A Clustering and Regression Analysis of Financial Health and Stock Performance

5579267

1 Introduction

The world of finance is evolving at a fast pace today. In such a market, it is crucial to evolve according to the market conditions. Analyzing corporate financial metrics and stock performance is gaining rapid interest. Prediction of stocks has been one great and interesting topic for investors. This report provides an insight into the comprehensive relationship between corporate financial health and historical stock prices. The analysis is performed on New York Stock Exchange (NYSE) data. One of the purposes of this analysis is to make investors aware of the important aspect of making risk-free investments. This aspect is the financial fundamentals of the company.

2 Abstract

Previous works have been focused mainly on performing either a qualitative or quantitative analysis. Back et al. (2001) clustered the companies based on the quantitative and qualitative information in the annual reports. They compared the resultant clusters and suggested that the performance of considering both quantitative and qualitative information is better than that of using just quantitative or qualitative information. One other such notable work is "Financial Ratios and the Probabilistic Prediction of Bankruptcy" by Ohlson (Ohlson 1980). Olhsen's work established a link between financial ratios and a company's risk of bankruptcy, which is a fundamental aspect of financial resilience.

In this analysis, we dive into two datasets from the New York Stock Exchange data fetched from Kaggle, namely the fundamentals and prices dataset (Gawlik 2017). A wide array of financial metrics for numerous companies is encompassed in the *fundamentals.csv* dataset and stock price movements over time are included in the *prices-split-adjusted.csv*.

This analysis follows a bilateral approach: clustering companies based on their financial fundamentals which reveal grouping that reflects various financial profiles according to financial metrics and conducting regression analysis on their stock prices to understand and predict stock prices. The relationship between financial ratios and stock performance provides a theoretical baseline for our clustering analysis as studied by Fama and French (1992) and Ohlson (1980). The use of time-series for predictive analysis in the work of Box and Jenkins (1976) provides the groundwork for our regression modeling. Clustering in the financial domain, particularly focusing on K-Means as explored by Huang (2011) aligns closely with our approach.

Key findings from the clustering of companies revealed that companies could be segmented into groups that share financial characteristics. These characteristics could range from high-growth companies to declining companies, which can help potential investors identify such companies and base their investment decisions on these insights. These clusters were further examined through a regression model which is applied to the stock prices dataset. The regression model is built on predicting the stock prices based on the previous day's closing values.

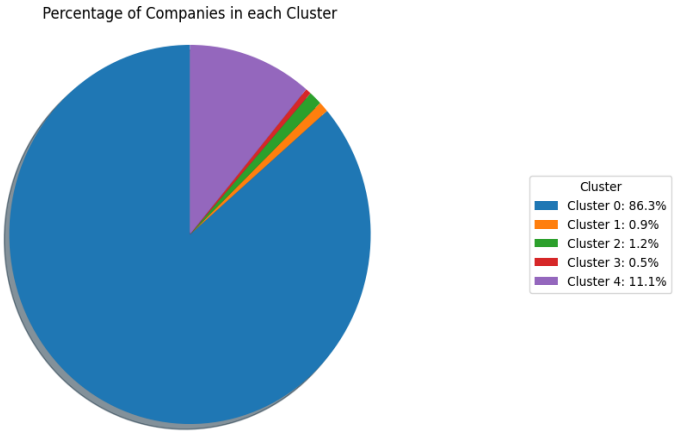

Figure 1: Percentage of Companies in Cluster

Furthermore, a detailed comparative analysis of two mega-cap technology companies, Apple (AAPL) and Nvidia (NVDA) provided great insights. The analysis is based on finding out the financial resilience to the

market downturns based on the clustering analysis. The main research question we are seeking to find out between these two companies is which company appears to be more financially resilient to market downturns.

3 Description

3.1 Source

The dataset that is being used in this analysis is sourced from the New York Stock Exchange (NYSE) dataset which is available on Kaggle. It can be found at the following URL: Kaggle NYSE Dataset.

Kaggle is a popular platform for data science and machine learning enthusiasts which is often known for hosting competitions and providing datasets for analysis and research. Kaggle datasets provide a diverse collection of datasets which is mostly uploaded by users and organizations. These datasets can range from small-scale data for beginner projects to large, complex datasets that are useful for in-depth research and machine learning model training.

3.2 Description of Dataset

The main motivation behind this dataset is to create a playground for fundamental and technical analysis of stocks. More and more trading and investment are being shifted towards automation and machines. If we want to fully automate our trading and make informed decisions, we need to learn from the historical data.

The dataset consists of four files - *prices.csv*, *prices-split-adjusted.csv*, *securities.csv*, and *fundamentals.csv* which in total is about 32MB in size. Amongst them, two are being used for our analysis - *fundamentals.csv*, and *prices-split-adjusted.csv*.

- **fundamentals.csv** :
  - Captures all the financial fundamentals and metrics of companies listed on the NYSE.
  - Contains 1781 entries and 79 attributes.
  - Includes various financial indicators such as 'Earnings Before Interest and Tax (EBIDTA)', 'Net Income', 'Earnings Per Share', 'Net Income', etc.
  - The fundamentals were fetched from Nasdaq Financials and the metrics were extracted from the annual SEC fillings (Gawlik 2017).
- **prices-split-adjusted.csv** :
  - Contains the daily stock prices (open, close, high, low, volume) of the companies listed on the NYSE ranging from 2010 to the end of 2016.
  - There have been around 140 splits between the range 2010 and 2016. The prices in this dataset have been adjusted for splits.
  - Contains 851263 entries and 7 columns.
  - Includes daily trading information such as opening price, closing price, lowest daily price, highest daily price and volume of the daily traded stock.
  - The prices were fetched from Yahoo Finance.

These two datasets were chosen particularly because of several important reasons:

The fundamentals file provides a variety of financial metrics, such as earnings, assets, operating income, etc. These indicators are important for evaluating a company's financial health and performance. The wide variety of data allows us to perform a thorough analysis of these financial fundamentals across several metrics such as liquidity, resilience, and profitability. The prices-split-adjusted file contains time-series data on stock prices such as opening price, closing price, etc. These are essential for performing stock market analysis or predictions. The prices-split-adjusted data accounts for stock splits to ensure the stock prices reflect the true value and reduce ambiguity. The dataset also spans multiple years, which allows us to interpret insights by performing analysis on this data. The dataset also covers a wide range of companies spanning across different sectors providing a rich information set for analysing a broader market view. The dataset reflects actual real-world data points and is representative of the actual market, which enables us to provide findings from the analysis that can be practically provided to investors or other stakeholders.

3.3 Data Pre-Processing

- The dataset had no duplicate records and relatively fewer missing values in some columns.
- These columns were left as it is until we chose the highly correlated columns pertaining to our analysis.
- These were then dropped if they were still in the list of highly correlated columns.
- Apart from that, there was a need to convert the date attribute in the datasets to date-time datatype from string.

This analysis can be replicated by downloading the dataset from the provided Kaggle link and following the above-mentioned pre-processing. For the fundamental analysis, it is essential to select the relevant financial fundamental indicators that correlate to the company's profitability and liquidity.

4 Result

4.1 Clustering Analysis

4.1.1 Methodology

Clustering analysis begins with selecting highly correlated metrics from the fundamentals dataset and filtering it to contain only these columns. The dataset is then cleaned to ensure missing values are handled. This is carried out using imputation which replaces missing data with estimates based on the present data. Then the data is standardized to bring all the values to a comparable scale. This step is crucial to ensure that one single attribute is not being heavily weighted in the model.

Clustering is carried out using the K-Means clustering algorithm with a number of clusters equal to 5 to segment the companies into clusters based on similarities in their financial fundamentals. After clustering, the cluster centers were scaled back to their original dimensions to interpret the mean values of financial metrics within each cluster.

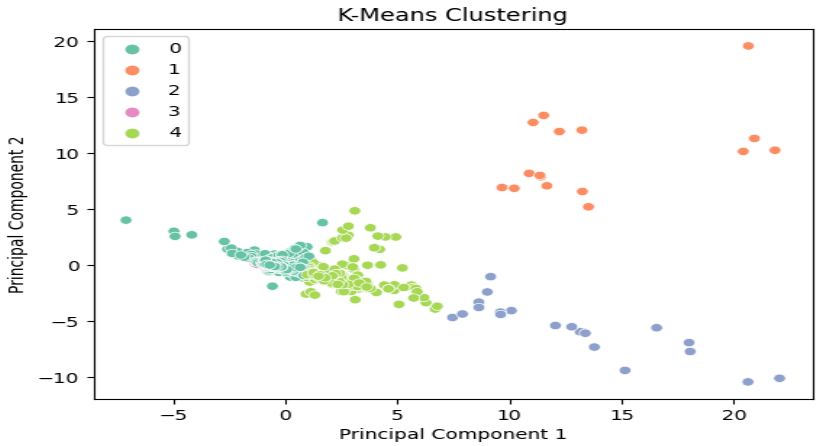

Figure 2: Clusters

4.1.2 Analysis of Clustering Results

The clustering has led to distinct groups of companies based on key financial metrics as evidenced by the pie chart in Figure 1. The majority of companies fall into a single cluster, which reveals a common financial pattern. The smaller clusters show the presence of companies with distinctive financial characteristics, possibly indicating unique markets or unique business strategies.

A particular cluster ***(Cluster 2)*** distinguished by a high 'After Tax ROE' and 'Capital Expenditures' points to companies with exceptional profitability relative to their equity and such companies making big strides in expanding their presence. This provides great incentives for investors seeking efficient capital use and capital expansion for higher growth in the future.

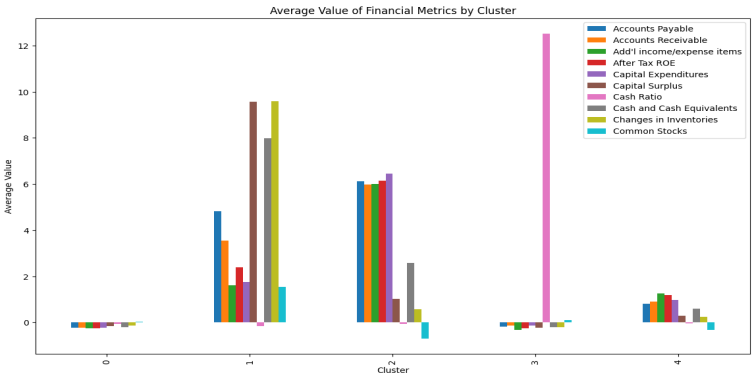

Figure 3: Financial Metrics by Cluster

The elevated 'Cash Ratio' cluster ***(Cluster 3)*** highlights a cluster that includes companies that can cover their short-term debts using its cash resources. Investing in such companies can provide the investors with a certain guarantee that the companies contained within this cluster are able to sort their obligations. Another cluster ***(Cluster 1)*** contains mainly non-operational activities, which highlights companies with significant investment activities.

Revenue analysis provides further insight into these clusters highlighting the stark differences in revenues across these clusters which can be seen in Figure 3. This can be accounted for by the fact that clusters are being differentiated in terms of company sizes or market demand for individual companies.

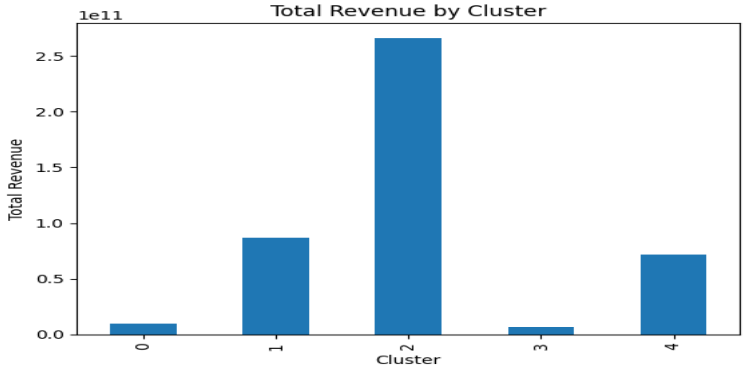

Figure 4: Revenue by Cluster

These analytical insights can potentially serve as a robust foundation for strategic planning, and risk management and for identifying opportunities for investment by investors. The application of the K-Means clustering algorithm has successfully divided companies into meaningful clusters.

4.2 Regression Analysis

4.2.1 Methodology

A Linear Regression model was utilized for the prediction of current stock prices based on the previous day's closing prices. For simplicity of the analysis, we focused on two mega-cap stocks, Apple *(AAPL)* and Nvidia *NVDA*. A rolling window prediction model was utilized for forecasting the stock prices. The model's was retrained at each step with the most recent data (within the date ranges of the dataset) to simulate a real-world trading scenario.

4.2.2 Analysis of Regression Results

The model captures the overall trend in stock prices and manages to align closely with the actual prices of the selected stocks with greater accuracy. The models performance was evaluated using RMSE which was 1.118. The alignment is particularly noteworthy in stable market situations, whereas the predictions are slightly erroneous in volatile market conditions. Such deviations show that there is a presence of external factors that might not be captured by the model.

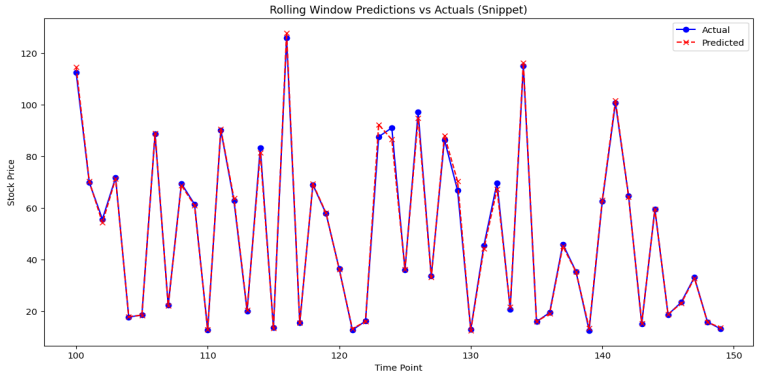

Figure 5: Regression predictions

These predictive insights can be instrumental in devising trading or investment strategies for investors. The model's accuracy in stable market conditions can provide base expectations for stock price movements. Combining this analysis with the clustering analysis can provide a deep insight into making the investment decision. The rolling window regression prediction model is a robust method for predicting stock price trends which can help in the investment decision making process.

4.3 Combined Analysis

4.3.1 Methodology

Cross-analysis between the financial clusters and the stock price predictions was conducted to examine how financial health is being reflected in stock market performance. This was achieved by integrating the stock price data with the clusters from the cluster analysis. The merging of the two datasets was achieved on the attribute 'Ticker Symbol' to associate each stock price observation with its corresponding financial cluster identified in the cluster analysis.

For each cluster, a regression analysis was performed to predict the current stock closing prices based on the previous day's close. The data was split into training and test subsets to validate the model's performance and calculate the prediction accuracy on Root Mean Squared Error (RMSE). This provided us insights into the stock price predictions within financial indicators defined by each cluster. Relatively stronger coefficients were extracted to understand the relationship between previous and current closing prices.

4.3.2 Analysis of Combined Results

The regression models provided different predictions across the clusters, with RMSE values ranging from 0.64 to 1.63. Cluster 1 consisting of companies with stable financial metrics showed the lowest RMSE. In contrast, the remaining clusters exhibited higher RMSE values highlighting more volatility within these clusters and thus less probability of predicting.

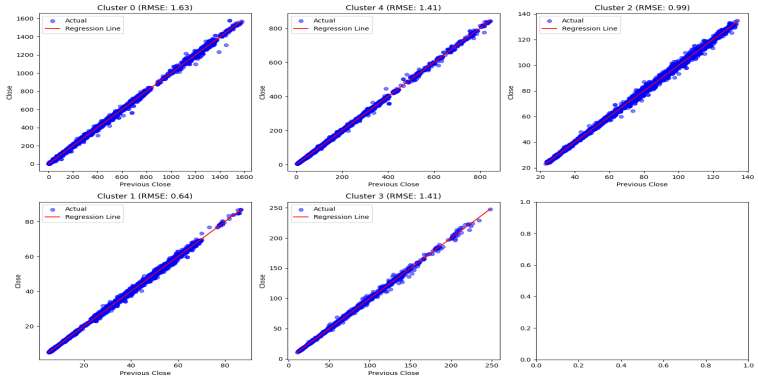

Figure 6: Combined Analysis

The scatter plots combined with regression lines for each cluster show the extent to which previous closing prices can predict current prices. From Figure 5, in clusters where the regression line closely follows the actual data, we can infer a strong linear relationship such as in Cluster 1. This suggests that historical prices can provide reliable future prices.

For investors looking for potential investments, the stocks within Cluster 1 could be considered more predictable and thus less risky. However, for the clusters with higher RMSE values, a more complex model or additional market metrics can be accounted for predicting the stock prices. Shen, Jiang, and Zhang (2012) use Machine learning models for predicting stock prices. They incorporate the correlation between global markets which might account for the volatility unpredictability.

These insights from the combined analysis can prove to be beneficial for investment strategies and risk management as well as resource allocations. This emphasizes the importance of considering both the financial fundamentals and stock market behavior in investment analysis.

4.4 Financial Resilience Analysis

4.4.1 Methodology

Financially resilient companies are those who earn enough income to stay operational and they also have capital to set aside which can get them through tough financial times.

Key financial indicators such as liquidity ratios (Quick Ratio and Current Ratio), profitability metrics (Net Income, Earnings Before Interest and Tax, Operating Income, and Profit Margin), and cash flow indicators were chosen from the fundamentals.csv dataset to capture the financial resilience. This analysis was performed on two companies Apple *(AAPL)* and Nvidia *(NVDA)* for the simplicity of the analysis results.

K-Means clustering was applied to these financial metrics to check the grouping of AAPL and NVDA within a broader market. This clustering helps in identifying companies with similar financial health and resilience characteristics. The cluster to which each company belonged was identified and the average values of their financial indicators were captured. This comparison provides insight into how Apple and Nvidia stand relative to each other in terms of financial resilience.

4.4.2 Analysis of the Resilience Results

Apple exhibits robust financial metrics which indicate strong profitability and super efficient operations highlighted by substantial Net Income and higher Operating Income. The company's Profit and Cash Flow further strengthen its financial strength. Apple is placed in the cluster which is characterized by the high-performing companies.

Nvidia while also being profitable, shows a lower value in profitability metrics as compared to Apple. However, the liquidity ratio of Nvidia was high which indicates a strong cash position and its ability to cover short-term obligations. The negative Net Cash Flow is a potential red flag for the company showcasing the liquidity concerns or any financial or investment activities that should be monitored to make any investment decision.

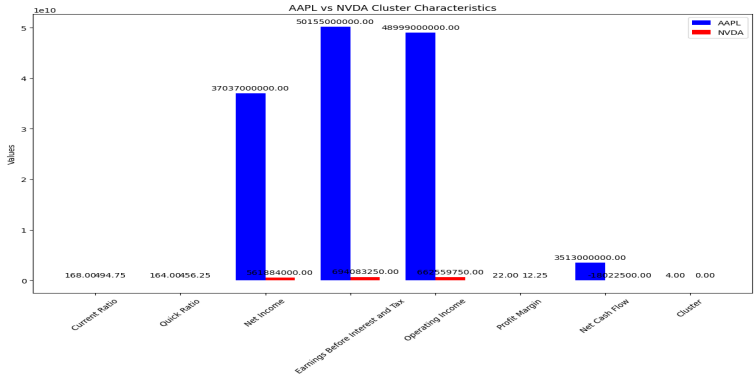

Figure 7: Financial Resilience Analysis

The bar chart compares Apple and Nvidia's financial characteristics side by side. The chart shows Apple's dominance in profitability whereas Nvidia maintains a higher liquidity position. This shows insight into the differences in their financial approaches and their implications on financial resilience within the broader market.

Resilience analysis can be useful for investors to judge a company's position within a broader market and its profitability. In the case of Apple and Nvidia, Apple makes for a potentially attractive option for investors who seek stronger growth and consistent returns. Whereas, Nvidia's higher liquidity ratios can be attractive to investors who seek moderate to higher risk and who look for stability and risk aversion. The direct contrasting Net Cash Flow needs to be monitored, as it shows differences in cash management and investment strategies. This could potentially impact the long-term financial health and resilience of the company.

The cluster characteristics for Apple show a high-performing and profitable business model, whereas Nvidia's cluster characteristics highlight its liquidity strength. For investors, this analysis can provide insights into making an informed decision that might align with their investment objectives and risk tolerance levels.

5 Conclusion

The comprehensive analysis using clustering and regression has provided great insight into the financial health and stock performance of companies.

Clustering analysis showed a dominant financial profile that captured the majority of companies and also identified clusters with unique financial characteristics. Regression modeling showed the predictability of stock prices based on historical closing prices highlighting the most accurate predictions in stable market conditions. The combined analysis then further strengthened the correlation between financial health and stock performance.

The resilience analysis provided deep insight into the company's efficient operations and profit-making frameworks. Apple is highlighted as an appealing option for investors seeking growth and consistent returns and Nvidia shows a strong liquidity position for investors having risk aversion.

5.1 Limitations

The presence of outliers and the spread of data points in other clusters highlight the limitations of using just a sole predictor attribute. There is an influence of external factors on stock price movements which needs to be captured for a more robust analysis in the future. The regression models in this analysis which solely rely on historical closing prices may not account for sudden shifts in markets or sudden events that can lead to an upturn or a downfall.

5.2 Further Work

The need for comprehensive analysis when making investment decisions and the limitations highlight the complexity of the stock market. To build upon this foundational work, future studies should consider including sentiment analysis or advanced machine learning techniques such as LSTM networks. Further analysis to explore the dynamic nature of the market and understanding the financial health changes over time can provide us better insights into making informed decisions for investment in any market conditions. Deep learning approaches can be utilized for price prediction and integrated with financial metrics to understand the dynamic nature of changing market conditions. Tetlock (2007) sentiment analysis and the predictive capabilities of the LSTM network as studied by Fischer and Krauss (2018) show methods for future research that can be utilized to enhance this analysis.

Incorporating real-time data and high-frequency trading data can prove to be essential and provide accurate insights. This analysis can also be conducted across different sectors and different market conditions to help generalize and validate the findings. This could potentially make the analysis more comprehensive and robust for investor decision making.

References

Back, B. et al. (2001). "Comparing numerical data and text information from annual reports using self-organizing maps". In: *International Journal of Accounting Information Systems* 2, pp. 249–269.

Box, G.E.P. and G.M. Jenkins (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day.

Fama, E.F. and K.R. French (1992). "The Cross-Section of Expected Stock Returns". In: *The Journal of Finance* 47.2, pp. 427–465.

Fischer, T. and C. Krauss (2018). "Deep learning with long short-term memory networks for financial market predictions". In: *European Journal of Operational Research* 270.2, pp. 654–669.

Gawlik, D. (2017). *New York Stock Exchange SP 500 companies historical prices with fundamental data*. `https://www.kaggle.com/datasets/dgawlik/nyse/data`. [Dataset].

Huang, Z. (2011). "Clustering Large Data Sets With Mixed Numeric and Categorical Values". In: *Proceedings of the Fifth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*.

Ohlson, James A. (1980). "Financial Ratios and the Probabilistic Prediction of Bankruptcy". In: *Journal of Accounting Research* 18.1, pp. 109–131. DOI: `10.2307/2490395`. URL: `https://doi.org/10.2307/2490395`.

Shen, Shunrong, Haomiao Jiang, and Tongda Zhang (2012). "Stock Market Forecasting Using Machine Learning Algorithms". In: URL: `https://api.semanticscholar.org/CorpusID:16643114`.

Tetlock, P.C. (2007). "Giving Content to Investor Sentiment: The Role of Media in the Stock Market". In: *The Journal of Finance* 62.3, pp. 1139–1168.