EDA Report

Team Lead: Amita Patil

Recorder: Hunter Huberdeau

Spokesperson: Eliel Polanco, Hayden Realmuto

# Background and Introduction

Understanding and anticipating the investment decisions of influential figures has long been a subject of interest in financial markets. Warren Buffett, renowned for his disciplined value investing approach, focuses on identifying fundamentally strong yet undervalued companies (Hagstrom, 2013). Investors and financial professionals closely track his moves through SEC filings and media reports, but these sources provide insights only in retrospect. Team Lambda aspires to bridge that gap by leveraging machine learning to predict which stocks Buffett is likely to buy next, based on historical data and company-level financial indicators. By deploying predictive modeling and analysis, we aim to create a decision-support tool that empowers individual investors with insights traditionally available only after the fact.

The intersection of behavioral finance and predictive analytics presents a compelling opportunity to model Buffett’s investment patterns. While previous research has explored stock price prediction and portfolio performance modeling, few studies have specifically attempted to emulate the buy decision-making process of a known value investor using explainable machine learning models (Fischer & Krauss, 2018). Our approach is novel in that it does not merely assess market trends but seeks to identify the financial characteristics that make a stock appealing to Buffett before he purchases.

We hypothesize that historical financial metrics such as a low price-to-earnings (P/E) ratio, strong revenue growth, and high dividend yield are significant predictors of his stock selections. This is because Buffett prioritizes companies that exhibit financial strength, are undervalued relative to their intrinsic worth, and demonstrate steady, long-term growth potential (Buffett & Cunningham, 2020). If successful, we predict this work could enhance traditional financial analysis by providing predictive insights into high-profile investment strategies, offering a practical tool for investors who seek to align their decisions with Buffett’s time-tested approach.

# Methods

## Dataset and Cleaning

Individual stock data, sourced from Bloomberg, was pulled as individual files representing quarterly snapshots from Q1 2007 to Q4 2024. These files were merged on column headers for a dataset containing all stock data. Warren Buffett’s portfolio activity is sourced from Dataroma, filtered on activity representing a buy (initial purchases and additional purchases of existing holdings), then a label was created to indicate 1 for all purchase activity. Economic data, sourced from Bloomberg, was pulled in three month increments to match the quarterly label for a merge. The dates were changed from short form (eg. 03/31/24) into two columns, Quarter (Q1) and Year (2024). These three datasets were then combined. Stock data and purchase activity merged on Ticker Quarter and Year. Stock data and economic data merged on Quarter and Year.

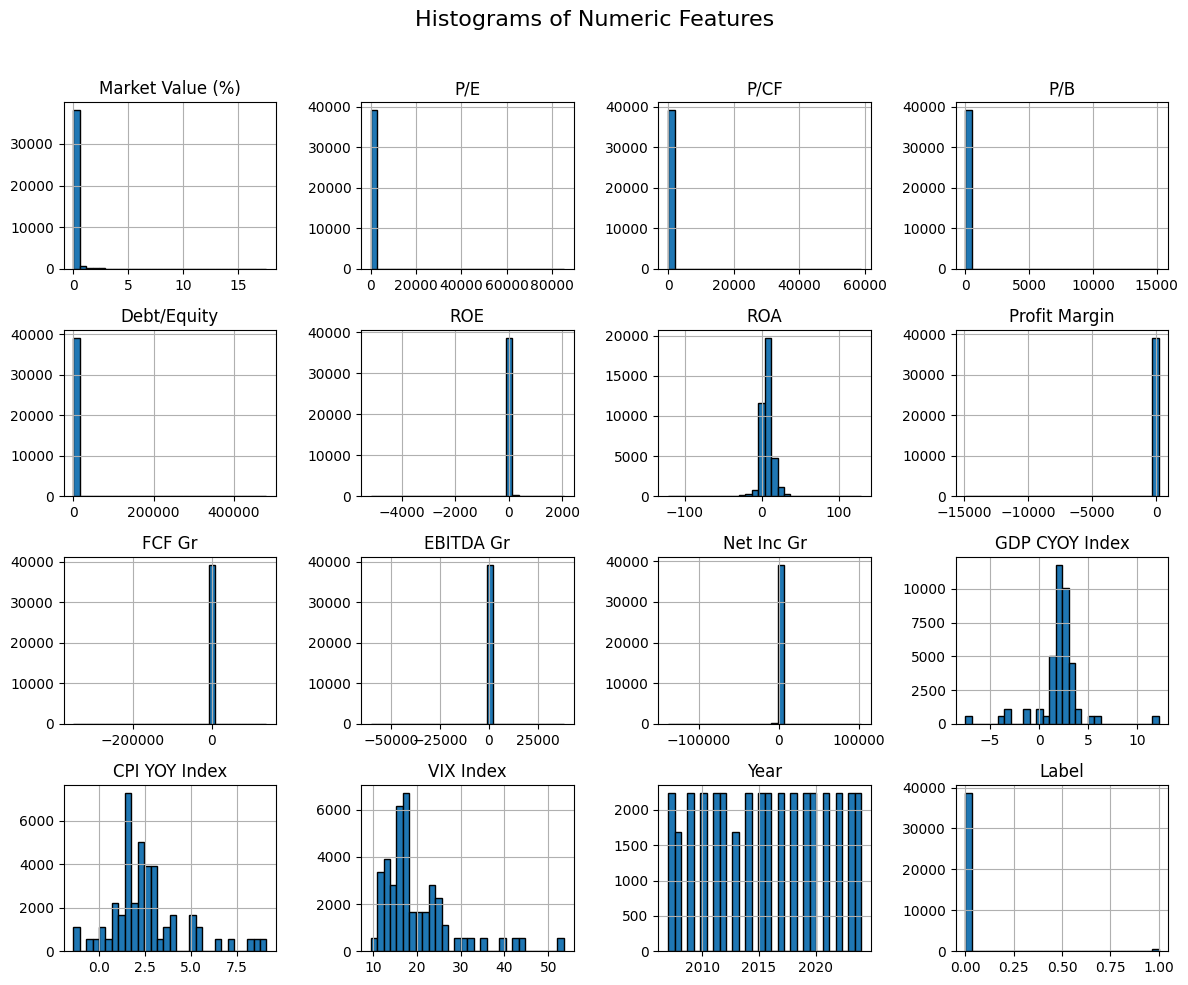
## Execution of Exploratory Data Analysis (EDA)

To perform Exploratory Data Analysis (EDA), we started by analyzing the dataset's structure, including checking for missing values, duplicates, and data types. We handled missing values by imputing numerical columns with the mean and categorical columns with the mode to ensure data completeness. We also removed duplicate rows to maintain data integrity. For feature selection, We focused on financial metrics (e.g., P/E, Profit Margin), macroeconomic indicators (e.g., GDP CYOY Index, VIX Index), and categorical/time-based features (e.g., Sector, Year) that are most relevant to predicting investor decisions.

We used visualizations such as histograms, scatter plots, and correlation heatmaps to understand the distribution of variables, relationships between features, and potential multicollinearity. For example, scatter plots helped identify trends between macroeconomic indicators like CPI YOY Index and PPI YOY Index, while the correlation matrix highlighted redundant features. These methods were chosen to ensure a comprehensive understanding of the dataset and to identify patterns and relationships that could inform feature engineering and predictive modeling.

# Results

Figure 1: Histograms of numeric features in the dataset.



These select features- P/E, P/CF, P/B, Debt/Equity, FCF Gr, , Profit Margin, Market value % are highly right skewed , with most values concentrated near zero and few extreme outliers on the higher end. We will consider handling outliers or applying transformations to reduce skewness.

Figure 2: Correlation matrix containing the first 17 selected variables which includes macro-economic indicators.

A screenshot of a graph

AI-generated content may be incorrect.

CPI YOY Index and PPI YOY Index are highly correlated. They may provide redundant information in your dataset. will drop one of the columns. Moderate negative correlation (~-0.40) between VIX Index and GDP CYOY Index, indicating that higher market volatility (VIX) is associated with lower GDP growth.

Figure 3: Scatter Plot of CPI YOY Index vs PPI YOY index.

A diagram of a graph

AI-generated content may be incorrect.

The points in the scatter plot form an upward trend, indicating a positive correlation between CPI YOY Index and PPI YOY Index.

Figure 4: Average P/E Ratio over years

A graph with a line and a line

AI-generated content may be incorrect.

This line plot shows the average P/E ratio over the years. The P/E ratio remained relatively stable between 2007 and 2019, fluctuating around 30–40. However, there was a sharp spike in 2020, likely due to market disruptions (COVID-19 pandemic), where stock prices increased disproportionately to earnings. After 2020, the P/E ratio dropped significantly and stabilized at lower levels in subsequent years.

Figure 5: Distribution of top sectors

A chart showing a distribution of top sectors

AI-generated content may be incorrect.

The above plot focuses on top sectors, by grouping smaller sectors into "other", the plot highlights the top 10 most frequent sectors. More analysis will be needed possibly breaking down the ‘other’ sector.

Table 1: Summary Statistics for Categorical Variables

| **Sector** | **Frequency (n)** | **Percentage (%)** |
| --- | --- | --- |
| Software & IT Services | 4,900 | 12.5 |
| Residential & Commercial REIT | 2,030 | 5.2 |
| Machinery, Equipment & Components | 1,890 | 4.8 |
| Healthcare Equipment & Supplies | 1,820 | 4.6 |
| Insurance | 1,610 | 4.1 |
| Investment Banking & Investment Services | 1,610 | 4.1 |
| Electrical Utilities & IPPs | 1,610 | 4.1 |
| Pharmaceuticals | 1,399 | 3.6 |
| Hotels & Entertainment Services | 1,330 | 3.4 |
| Food & Tobacco | 1,260 | 3.2 |
| Media & Publishing | 1,260 | 3.2 |
| Professional & Commercial Services | 1,260 | 3.2 |
| Banking Services | 1,260 | 3.2 |
| Semiconductors & Semiconductor Equipment | 1,260 | 3.2 |
| Oil & Gas | 1,190 | 3.0 |
| Chemicals | 1,120 | 2.9 |
| Aerospace & Defense | 910 | 2.3 |
| Healthcare Providers & Services | 910 | 2.3 |
| Specialty Retailers | 840 | 2.1 |
| Telecommunications Services | 630 | 1.6 |
| Automobiles & Auto Parts | 630 | 1.6 |
| Freight & Logistics Services | 560 | 1.4 |
| Beverages | 560 | 1.4 |
| Oil & Gas Related Equipment and Services | 560 | 1.4 |
| Multiline Utilities | 490 | 1.3 |
| Computers, Phones & Household Electronics | 490 | 1.3 |
| Personal & Household Products & Services | 490 | 1.3 |
| Electronic Equipment & Parts | 490 | 1.3 |
| Metals & Mining | 420 | 1.1 |
| Homebuilding & Construction Supplies | 420 | 1.1 |
| Communications & Networking | 420 | 1.1 |
| Containers & Packaging | 420 | 1.1 |
| Biotechnology & Medical Research | 420 | 1.1 |
| Diversified Retail | 420 | 1.1 |
| Textiles & Apparel | 280 | 0.7 |
| Passenger Transportation Services | 280 | 0.7 |
| Consumer Goods Conglomerates | 280 | 0.7 |
| Food & Drug Retailing | 280 | 0.7 |
| Financial Technology (Fintech) & Infrastructure | 210 | 0.5 |
| Construction Materials | 140 | 0.4 |
| Leisure Products | 140 | 0.4 |
| Renewable Energy | 140 | 0.4 |
| Water Utilities | 70 | 0.2 |
| Natural Gas Utilities | 70 | 0.2 |
| Integrated Hardware & Software | 70 | 0.2 |
| Real Estate Operations | 70 | 0.2 |
| Paper & Forest Products | 70 | 0.2 |
| Office Equipment | 70 | 0.2 |
| Household Goods | 70 | 0.2 |
| Construction & Engineering | 70 | 0.2 |
| Software & IT Services | 4,900 | 12.5 |

The Sector variable is dominated by Software & IT Services, which accounts for 12.5% of the dataset, followed by Residential & Commercial REIT (5.18%) and Machinery, Equipment & Components (4.82%). The remaining sectors are more evenly distributed, with many contributing less than 4% each. This indicates a concentration of data in a few key sectors.

Table 2: Quarter Summary: The Quarter variable is evenly distributed across all four quarters

|  |  |  |
| --- | --- | --- |
| **Quarter** | **Frequency** | **Percentage (%)** |
| Q3 | 10080 | 25.7149 |
| Q2 | 10080 | 25.7149 |
| Q4 | 9520 | 24.2863 |
| Q1 | 9519 | 24.2838 |

Q3 and Q2 each accounting for 25.71% of the data, while Q4 and Q1 contribute slightly less at 24.29% and 24.28%, respectively. This suggests that the dataset captures data consistently across time periods.

Table 3: Summary statistics for numerical variables

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **mean** | **std** | **min** | **0.25** | **0.5** | **0.75** | **max** | **Skewness** | **Kurtosis** |
| **Market Value (%)** | 39199 | 0.1995 | 0.3808 | 0.0010 | 0.0851 | 0.1515 | 0.2054 | 17.5487 | 22.3221 | 788.8478 |
| **P/E** | 39199 | 40.3603 | 546.4104 | 0.3024 | 15.5430 | 22.8503 | 40.0490 | 85278.5046 | 118.0390 | 16617.0112 |
| **P/CF** | 39199 | 22.8484 | 352.9619 | 0.0576 | 9.2398 | 15.5651 | 23.1049 | 58987.6499 | 135.6257 | 21018.5483 |
| **P/B** | 39199 | 9.1153 | 110.4801 | 0.1136 | 2.0544 | 3.8230 | 9.1975 | 15144.2615 | 88.3260 | 10175.9129 |
| **Debt/Equity** | 39199 | 211.1947 | 3396.1508 | 0.0000 | 39.9607 | 88.4201 | 211.0715 | 481004.9332 | 117.7311 | 15441.8814 |
| **ROE** | 39199 | 20.5654 | 54.8810 | -5167.3996 | 9.5613 | 18.0599 | 23.0435 | 2065.2670 | -11.7281 | 2306.1249 |
| **ROA** | 39199 | 6.5920 | 8.6610 | -121.6287 | 2.7654 | 6.6202 | 9.5230 | 129.4297 | -0.0729 | 30.8666 |
| **Profit Margin** | 39199 | 7.9526 | 121.9557 | -14870.6981 | 5.9438 | 9.0841 | 15.9124 | 179.8998 | -76.8068 | 7459.0506 |
| **FCF Gr** | 39199 | 35.3141 | 2656.3780 | -349956.3125 | -11.8360 | 34.8348 | 34.8348 | 137717.8125 | -56.4565 | 8336.7584 |
| **EBITDA Gr** | 39199 | 31.0364 | 604.6570 | -60612.1622 | -0.0036 | 13.7099 | 31.3041 | 38525.8352 | -9.2304 | 4227.8228 |

The table provides descriptive statistics for all numeric variables, including measures of central tendency (mean, median), spread (standard deviation, min, max), and distribution shape (skewness, kurtosis). Variables like  P/E, P/CF, and Debt/Equity are highly right-skewed with extreme outliers, as indicated by their high skewness and kurtosis values. In contrast, variables like GDP CYOY Index and CPI YOY Index exhibit relatively normal distributions with low skewness and kurtosis.

# Discussion

Exploring sector-specific trends were revealed through our visualizations and summary statistics. Examining the correlation matrix and presented metrics provided initial insights into investor behavior during key market events and uncover patterns that may inform future investment preferences.

Figure 4: Average P/E Ratio over years exhibits a technology Sector Spike in P/E Ratios (2020). During the pandemic, the technology sector experienced a surge in demand due to remote work, e-commerce, and digital transformation. Implication for Predictions: Investors may continue to favor technology stocks during periods of innovation or economic uncertainty, especially if similar conditions (e.g., increased reliance on technology) arise in the future.

The Software & IT Services sector is the most popular in Figure 5: Distribution of top sectors, reflecting strong investor preference for technology stocks, especially in growth-driven markets. Additionally, the dominance of the "Other" category suggests diverse investor interest in smaller or less-defined sectors, warranting further analysis to uncover specific trends or niche opportunities. This aligns with Table 1: Summary Statistics for Categorical Variables, reflecting the market’s preference for technology stocks. Additionally, the dominance of the "Other" category suggests diverse interest in less-defined sectors or limitations in the thoroughness of our data.

## Next Steps

This workflow begins with finalizing the dataset and addressing any missing data through imputation as needed. Numerical variables will be normalized to ensure consistent scaling for modeling. Outliers will be handled to reduce their impact on the model (e.g., using winsorization or capping). Encoding categorical variables (e.g., Sector, Quarter) using one-hot encoding or label encoding. The data will then be split into training and testing sets using an 80/20 ratio.

For modeling, we will start with a logistic regression model as our baseline due to its interpretability and effectiveness as a benchmark. Next, we will implement a random forest classifier, which is well-suited for capturing non-linear relationships and offers feature importance measures. Finally, we will explore gradient boosted trees using either XGBoost or LightGBM to optimize performance. Model evaluation will include confusion matrices, ROC-AUC curves, and SHAP values to ensure both performance and interpretability are addressed throughout the process.

Team Lambda has decided to stay on course with our current analysis plan. We are confident that the outlined steps and modeling approach provide a solid foundation for addressing our research question.

# Appendix

References:

Buffett, W., & Cunningham, L. A. (2020). *The Essays of Warren Buffett: Lessons for Corporate America*. The Cunningham Group.

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research, 270(2)*, 654-669.

Hagstrom, R. G. (2013). *The Warren Buffett Way*. Wiley.

## Data Dictionary

|  |  |
| --- | --- |
| Column Name | Description |
| Market Value (%) | The stock's weight in the portfolio, indicating its relative importance. |
| P/E | Price-to-Earnings ratio, a valuation metric showing how much investors pay per unit of earnings. |
| P/CF | Price-to-Cash Flow ratio, a valuation metric showing the stock price relative to cash flow. |
| P/B | Price-to-Book ratio, a valuation metric comparing the stock price to book value. |
| Debt/Equity | Financial leverage ratio, showing the proportion of debt relative to equity. |
| ROE | Return on Equity, a profitability metric indicating how efficiently equity generates returns. |
| ROA | Return on Assets, a profitability metric showing how efficiently assets generate returns. |
| Profit Margin | The percentage of revenue that turns into profit after expenses. |
| FCF Gr | Free Cash Flow Growth, indicating the growth rate of free cash flow over time. |
| EBITDA Gr | EBITDA Growth, showing the growth rate of earnings before interest, taxes, depreciation, and amortization. |
| Net Inc Gr | Net Income Growth, indicating the growth rate of net income over time. |
| GDP CYOY Index | Year-over-Year GDP growth, reflecting economic growth trends. |
| CPI YOY Index | Year-over-Year Consumer Price Index, a measure of inflation. |
| PPI YOY Index | Year-over-Year Producer Price Index, another measure of inflation. |
| VIX Index | Volatility Index, reflecting market uncertainty and investor sentiment. |
| Sector | The industry or sector to which the stock belongs. |
| Year | The year of the data record, used for time-based analysis. |
| Quarter | The quarter of the year (Q1, Q2, Q3, Q4), capturing seasonality. |
| Label | The target variable indicating whether the stock was purchased (1) or not (0). |

## Codebook