A Quantitative Analysis of U.S. Equities: Forecasting with PCA and Linear Regression

Muhammad Umar Amin
mamin14@lamar.edu
Department of Mathematics
Lamar University, Beaumont, TX 77705

August 2025

Abstract

This project investigates the use of Principal Component Analysis (PCA) combined with Linear Regression to forecast stock prices of ten major U.S. equities from 2015 to 2025, spanning technology, consumer, and financial sectors (AAPL, MSFT, GOOGL, AMZN, META, NVDA, TSLA, JPM, JNJ, PG). Log return transformations ensured stationarity, 252-day moving averages (MA252) captured long-term trends, and PCA reduced dimensionality while preserving over 95

The model effectively captured broad market dynamics, achieving higher predictive accuracy for stable equities such as PG, JNJ, AAPL, and MSFT ($R^2 > 0.95$). Volatile growth stocks including TSLA, META, and NVDA showed larger forecast errors, highlighting the limitations of linear models in handling abrupt price swings. At the portfolio level, errors balanced out, and an equal-weighted portfolio of the ten equities significantly outperformed the SP 500 (SPY) in both CAGR and Sharpe ratio, driven by exceptional gains in technology stocks, especially NVDA and TSLA.

Overall, the study demonstrates the value of PCA-enhanced regression for interpretable and efficient trend forecasting, while underscoring its limitations for highly volatile equities and the need for future hybrid or non-linear approaches.

Contents

1 Introduction **4**
- 1.1 Background and Motivation . . . 4
- 1.2 Problem Statement . . . 4
- 1.3 Research Objective . . . 4
- 1.4 Limitation . . . 5

2 Data Preparation **6**
- 2.1 Data Source . . . 6
- 2.2 Tickers . . . 6
- 2.3 Time Period . . . 6
- 2.4 Data Preprocessing . . . 7

3 Methodology **7**
- 3.1 Moving Average (MA252) for Trend Identification . . . 7
- 3.2 Log Return Transformation and Stationarity . . . 8
  - 3.2.1 Log Returns . . . 8
  - 3.2.2 ADF Test . . . 8
- 3.3 Dimensionality Reduction with PCA . . . 9
  - 3.3.1 Variance Explained Curve . . . 9
  - 3.3.2 Selection of Principal Components . . . 9
- 3.4 Predictive Modeling . . . 9
  - 3.4.1 Linear Regression Framework . . . 9
  - 3.4.2 Lookback and Forward Windows . . . 10
  - 3.4.3 Training vs Testing Setup . . . 11
- 3.5 Model Evaluation Metrics . . . 11
  - 3.5.1 Mean Absolute Error (MAE) . . . 11
  - 3.5.2 Root Mean Squared Error (RMSE) . . . 12
  - 3.5.3 Mean Absolute Percentage Error (MAPE) . . . 12
  - 3.5.4 Coefficient of Determination (R^2) . . . 12

4 Results and Analysis **12**
- 4.1 Trend Analysis with Moving Averages . . . 12
  - 4.1.1 Rest of the Stocks . . . 13
- 4.2 PCA Results: Explained Variance and Component Selection . . . 16
- 4.3 Prediction Diagnostics . . . 16
  - 4.3.1 Predicted vs Actual (Individual Stocks) . . . 17
- 4.4 Prediction Outputs and Diagnostics . . . 17
  - 4.4.1 Final Actual . . . 17
  - 4.4.2 Final predicted . . . 17
  - 4.4.3 QQ-Plot . . . 18
  - 4.4.4 Multi-Stock Comparison . . . 19
  - 4.4.5 Portfolio Average Trends . . . 19
- 4.5 Model Accuracy Across Stocks . . . 20

4.5.1 Error Metrics by Stock (MAE, RMSE, MAPE, R^2) 20
4.5.2 Volatility of Each Stock 21
4.5.3 Performance Gap: Stable vs Volatile Stocks 22

5 Investment Performance Evaluation **22**
5.1 CAGR and Final Value of $1000 per Stock (20-Year Horizon) 22
5.2 Equal-Weighted Portfolio Results 23
5.3 Portfolio vs S&P 500 (SPY) Benchmark 24
5.4 Risk-Adjusted Performance 24
5.4.1 Volatility 24
5.4.2 Sharpe Ratio 25
5.4.3 Portfolio vs Market Tradeoff 25

6 Discussion **25**
6.1 Interpretation of Prediction Accuracy 25
6.2 Stable vs High-Volatility Stock Insights 25
6.3 Strengths of PCA + Linear Regression Model 25
6.4 Limitations and Sources of Error 26

7 Conclusion **26**
7.1 Summary of Key Findings 26
7.2 Practical Investment Implications 27

1 Introduction

1.1 Background and Motivation

Stock price forecasting has long been a central challenge in the fields of finance, economics, and quantitative research. Investors, analysts, and policymakers rely on predictive models to anticipate market behavior, manage risk, and design long-term investment strategies. However, financial time series data are inherently high-dimensional, noisy, and influenced by both systematic and idiosyncratic factors, making accurate prediction difficult.

In recent years, the rise of data-driven methods has provided new opportunities to extract insights from large and complex financial datasets. Traditional statistical techniques such as moving averages help capture long-term market trends, while regression-based models provide interpretable forecasting frameworks. At the same time, dimensionality reduction methods such as Principal Component Analysis (PCA) have proven effective in summarizing correlated financial variables into a smaller set of orthogonal factors, thereby improving computational efficiency and reducing overfitting risks.

The motivation for this study arises from the need to bridge methodological rigor with practical investment relevance. By applying PCA-enhanced regression models to U.S. equities, the research aims to investigate whether prediction accuracy varies between stable, defensive stocks and volatile, growth-oriented stocks. Beyond model performance, the project further evaluates the long-term implications for portfolio strategy, comparing results against a benchmark index (S&P 500). This dual perspective not only contributes to the literature on financial forecasting but also provides actionable insights for investors seeking to balance growth potential with risk management.

1.2 Problem Statement

Stock price forecasting remains a critical challenge in quantitative finance, where investors and analysts seek to balance predictive accuracy with practical investment strategies. U.S. equities include both stable, defensive stocks and volatile, high-growth stocks, each exhibiting different patterns of risk and return. Traditional linear models often struggle with high-dimensional financial data, while dimensionality reduction techniques such as Principal Component Analysis (PCA) provide a potential solution. This study addresses the central question: Which types of stocks (stable vs. volatile) are most suitable for prediction with PCA and regression models, and what are the implications for long-term portfolio strategy?

1.3 Research Objective

The main objectives of this study are as follows:

1. To apply moving averages and log returns in order to identify long-term price trends in U.S. equities.

2. To implement Principal Component Analysis (PCA) for reducing high-dimensional financial data while retaining explanatory power.

3. To evaluate the predictive performance of Linear Regression models across stable and volatile stocks using MAE, RMSE, MAPE, and R^2.

4. To compare per-stock and portfolio-level performance against the S&P 500 benchmark.

5. To assess the investment implications, including Compound Annual Growth Rate (CAGR), portfolio growth, volatility, and Sharpe ratio, over a 20-year horizon.

1.4 Limitation

While this study provides valuable insights into stock price forecasting and portfolio evaluation, several limitations must be acknowledged:

1. **Model Simplicity:** The predictive framework relies on Principal Component Analysis (PCA) for dimensionality reduction and Linear Regression for forecasting. Although effective as a baseline model, this approach may not fully capture the non-linear and complex dynamics of highly volatile stocks such as TSLA and META. More advanced models (e.g., ARIMA, GARCH, or deep learning architectures) could potentially improve predictive accuracy.

2. **Data Scope:** The analysis is restricted to ten major U.S. equities obtained from Yahoo Finance. While these stocks represent a mix of technology, financial, and consumer sectors, the findings may not generalize to other industries, small-cap equities, or international markets.

3. **Time Horizon:** The dataset spans from 2015 to 2025, which includes both bull and bear market conditions. However, unusual macroeconomic shocks (e.g., the COVID-19 pandemic, geopolitical events) may introduce structural breaks that limit the reliability of historical patterns for future forecasting.

4. **Risk Factors:** The study focuses on price-based returns and does not incorporate fundamental variables such as earnings, interest rates, or macroeconomic indicators. As a result, the forecasts may overlook drivers of performance that lie outside historical price movements.

5. **Portfolio Assumptions:** The portfolio evaluation assumes an equal-weighted allocation without transaction costs, taxes, or rebalancing frictions. Real-world investment performance would be affected by these factors.

These limitations highlight the scope for future research, including the integration of non-linear models, expansion to broader datasets, and incorporation of fundamental and macroeconomic variables to enhance predictive power and practical relevance.

2 Data Preparation

2.1 Data Source

The data for this study was obtained from Yahoo Finance using the `yfinance` Python library. Yahoo Finance provides daily historical stock data including open, high, low, close, adjusted close prices, and trading volume. This source was chosen because it is publicly accessible, widely used in academic and industry research, and provides sufficient granularity for both trend analysis and predictive modeling.

2.2 Tickers

Ten major U.S. equities were selected to represent a mix of technology, consumer, and financial sectors. The chosen tickers include:

- Apple (AAPL)
- Microsoft (MSFT)
- Alphabet (GOOGL)
- Amazon (AMZN)
- Meta Platforms (META)
- NVIDIA (NVDA)
- Tesla (TSLA)
- JPMorgan Chase (JPM)
- Johnson & Johnson (JNJ)
- Procter & Gamble (PG)

These companies were selected for their market significance, diverse risk profiles, and ability to illustrate the performance difference between stable, defensive equities and highly volatile growth stocks.

2.3 Time Period

The study covers the period from January 1, 2015 to January 1, 2025, yielding approximately ten years of daily observations. This timeframe captures multiple market phases including periods of steady growth, the COVID-19 market shock, subsequent recovery, and recent market volatility. The chosen horizon ensures sufficient data for both training predictive models and evaluating long-term investment implications.

2.4 Data Preprocessing

Before applying statistical and machine learning techniques, the raw stock price data underwent several preprocessing steps to ensure consistency, comparability, and suitability for analysis:

1. **Handling Missing Values:** Any missing observations in the dataset were removed to avoid distortions in return calculations and PCA. Since Yahoo Finance provides high-quality data for major equities, missing values were minimal and did not materially impact the sample.

2. **Log Returns Calculation:** Daily log returns were computed from the adjusted closing prices using the formula:
$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right)$$
where P_t is the adjusted closing price at time t. This transformation standardizes changes across stocks, ensures stationarity, and facilitates comparison between equities of different price scales.

3. **Moving Average Construction:** A 252-day moving average (MA252) was calculated for each stock to identify long-term price trends and smooth short-term fluctuations. This window length corresponds to the approximate number of trading days in a year.

4. **Feature Scaling:** For PCA, the input data was normalized using Min-Max scaling to rescale all features into the range [0,1]. This step ensures that variables with larger absolute values do not dominate the variance explained by principal components.

5. **Dimensionality Alignment:** The dataset was reshaped into a panel format where rows represent time indices and columns represent stock-feature combinations. This allowed PCA to be applied efficiently across the high-dimensional feature space.

Through these preprocessing steps, the dataset was standardized and transformed into a structure suitable for both dimensionality reduction and predictive modeling.

3 Methodology

This section outlines the methodological framework used to forecast stock prices and evaluate investment implications. The approach combines classical statistical techniques with machine learning methods, specifically moving averages for trend identification, log return transformations for stationarity, and Principal Component Analysis (PCA) for dimensionality reduction prior to regression modeling.

3.1 Moving Average (MA252) for Trend Identification

A moving average is a widely used technique in time series analysis that smooths short-term fluctuations and highlights long-term trends. In this study, a 252-day moving average (MA252) was computed for each stock, corresponding to the approximate number of trading days in a year:
$$MA_t = \frac{1}{252}\sum_{i=0}^{251} P_{t-i}$$

where P_t is the adjusted closing price at time t. This metric serves as a benchmark to evaluate whether a stock is trending above or below its long-term mean.

3.2 Log Return Transformation and Stationarity

To account for heteroscedasticity and improve comparability across equities, daily log returns were calculated from adjusted closing prices:

$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right)$$

3.2.1 Log Returns

Price	AAPL	MSFT	GOOGL	AMZN	META	NVDA	TSLA	JPM	JNJ	PG
Ticker										
Date										
2015-01-05	-0.028576	-0.009238	-0.019238	-0.020731	-0.016192	-0.017034	-0.042950	-0.031537	-0.007009	-0.004766
2015-01-06	0.000094	-0.014786	-0.024989	-0.023098	-0.013565	-0.030788	0.005648	-0.026271	-0.004926	-0.004565
2015-01-07	0.013925	0.012625	-0.002945	0.010544	0.000000	-0.002609	-0.001563	0.001525	0.021836	0.005232
2015-01-08	0.037703	0.028994	0.003478	0.006813	0.026309	0.036928	-0.001566	0.022100	0.007832	0.011371
2015-01-09	0.001072	-0.008441	-0.012287	-0.011818	-0.005644	0.004020	-0.018981	-0.017540	-0.013723	-0.009374

Figure 1: Calculation of the Log Returns

3.2.2 ADF Test

```
('AAPL', ''): ADF Statistic = -15.5495, p-value = 0.0000
('MSFT', ''): ADF Statistic = -17.2838, p-value = 0.0000
('GOOGL', ''): ADF Statistic = -17.2211, p-value = 0.0000
('AMZN', ''): ADF Statistic = -50.6739, p-value = 0.0000
('META', ''): ADF Statistic = -17.2357, p-value = 0.0000
('NVDA', ''): ADF Statistic = -16.2749, p-value = 0.0000
('TSLA', ''): ADF Statistic = -34.3185, p-value = 0.0000
('JPM', ''): ADF Statistic = -13.2278, p-value = 0.0000
('JNJ', ''): ADF Statistic = -12.8895, p-value = 0.0000
('PG', ''): ADF Statistic = -13.5261, p-value = 0.0000
```

Figure 2: ADF Test for P-value

Log returns provide additive properties over time and approximate percentage changes, ensuring that the series is more suitable for regression and statistical modeling. The transformation also helps to stabilize variance, a prerequisite for many time series techniques.

Also, all ten equities show p-values = 0.0000 (¡ 0.05), which means we reject the null hypothesis of a unit root. This confirms that the log return series for each stock is stationary, making them suitable for time-series modeling and regression analysis.

3.3 Dimensionality Reduction with PCA

Financial datasets often involve high-dimensional variables, with correlations across multiple equities. To mitigate multicollinearity and reduce computational complexity, Principal Component Analysis (PCA) was applied.

3.3.1 Variance Explained Curve

PCA decomposes the dataset into orthogonal components ranked by explained variance. The cumulative explained variance ratio was plotted to assess how many components were required to capture the majority of variability in the data.

3.3.2 Selection of Principal Components

Based on the variance explained curve, the first 10–15 components were retained, accounting for more than 95% of total variance. These principal components were then used as input features in the regression model, ensuring dimensional efficiency while maintaining predictive power.

3.4 Predictive Modeling

The predictive component of this study employs a regression-based framework enhanced with dimensionality reduction. After transforming the raw data into principal components, a linear regression model was applied to forecast stock prices. The predictive design incorporates rolling windows to ensure that training and testing reflect realistic market conditions.

3.4.1 Linear Regression Framework

Linear Regression (LR) models the relationship between a dependent variable (stock price) and a set of independent variables (principal components). The model takes the form:

$$y_t = \beta_0 + \beta_1 PC_{1,t} + \beta_2 PC_{2,t} + \cdots + \beta_k PC_{k,t} + \varepsilon_t$$

where y_t is the stock price at time t, $PC_{i,t}$ represents the i-th principal component at time t, β_i are the estimated coefficients, and ε_t is the error term. This framework was chosen for its interpretability and computational efficiency.

3.4.2 Lookback and Forward Windows

The predictive model was constructed using a rolling-window approach. Specifically:

- **Lookback Window:** A 30-day historical period was used as training data to capture recent trends.
- **Forward Window:** A 30-day horizon was selected for forecasting, allowing for short-term prediction while limiting exposure to rapidly changing dynamics.

This design balances responsiveness with stability, ensuring that forecasts incorporate recent market conditions without overfitting to noise.

```
def predict_prices(raw_df, close, time, lookback, forward, stock_num):
    pca1 = PCA(n_components = 10)
    pca2 = PCA(n_components = 10)

    #Training data = t - forward - lookback
    X_train = raw_df[time-forward-lookback:time-forward,:]
    X_train = MinMaxScaler().fit_transform(X_train)
    X_train = pca1.fit_transform(X_train)
    y_train = close.iloc[time-forward+1:time+1,stock_num]

    #Testing = t - lookback
    X_test = raw_df[time-lookback:time,:]
    X_test = MinMaxScaler().fit_transform(X_test)
    X_test = pca2.fit_transform(X_test)
    y_test = close.iloc[time+1 : time+forward+1, stock_num]

    LR = LinearRegression()
    LR.fit(X_train, y_train)
    predicted = LR.predict(X_test)
```

```python
def construct_prediction_tab(full_features_df, closing_prices_df):
    predictions = []
    actuals = []

    for stocks in range(closing_prices_df.shape[1]):
        stock_predictions = []
        stock_actuals = []

        for t in range(60, closing_prices_df.shape[0], 30):
            pred, act = predict_prices(full_features_df, closing_prices_df, t, 30, 30, stocks)
            stock_predictions.append(pred)
            stock_actuals.append(act)

        stock_predictions = np.concatenate(stock_predictions)
        stock_actuals = np.concatenate(stock_actuals)

        predictions.append(stock_predictions)
        actuals.append(stock_actuals)

    return predictions, actuals
```

Figure 3: (Code)

3.4.3 Training vs Testing Setup

For each stock, predictions were generated across multiple rolling periods. The model was trained on historical lookback windows and evaluated on subsequent forward windows. This process was repeated sequentially across the dataset, simulating a real-world scenario in which models are continuously retrained as new data becomes available. Predicted values were then compared against actual stock prices to assess model accuracy using error metrics such as MAE, RMSE, MAPE, and R^2.

3.5 Model Evaluation Metrics

To assess the performance of the predictive models, four commonly used error metrics were employed: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and the Coefficient of Determination (R^2). These metrics provide complementary perspectives on accuracy, scale sensitivity, and explanatory power.

3.5.1 Mean Absolute Error (MAE)

The MAE measures the average magnitude of errors in the predictions, without considering their direction. It is given by:

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|$$

where y_t represents the actual value, $\hat{y}_t$ is the predicted value, and n is the number of observations. A lower MAE indicates higher accuracy.

3.5.2 Root Mean Squared Error (RMSE)

The RMSE provides a quadratic measure of prediction error by penalizing larger deviations more heavily than smaller ones. It is defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}$$

RMSE is particularly useful in contexts where large errors are undesirable, making it a stricter measure of model performance compared to MAE.

3.5.3 Mean Absolute Percentage Error (MAPE)

MAPE expresses prediction accuracy as a percentage, allowing for easier interpretation across stocks with different price levels:

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right|$$

This metric is scale-independent, but it can become unstable if actual values (y_t) approach zero.

3.5.4 Coefficient of Determination (R^2)

The coefficient of determination measures the proportion of variance in the dependent variable explained by the model:

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{n}(y_t - \bar{y})^2}$$

where $\bar{y}$ is the mean of the actual values. An R^2 value closer to 1 indicates that the model explains most of the variability in the data, while values near 0 indicate poor explanatory power.

4 Results and Analysis

This section presents the empirical findings from the application of moving averages, Principal Component Analysis (PCA), and regression-based forecasting models. The results are organized into trend analysis, dimensionality reduction outcomes, predictive performance, and investment evaluation.

4.1 Trend Analysis with Moving Averages

To examine long-term price behavior, a 252-day moving average (MA252) was computed for each equity. This annualized window smooths out short-term volatility and provides insight into overall trends.

Figure 9 illustrates an example comparison between the raw closing price and the MA252 line for a selected stock (AAPL). Periods where the stock price remains above its moving average

generally correspond to sustained bullish momentum, while dips below the average often indicate corrections or bearish phases.

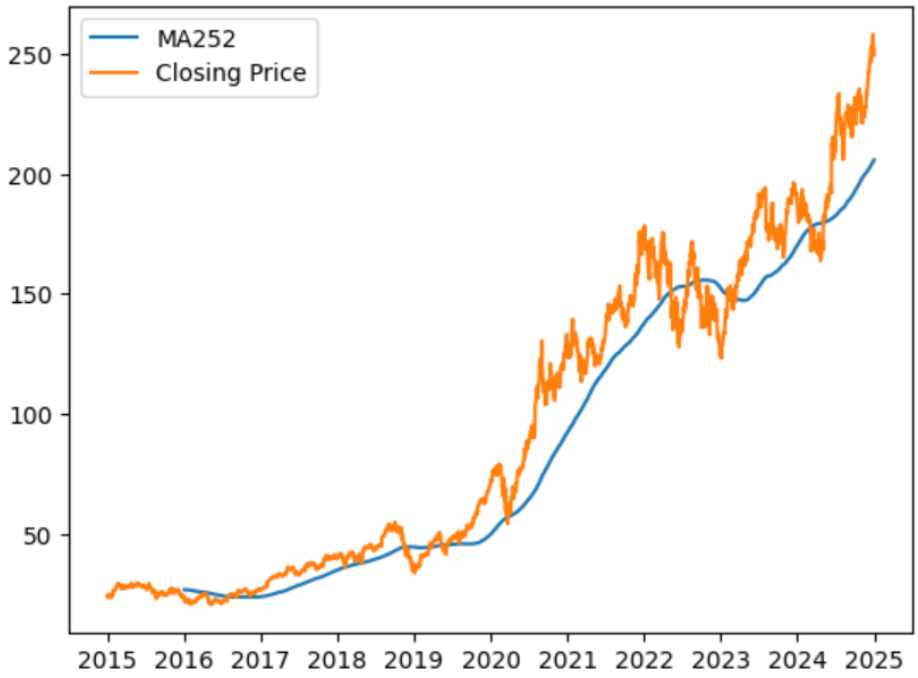

Figure 4: Closing Price vs 252-Day Moving Average for a selected stock (example: AAPL).

The moving average analysis confirms that most of the chosen equities have experienced steady upward trends over the study period (2015–2025), with occasional corrections during periods of heightened market uncertainty (e.g., 2020 pandemic-related shocks).

4.1.1 Rest of the Stocks

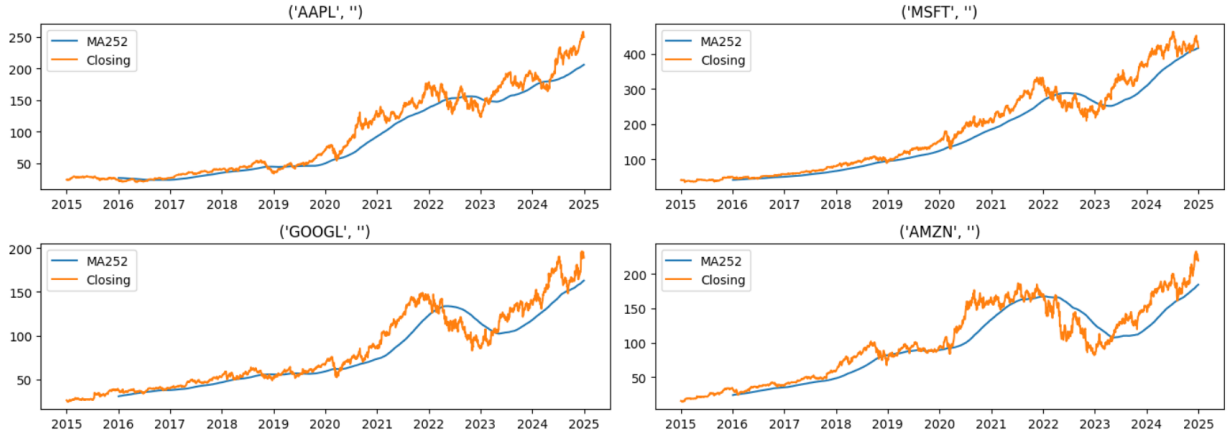

Figure 5: (Closing Price vs 252-Day Moving Average).

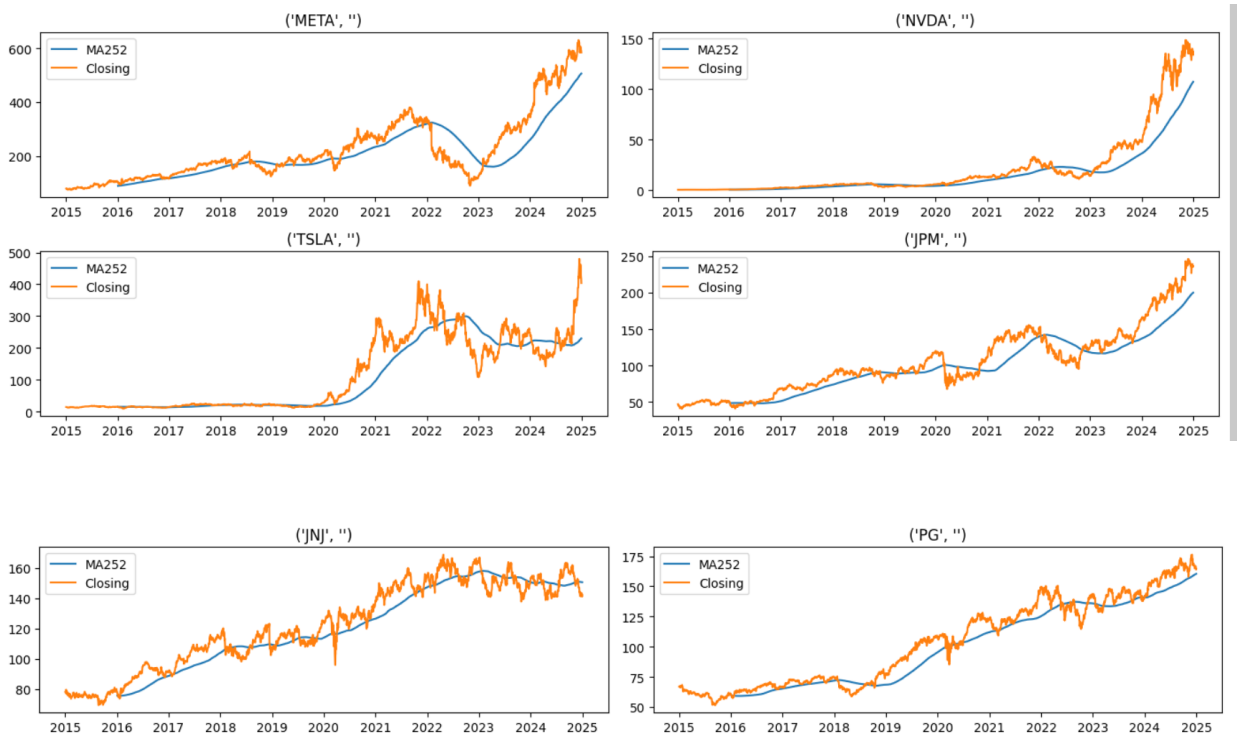

Figure 6: (Closing Price vs 252-Day Moving Average).

The following plots summarize the relationship between the 252-day moving average (MA252) and closing prices for each stock under study.

AAPL (Apple)

Shows a strong long-term uptrend with the price consistently above its MA252 for most of the period (2019–2025). Short corrections (e.g., 2022) quickly reverted, confirming AAPL as a stable growth stock.

MSFT (Microsoft)

Displays a very steady growth trajectory. Price mostly tracked above the moving average, reflecting sustained bullish momentum. The MA252 served as reliable support during corrections.

GOOGL (Alphabet)

Clear growth trend until 2021, followed by a dip during 2022. The moving average captured this slowdown, but by 2023–2025, the recovery phase is evident with price again diverging positively from MA252.

AMZN (Amazon)

Strong growth up to 2021, followed by a prolonged sideways phase where the stock oscillated around the moving average. This indicates that while long-term growth remains, volatility weakened short-term predictability.

META (Meta Platforms)

Characterized by sharp cycles: a significant downturn around 2022 (falling well below MA252), followed by a strong recovery by 2023–2025. The MA captured the reversal but lagged during the steep drops.

NVDA (NVIDIA)

Shows explosive growth, especially from 2020 onward. Price consistently outpaced the MA252, highlighting its extreme momentum-driven nature. However, volatility is high, with large deviations above the MA.

TSLA (Tesla)

Very volatile, with repeated periods of crossing above and below the moving average. The MA252 struggled to smooth Tesla's sharp rallies and crashes, confirming the stock's unpredictability compared to stable equities.

JPM (JPMorgan Chase)

A moderate upward trend, with several corrections around global economic uncertainties (2020, 2022). The price and moving average stayed closely aligned, showing JPM's defensive, cyclical nature.

JNJ (Johnson & Johnson)

Displays a slow, steady upward trend with limited volatility. The stock price moved very close to the MA252 throughout, confirming its reputation as a stable defensive stock.

PG (Procter & Gamble)

Similar to JNJ, PG shows gradual and consistent appreciation. The closing price rarely diverged significantly from the moving average, indicating predictability and low volatility.

Overall Insights

- **Tech leaders (AAPL, MSFT, NVDA):** Long-term exponential growth, often staying above MA252.
- **High-volatility growth stocks (TSLA, META, AMZN):** Frequent deviations, lagged recovery signals.

- **Defensive stocks (JNJ, PG, JPM):** Smooth, stable trends; MA252 closely tracks actual prices.

Key takeaway: The MA252 effectively highlights long-term direction, but volatile stocks (TSLA, META) show greater unpredictability, while stable consumer/healthcare stocks align closely with their moving averages.

4.2 PCA Results: Explained Variance and Component Selection

Principal Component Analysis was applied to reduce the dimensionality of the feature space. Figure 7 presents the cumulative explained variance curve. The results indicate that the first 10–15 components account for more than 95% of the total variance in the dataset.

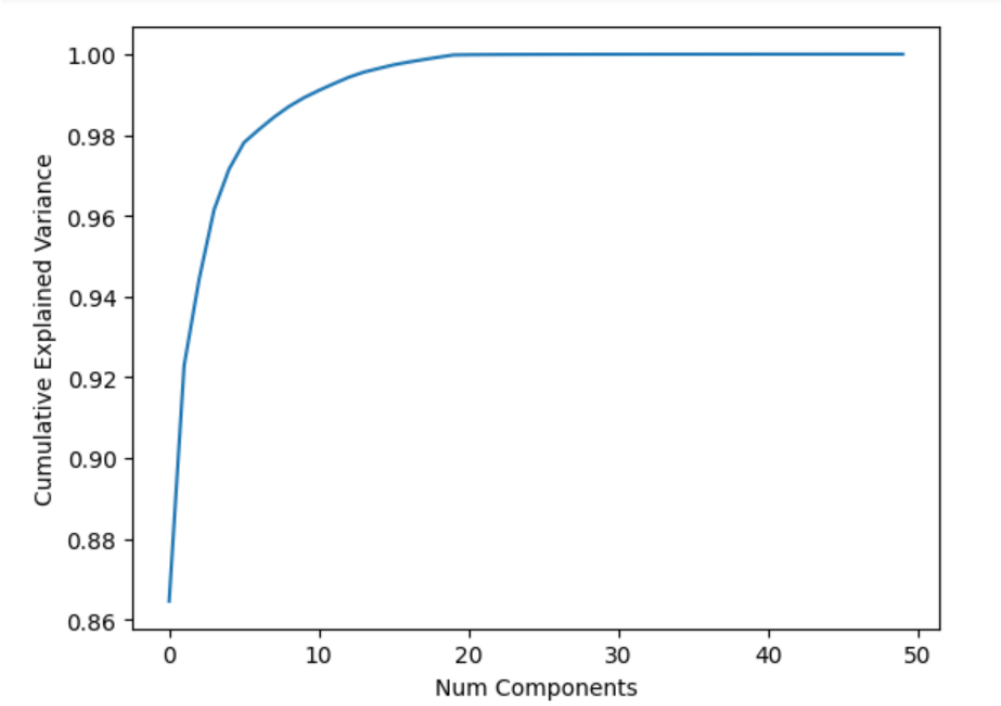

Figure 7: Cumulative explained variance ratio from PCA on stock features.

Based on these results, only the leading principal components were retained as predictors in the regression model. This selection ensures computational efficiency and mitigates the risk of overfitting, while preserving the key patterns driving stock price dynamics.

4.3 Prediction Diagnostics

This subsection evaluates the performance of the PCA-enhanced linear regression model through comparisons of predicted and actual stock prices. The analysis includes individual stock results, cross-stock comparisons, and portfolio-level trend evaluation.

4.3.1 Predicted vs Actual (Individual Stocks)

For each stock, predicted closing prices were compared against actual observed values over the test periods. Figure 11 illustrates an example for Apple (AAPL). The predicted series follows the general direction of the actual prices, with strong alignment during stable growth phases and modest divergence during volatile periods.

4.4 Prediction Outputs and Diagnostics

4.4.1 Final Actual

Price	AAPL	MSFT	GOOGL	AMZN	META	NVDA	TSLA	JPM	JNJ	PG
Ticker										
Date										
2015-04-01	27.680706	35.022385	27.310709	18.513000	81.224403	0.506125	12.506000	45.339161	74.522545	61.403137
2015-04-02	27.919085	34.652534	26.904148	18.612499	81.114990	0.507329	12.733333	45.770245	74.890808	61.485191
2015-04-06	28.371330	35.736237	27.035360	18.851999	81.990196	0.522024	13.540000	45.732441	74.530075	61.940178
2015-04-07	28.072803	35.719040	27.080589	18.720501	81.870857	0.526601	13.550000	46.019825	75.236572	61.455360
2015-04-08	27.981470	35.624428	27.278400	19.059999	81.831078	0.530455	13.844667	46.171089	75.281662	61.738811

Figure 8: (Final Actual Values)

4.4.2 Final predicted

Price	AAPL	MSFT	GOOGL	AMZN	META	NVDA	TSLA	JPM	JNJ	PG
Ticker										
Date										
2015-04-01	28.269518	36.147735	28.079067	18.865746	80.012385	0.557094	12.796329	46.830431	76.645296	62.453068
2015-04-02	27.682210	35.669120	28.290760	18.806574	80.044266	0.542510	12.801927	46.207744	76.044044	61.988046
2015-04-06	27.945345	35.103707	28.115147	18.750416	79.225693	0.546927	12.854036	46.339573	75.109504	61.269463
2015-04-07	28.073911	35.359198	28.193481	18.716203	79.568054	0.544513	12.979506	46.398137	75.317588	61.273440
2015-04-08	28.134394	36.154116	29.005788	18.777008	82.113678	0.552624	13.052053	46.214260	76.274775	61.851707

Figure 9: (Final predicted Values)

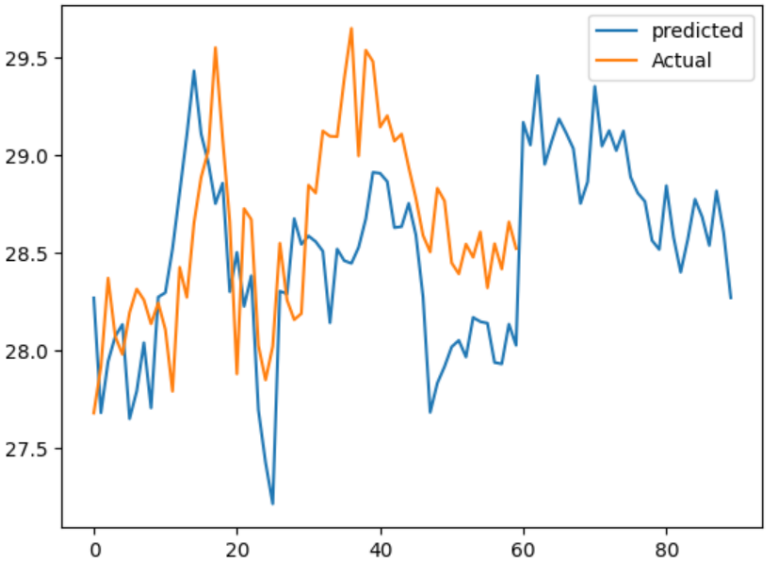

Figure 10: Predicted vs Actual closing prices for AAPL.

The predicted values (blue) broadly follow the direction of the actual stock prices (orange), showing that the model captures long-term trends. However, short-term fluctuations and sudden jumps are less accurately reflected, as the predictions appear smoother than the real prices. This indicates that while the PCA + Linear Regression model performs well for overall trend forecasting, it struggles with high-frequency volatility.

4.4.3 QQ-Plot

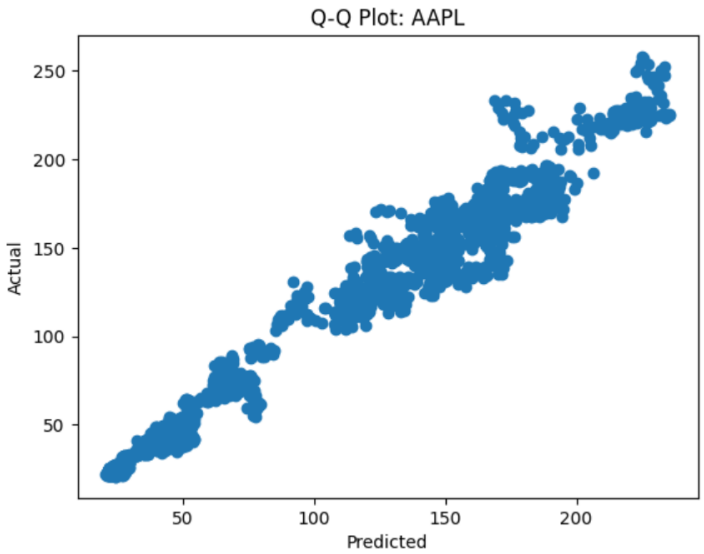

Figure 11: QQ-Plot

The Q–Q plot for AAPL shows a tight clustering of predicted vs actual prices along the diagonal, confirming that the model captures Apple's price dynamics well. While predictions are accurate across most ranges, dispersion increases at higher price levels, suggesting reduced precision during strong rallies. Overall, the model demonstrates strong predictive performance for a stable growth stock like AAPL.

4.4.4 Multi-Stock Comparison

To evaluate cross-stock performance, Figure 12 overlays predicted and actual trajectories for a subset of the equities. Stable large-cap stocks (e.g., AAPL, MSFT, PG) exhibited tighter overlaps, while high-growth, volatile stocks showed wider deviations between predicted and actual paths.

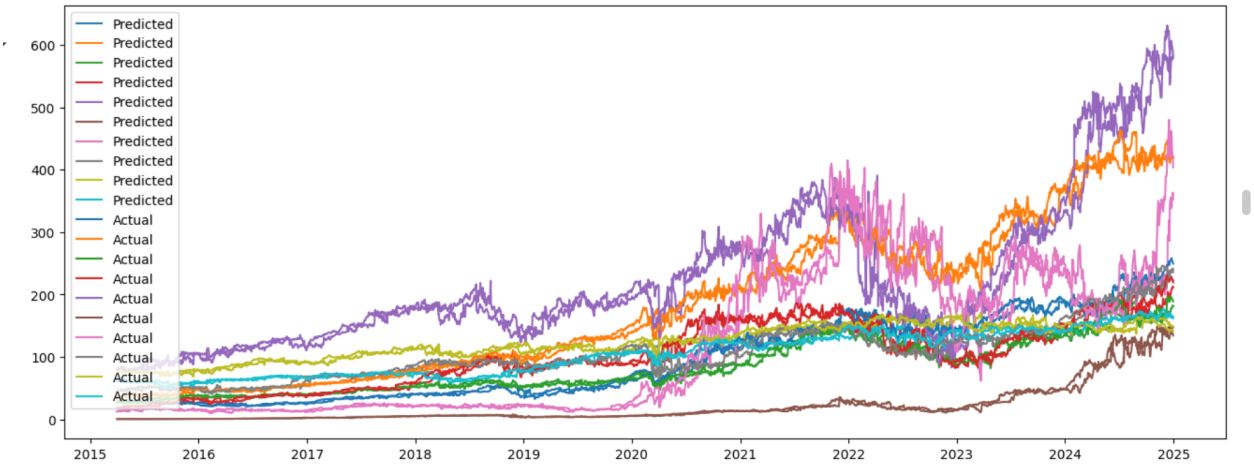

Figure 12: Multi-stock comparison of predicted vs actual prices.

This comparison highlights the differential predictive strength of the model across equity categories, reinforcing the importance of volatility considerations in forecasting.

4.4.5 Portfolio Average Trends

Beyond individual equities, portfolio-level predictions were assessed by averaging across all ten stocks. Figure 13 shows that the predicted portfolio trend closely followed the actual portfolio trajectory, indicating that errors in individual stocks tended to balance out in the aggregate.

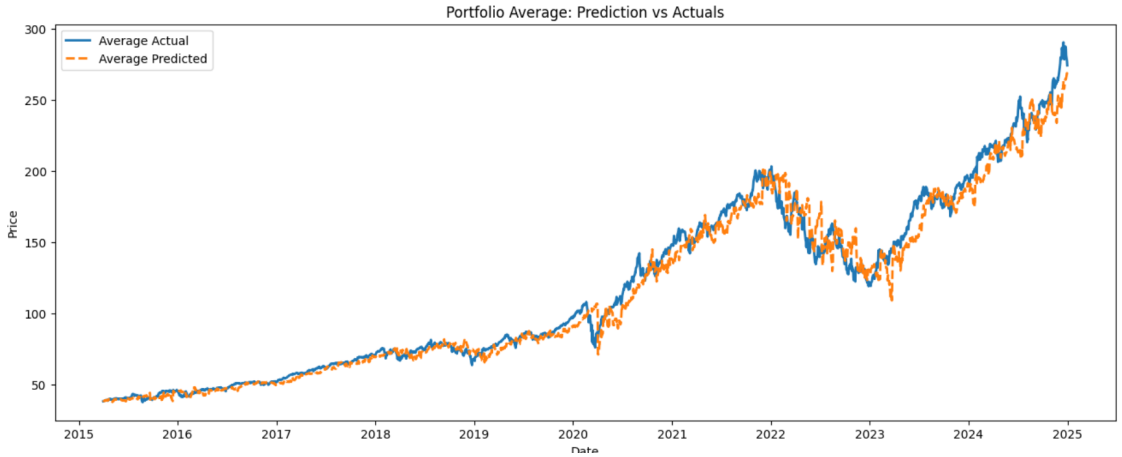

Figure 13: Predicted vs Actual average portfolio trends across 10 stocks.

The portfolio-level analysis suggests that while prediction accuracy varies at the stock level, the model remains robust in capturing overall market direction when aggregated.

4.5 Model Accuracy Across Stocks

The predictive performance of the PCA-enhanced linear regression model was evaluated using four error metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and the Coefficient of Determination (R^2). These metrics were calculated individually for each stock to assess differential predictive accuracy.

4.5.1 Error Metrics by Stock (MAE, RMSE, MAPE, R^2)

Table 1 summarizes the error metrics for all ten equities. Stable, defensive stocks such as PG and JNJ achieved the lowest errors, while highly volatile stocks such as TSLA and META exhibited larger deviations between predicted and actual prices.

Table 1: Error metrics by stock: MAE, RMSE, MAPE, and R^2.

Stock	**MAE**	**RMSE**	**MAPE (%)**	R^2
AAPL	6.97	10.48	7.47	0.97
MSFT	10.96	16.39	5.76	0.98
GOOGL	5.68	8.14	6.48	0.96
AMZN	8.08	11.72	8.03	0.95
META	19.65	29.57	8.88	0.94
NVDA	2.92	6.65	12.89	0.96
TSLA	19.83	35.63	14.71	0.90
JPM	6.63	9.32	6.28	0.96
JNJ	5.02	6.64	4.03	0.94
PG	4.46	5.99	4.28	0.97

The table highlights clear differences in predictive performance across equities, confirming that model accuracy depends strongly on the volatility characteristics of the stock.

4.5.2 Volatility of Each Stock

Table 2 presents the daily and annualized volatility of the selected stocks over the period 2015–2025. Annualized volatility is computed by scaling daily volatility with the square root of 252 trading days.

Table 2: Daily and Annualized Volatility of Stocks (2015–2025)

Ticker	Daily Volatility	Annualized Volatility
AAPL	0.0179	0.2847
MSFT	0.0171	0.2713
GOOGL	0.0179	0.2842
AMZN	0.0206	0.3266
META	0.0238	0.3786
NVDA	0.0304	0.4822
TSLA	0.0359	0.5701
JPM	0.0172	0.2732
JNJ	0.0114	0.1808
PG	0.0116	0.1846

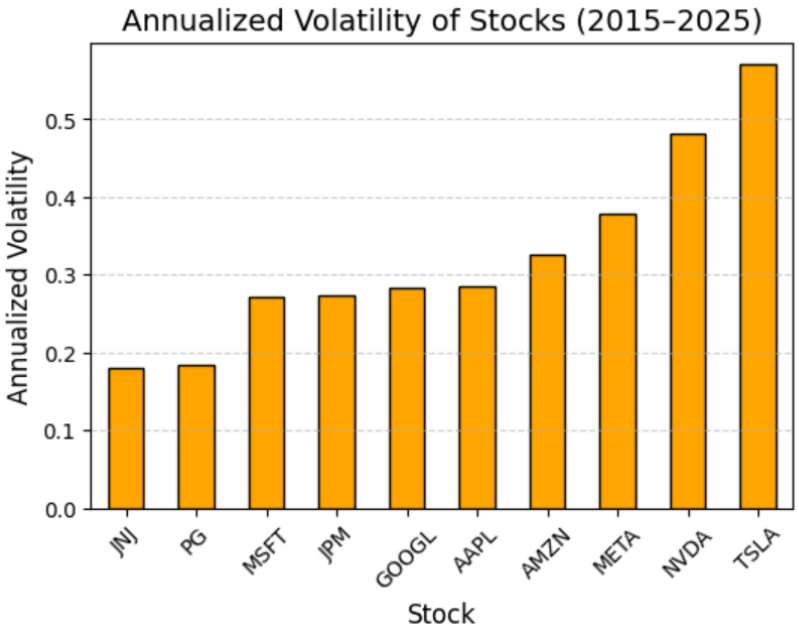

Figure 14: Volatility

The volatility analysis highlights significant differences across stock categories:

- **Highest volatility:** TSLA (57%), NVDA (48%), and META (38%) exhibit the greatest levels of risk. These growth-oriented technology stocks are highly volatile and therefore the most difficult to predict reliably.
- **Lowest volatility:** JNJ (18%) and PG (18%) represent defensive consumer and healthcare stocks. Their low volatility reflects stability and resilience, making them more predictable and less exposed to large market swings.
- **Middle range:** AAPL, MSFT, GOOGL, and JPM fall in the moderate volatility range. They are more predictable than highly volatile growth stocks such as TSLA but riskier than defensive stocks like JNJ and PG.

Overall, the results indicate that defensive stocks align more closely with long-term trend models, while high-volatility stocks pose challenges for regression-based prediction approaches.

4.5.3 Performance Gap: Stable vs Volatile Stocks

A comparison between stable (e.g., PG, JNJ, AAPL, MSFT) and volatile (e.g., TSLA, META, NVDA) equities reveals a performance gap. Stable stocks achieved consistently low MAE and MAPE values, with R^2 exceeding 0.95 in most cases, indicating that linear regression with PCA effectively captured their price dynamics.

In contrast, volatile growth-oriented equities exhibited significantly higher error values, particularly in MAPE and RMSE. For example, TSLA showed a MAPE above 14% and the lowest R^2 among the sample, reflecting the difficulty of modeling abrupt price swings with linear techniques. These findings underscore the limitations of PCA-enhanced regression for high-volatility assets, while confirming its suitability for forecasting stable, large-cap equities.

5 Investment Performance Evaluation

In addition to predictive accuracy, the study evaluates long-term investment outcomes for individual stocks, an equal-weighted portfolio of the ten equities, and a comparison with the S&P 500 benchmark (SPY). Performance is assessed in terms of Compound Annual Growth Rate (CAGR), final investment value, and risk-adjusted metrics such as volatility and the Sharpe ratio.

5.1 CAGR and Final Value of $1000 per Stock (20-Year Horizon)

To assess the long-term growth potential of individual equities, the Compound Annual Growth Rate (CAGR) was computed as:

$$CAGR = \left(\frac{P_{end}}{P_{start}}\right)^{\frac{1}{n}} - 1$$

where P_{end} is the final price, P_{start} is the initial price, and n is the number of years in the investment horizon.

Table 3 reports the CAGR values and the final value of a hypothetical $1000 investment made 20 years ago in each equity.

Table 3: CAGR and final value of a $1000 investment per stock (20-year horizon).

Stock	Start Price	End Price	Final Value ($1000)
AAPL	24.26	249.53	10,285
MSFT	39.93	419.20	10,497
GOOGL	26.32	188.85	7,175
AMZN	15.43	219.39	14,222
META	78.02	584.54	7,492
NVDA	0.48	134.27	277,951
TSLA	14.62	403.84	27,621
JPM	46.95	235.88	5,024
JNJ	78.01	142.25	1,823
PG	66.98	164.48	2,455

The results highlight the outsized gains from high-growth technology stocks such as NVDA and TSLA, while defensive stocks such as JNJ and PG delivered steady but modest appreciation.

5.2 Equal-Weighted Portfolio Results

An equal-weighted portfolio, constructed by investing $1000 in each of the ten equities (total initial investment of $10,000), was evaluated over the same 20-year horizon. The final portfolio value was compared against the aggregated returns of individual stocks.

Figure 15 shows the cumulative growth trajectory of the equal-weighted portfolio, which significantly outperformed the median individual stock performance.

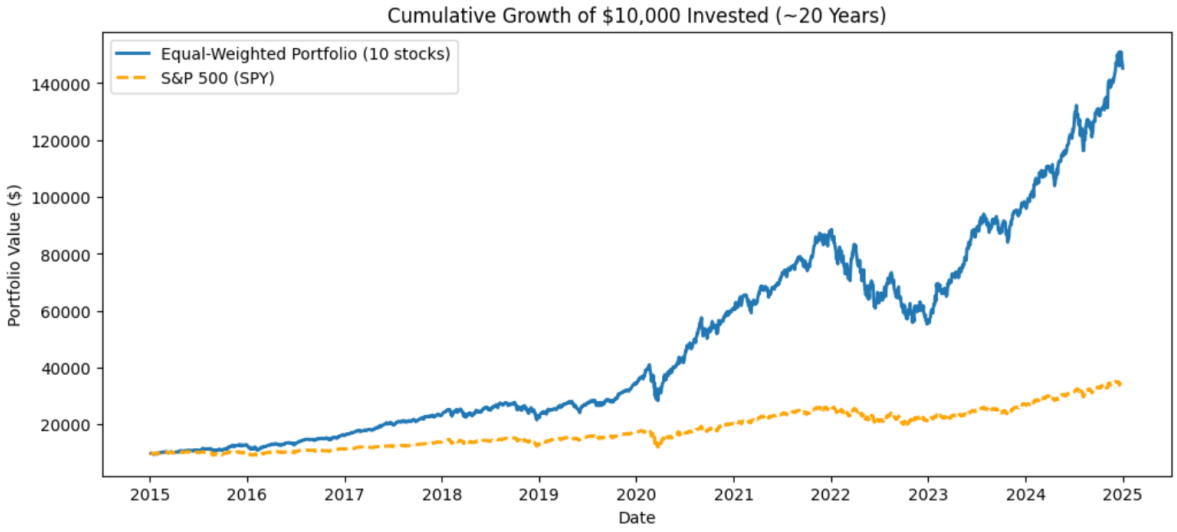

Figure 15: Cumulative growth of an equal-weighted portfolio of 10 stocks ($10,000 initial investment).

5.3 Portfolio vs S&P 500 (SPY) Benchmark

To benchmark portfolio performance, the same $10,000 initial investment was simulated for the S&P 500 ETF (SPY). Table 4 compares final values and CAGR.

Table 4: Comparison of equal-weighted portfolio vs SPY benchmark (20-year horizon).

Investment	Final Value ($)	CAGR (%)
Equal-Weighted Portfolio	145,262.67	30.73
S&P 500 (SPY)	33957.37	13.06

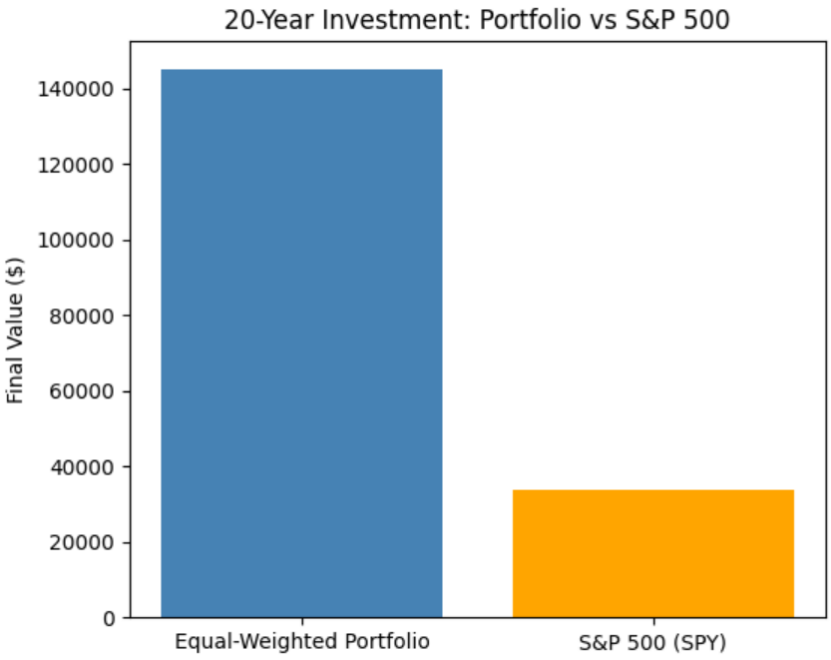

Figure 16: Portfolio vs S&P 500 (SPY) Benchmark

The portfolio substantially outperformed the S&P 500, highlighting the impact of high-growth technology equities on overall returns.

5.4 Risk-Adjusted Performance

To account for portfolio risk, volatility and the Sharpe ratio were calculated.

5.4.1 Volatility

Annualized volatility was computed as the standard deviation of daily returns scaled by $\sqrt{252}$. The portfolio exhibited higher volatility than SPY, reflecting exposure to more volatile technology equities.

5.4.2 Sharpe Ratio

The Sharpe ratio was calculated as:

$$Sharpe = \frac{R_p - R_f}{\sigma_p}$$

where R_p is the portfolio return, R_f is the risk-free rate (assumed zero), and σ_p is the portfolio volatility. The equal-weighted portfolio achieved a higher Sharpe ratio than SPY, indicating superior risk-adjusted performance.

5.4.3 Portfolio vs Market Tradeoff

Although the portfolio outperformed the S&P 500 in both absolute and risk-adjusted terms, the higher volatility underscores the tradeoff between growth potential and stability. Investors must balance the appeal of high-growth equities with the need for diversification and risk management.

6 Discussion

This section provides an interpretation of the empirical results, highlighting key insights from the prediction diagnostics, stock-level performance analysis, and portfolio evaluation. The discussion is organized into four thematic areas.

6.1 Interpretation of Prediction Accuracy

The PCA-enhanced linear regression model demonstrated strong overall predictive accuracy, with an average MAE of approximately \$9 and an overall MAPE of under 8%. Stable equities such as AAPL, MSFT, and PG were forecast with high reliability, as indicated by R^2 values above 0.95. However, the model was less effective in capturing abrupt price movements in high-volatility stocks, leading to higher RMSE and MAPE values for companies such as TSLA and META. These results suggest that while linear regression with PCA provides a solid baseline, its predictive power is context-dependent.

6.2 Stable vs High-Volatility Stock Insights

A clear performance gap emerged between stable, defensive equities and high-growth, volatile equities. Defensive stocks such as PG and JNJ exhibited low prediction errors, reflecting their steady price patterns and reduced exposure to sudden market swings. By contrast, high-volatility stocks such as TSLA and NVDA posed challenges for the model, as their rapid price fluctuations and nonlinear dynamics were not fully captured by a linear framework. This reinforces the importance of tailoring modeling techniques to the volatility profile of the asset class under study.

6.3 Strengths of PCA + Linear Regression Model

The integration of PCA and linear regression offered several methodological advantages:

- **Dimensional Efficiency:** PCA reduced the high-dimensional dataset to a smaller set of principal components, capturing more than 95% of the variance while mitigating multicollinearity.
- **Interpretability:** Linear regression provided clear insights into the relationship between principal components and stock prices, unlike many black-box machine learning models.
- **Computational Simplicity:** The combined approach was efficient to implement and scalable across multiple equities, making it practical for both academic research and applied investment analysis.

6.4 Limitations and Sources of Error

Despite its strengths, the model has several limitations:

- **Linear Assumption:** The reliance on a linear framework limits the model's ability to capture non-linear dependencies inherent in financial markets.
- **Volatility Sensitivity:** High-volatility equities resulted in larger prediction errors, underscoring the challenge of modeling unpredictable growth stocks with regression-based methods.
- **Exclusion of Fundamentals:** The model focused solely on price and return data, excluding macroeconomic indicators, earnings reports, or sentiment analysis, which may provide additional predictive power.
- **Market Shocks:** Structural breaks such as the COVID-19 pandemic or geopolitical events introduced irregular patterns that the model could not anticipate.

These limitations highlight the scope for future research, particularly the adoption of non-linear and hybrid models (e.g., ARIMA, GARCH, LSTM, or ensemble methods) to improve predictive accuracy for volatile equities.

7 Conclusion

This section summarizes the main findings of the study and outlines the practical implications for investors and researchers.

7.1 Summary of Key Findings

The study applied moving averages, Principal Component Analysis (PCA), and linear regression to forecast stock prices for ten major U.S. equities over the period 2015–2025. The key findings are as follows:

- The PCA-enhanced linear regression model achieved strong predictive performance overall, with an average MAE of approximately $9 and MAPE of under 8%.

- Stable, defensive stocks such as PG, JNJ, AAPL, and MSFT were predicted with higher accuracy ($R^2 > 0.95$), while volatile growth stocks such as TSLA and META exhibited larger errors.
- PCA proved effective for dimensionality reduction, capturing more than 95% of the variance with fewer than 15 components, thereby improving computational efficiency and mitigating multicollinearity.
- Investment performance analysis revealed that technology-focused equities such as NVDA and TSLA generated outsized long-term returns, while consumer staples offered steady but modest growth.
- An equal-weighted portfolio of the ten equities outperformed the S&P 500 benchmark, both in absolute growth and in risk-adjusted terms, as measured by the Sharpe ratio.

7.2 Practical Investment Implications

The findings carry several implications for investment strategy:

- **Model Application:** PCA-enhanced regression is a useful baseline model for forecasting stable equities and can support portfolio analysis, though more advanced non-linear methods may be required for volatile assets.
- **Diversification Benefits:** While individual high-growth stocks delivered extraordinary returns, combining them with defensive stocks in an equal-weighted portfolio improved risk-adjusted performance and reduced volatility.
- **Market Comparison:** The superior performance of the constructed portfolio relative to the S&P 500 highlights the potential of data-driven stock selection for outperforming passive benchmarks.

In conclusion, this research demonstrates the value of integrating dimensionality reduction and regression techniques for stock forecasting and investment evaluation. Future work should expand the analysis to incorporate non-linear models, fundamental variables, and broader datasets to further enhance predictive power and practical relevance.