth of Microsoft, reflecting both the company's historical performance and forecasted future trends.

b) Investment Implications
  - These insights have substantial implications for investors, providing a data-driven basis for investment decisions. The tier classification system and the 'Final Score' serve as valuable tools in evaluating the attractiveness of an investment opportunity in the context of the broader market and economic conditions.

5. Data Visualization

This section showcases the data visualization techniques employed to interpret and present the complex data collected and analyzed in this report. Two primary visualizations were created: 1) Correlation Matrix and 2) Line Chart of Final Scores. For stakeholders, these visualizations serve as tools for quickly assessing the investment potential and risks associated with these top companies. They distill complex datasets into understandable formats, enabling informed decision-making.

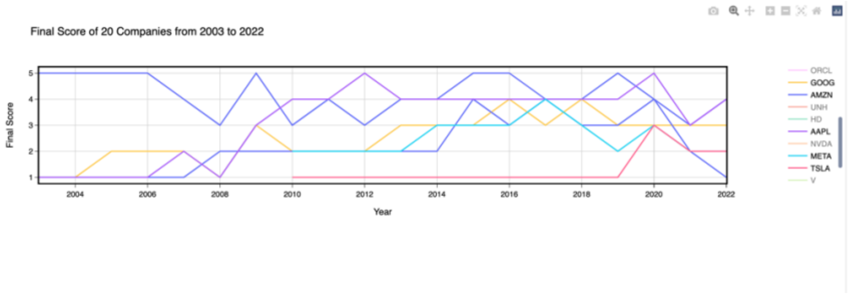
1) Correlation Matrix of Financial and Economic Indicators:
   a) Purpose and Design:
      - The Correlation Matrix was designed to identify and illustrate the relationships between various financial and economic indicators. Each cell in the matrix provides the correlation coefficient between two specific indicators, ranging from -1 to 1.
   b) Interpretation of Results:
      - Values close to 1 indicate a strong positive correlation, suggesting that as one indicator increases, the other tends to increase as well. Conversely, values near -1 indicate a negative correlation.
        Example: The 'Aggregate Score', a composite metric derived from the analyzed indicators, shows a strong positive correlation with the 'Final Score' (0.96). This strong relationship underscores the 'Aggregate Score's effectiveness in summarizing overall investment potential.



Key Observations:

1) Stock Price Correlation: There is a notable positive correlation between 'Stock Price' and 'Market Cap' (0.65), indicating that as the market capitalization of a company increases, the stock price tends to rise.
2) Revenue and Market Cap: 'Revenue' shows a moderate positive correlation with 'Market Cap' (0.34), suggesting that higher revenue is often associated with a larger market capitalization.
3) Economic Indicators:
   a) ‘GDP Growth' and 'Inflation Rate' show a moderate positive correlation with 'Market Cap' and 'Stock Price', indicating that economic growth and inflation may influence market values.
   b) 'Unemployment Rate' exhibits a negative correlation with 'Stock Price' (-0.31), which could imply that higher unemployment rates might negatively affect stock prices.

2) Line Chart of Final Scores from 2003 to 2022:
   a) Purpose and Design:
      - The Line Chart visualizes the 'Final Score' of the top 20 companies over the 20-year period. Each company is represented by a line, with the x-axis denoting the year and the y-axis indicating the 'Final Score'. This chart provides a clear visual representation of how each company's score has evolved over time.
   b) Interpretation of Results:
      - By examining the trajectories of the lines, stakeholders can discern trends and patterns in company performance. For example, a consistently high 'Final Score' over several years may indicate a strong and stable investment opportunity, while fluctuating scores may suggest volatility or a change in company fortunes.



Key Observations:

1) Companies in the technology sector, such as Google, Amazon, Apple, NVIDIA, Meta, and Tesla, show strong performance and an increasing trend in 'Final Score', signifying the sector's growth and its appeal to investors.
2) The post-2008 period demonstrates recovery and improvement for most companies, suggesting resilience and adaptability in the face of economic challenges.
3) High-performing companies like Microsoft, Apple, and Google could be considered reliable anchors for long-term investment strategies.
4) The rise of Tesla's score in recent years could indicate a shifting landscape with new opportunities in emerging markets like electric vehicles.

6. Future Work

Given more time, we would like to concentrate on analyzing and improving the model to achieve the highest possible accuracy.