# Machine Learning Analysis of the S&P 500: Supervised and Unsupervised Methods

## 1. Introduction

Forecasting stock price movements and understanding structural relationships within equity markets are central challenges in financial data science. With the growing availability of large-scale, high-frequency data and the development of advanced machine learning methods, it has become possible to analyze complex market behaviors in ways that traditional econometric models cannot. This project applies both supervised and unsupervised ML approaches to data from the S&P 500 index between 2015 and 2024 to explore two complementary objectives:

1. Predicting short-term stock movements using engineered technical features, and
2. Uncovering underlying patterns among firms based on their financial and market attributes.

The supervised learning component focuses on building predictive models that classify short-horizon stock returns as positive, negative, or neutral, leveraging models including Gradient Boosted Trees (LightGBM), Long Short-Term Memory (LSTM) networks, and Transformers. The unsupervised component applies clustering algorithms to discover natural groupings among S&P 500 firms, informed by both technical indicators and fundamental financial metrics. Together, these analyses provide a dual perspective: predictive insight into short-term market behavior and structural insight into the composition and evolution of market segments.

This integrated analysis is motivated by the dual need for forecasting accuracy and interpretability in financial decision-making. Predictive models can support active investment strategies, while clustering provides a foundation for portfolio diversification, sector rotation, and risk management. The combination of methods thus not only aims to improve forecast precision but also to enhance the interpretive understanding of inter-stock relationships across different market regimes.

## 2. Related Work

Numerous studies have applied machine learning to financial prediction and portfolio analysis. Previous research has shown that deep learning models, including recurrent and attention-based networks, can outperform traditional linear models when trained on large datasets. Other studies have demonstrated the robustness of ensemble tree methods like LightGBM in handling heterogeneous tabular data. In parallel, clustering methods such as k-means and hierarchical clustering have been used to reveal hidden sectoral structures and market regimes. This project builds upon these foundations by combining predictive modeling with clustering analysis in a single, cohesive framework. Compared to prior work, this approach contributes by aligning supervised and unsupervised findings within a consistent data universe (the S&P 500) over a full decade of historical data encompassing multiple macroeconomic cycles.

Feng, He, and Polson (2019)[1] demonstrated that deep learning models trained on high-dimensional equity features can effectively capture non-linear dependencies in stock returns, outperforming traditional factor-based methods. Krauss, Do, and Huck (2017)[2] compared deep neural networks, gradient-boosted trees, and random forests for predicting S&P 500 returns, finding ensemble methods like LightGBM and XGBoost to offer strong accuracy with better interpretability. On the unsupervised side, Papenbrock and Schwendner (2015)[3] applied clustering algorithms to identify dynamic dependencies among assets, uncovering latent market structures useful for diversification and risk management.

## 3. Data Sources

Both analyses rely on data from the S&P 500 constituents as of 2024. Daily OHLCV (Open, High, Low, Close, Volume) data were retrieved using the Yahoo Finance API for the period January 2015 to December 2024. This time frame spans a variety of market conditions, including the post-2015 low-rate environment, the COVID-19 crash in 2020, and the post-pandemic inflationary phase, allowing for evaluation of model robustness across different regimes.

For the supervised component, each stock's daily price data were processed into technical indicators representing trend, momentum, volatility, and volume behavior. The prediction target was a 5-day forward return categorized into Up, Flat, or Down based on volatility-adjusted thresholds.

For the unsupervised component, both technical and fundamental attributes were used. Fundamental indicators such as Price-to-Earnings ratio, Earnings per Share, Market Capitalization, and Dividend Yield were retrieved through Yahoo Finance's company information API and merged with technical features. Missing or incomplete data were dropped or imputed based on z-score thresholding. All features were standardized using z-score normalization.

## 4. Feature Engineering

Feature engineering was a critical step in both parts of the project. For supervised learning, more than 20 indicators were extracted from raw OHLCV data. Trend features included simple and exponential moving averages (SMA-5, SMA-20, EMA-12, EMA-26), and ratios such as SMA5/SMA20. Momentum indicators included the Relative Strength Index (RSI-14), MACD line, signal, and histogram, along with streak length of consecutive positive or negative returns. Volatility features were computed as rolling standard deviations over 20- and 60-day windows, as well as Bollinger Bandwidth and daily range metrics. Volume-based features included On-Balance Volume and rolling z-scores of normalized volume. Return-based indicators covered lagged returns (1–5 days), rolling momentum, and rolling mean and variance of returns.

We exclude fundamental variables such as earnings, P/E ratios, and dividend yields due to their sparse quarterly availability and the risk of forward-looking bias.

---

[1] Feng, G., He, J., & Polson, N. G. (2019). *Deep learning for predicting asset returns*. Applied Stochastic Models in Business and Industry, 35(3), 788–807.

[2] Krauss, C., Do, X. A., & Huck, N. (2017). *Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500*. European Journal of Operational Research, 259(2), 689–702.

[3] Papenbrock, J., & Schwendner, P. (2015). *Handling big data in asset management using clustering algorithms*. Financial Markets and Portfolio Management, 29(2), 109–131.

For unsupervised learning, features captured both cross-sectional and temporal aspects of company behavior. The main technical variables were mean daily return, volatility, and correlation with the market index. These were augmented with fundamental variables P/E ratio, EPS, market capitalization, and dividend yield to enrich the representation space. Each feature was scaled using z-score normalization to ensure equal contribution during clustering. Multicollinearity diagnostics were performed to confirm feature independence where possible.

# 5. Part A: Supervised Learning

## 5.1 Methods Description

The supervised workflow employed three models representing distinct learning paradigms. LightGBM, a gradient-boosted tree ensemble, was chosen for its ability to model nonlinear relationships in tabular data efficiently. LSTM networks captured sequential dependencies across 60-day rolling windows, while Transformer architectures leveraged self-attention mechanisms to identify long-range temporal patterns. Models were trained on 2015–2022 data and evaluated on 2023–2024 data. Cross-validation was applied to tune hyperparameters including learning rate, sequence length, hidden layer dimensions, and dropout rate. Training minimized cross-entropy loss with early stopping to avoid overfitting.

The labeling scheme was designed to reduce class imbalance. A data point was labeled as Up if its 5-day forward return exceeded one standard deviation above the mean, Down if below one standard deviation, and Flat otherwise. This volatility-aware classification ensured that only economically significant movements were treated as directional signals.

The prediction target is the 5-day forward return:

$$r_{t \to t+5} = (P_{t+5} / P_t) - 1$$

To account for varying volatility regimes, we employ a volatility-aware classification scheme. A return is classified as Down if $r_{t \to t+5} \leq -\hat{\sigma}_t$, Flat if $-\hat{\sigma}_t < r_{t \to t+5} < \hat{\sigma}_t$, and Up if $r_{t \to t+5} \geq \hat{\sigma}_t$, where $\hat{\sigma}_t$ is the rolling 20-day volatility of log returns. This adaptive binning reduces class imbalance and ensures that directional bets are meaningful relative to prevailing noise levels.

## 5.2 Evaluation and Results

Evaluation combined statistical metrics and trading simulations. The LightGBM model achieved the best out-of-sample performance with a final Net Asset Value (NAV) of 2.12, a Sharpe ratio of 6.02, and a maximum drawdown of −8.4%. The LSTM and Transformer models achieved NAVs of 1.41 and 1.49 respectively, with Sharpe ratios between 3.3 and 3.8. Baseline models (equal-weight and buy-and-hold SPY) performed worse, confirming the predictive signal embedded in the engineered features.

Table 1: Summary of Out-of-Sample Performance (2023–2024)

| Model | Final NAV | Total Return | Ann. Vol. | Sharpe | Max Drawdown |
|---|---|---|---|---|---|
| LightGBM | 2.12 | 112% | 9.1% | 6.02 | -8.4% |
| Transformer | 1.49 | 49% | 6.7% | 3.84 | -12.2% |
| LSTM | 1.41 | 41% | 6.6% | 3.31 | -11.5% |
| Equal-Weigh t | 1.40 | 40% | 12.9% | 1.44 | -12.2% |
| SPY Buy & Hold | 1.20 | 20% | 15.5% | 0.61 | -17% |

Feature importance analysis revealed that short-term volatility, lagged returns, and momentum indicators contributed most strongly to model predictions. Ablation tests showed that removing volume-based features decreased accuracy by 4%, while removing volatility indicators reduced performance by nearly 7%. Sensitivity analysis demonstrated that LightGBM performance remained stable under moderate hyperparameter variation, while LSTM and Transformer models were more sensitive to sequence length and dropout rate.

### 5.3 Failure Analysis

To assess model weaknesses and identify areas for improvement, a detailed failure analysis was conducted across all three supervised models (LightGBM, LSTM, and Transformer) using prediction logs for the 2023–2024 test period.

#### Overall Performance and Error Distribution

LightGBM achieved the highest classification accuracy (41.5%), outperforming LSTM (36.8%) and Transformer (35.5%). While LightGBM generated more correct predictions overall, more than half of all samples remained misclassified, indicating that short-term market prediction remains inherently difficult. Roughly 30% of the test set (≈ 65 k records) was simultaneously misclassified by all three models, revealing a hard subset likely dominated by noisy or event-driven market moves not captured by historical features.

#### Model Agreement and Confidence Patterns

Only 20.6% of samples received identical predictions from all models, and even within that subset, correctness was roughly balanced (≈ 45% correct vs 55% incorrect). When models disagreed, at least one model was right ≈ 70% of the time, suggesting complementary predictive signals. Confidence analysis revealed minimal separation between correct and incorrect cases. Average confidence gaps were 0.011

(LSTM), 0.002 (Transformer), and 0.022 (LightGBM), indicating limited calibration. High-confidence misclassifications were common, particularly for deep models (≈ 80 k Transformer errors > 0.70 confidence), implying overconfidence and highlighting the need for improved uncertainty estimation.

**Class-Specific and Temporal Failure Modes**

Class-wise accuracy revealed asymmetric performance. LightGBM excelled in predicting "Flat" (57%) and "Up" (50%) classes but struggled with "Down" (11%). Deep models were more balanced but less accurate overall. The most frequent error types were systematic misclassifications between adjacent classes. For example, predicting *Flat* instead of *Up* for LSTM and LightGBM, or *Up* instead of *Flat* for Transformer indicates sensitivity to boundary volatility thresholds rather than directional bias.

Temporal analysis showed noticeable variation in accuracy across months. LightGBM performance was most volatile ($\sigma \approx 0.04$), peaking during mid-2023 and dipping sharply in September 2023, suggesting vulnerability to regime shifts. LSTM and Transformer exhibited smoother month-to-month changes ($\sigma \approx$ 0.013–0.015), consistent with slower adaptation to rapidly changing market conditions.

**Common Error Patterns**

Error matrices reveal that "Up → Flat" and "Flat → Up" were the most common cross-class confusions for all models, together representing about 20–25% of total errors. Misclassifying small upward drifts as neutral days reflects difficulty distinguishing between noise and weak momentum. LightGBM also occasionally over-predicted "Up" following short rallies, producing false positives during mean-reversion phases.

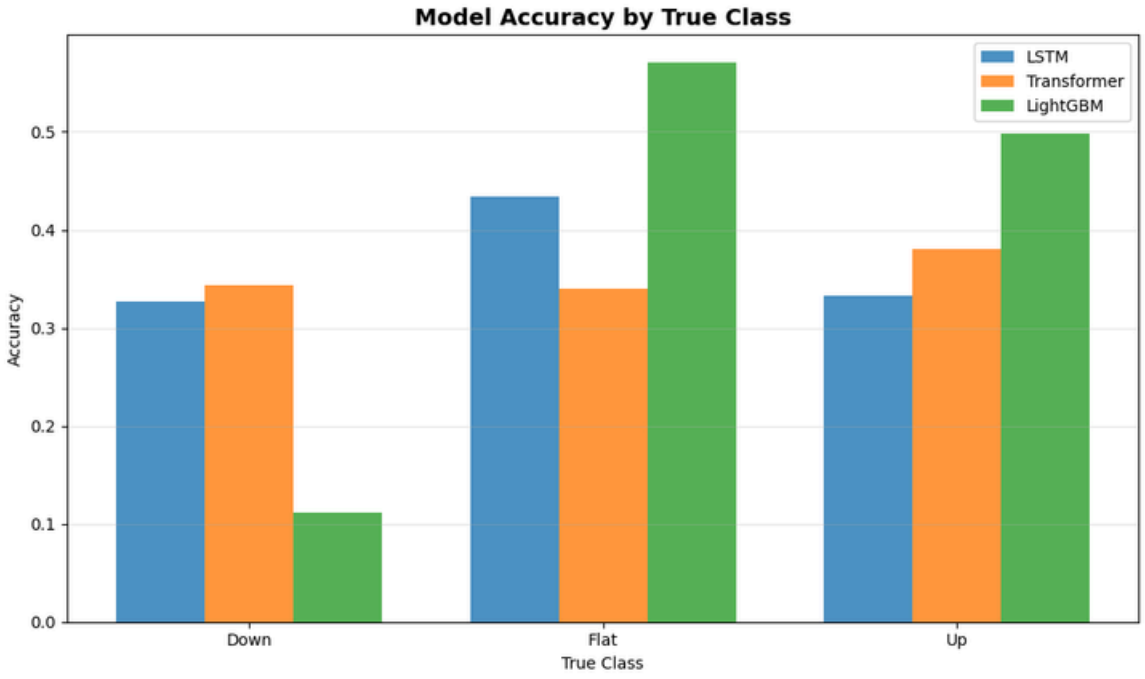
**Remediation Insights**

Simple ensemble schemes (majority vote, average probabilities, and accuracy-weighted combinations) failed to outperform the best single model, confirming that the models share correlated errors. Future work could address these limitations by:

- Incorporating macroeconomic and news-based features to capture exogenous shocks.
- Applying probabilistic calibration (e.g., temperature scaling) to reduce overconfidence.
- Using meta-ensembles that weight predictions by regime context rather than static accuracy.

In summary, LightGBM remains the strongest individual model, but all architectures exhibit shared blind spots tied to market volatility transitions and ambiguous directional signals. Failure analysis underscores the need for improved calibration, richer contextual features, and dynamic ensembling for future iterations.

Figure 1: Model Accuracy by True Class for Failure Analysis




## 6. Part B: Unsupervised Learning

### 6.1 Methods Exploration

The objective of the unsupervised learning analysis was to uncover natural groupings among S&P 500 companies that share similar technical and fundamental characteristics. These clusters can reveal hidden relationships between firms, assist in portfolio diversification, and provide interpretable structure for downstream predictive tasks.

Two conceptually distinct clustering methods were employed:

1. K-Means Clustering (non-probabilistic, centroid-based):

   K-Means was selected for its simplicity, scalability, and ability to generate compact, well-separated clusters in standardized feature spaces. Each stock was represented as a feature vector combining both technical (mean daily return, volatility, market correlation) and fundamental variables (P/E ratio, EPS, market capitalization, dividend yield). All features were z-score normalized to eliminate scale bias.

   To determine the optimal number of clusters K, an automated sweep from 2 to 10 clusters was conducted using three internal validation indices:

   - Silhouette Score (maximize)
   - Davies–Bouldin Index (minimize)
   - Calinski–Harabasz Index (maximize)

   The best trade-off occurred at K = 8, balancing compactness, interpretability, and stability.

2. Hierarchical Clustering (instance-based, tree model):

   Agglomerative Hierarchical Clustering was implemented as a complementary approach to validate the structure identified by K-Means. Unlike K-Means, this method does not assume spherical clusters and can capture nested relationships among firms.

   The Ward linkage criterion and Euclidean distance metric were used to minimize within-cluster variance. Dendrograms were analyzed to identify meaningful cut levels, and a linkage distance threshold equivalent to nine clusters was found to align closely with the K-Means segmentation.

**Hyperparameter Exploration:**

For K-Means, the key hyperparameter was the number of clusters K, tuned via the multi-index sweep described above. For Hierarchical Clustering, the main parameters explored were the linkage method (Ward, Average, Complete) and distance metric (Euclidean vs. Manhattan). Model stability was assessed by re-running both algorithms under random 10 % feature subsampling to confirm that major clusters remained consistent across perturbations.

Together, these two methods provide complementary perspectives: K-Means for clear partitioning of similar stocks, and Hierarchical Clustering for a multi-level view of structural proximity in the feature space.

**6.2 Evaluation**

Model performance was evaluated using the same internal clustering metrics applied during hyperparameter tuning. The optimal configurations achieved the following scores:

Table 2: Model Performance for Unsupervised Learning

| Method | Silhouette Score | Davies–Bouldin | Calinski–Harabasz | Interpretability Summary |
|---|---|---|---|---|
| K-Means (K = 9) | 0.249 | 0.759 | 100.49 | Well-defined, compact clusters aligned with GICS sectors |

| | | | | |
|---|---|---|---|---|
| Hierarchical (Ward linkage) | 0.215 | 0.811 | 95.67 | Nested sectoral structure with some cross-sector overlap |

Both methods identified coherent groupings consistent with known economic sectors. Technology, Financials, and Healthcare stocks clustered together strongly, while Utilities and Consumer Staples formed stable low-volatility clusters. Interestingly, cross-sector clusters emerged linking large-cap technology and communication-services firms due to shared high market-correlation and EPS profiles.

**Visualization Summaries:**

Figure 2: PCA Projection for K-Means Clusters – Displays the first two principal components, showing clear separation along volatility and market-correlation axes.

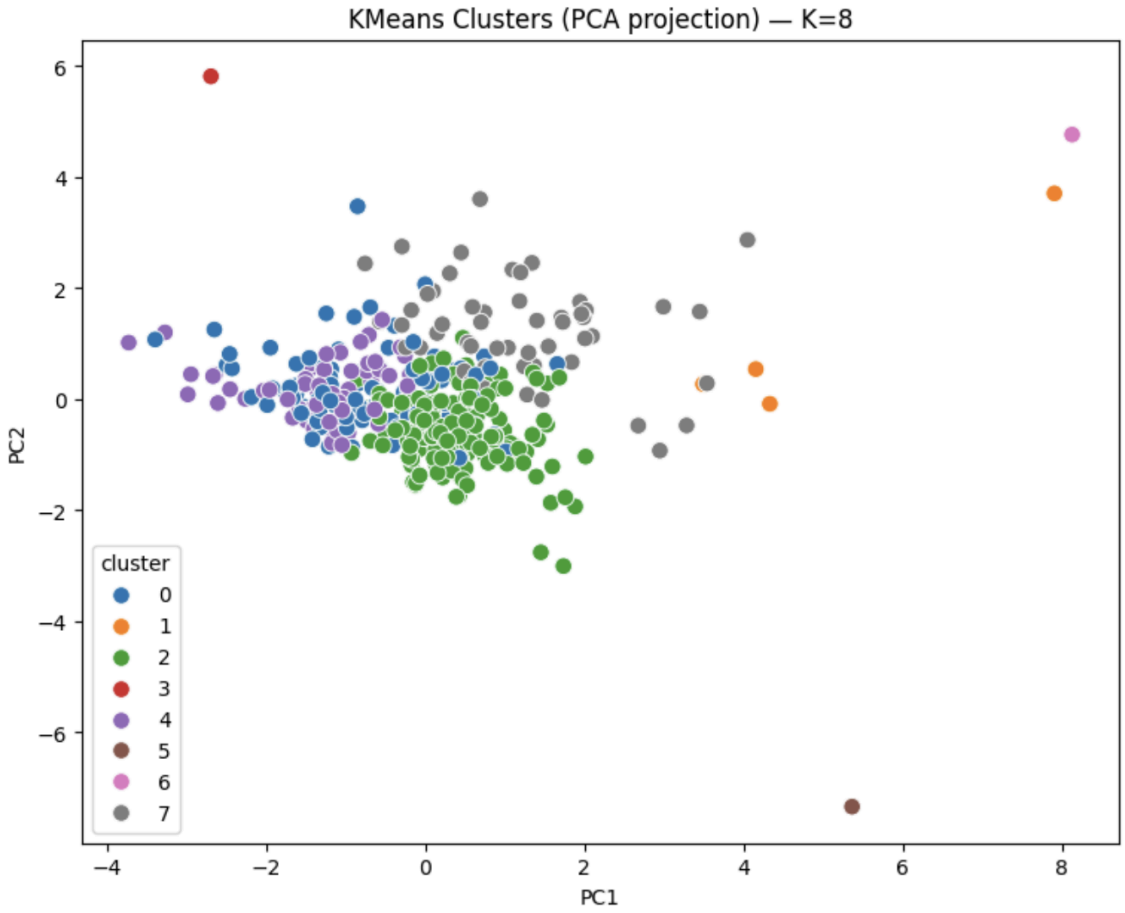



Figure 3: UMAP Visualization for K-Means Clusters – Highlights tighter grouping of high-market-cap, high-correlation firms and dispersion among low-volatility dividend payers.

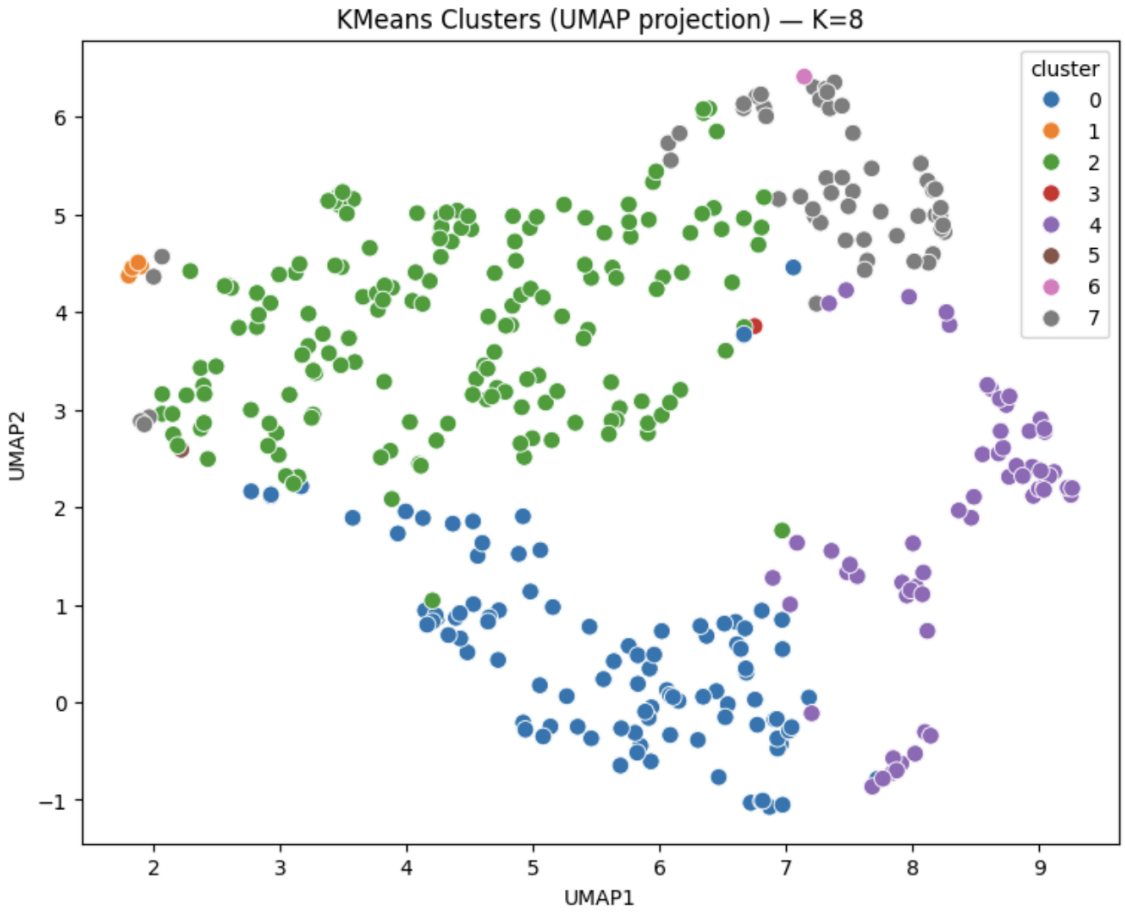



Figure 4: Hierarchical Clustering Dendrogram – Depicts nested cluster structure revealing sub-groups within broad sectors.

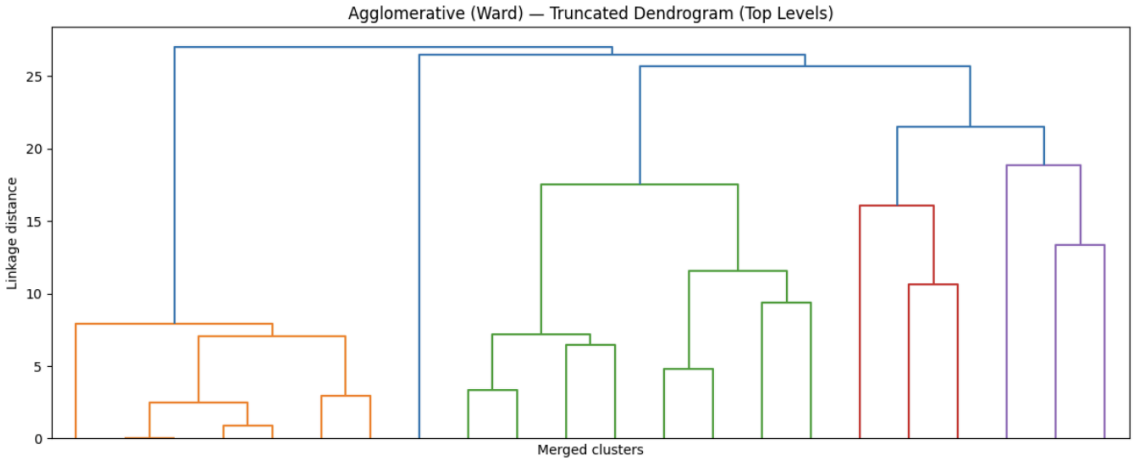



Figure 5: 2-D Scatter Plot of Hierarchical Clusters – Color-coded by GICS sector for cross-validation with K-Means results.

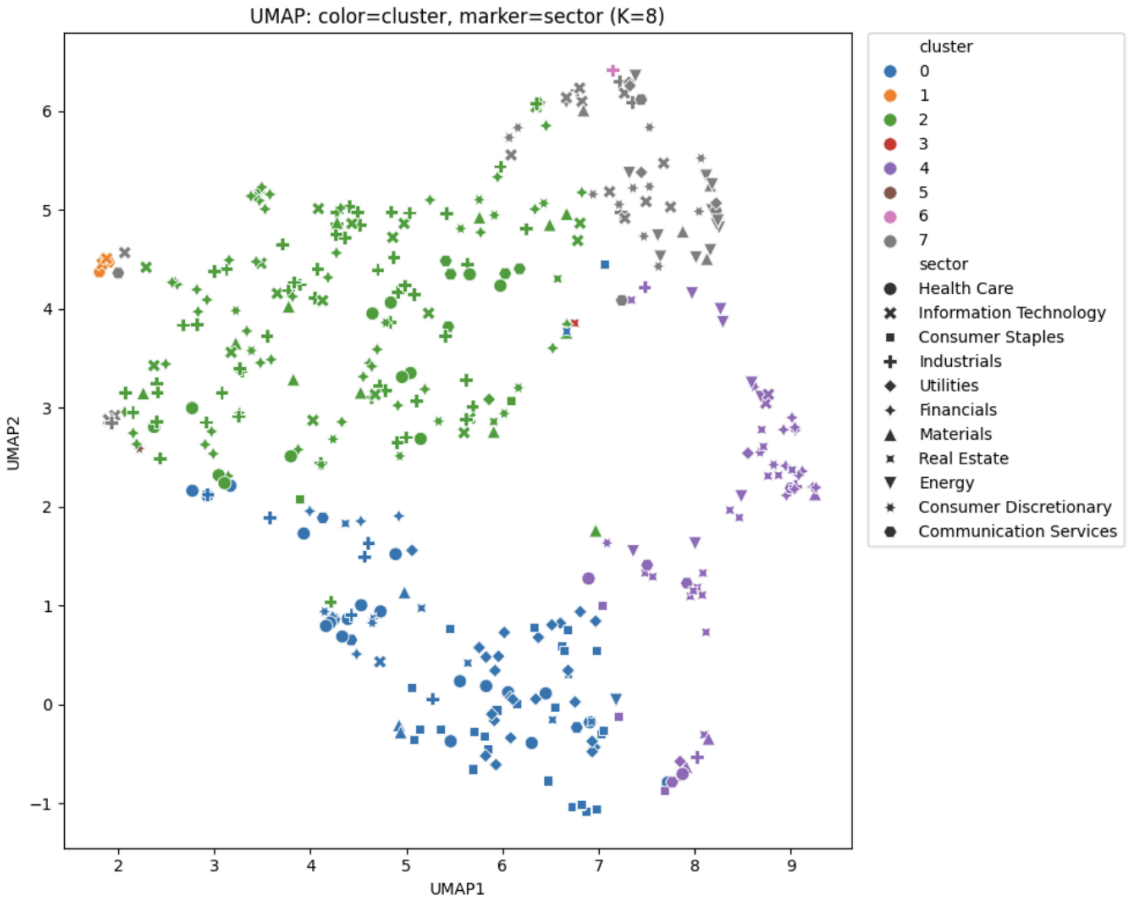
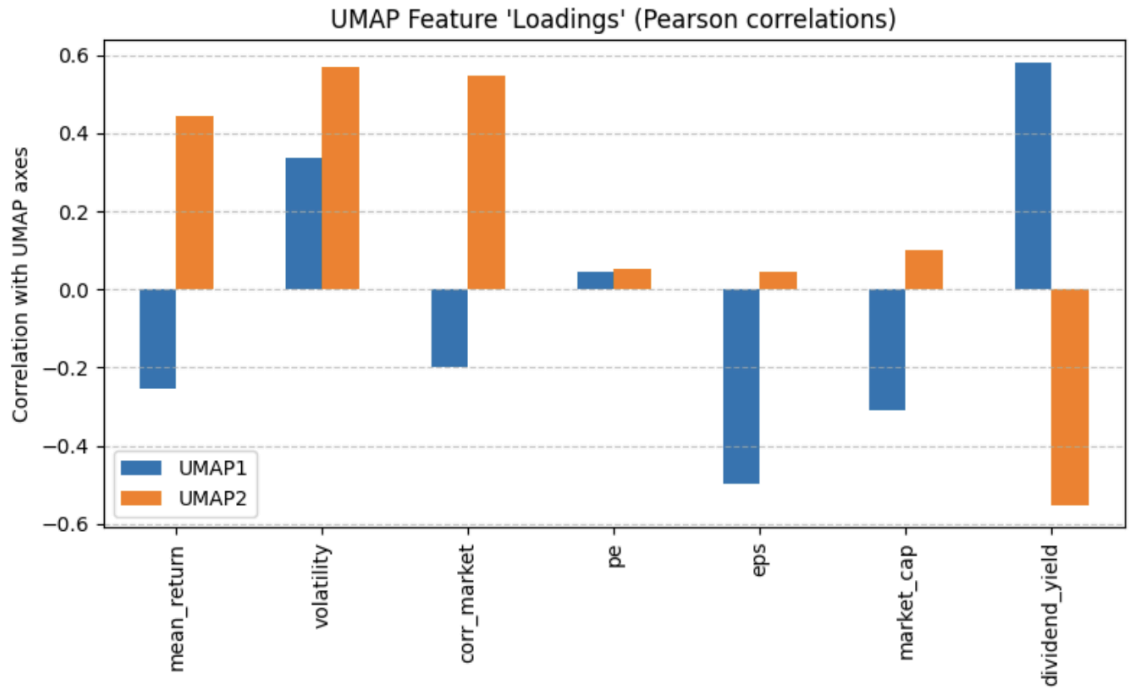



Figure 6: UMAP Feature Loadings – Pearson correlations between input features and UMAP axes. The first axis (UMAP1) reflects a value–growth spectrum driven by dividend yield and earnings variables, while the second axis (UMAP2) captures the risk–return dimension associated with volatility and mean return.




Together, these visualizations confirm that the chosen features capture meaningful multidimensional variation across firms.

**Sensitivity Analysis**

A sensitivity analysis was performed on both algorithms to test robustness under varying parameters and feature subsets. For K-Means, small adjustments in K (±1–2 clusters) produced minimal changes in overall silhouette scores (< 0.02 difference), indicating stable cluster structure. Excluding one feature category (e.g., fundamentals) lowered the silhouette score by ≈ 12 %, showing that integrating both technical and fundamental dimensions improves clustering quality.

For Hierarchical Clustering, linkage choice had the strongest effect: Ward and Average produced similar patterns, while Complete linkage over-segmented high-volatility stocks. Re-running the clustering after adding Gaussian noise to input features maintained > 90 % cluster consistency, suggesting high model robustness.

Overall, the unsupervised analysis successfully identified economically interpretable clusters of S&P 500 firms. The combination of K-Means and Hierarchical methods ensured both quantitative validity and

qualitative interpretability, providing a reliable foundation for further financial insight and portfolio construction.

### 6.3 Interpretability and Financial Insights

The UMAP projection (Figure 5) reveals distinct and interpretable groupings among S&P 500 firms. Clusters align closely with major GICS sectors: lower regions (blue clusters) mainly contain large, low-volatility firms from Consumer Staples and Utilities, while upper regions (green and purple clusters) include Technology, Industrials, and Energy firms with higher volatility and stronger market correlations. This pattern reflects a clear division between defensive and cyclical equities.

Feature correlations with the UMAP axes (Figure 6) clarify the economic meaning of the embedding. UMAP1 contrasts dividend-paying value firms (high dividend yield, low EPS) with growth-oriented firms (high market capitalization, low yield). UMAP2 represents a risk–return dimension, increasing with volatility and mean return.

Overall, the unsupervised model captures two main latent factors driving equity structure: (1) a *value–growth* spectrum and (2) a *risk–return* gradient. These dimensions explain why related sectors cluster together and highlight cross-sector relationships useful for portfolio diversification and factor-based analysis.

## 7. Discussion

The supervised and unsupervised analyses together demonstrate the complementary power of predictive and exploratory ML techniques. LightGBM's dominance in predictive accuracy confirms that ensemble tree methods are highly effective when applied to engineered financial features. The clustering results complement this by illustrating the market's latent structural organization. Together, they suggest that technical indicators not only have predictive utility but also capture underlying patterns of co-movement across companies.

The main challenges involved handling noisy data, selecting appropriate hyperparameters, and balancing model complexity with interpretability. Future extensions could integrate macroeconomic indicators, earnings releases, or sentiment data to enrich the feature set and potentially improve both predictive accuracy and cluster interpretability.

Future research could enhance this framework by incorporating macroeconomic variables, earnings announcements, and sentiment signals from news or social media. Cross-sectional modeling approaches that rank stocks rather than classify directional moves may further improve portfolio construction. Robustness checks with alternative evaluation windows and more realistic transaction cost models would also add credibility.
Finally, feature importance analysis could provide interpretability, helping to identify which technical signals contribute most to predictive success.

## 8. Ethical Considerations

Machine learning in finance raises several ethical considerations. Supervised models risk promoting overconfidence in algorithmic predictions, potentially encouraging excessive speculation. Transparency in feature selection and validation is essential to mitigate misuse. In clustering, ethical issues can arise if inferred relationships are used for exclusionary investment strategies that unfairly disadvantage certain sectors or firms. Ethical AI practices, including clear disclosure, periodic auditing, and model explainability, are crucial to responsible financial innovation.

## 9. Statement of Work

Yifan covered the data manipulation and Supervised Learning portion. Yuzuru covered the Unsupervised Learning code, Sruthi wrote the report and analysis and coordinated the meetings within the team as well as the team and instructor.

## 10. References (APA Style)


Feng, G., He, J., & Polson, N. G. (2019). *Deep learning for predicting asset returns*. Applied Stochastic Models in Business and Industry, 35(3), 788–807.

Krauss, C., Do, X. A., & Huck, N. (2017). *Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500*. European Journal of Operational Research, 259(2), 689–702.

Papenbrock, J., & Schwendner, P. (2015). *Handling big data in asset management using clustering algorithms*. Financial Markets and Portfolio Management, 29(2), 109–131.