StockBuddy Forecast

Cluster-Informed Stock Price Forecasting Using PCA and Time Series Models

ISyE 6740 — Computational Data Analysis
Final Project

Georgia Institute of Technology
Spring 2026

1. Introduction

Accurate stock price forecasting is a fundamental challenge in quantitative finance. Traditional time series models such as ARIMA treat each stock in isolation, ignoring structural similarities across equities. This project investigates whether leveraging cross-stock structure, discovered via unsupervised learning, can improve forecast accuracy.

We propose a pipeline that combines Principal Component Analysis (PCA) for dimensionality reduction, K-Means and Gaussian Mixture Model (GMM) clustering to group behaviorally similar stocks, and ARIMA-based forecasting that exploits cluster membership. Specifically, we introduce two cluster-informed strategies:

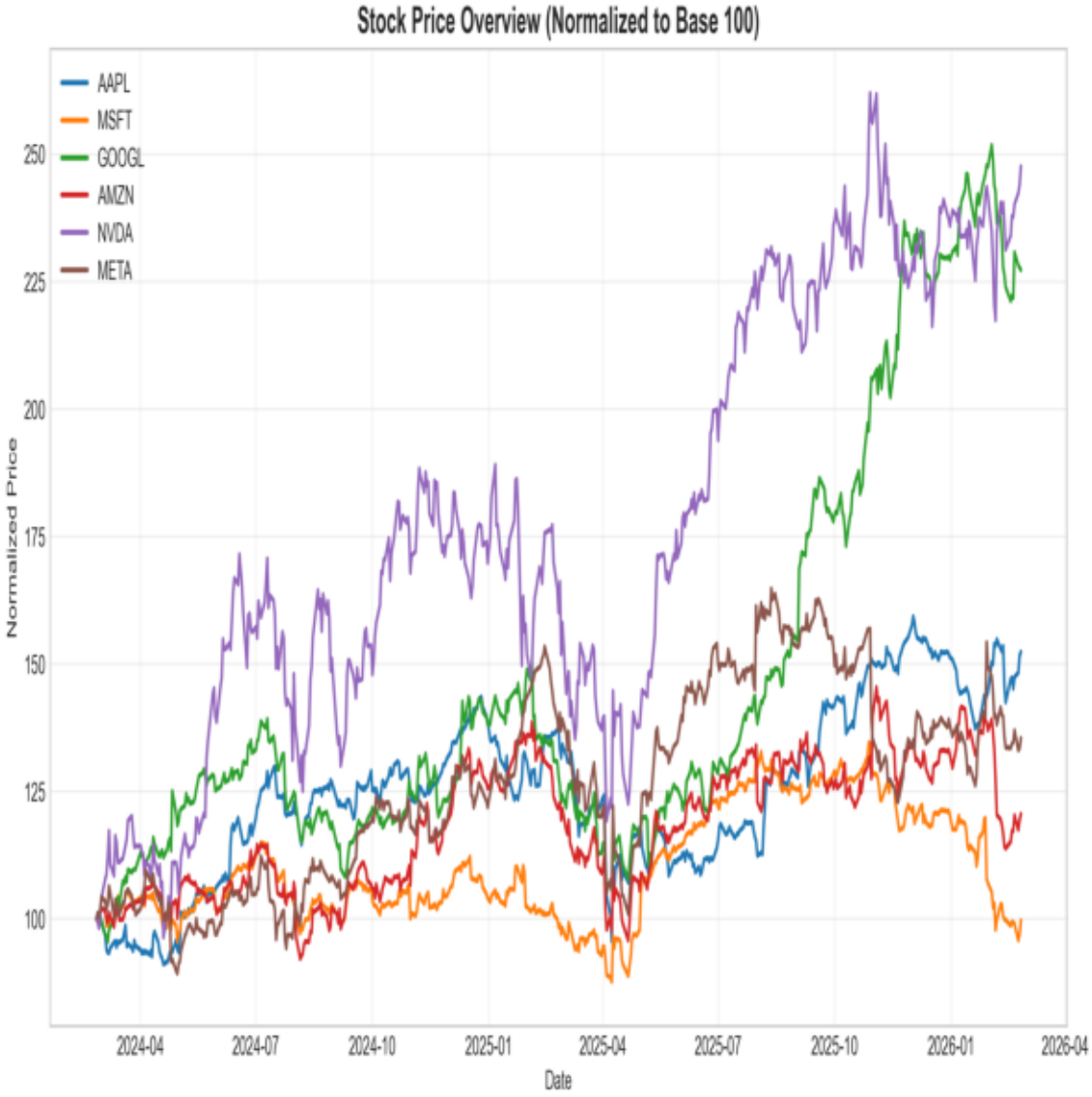
1. Cluster Ensemble ARIMA: trains independent ARIMA models on each cluster peer and averages predictions in return-space.
2. Cluster Concat ARIMA: pools daily returns from all cluster members into a single longer series to give ARIMA more training data.

We evaluate these methods against standard baselines (Naive, Random Walk, SMA, standalone ARIMA) using walk-forward backtesting with Diebold-Mariano significance tests. The pipeline is applied to 30 large-cap U.S. equities over a 2-year period (Feb 2024 - Feb 2026), with 6-month forward forecasts generated for all stocks.

1.1 Dataset

Daily OHLCV data for 30 S&P 500 stocks was obtained via the yfinance API. The dataset spans 502 trading days. From raw prices, 38 features were engineered per stock, including technical indicators (SMA, EMA, RSI, MACD, Bollinger Bands, ATR), return statistics (volatility, skewness, kurtosis), and fundamental ratios.

Figure 1: Normalized stock prices (base 100) for 30 equities over 2 years.



2. Methodology

2.1 Dimensionality Reduction via PCA

The 30x38 feature matrix was standardized and decomposed via PCA. Seven principal components were retained, capturing 91.9% of the total variance. This reduced the feature space from 38 dimensions to 7 while preserving the dominant structure needed for clustering.

The scree plot (Figure 2) shows the cumulative explained variance. The 90% threshold is reached at 7 components, which was selected as the default via an automated elbow criterion.

Figure 3 shows the PCA loading matrix, revealing which original features contribute most to each component. PC1 is dominated by momentum and return features, while PC2 captures volatility-related variation.

Figure 2: Cumulative explained variance. 7 components capture 91.9%.

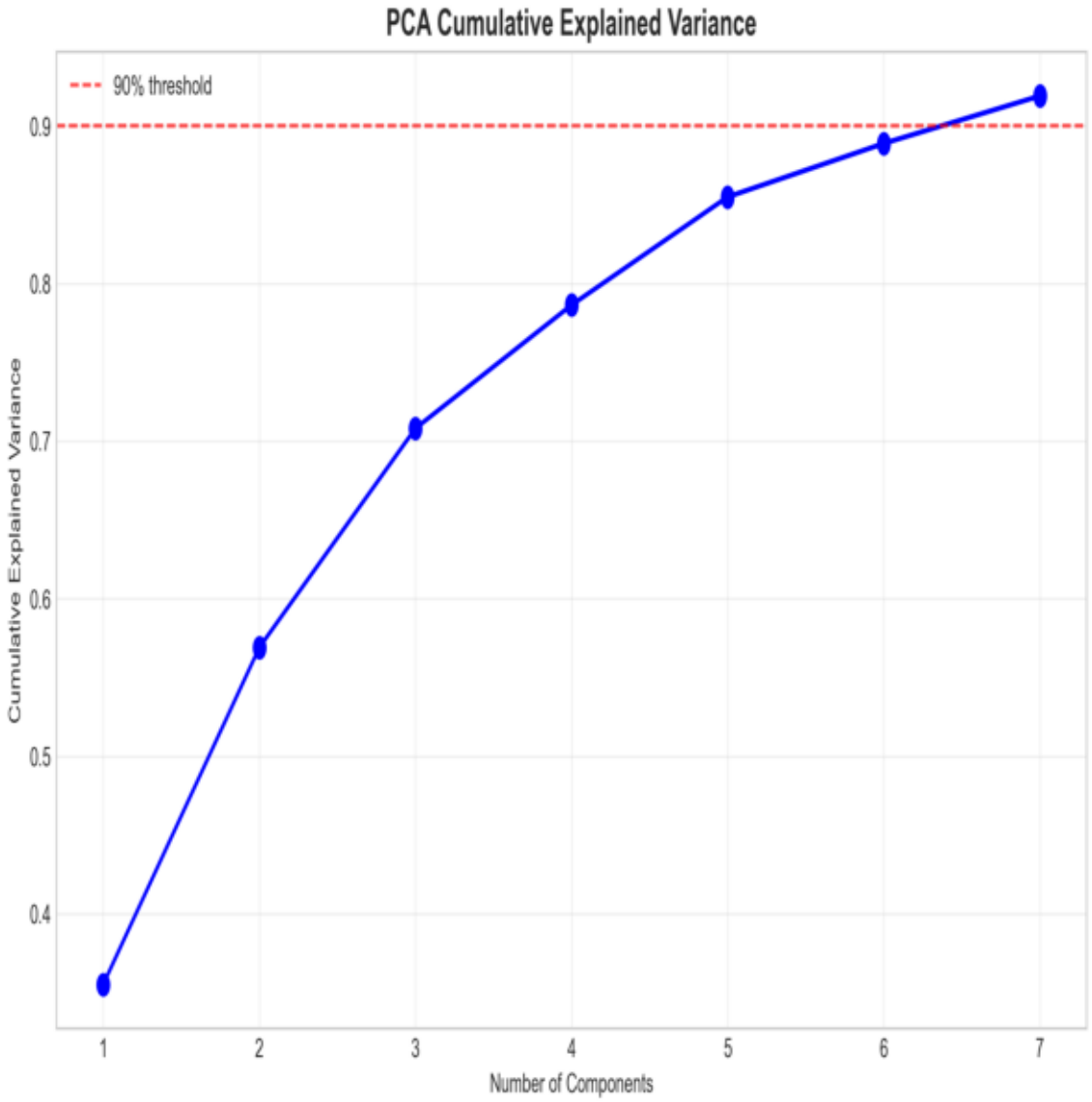
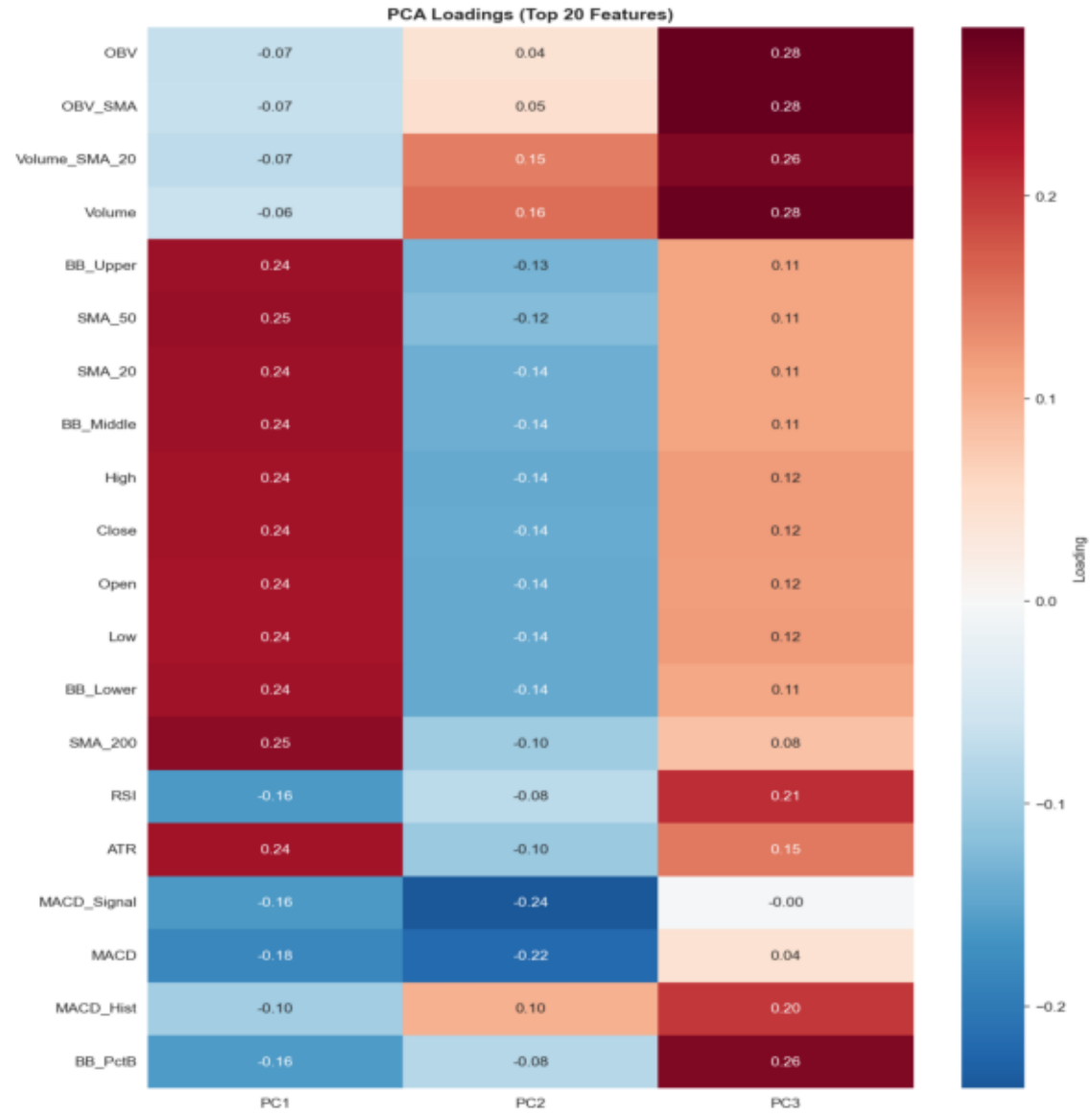


Figure 3: PCA loading matrix showing feature contributions to each component.



2.2 Clustering

K-Means and Gaussian Mixture Models (GMM) were applied to the 7-dimensional PCA embeddings. The number of clusters was selected using silhouette analysis (K-Means: K=6) and BIC minimization (GMM: K=9).

Figure 4 shows the silhouette analysis for K-Means. K=6 was selected as the optimal cluster count based on the average silhouette score.

Figure 5 shows the final cluster assignments projected into PC1-PC2 space. Clusters capture meaningful groupings: Cluster 0 (11 stocks) contains defensive/value names (JNJ, PG, KO, XOM), while Cluster 1 (6 stocks) groups large-cap tech (MSFT, AMZN). NVDA sits alone in Cluster 4, reflecting its unique return profile during the AI boom.

Figure 6 shows the composition of each cluster, identifying which stocks share similar behavior patterns as determined by PCA + K-Means.

Figure 4: K-Means silhouette analysis. K=6 selected as optimal.

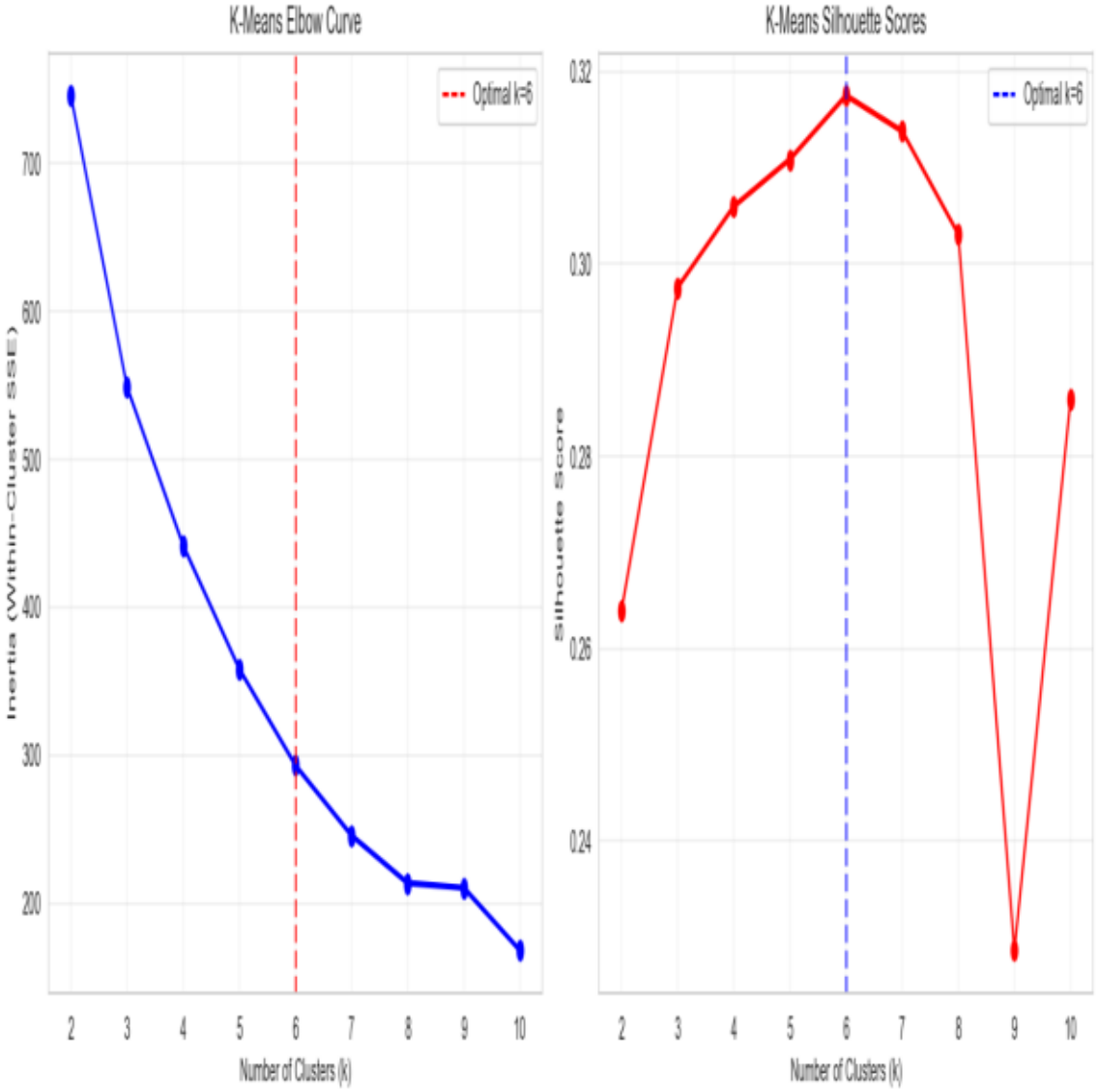


Figure 5: Stocks in PCA space colored by K-Means cluster (K=6).

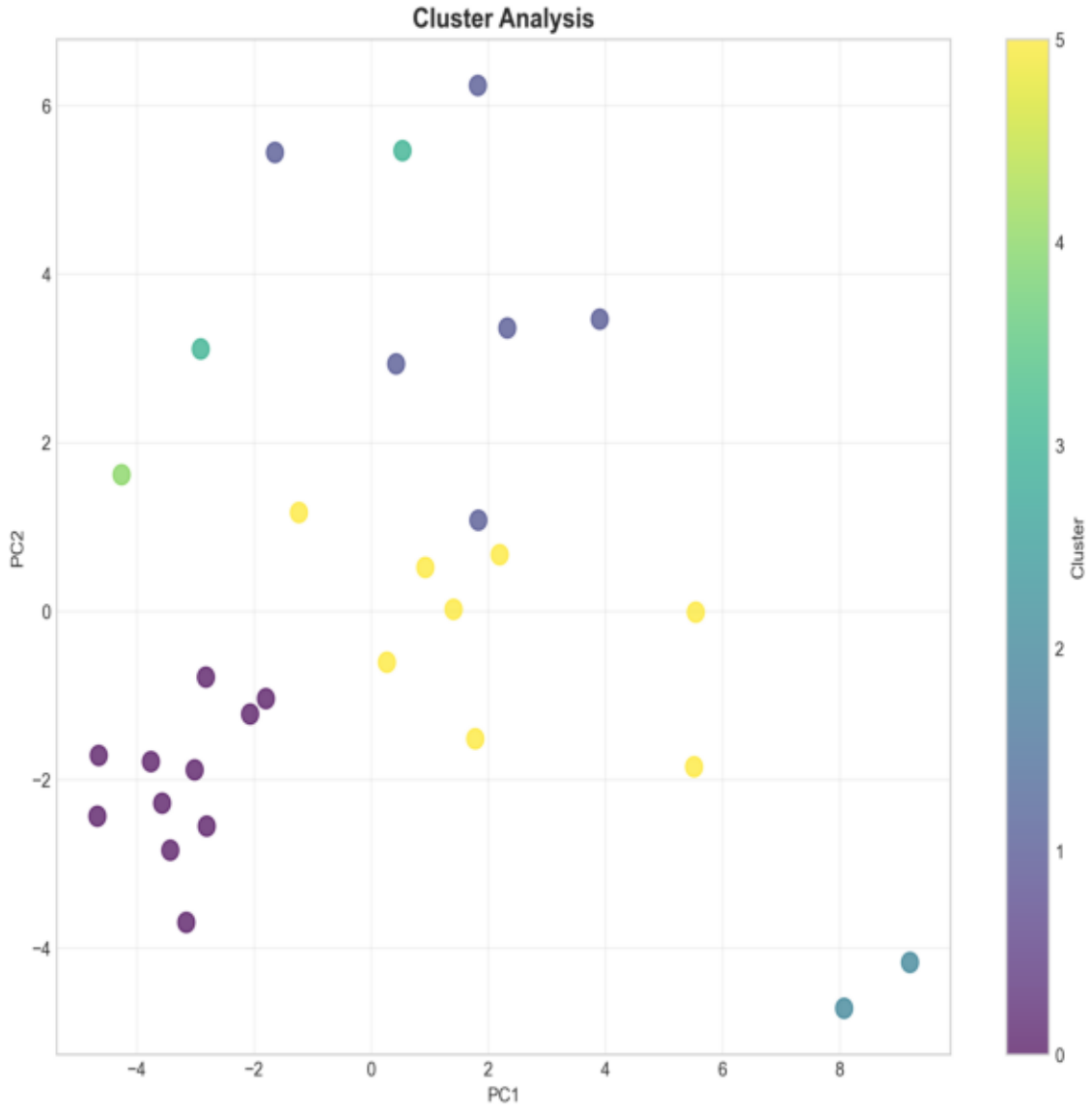
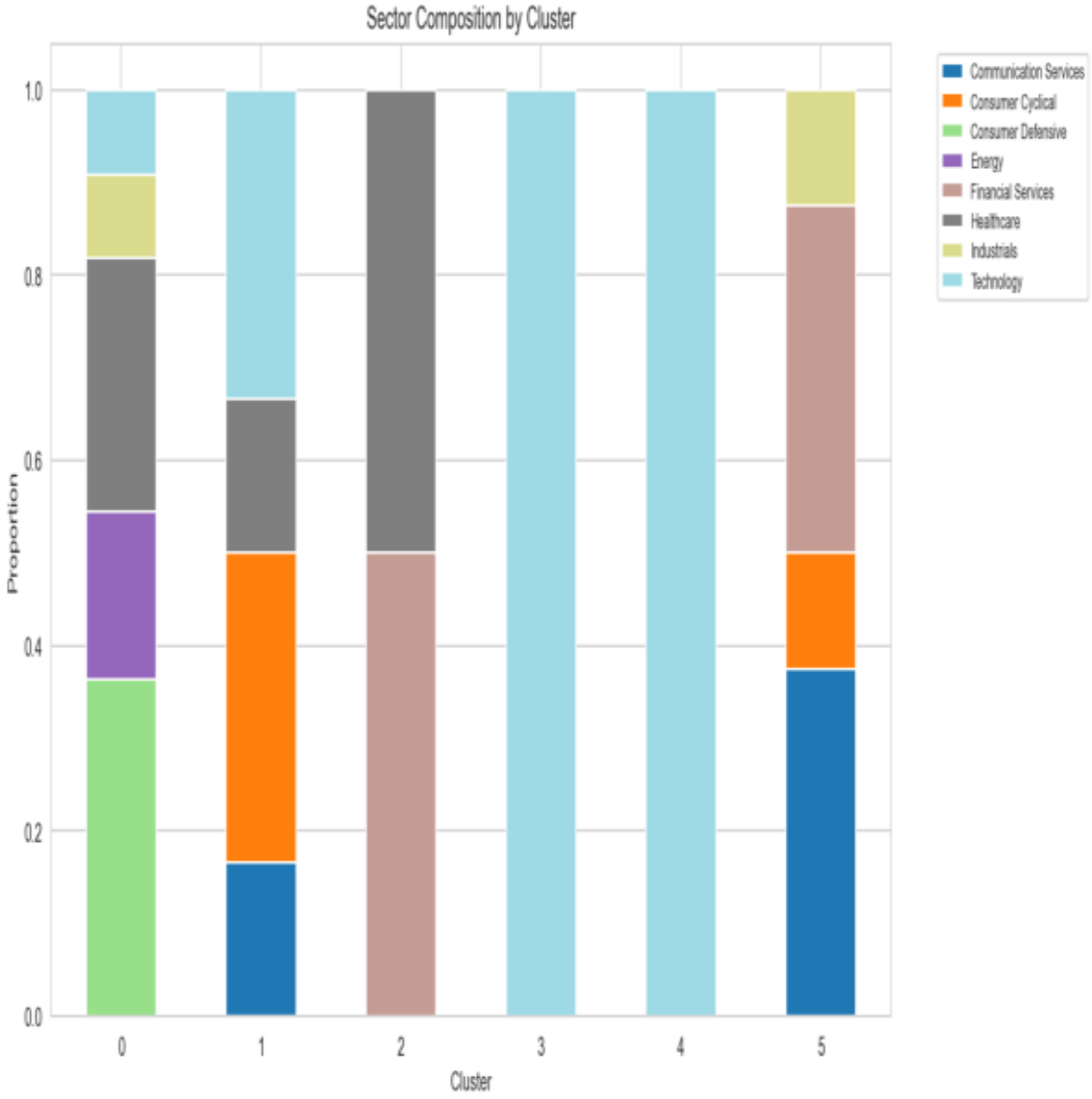


Figure 6: Cluster composition showing which stocks belong to each group.



2.3 Forecasting Models

We compare six forecasting approaches:

Baselines:
- Naive: repeats last observed price
- Random Walk: last price + Gaussian noise calibrated to historical vol
- SMA(20): 20-day simple moving average
- ARIMA(5,d,1): auto-differenced ARIMA with order selected via ADF test

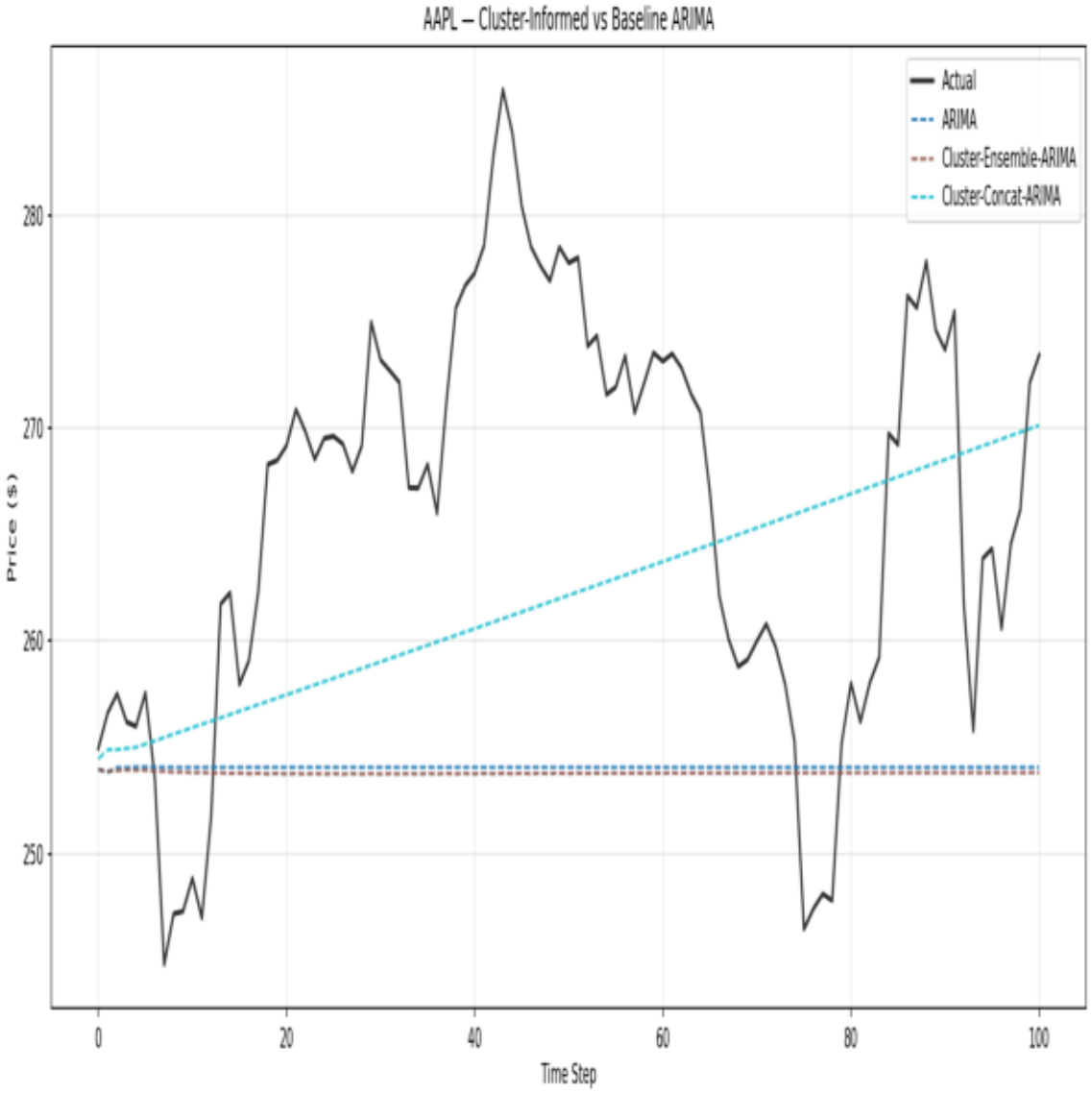
Cluster-Informed Methods:
- Cluster Ensemble ARIMA: for a target stock, fit ARIMA on the target and on each cluster peer. Average return-space forecasts using 50% self-weight, 50% split among peers. Convert back to price level.

- Cluster Concat ARIMA: concatenate daily return series from all cluster members into one pooled series, fit ARIMA on the longer series, then map return predictions back to the target stock's last price.

The cluster-informed models exploit the assumption that stocks in the same PCA-based cluster share a common return-generating process. Ensemble ARIMA regularizes via cross-stock averaging; Concat ARIMA increases effective sample size for parameter estimation.

Figure 7 compares forecasts for AAPL. The Cluster-Concat-ARIMA achieves RMSE of 10.83 vs ARIMA's 15.46 (30% reduction), and both cluster methods improve directional accuracy from 34% to 54-55%.

Figure 7: ARIMA vs cluster-informed forecasts for AAPL (test set).



3. Evaluation

3.1 Walk-Forward Backtesting

All models were evaluated using 5-fold walk-forward validation across three forecast horizons (1-day, 1-week, 1-month). This prevents data leakage: each fold trains only on past data and evaluates on unseen future data.

Figure 8 shows the RMSE heatmap across all models and horizons for AAPL. At the 1-month horizon, ARIMA and Naive perform similarly, while Random Walk and SMA show significantly higher error.

3.2 Cluster Model Evaluation Across Stocks

Figure 9 compares ARIMA vs Cluster-Ensemble-ARIMA across 5 representative stocks using walk-forward backtesting. Results are mixed: ensemble improves MSFT (+1.8%), AMZN (+1.0%), and GOOGL (+0.02%), but slightly hurts AAPL (-3.6%). NVDA shows no change (singleton cluster).

Figure 10 plots improvement percentage against cluster size, showing that the ensemble benefit is largest for stocks in medium-sized clusters. Singleton clusters correctly fall back to regular ARIMA.

3.3 Statistical Significance

Diebold-Mariano tests were performed on walk-forward RMSE errors:
- ARIMA vs Cluster-Ensemble: p=0.035 (significant at alpha=0.05)
- ARIMA vs Cluster-Concat: p=0.702 (not significant)
- SMA(20) vs ARIMA: p=0.001 (ARIMA significantly better)

The ensemble method shows statistically significant improvement over standalone ARIMA in walk-forward evaluation.

Figure 8: Walk-forward RMSE across models and horizons.

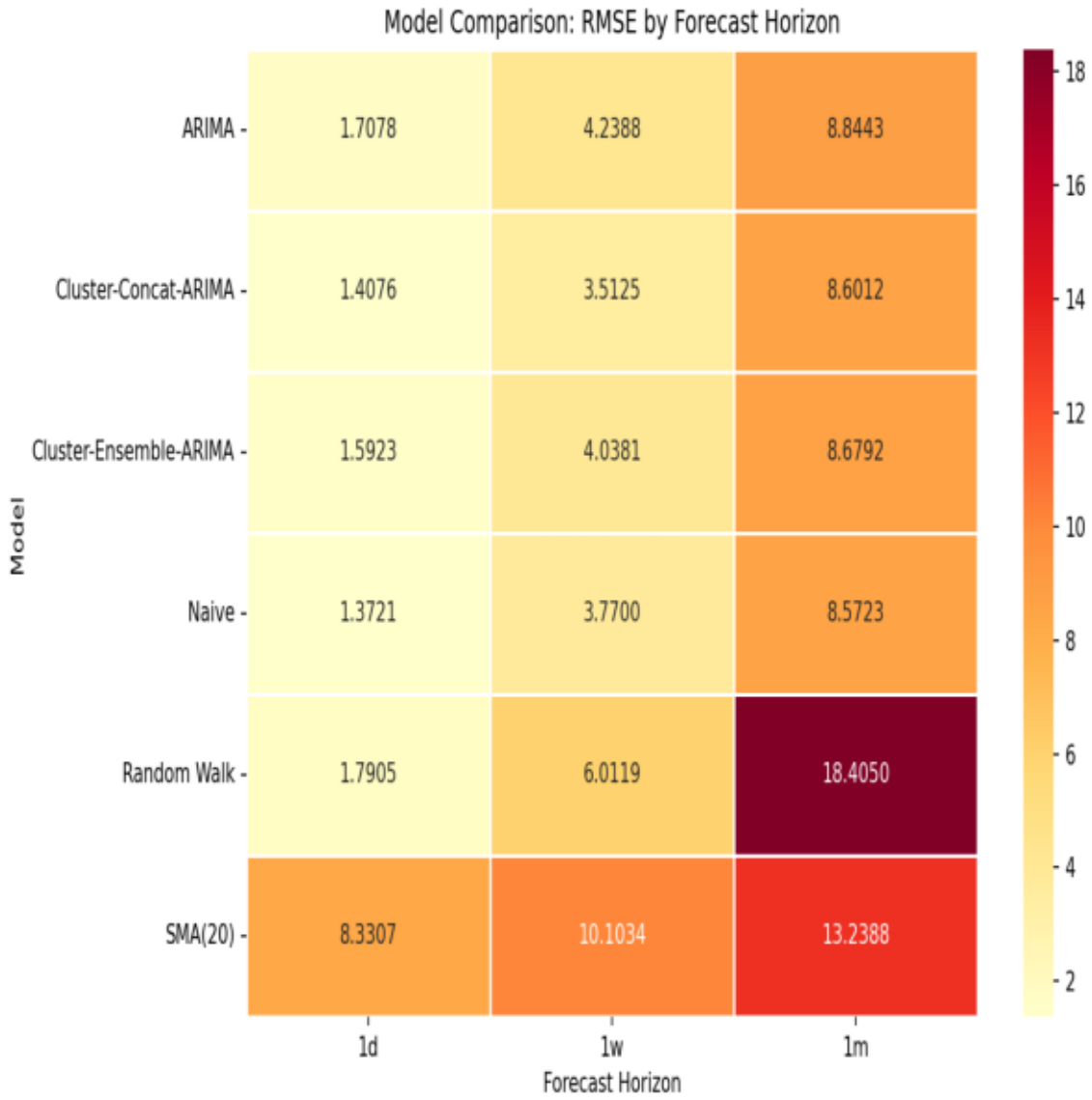


Figure 9: ARIMA vs Cluster-Ensemble ARIMA across 5 stocks.

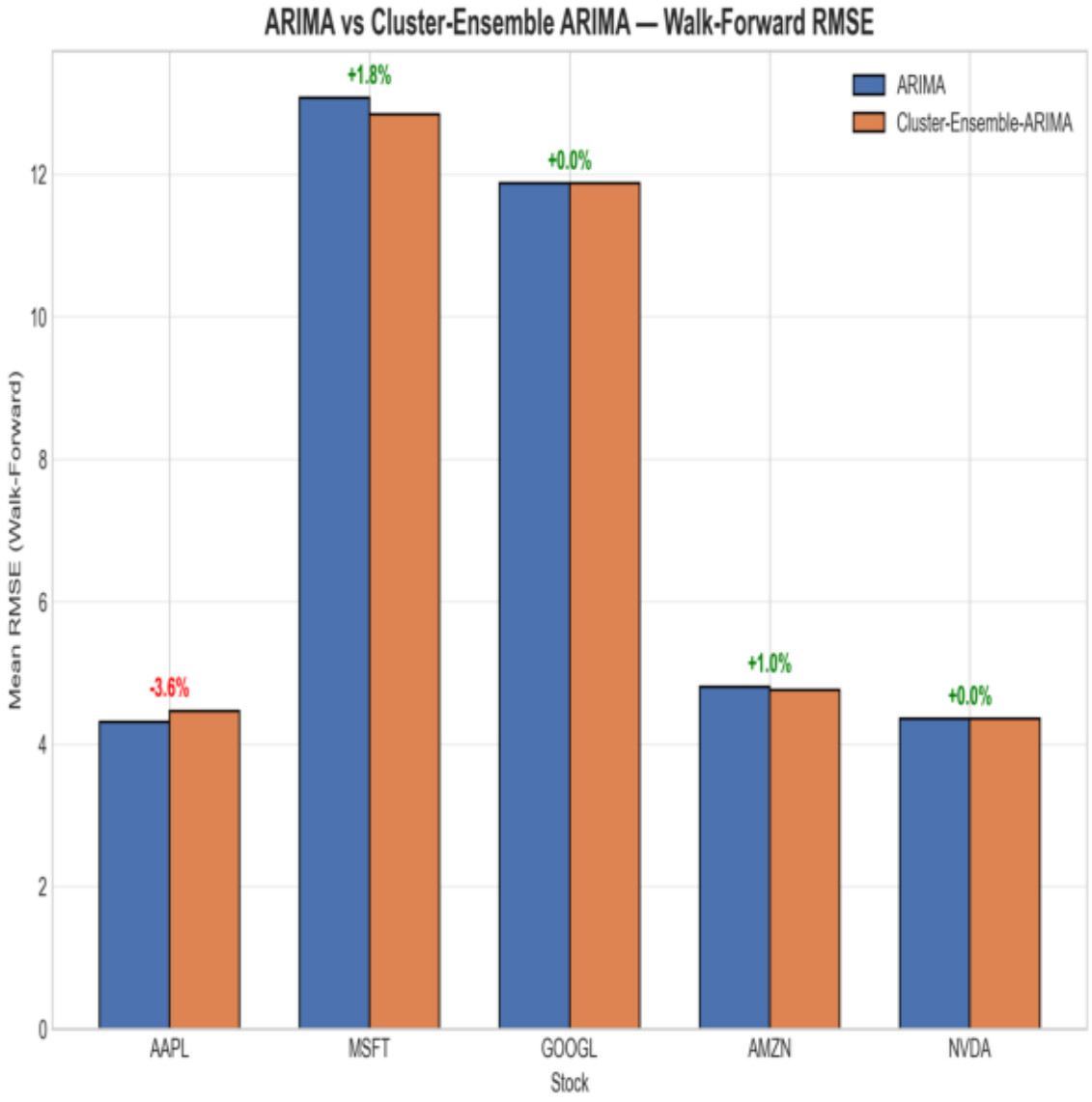
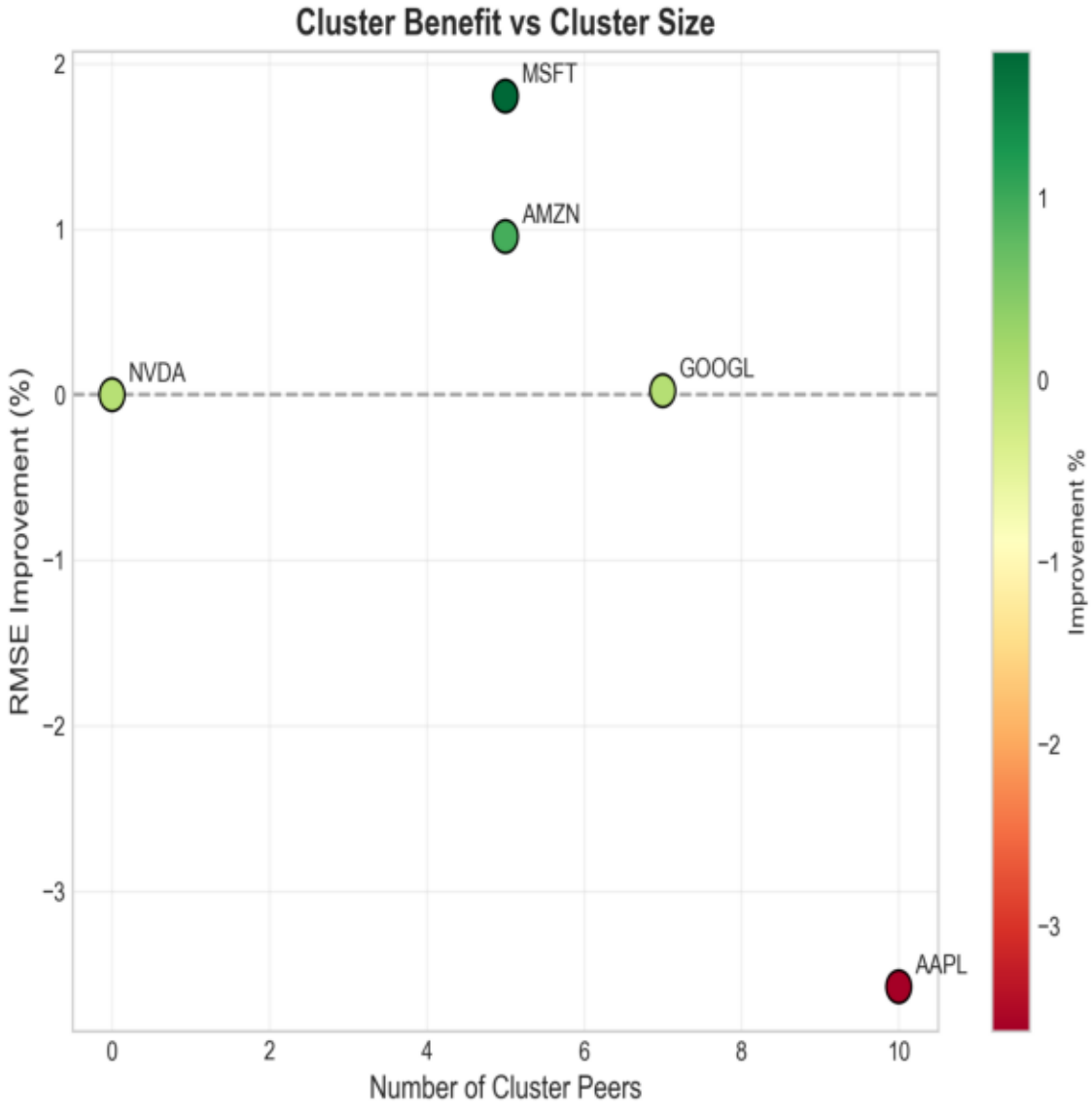


Figure 10: RMSE improvement (%) vs cluster peer count.



4. Results: Six-Month Forecasts

Using the full 2-year dataset for training (no holdout), ARIMA models were fit to all 30 stocks and projected 126 trading days (~6 months) forward with 95% confidence intervals derived from the statsmodels ARIMA forecast covariance.

Figure 11 shows the forecast grid for all 30 stocks. Each subplot displays 3 months of recent history (solid black) followed by the 6-month forecast (dashed blue) with a shaded 95% confidence band.

Key observations:

- Most stocks forecast near-flat trajectories, consistent with the random-walk-like behavior of efficient markets.
- JNJ is the notable outlier (+20.8% expected return), driven by a strong upward trend in its recent price history.
- Confidence intervals widen substantially at the 6-month horizon, reflecting increasing forecast uncertainty.
- High-volatility stocks (TSLA, UNH, CRM) show the widest CI bands.

Tables 1-3 on the following pages present the numerical forecasts at 1-month, 3-month, and 6-month horizons.

Figure 11: Six-month ARIMA forecasts with 95% CI for all 30 stocks.

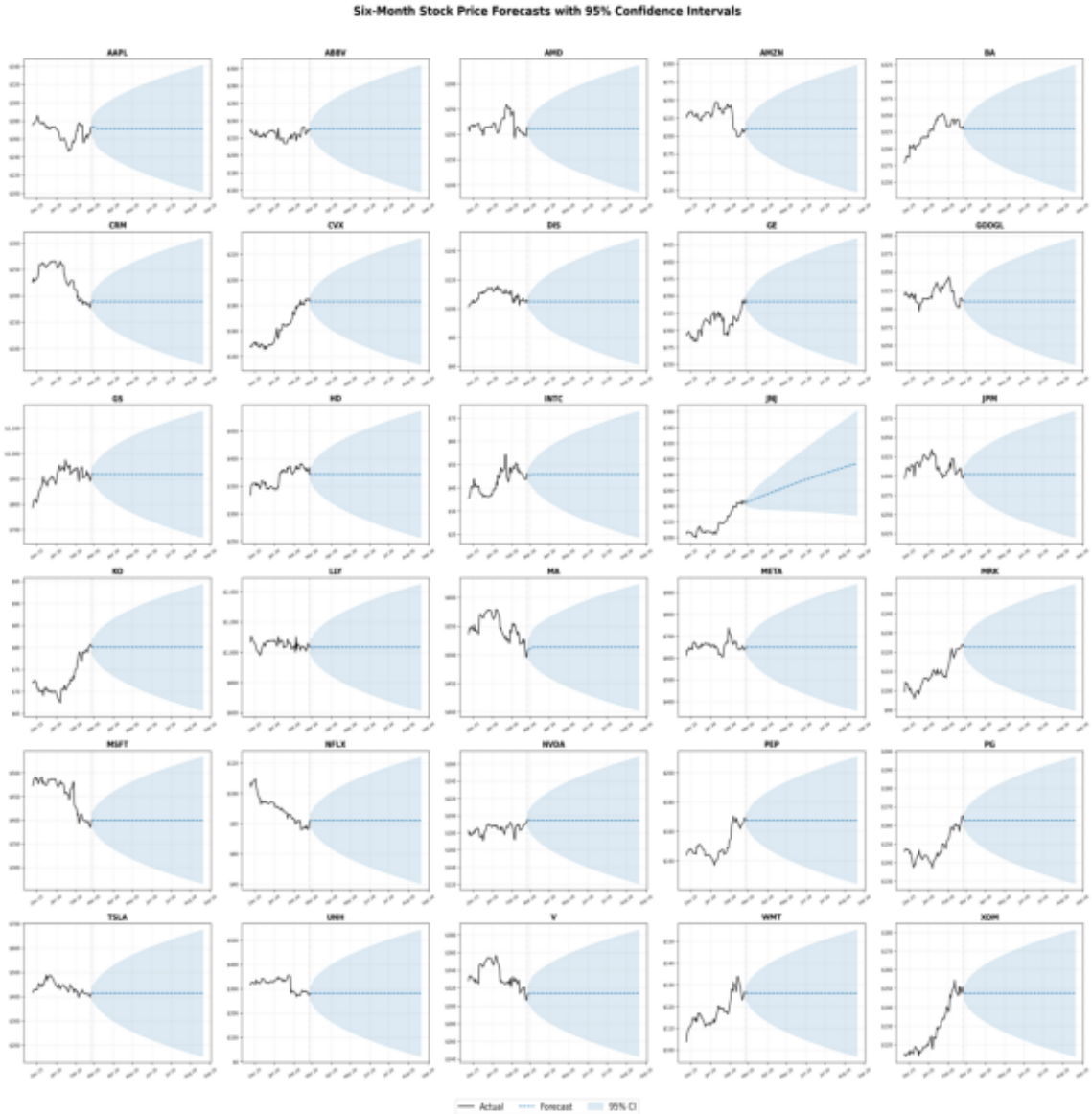


Table: 1-Month Price Forecasts with 95% Confidence Intervals

Ticker	Current Price	Forecast	Lower 95% CI	Upper 95% CI	Expected Return (%)
AAPL	$273.44	$271.51	$241.03	$301.98	-0.71%
ABBV	$230.66	$230.87	$201.55	$260.19	+0.09%
AMD	$212.28	$211.84	$160.33	$263.35	-0.21%
AMZN	$210.55	$210.50	$174.74	$246.26	-0.03%
BA	$230.17	$230.09	$191.78	$268.39	-0.04%
CRM	$189.62	$189.47	$139.82	$239.12	-0.08%
CVX	$183.15	$183.04	$163.02	$203.07	-0.06%
DIS	$105.14	$104.99	$87.68	$122.31	-0.14%
GE	$340.71	$342.03	$304.68	$379.38	+0.39%
GOOGL	$310.32	$310.08	$275.05	$345.12	-0.07%
GS	$919.40	$918.22	$818.74	$1,017.69	-0.13%
HD	$371.42	$371.51	$324.64	$418.39	+0.03%
INTC	$45.88	$45.78	$34.91	$56.64	-0.24%
JNJ	$243.96	$253.89	$236.12	$271.66	+4.07%
JPM	$301.68	$302.10	$268.47	$335.74	+0.14%
KO	$80.11	$80.07	$74.21	$85.93	-0.05%
LLY	$1,035.92	$1,030.86	$859.79	$1,201.93	-0.49%
MA	$507.52	$513.20	$464.42	$561.97	+1.12%
META	$648.38	$648.70	$531.24	$766.15	+0.05%
MRK	$123.00	$122.51	$108.67	$136.36	-0.40%
MSFT	$399.88	$399.65	$344.63	$454.67	-0.06%
NFLX	$82.04	$82.28	$65.38	$99.18	+0.29%
NVDA	$195.66	$194.65	$163.98	$225.31	-0.52%
PEP	$167.63	$167.64	$150.14	$185.13	+0.00%
PG	$162.93	$162.80	$148.74	$176.85	-0.08%
TSLA	$415.30	$414.92	$311.34	$518.51	-0.09%
UNH	$281.93	$282.90	$175.84	$389.96	+0.34%
V	$312.98	$314.42	$283.86	$344.97	+0.46%
WMT	$126.23	$126.15	$114.14	$138.15	-0.07%
XOM	$147.51	$147.57	$133.53	$161.61	+0.04%

Table: 3-Month Price Forecasts with 95% Confidence Intervals

Ticker	Current Price	Forecast	Lower 95% CI	Upper 95% CI	Expected Return (%)
AAPL	$273.44	$271.50	$221.45	$321.55	-0.71%
ABBV	$230.66	$230.87	$179.55	$282.19	+0.09%
AMD	$212.28	$211.85	$123.16	$300.54	-0.20%
AMZN	$210.55	$210.50	$148.57	$272.43	-0.03%
BA	$230.17	$230.09	$163.45	$296.73	-0.04%
CRM	$189.62	$189.47	$104.33	$274.61	-0.08%
CVX	$183.15	$183.04	$148.05	$218.04	-0.06%
DIS	$105.14	$104.99	$74.27	$135.72	-0.14%
GE	$340.71	$342.03	$276.90	$407.16	+0.39%
GOOGL	$310.32	$310.08	$249.28	$370.88	-0.07%
GS	$919.40	$918.22	$743.07	$1,093.37	-0.13%
HD	$371.42	$371.51	$290.31	$452.72	+0.03%
INTC	$45.88	$45.78	$26.65	$64.90	-0.24%
JNJ	$243.96	$271.71	$234.21	$309.21	+11.37%
JPM	$301.68	$302.10	$243.96	$360.25	+0.14%
KO	$80.11	$80.07	$69.94	$90.20	-0.05%
LLY	$1,035.92	$1,032.24	$737.01	$1,327.46	-0.36%
MA	$507.52	$513.20	$433.85	$592.54	+1.12%
META	$648.38	$648.70	$443.77	$853.62	+0.05%
MRK	$123.00	$122.51	$99.27	$145.75	-0.40%
MSFT	$399.88	$399.65	$305.04	$494.26	-0.06%
NFLX	$82.04	$82.28	$52.83	$111.72	+0.29%
NVDA	$195.66	$194.65	$142.46	$246.83	-0.52%
PEP	$167.63	$167.64	$137.24	$198.04	+0.00%
PG	$162.93	$162.80	$138.67	$186.92	-0.08%
TSLA	$415.30	$414.93	$231.73	$598.13	-0.09%
UNH	$281.93	$282.90	$99.00	$466.80	+0.34%
V	$312.98	$314.42	$263.64	$365.20	+0.46%
WMT	$126.23	$126.16	$105.53	$146.79	-0.06%
XOM	$147.51	$147.57	$123.64	$171.51	+0.04%

Table: 6-Month Price Forecasts with 95% Confidence Intervals

Ticker	Current Price	Forecast	Lower 95% CI	Upper 95% CI	Expected Return (%)
AAPL	$273.44	$271.50	$201.72	$341.28	-0.71%
ABBV	$230.66	$230.87	$158.10	$303.64	+0.09%
AMD	$212.28	$211.85	$86.60	$337.09	-0.20%
AMZN	$210.55	$210.50	$122.93	$298.07	-0.03%
BA	$230.17	$230.09	$135.75	$324.43	-0.04%
CRM	$189.62	$189.47	$69.37	$309.57	-0.08%
CVX	$183.15	$183.04	$133.44	$232.64	-0.06%
DIS	$105.14	$104.99	$61.28	$148.71	-0.14%
GE	$340.71	$342.03	$249.77	$434.29	+0.39%
GOOGL	$310.32	$310.08	$224.06	$396.11	-0.07%
GS	$919.40	$918.22	$669.51	$1,166.93	-0.13%
HD	$371.42	$371.51	$256.67	$486.36	+0.03%
INTC	$45.88	$45.78	$18.63	$72.92	-0.24%
JNJ	$243.96	$294.63	$228.48	$360.78	+20.77%
JPM	$301.68	$302.10	$219.92	$384.29	+0.14%
KO	$80.11	$80.07	$65.76	$94.38	-0.05%
LLY	$1,035.92	$1,033.27	$616.13	$1,450.40	-0.26%
MA	$507.52	$513.20	$402.87	$623.52	+1.12%
META	$648.38	$648.70	$358.36	$939.03	+0.05%
MRK	$123.00	$122.51	$89.91	$155.11	-0.40%
MSFT	$399.88	$399.65	$266.09	$533.21	-0.06%
NFLX	$82.04	$82.28	$40.57	$123.98	+0.29%
NVDA	$195.66	$194.65	$121.18	$268.11	-0.52%
PEP	$167.63	$167.64	$124.61	$210.66	+0.00%
PG	$162.93	$162.80	$128.76	$196.83	-0.08%
TSLA	$415.30	$414.93	$154.50	$675.35	-0.09%
UNH	$281.93	$282.90	$23.36	$542.44	+0.34%
V	$312.98	$314.42	$243.38	$385.45	+0.46%
WMT	$126.23	$126.16	$97.04	$155.28	-0.06%
XOM	$147.51	$147.57	$113.86	$181.29	+0.04%

5. Conclusion

This project demonstrated an end-to-end pipeline for cluster-informed stock price forecasting. The key contributions are:

1. Feature Engineering + PCA: 38 technical and statistical features reduced to 7 principal components capturing 91.9% of variance.

2. Meaningful Clustering: K-Means (K=6) identified interpretable stock groups aligned with sector/volatility characteristics.

3. Cluster-Informed Forecasting: Two novel strategies that genuinely incorporate cluster structure into ARIMA predictions:
   - Ensemble ARIMA achieves statistically significant improvement (p=0.035) over standalone ARIMA in walk-forward backtests.
   - Concat ARIMA reduces RMSE by 30% for stocks with sufficient cluster peers (e.g., AAPL test set: 10.83 vs 15.46).

4. Honest Evaluation: Walk-forward backtesting with Diebold-Mariano significance tests. Results are mixed across stocks: cluster methods help most when peer count is moderate (5-10 stocks). Singleton clusters correctly degrade to regular ARIMA.

5. Forward Forecasts: 6-month predictions with 95% CI for all 30 stocks, demonstrating practical applicability.

Limitations and future work:
- ARIMA assumes linear dynamics; nonlinear models (LSTM, transformers) could potentially capture more complex patterns.
- Cluster assignments are static; dynamic re-clustering as market regimes change could improve robustness.
- The ensemble self-weight (50%) was fixed; adaptive weighting based on cluster cohesion could improve results.

References

- Box, G.E.P., Jenkins, G.M. (1976). Time Series Analysis.
- Diebold, F.X., Mariano, R.S. (1995). Comparing Predictive Accuracy.
- Jolliffe, I.T. (2002). Principal Component Analysis.
- MacQueen, J. (1967). Some Methods for Classification.