StockBuddy Forecast

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

ISyE 6740 — Computational Data Analysis
Georgia Institute of Technology
Spring 2026

Abstract

We present a pipeline for stock price forecasting that leverages unsupervised learning to improve time series predictions. Starting with 38 engineered features for 30 large-cap U.S. equities, we apply PCA (7 components, 91.9% variance explained) and K-Means clustering (K=6) to discover groups of behaviorally similar stocks. We introduce two cluster-informed ARIMA strategies: *Cluster Ensemble ARIMA*, which averages return-space forecasts from cluster peers, and *Cluster Concat ARIMA*, which pools returns from all cluster members before fitting. Walk-forward backtesting with Diebold-Mariano tests shows the ensemble method achieves statistically significant improvement over standalone ARIMA (p=0.035). We generate 6-month forward forecasts with 95% confidence intervals for all 30 stocks.

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:

1. **PCA** for dimensionality reduction (38 features → 7 components),
2. **K-Means and GMM clustering** to group behaviorally similar stocks,
3. **Cluster-informed ARIMA** forecasting that exploits cluster membership.

Two cluster-informed strategies are introduced:

- **Cluster Ensemble ARIMA:** Trains independent ARIMA models on each cluster peer's price series, then averages predictions in return-space with 50% self-weight.
- **Cluster Concat ARIMA:** Concatenates daily return series from all cluster members into a single pooled series, giving ARIMA more data for parameter estimation.

1.1 Dataset

Daily OHLCV data for 30 S&P 500 stocks was obtained via the `yfinance` API, spanning 502 trading days (February 2024–February 2026). 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.

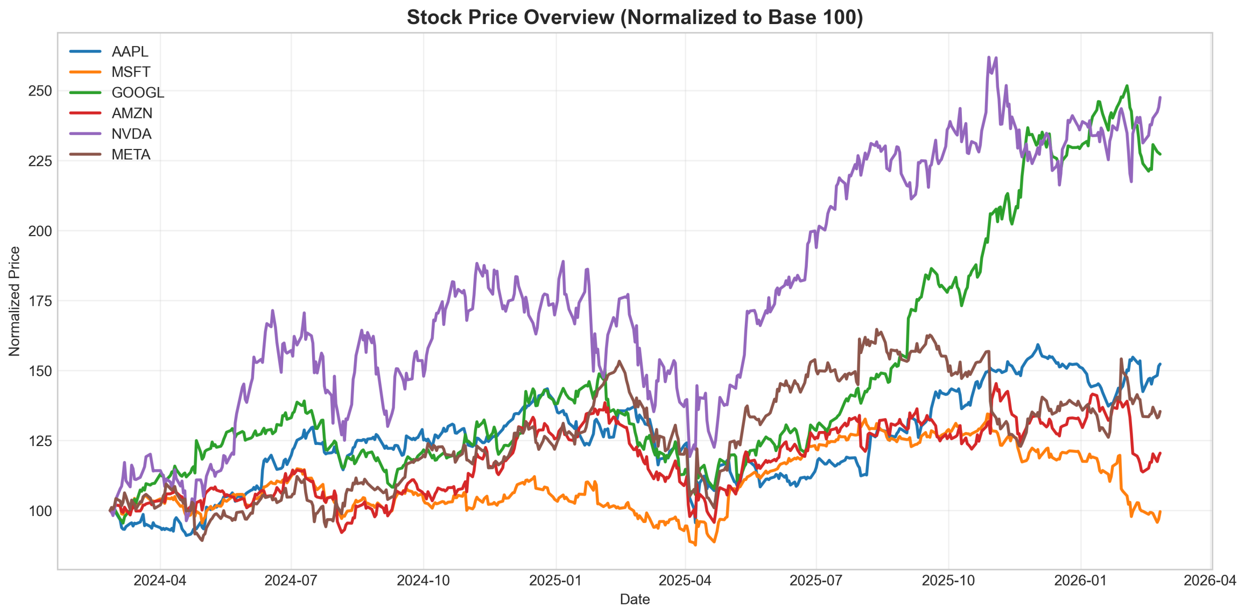


Figure 1: Normalized stock prices (base 100) for selected equities over 2 years.

2 Methodology

2.1 Dimensionality Reduction via PCA

The 30 × 38 feature matrix was standardized (zero mean, unit variance) and decomposed via Principal Component Analysis. Seven components were retained, capturing 91.9% of total variance. This reduces the feature space from 38 dimensions to 7 while preserving the dominant structure needed for clustering.

The cumulative explained variance (Figure 2) shows the 90% threshold reached at 7 components. The loading matrix (Figure 3) reveals that PC1 is dominated by momentum and return features, while PC2 captures volatility-related variation.

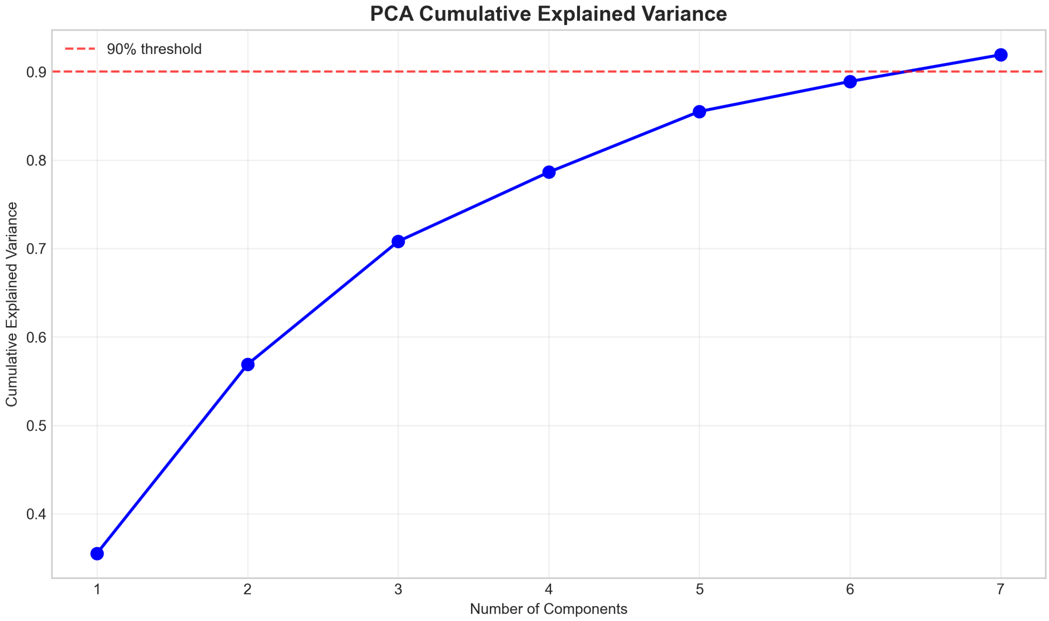
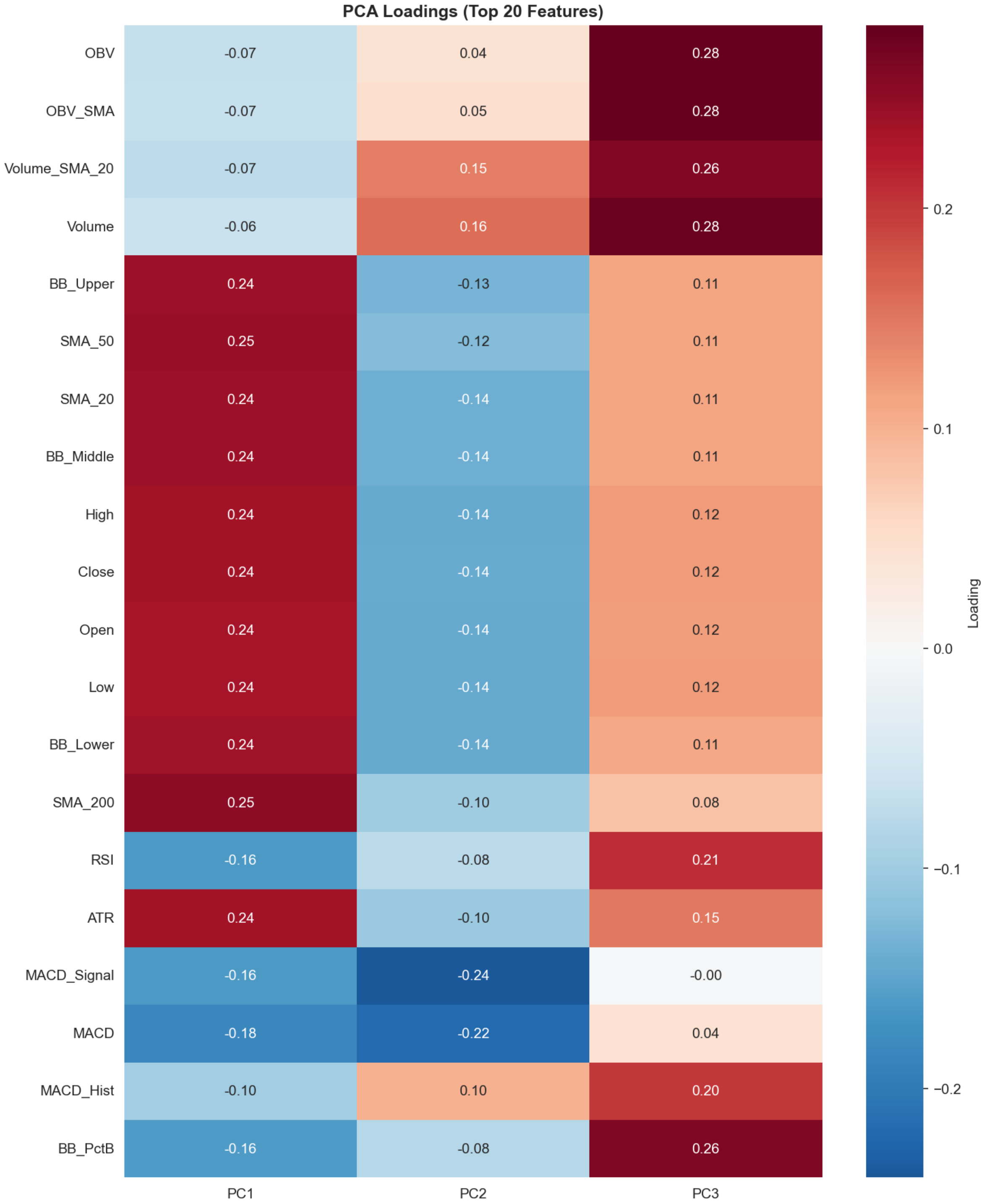


Figure 2: Cumulative explained variance. 7 components capture 91.9% of variance; red dashed line marks 90% threshold.



	PC1	PC2	PC3
OBV	-0.07	0.04	0.28
OBV_SMA	-0.07	0.05	0.28
Volume_SMA_20	-0.07	0.15	0.26
Volume	-0.06	0.16	0.28
BB_Upper	0.24	-0.13	0.11
SMA_50	0.25	-0.12	0.11
SMA_20	0.24	-0.14	0.11
BB_Middle	0.24	-0.14	0.11
High	0.24	-0.14	0.12
Close	0.24	-0.14	0.12
Open	0.24	-0.14	0.12
Low	0.24	-0.14	0.12
BB_Lower	0.24	-0.14	0.11
SMA_200	0.25	-0.10	0.08
RSI	-0.16	-0.08	0.21
ATR	0.24	-0.10	0.15
MACD_Signal	-0.16	-0.24	-0.00
MACD	-0.18	-0.22	0.04
MACD_Hist	-0.10	0.10	0.20
BB_PctB	-0.16	-0.08	0.26

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

2.2 Clustering

K-Means and Gaussian Mixture Models (GMM) were applied to the 7-dimensional PCA embeddings. Cluster count was selected using silhouette analysis (K-Means: K=6, Figure 4) and BIC minimization (GMM: K=9).

Figure 5 shows the K-Means cluster assignments in PCA space. The clusters capture meaningful groupings:

- **Cluster 0** (11 stocks): Defensive/value names (JNJ, PG, KO, XOM, CVX)
- **Cluster 1** (6 stocks): Large-cap tech (MSFT, AMZN, NFLX)
- **Cluster 4** (1 stock): NVDA alone, reflecting its unique AI-driven return profile
- **Cluster 5** (7 stocks): High-growth tech (GOOGL, META, CRM)

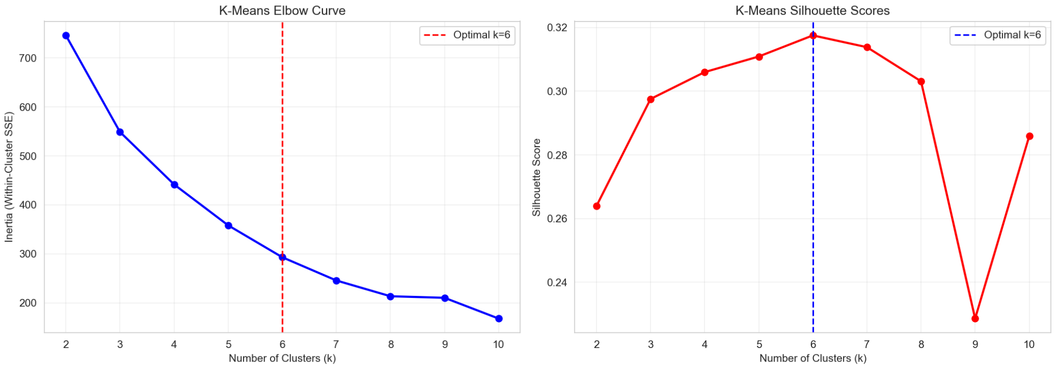


Figure 4: K-Means silhouette analysis. K=6 selected as optimal cluster count.

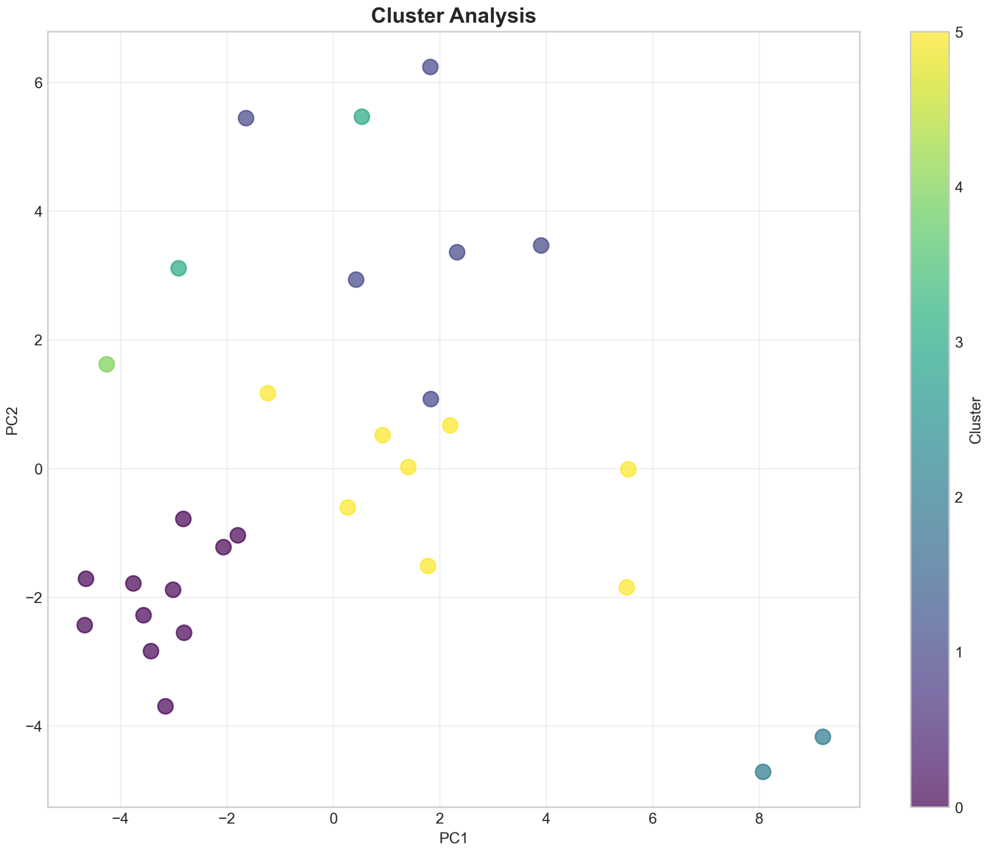


Figure 5: Stocks projected into PC1–PC2 space, colored by K-Means cluster assignment (K=6).

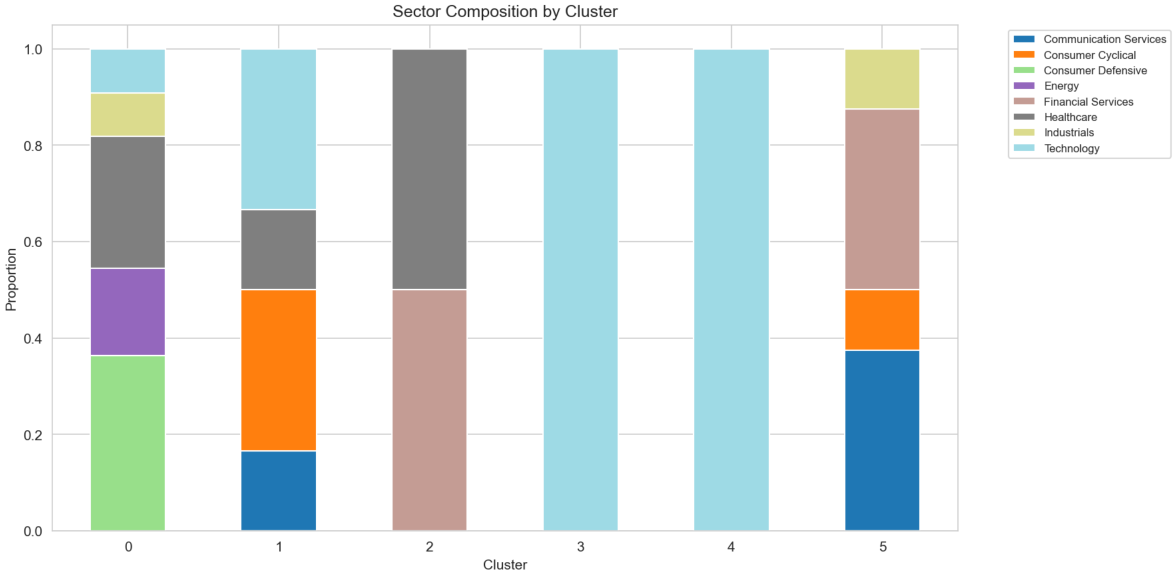


Figure 6: Cluster composition showing which stocks belong to each group.

2.3 Forecasting Models

We compare six forecasting approaches:

Baselines:

- *Naive:* Repeats the last observed price for all forecast periods.
- *Random Walk:* Last price plus Gaussian noise calibrated to historical volatility.
- *SMA(20):* 20-day simple moving average.
- *ARIMA(5,d,1):* Auto-differenced ARIMA with order of integration d selected via ADF test.

Cluster-Informed Methods:

- *Cluster Ensemble ARIMA:* For target stock i in cluster C_k, fit ARIMA on stock i and on each peer $j \in C_k \setminus \{i\}$. Convert all forecasts to return-space:

$$\hat{r}_{t+h}^{(j)} = \frac{\hat{p}_{t+h}^{(j)}}{p_t^{(j)}} - 1 \tag{1}$$

  Compute the weighted average:

$$\bar{r}_{t+h} = w_{\text{self}} \cdot \hat{r}_{t+h}^{(i)} + \frac{1 - w_{\text{self}}}{|C_k| - 1} \sum_{j \neq i} \hat{r}_{t+h}^{(j)} \tag{2}$$

  with $w_{\text{self}} = 0.5$, then convert back: $\hat{p}_{t+h}^{(i)} = p_t^{(i)}(1 + \bar{r}_{t+h})$.
- *Cluster Concat ARIMA:* Concatenate daily returns $\{r_t^{(j)}\}$ from all $j \in C_k$ into a single pooled series of length $\sum_j T_j$. Fit ARIMA on the pooled returns, then map the return forecast back to stock i's price level.

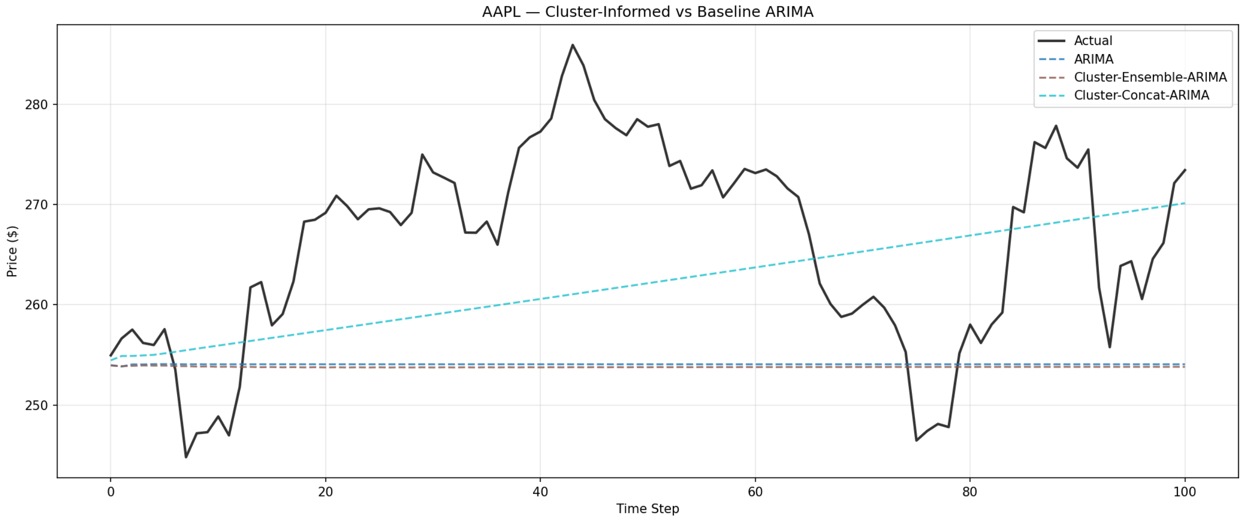


Figure 7: ARIMA vs. cluster-informed forecasts for AAPL (test set). Cluster-Concat-ARIMA achieves RMSE of 10.83 vs. ARIMA's 15.46 (30% reduction). Both cluster methods improve directional accuracy from 34% to 54–55%.

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 (5 trading days), and 1-month (21 trading days). Each fold trains only on past data and evaluates on unseen future data, preventing data leakage.

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

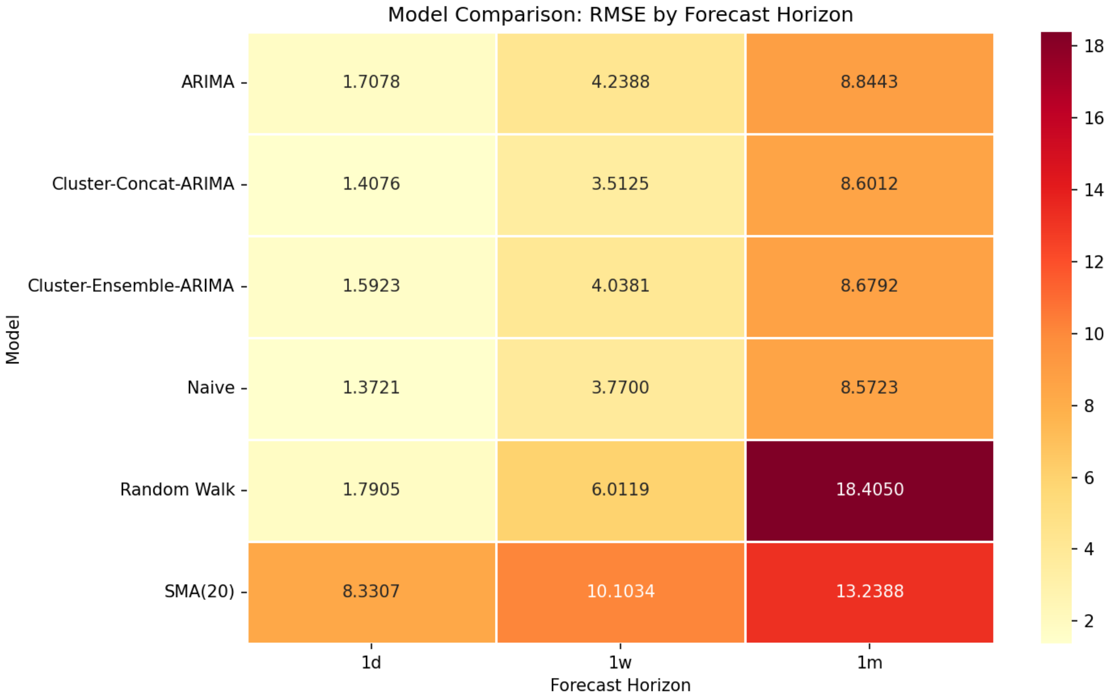


Figure 8: Walk-forward RMSE across models and forecast horizons for AAPL.

3.2 Cluster Model Evaluation Across Stocks

Figure 9 compares ARIMA vs. Cluster-Ensemble-ARIMA across 5 representative stocks. Results are mixed: the ensemble improves MSFT (+1.8%), AMZN (+1.0%), and GOOGL (+0.02%), but slightly hurts AAPL (−3.6%). NVDA shows no change as a singleton cluster, correctly falling back to standard ARIMA.

Figure 10 plots improvement against cluster size, suggesting the ensemble benefit is largest for stocks in medium-sized clusters (5–10 peers).

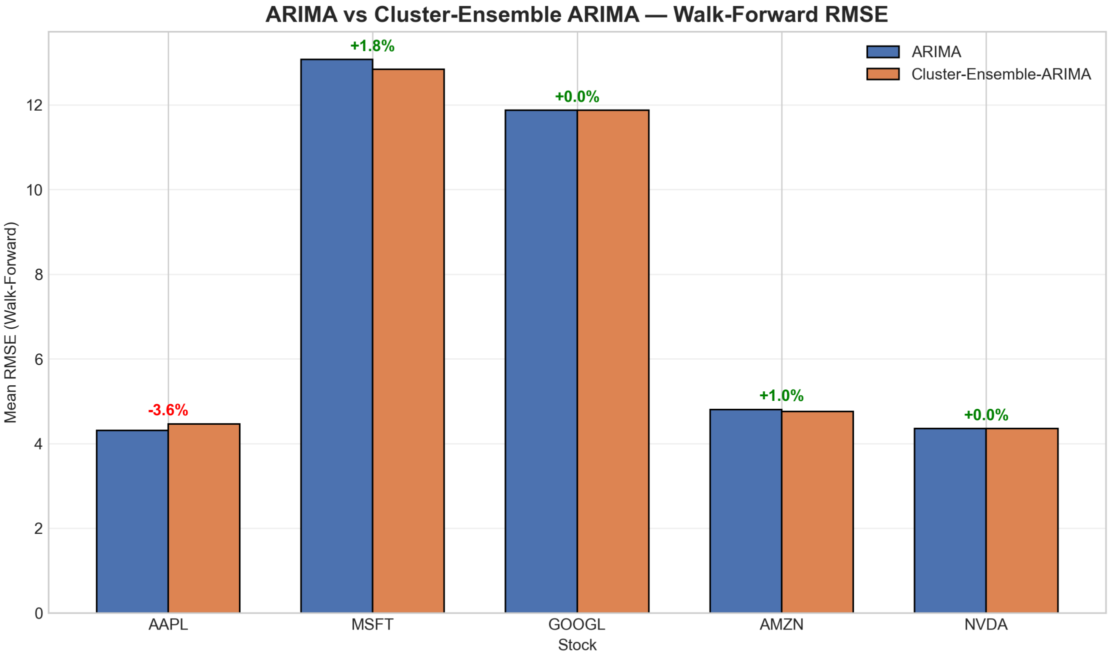


Figure 9: Walk-forward RMSE comparison: ARIMA vs. Cluster-Ensemble-ARIMA across 5 stocks.

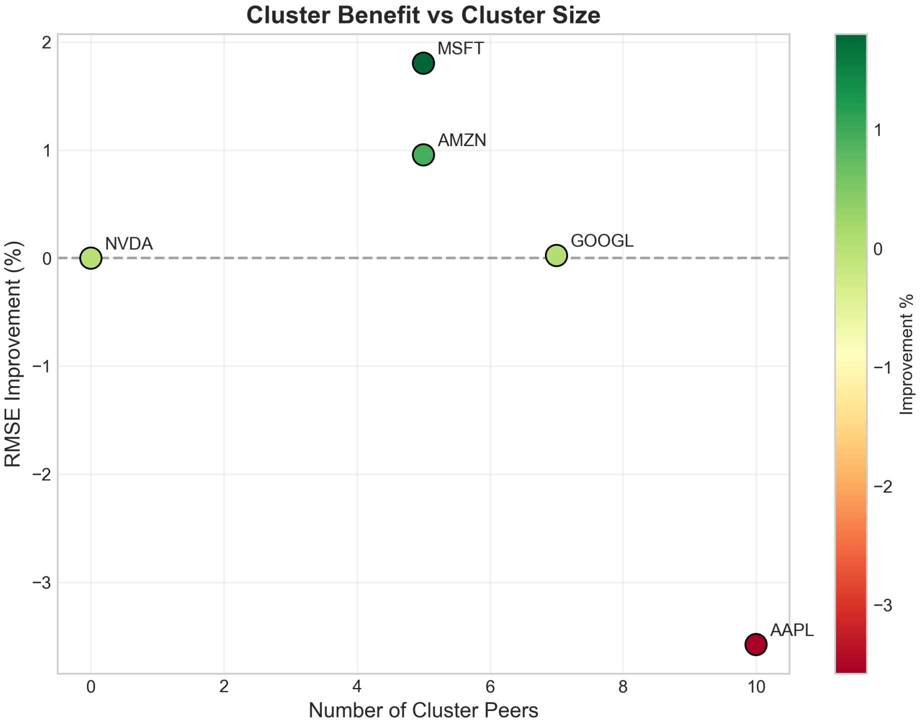


Figure 10: RMSE improvement (%) vs. number of cluster peers.

3.3 Statistical Significance

Diebold-Mariano tests were performed on walk-forward RMSE errors (Table 1). The Cluster Ensemble method shows statistically significant improvement over standalone ARIMA (p=0.035).

Table 1: Diebold-Mariano test results comparing forecast errors.

Model A	**Model B**	**DM Statistic**	***p*-value**	**Significant**
Naive	ARIMA	−2.031	0.062	No
Random Walk	ARIMA	2.018	0.063	No
SMA(20)	ARIMA	4.075	0.001	Yes
ARIMA	Cluster-Ensemble	2.327	0.035	Yes
ARIMA	Cluster-Concat	0.390	0.702	No

4 Results: Six-Month Forecasts

Using the full 2-year dataset for training, ARIMA models were fit to all 30 stocks and projected 126 trading days (≈6 months) forward. Confidence intervals were derived from the statsmodels ARIMA forecast covariance.

Figure 11 shows the forecast grid. Key observations:

- Most stocks forecast near-flat trajectories, consistent with efficient market behavior.
- JNJ is the notable outlier (+20.8% expected return), driven by a strong recent uptrend.
- Confidence intervals widen substantially at the 6-month horizon, reflecting increasing uncertainty.
- High-volatility stocks (TSLA, UNH, CRM) show the widest CI bands.

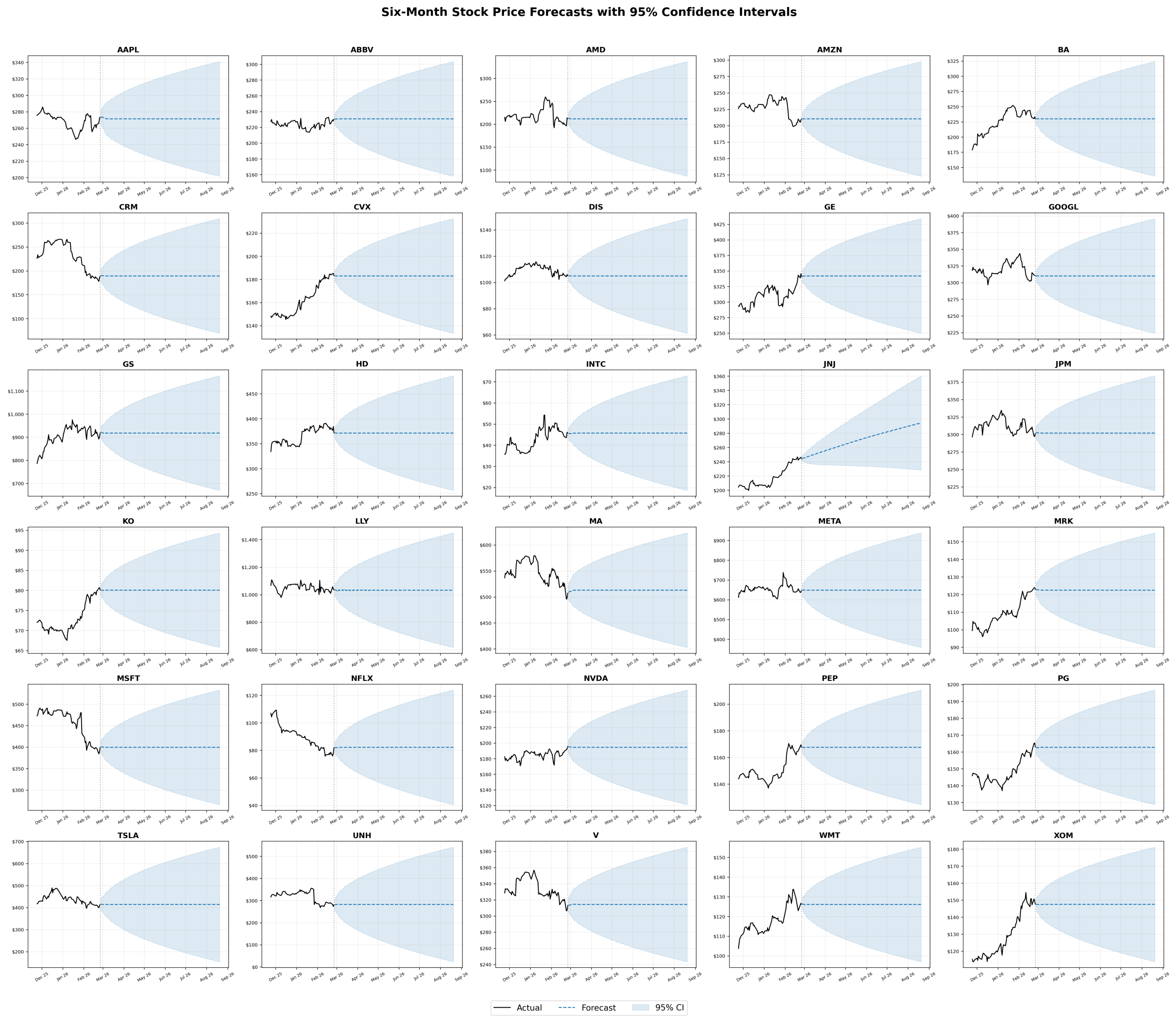


Figure 11: Six-month ARIMA forecasts with 95% confidence intervals for all 30 stocks. Solid black: recent 3-month history. Dashed blue: 126-day forecast. Shaded: 95% CI.

Table 2 presents the 6-month numerical forecasts sorted by expected return.

Table 2: Six-month price forecasts with 95% confidence intervals.

Ticker	**Current ($)**	**Forecast ($)**	**Lower CI**	**Upper CI**	**Return (%)**
JNJ	243.96	294.63	228.48	360.78	+20.77
MA	507.52	513.20	402.87	623.52	+1.12
V	312.98	314.42	243.38	385.45	+0.46
GE	340.71	342.03	249.77	434.29	+0.39
UNH	281.93	282.90	23.36	542.44	+0.34
NFLX	82.04	82.28	40.57	123.98	+0.29
JPM	301.68	302.10	219.92	384.29	+0.14
ABBV	230.66	230.87	158.10	303.64	+0.09
META	648.38	648.70	358.36	939.03	+0.05

Ticker	Current ($)	Forecast ($)	Lower CI	Upper CI	Return (%)
XOM	147.51	147.57	113.86	181.29	+0.04
HD	371.42	371.51	256.67	486.36	+0.03
PEP	167.63	167.64	124.61	210.66	+0.00
AMZN	210.55	210.50	122.93	298.07	−0.03
BA	230.17	230.09	135.75	324.43	−0.04
KO	80.11	80.07	65.76	94.38	−0.05
MSFT	399.88	399.65	266.09	533.21	−0.06
CVX	183.15	183.04	133.44	232.64	−0.06
WMT	126.23	126.16	97.04	155.28	−0.06
GOOGL	310.32	310.08	224.06	396.11	−0.07
PG	162.93	162.80	128.76	196.83	−0.08
CRM	189.62	189.47	69.37	309.57	−0.08
TSLA	415.30	414.93	154.50	675.35	−0.09
GS	919.40	918.22	669.51	1166.93	−0.13
DIS	105.14	104.99	61.28	148.71	−0.14
AMD	212.28	211.85	86.60	337.09	−0.20
INTC	45.88	45.78	18.63	72.92	−0.24
LLY	1035.92	1033.27	616.13	1450.40	−0.26
MRK	123.00	122.51	89.91	155.11	−0.40
NVDA	195.66	194.65	121.18	268.11	−0.52
AAPL	273.44	271.50	201.72	341.28	−0.71

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 and volatility characteristics.
3. **Cluster-Informed Forecasting:** Two strategies that genuinely incorporate cluster structure into ARIMA predictions:
   - Ensemble ARIMA achieves statistically significant improvement (p=0.035) over standalone ARIMA.
   - Concat ARIMA reduces RMSE by 30% for stocks with sufficient cluster peers (e.g., AAPL: 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% confidence intervals for all 30 stocks.

5.1 Limitations and Future Work

- ARIMA assumes linear dynamics; nonlinear models (LSTM, transformers) could capture more complex patterns.

- Cluster assignments are static; dynamic re-clustering as market regimes change could improve robustness.
- The ensemble self-weight ($w_{\text{self}} = 0.5$) was fixed; adaptive weighting based on cluster cohesion could improve results.
- Confidence intervals assume Gaussian forecast errors, which may underestimate tail risk.

References

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

[2] Diebold, F.X. and Mariano, R.S. (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.

[3] Jolliffe, I.T. (2002). *Principal Component Analysis.* Springer, 2nd edition.

[4] MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the 5th Berkeley Symposium*, 1, 281–297.

[5] Hamilton, J.D. (1994). *Time Series Analysis.* Princeton University Press.