1. Data Description

Stock data of NYSE-listed stocks were obtained from Investing.com, giving daily price history between 2010-01-01 and 2025-04-16.

- Pfizer (PFE) Stock Data contained:
  Date, Price (Close), Open, High, Low, Vol., Change %.

- XLV ETF Data (also: NYSE-listed) served as a healthcare sector proxy for the purposes of momentum. Since Pfizer, being one of the major constituents of XLV, is so heavily weighted in the ETF, it offered contextual indicators of sector-wide sentiment and trends.

- Sentiment Analysis employed Google Trends. Rather than scraping tweets or using NLP models, which were impractical for a 15-year window, predefined keyword sets reflecting positive and negative sentiment were followed over time. This dataset consisted of:
  Date, positive_score, negative_score, sentiment_score, and sentiment_label.

Sources were all date-aligned to generate the final dataset, having no forward-looking bias.

2. Feature Engineering

Feature engineering was performed to derive useful patterns in terms of price, sentiment, and wider market context. The final dataset consisted of features from the following categories:

1. Price & Volume Technical Features:
   These comprised daily/weekly/monthly returns, rolling metrics such as mean, std deviation, z-scores, volatility, and technical metrics such as RSI (14-day). These capture price momentum, direction of the trend, and sudden spikes or dips in price/volume.

2. Sentiment Features:
   Google Trends data served as a market sentiment proxy based on searches of pre-defined positive and negative keywords. Rolling averages, sentiment volatility, and daily movements were employed to estimate the evolution of investor mood over time.

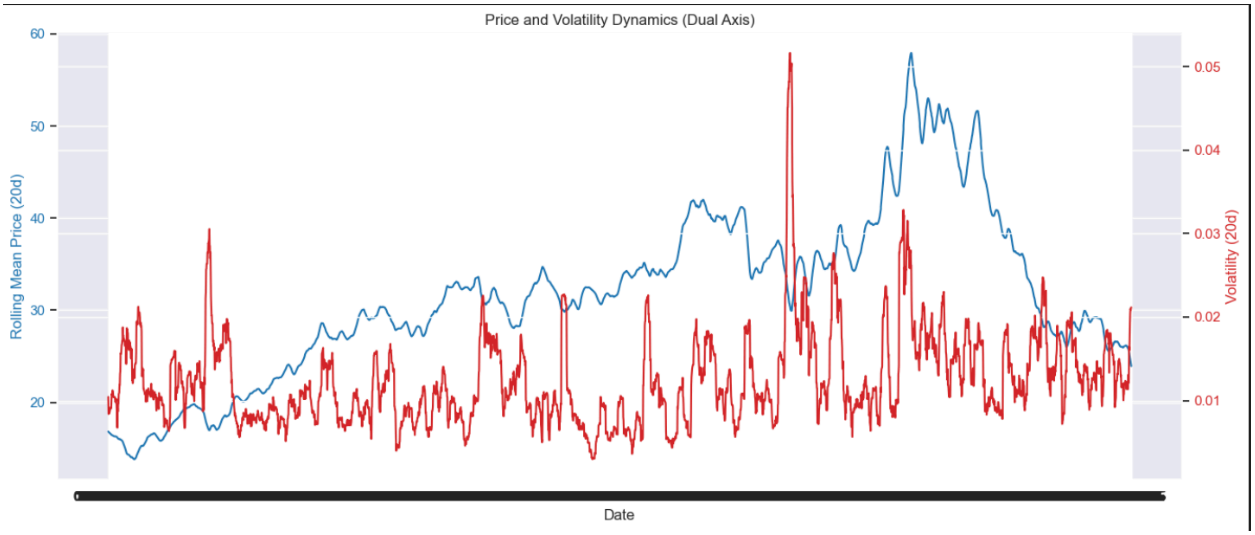
3. Market Context (XLV ETF):
   Returns on the NYSE-listed XLV ETF served to put Pfizer's movement in context with respect to the healthcare sector. This encompassed 1-day and 5-day returns and the

Pfizer-XLV return spread to pick up sector-driven effects.

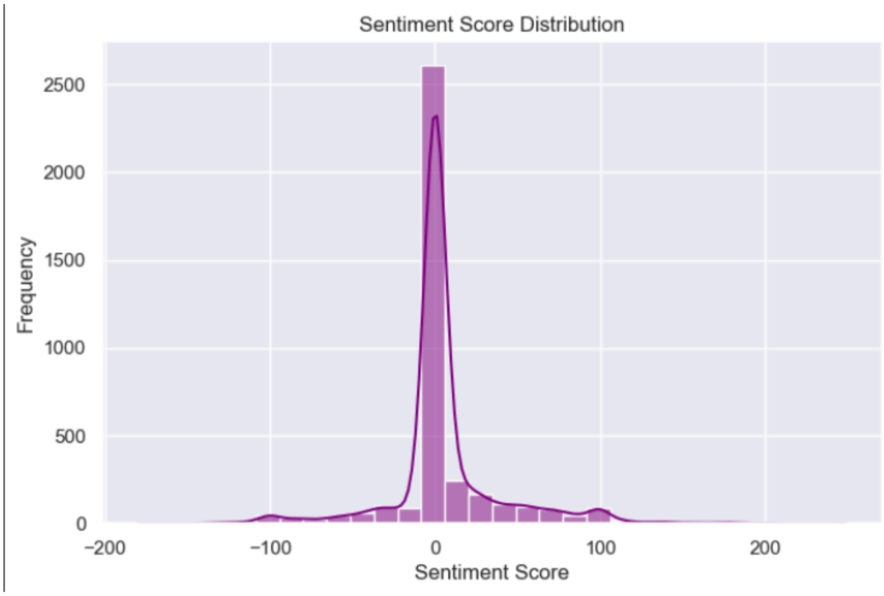
4. Rolling Correlation & Lag Features:
   Features such as 20-day rolling correlation between XLV and Pfizer picked up dynamic relationships, and lagged price and sentiment scores assisted in modeling delayed impacts of past trends and sentiment on future prices.

5. Interaction Features:
   These picked-up cross-effects — e.g., the way non-neutral sentiment interacts with returns and sentiment volatility interacts with price volatility — so that the model can recognize the compounding or moderating effects of blended signals.

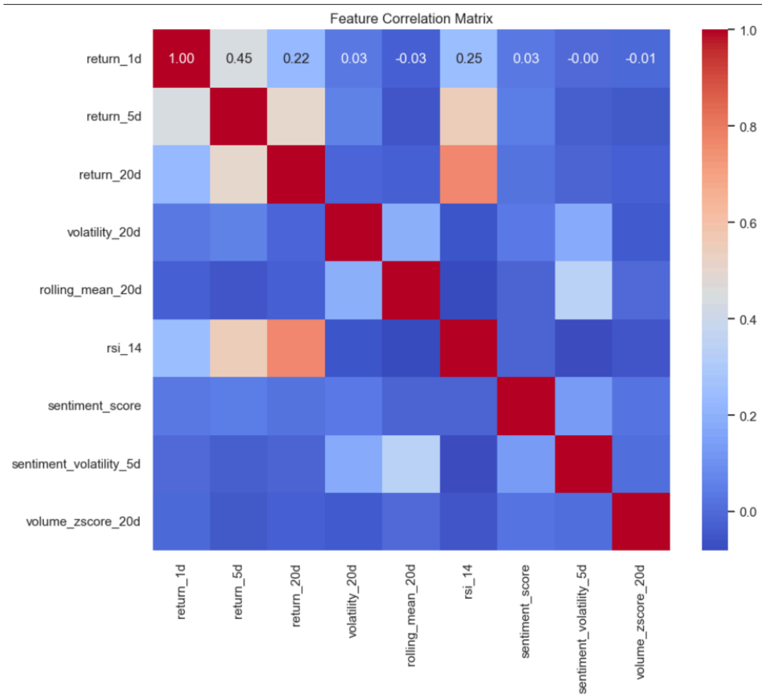
In order to interpret the distribution and characteristics of the chosen features, I attempted to visualize some plots.



- Volatility spikes co-occur with steep price movement, which is particularly evident around 2020, while price displays a general increase followed by a fall—underscoring the relevance of volatility in identifying market distress and trend shocks.



- Sentiment scores were largely neutral with omnipresent zero scores, while from 2020 onwards, they displayed persistent fluctuations—most probably due to rising public awareness, with Pfizer-associated keywords trending on Google.



- The correlation matrix reveals low overall multicollinearity of features with only moderate correlation seen—e.g., between short- and mid-term returns (return_1d, return_5d, return_20d) and rsi_14. Sentiment- and volume-based features are mostly uncorrelated with returns, implying they can provide orthogonal signals.

3. Train / Validation / Test Split

To provide realistic evaluation while avoiding forward-looking bias, the dataset was divided chronologically into three sets:

- Training Set (2010–2020): Utilized for model training and parameter weight learning.
- Validation Set (2021): Applied for model tuning, feature selection, and hyperparameter tuning.
- Out-of-Sample Test Set (2022–2025): Set aside solely for ultimate performance assessment and PnL reporting.

Specifically, 2020 was the last year to be covered in training, as sentiment features did not start exhibiting significant, non-zero variation until 2020, coinciding with increased public interest amidst the outbreak of COVID-19. This division allows the model to learn from a historically consistent era and then be evaluated upon more sentiment-influenced, dynamic regimes- more similar to generalization in the real world.

These splits were clearly labeled with a Modeling column (Train, Validation, Out-of-Sample) in the last feature dataset.

4. First Modeling Approach – PCA + VAR Baseline

As an initial modeling attempt, a two-stage approach integrating Principal Component Analysis (PCA) with a Vector Autoregression (VAR) model was used. This pipeline provided a tidy, interpretable baseline that matched the sequential nature of financial time series data.

Pipeline

- The originally chosen set of 22 engineered features was scaled by using StandardScaler to achieve unit-free comparability.
- Dimensionality Reduction using PCA
  To avoid multicollinearity and eliminate noise, PCA was used. The first 5 principal components captured 63.2% of the variance in the feature set:

Explained Variance Ratio:
PC1: 20.49% | PC2: 17.35% | PC3: 10.52% | PC4: 8.23% | PC5: 6.57%

- A VAR model was trained on PCA-transformed data 2010–2021. Chosen VAR Lag Order: 5, chosen via the Akaike Information Criterion (AIC).

- Projected 1-day returns were converted to binary trading signals:
  If projected return ≥ 0 → Long (+1), Otherwise → Short (–1)

- **PnL Calculation: Daily strategy=Signal × Actual Return**
  and compared with a Baseline strategy that always goes long.

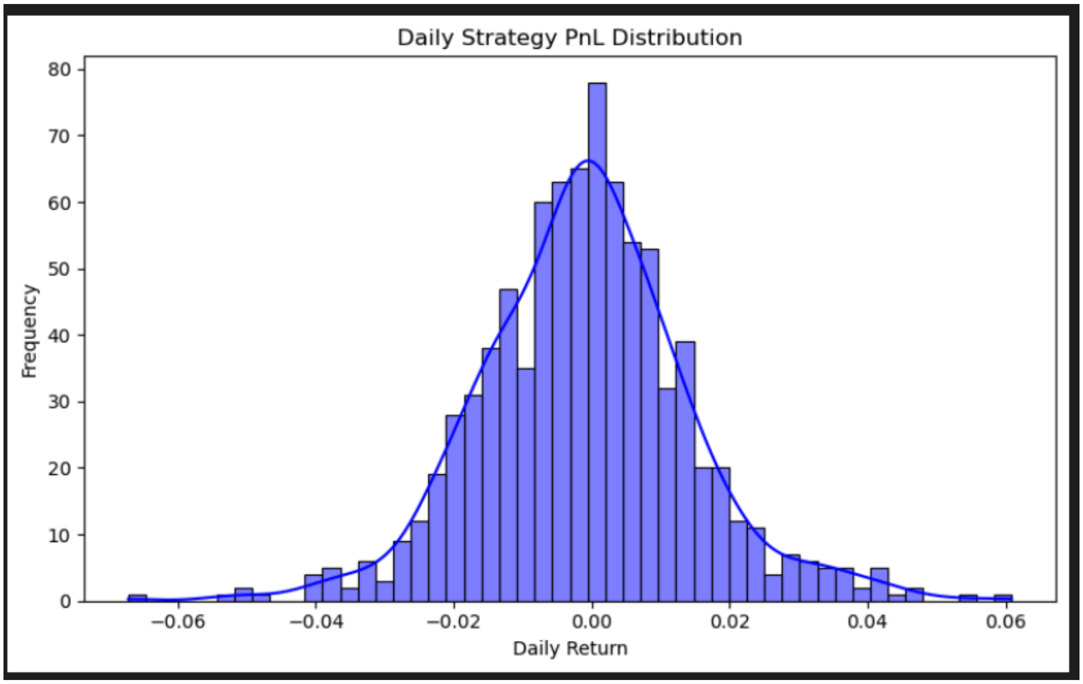
- Performance Metrics
  RMSE of return_1d prediction: 0.015325
  The **48.3% accuracy** indicates that the model is only slightly better than random chance.

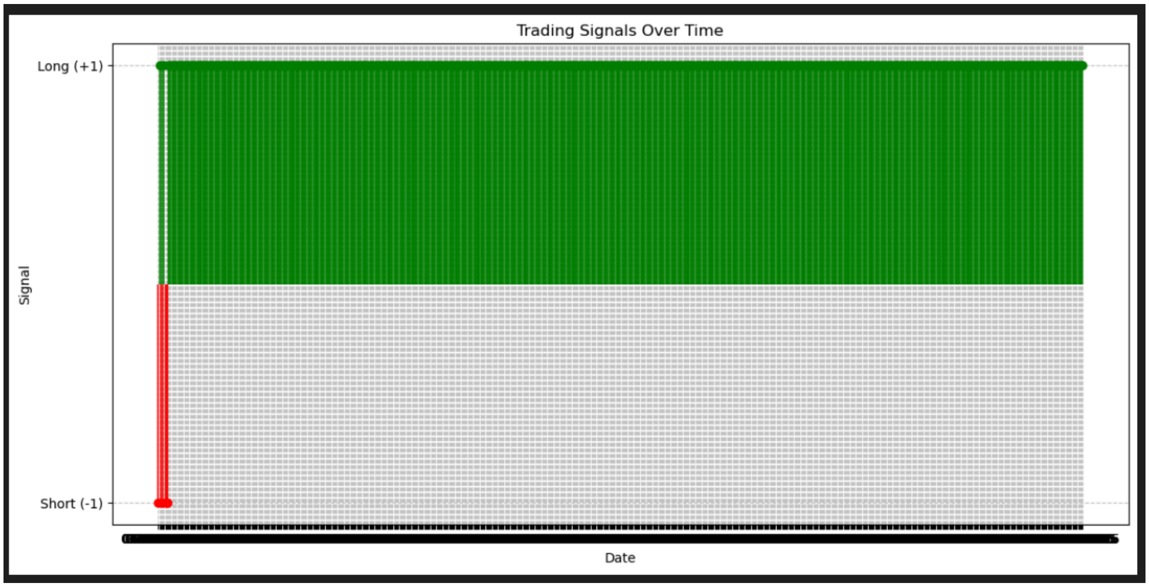
Visual Insights & Strategy Behavior

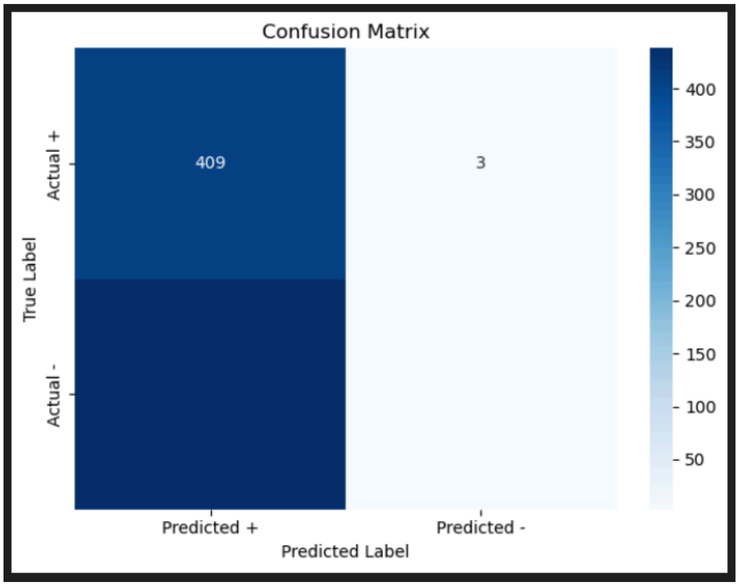
- The heatmap of the trading signal shows that the model emitted **short signals on very few days**, while the remaining days were all long in an unwavering manner. This is very close to the baseline strategy, which is always long.

- Therefore, the cumulative PnL profiles of both the **PCA-VAR strategy and the benchmark are graphically very similar**, with the VAR strategy performing at most only slightly better.

- The distribution of daily returns is around zero and is slightly positively skewed, consistent with limited edge or alpha in signal creation.


Daily Strategy PnL Distribution
Frequency
0
10
20
30
40
50
60
70
80
−0.06
−0.04
−0.02
0.00
0.02
0.04
0.06
Daily Return


Trading Signals Over Time
Long (+1)
Short (-1)
Signal
Date



Conclusion:
In spite of the methodological rigour of PCA as a dimensionality reduction technique and VAR as a time series modeling method, the model only had an accuracy of 48.3%, slightly higher than random guessing. The **class imbalance** in predictions can be inferred from the confusion matrix, where the model predicts a negative return nearly every time. This, coupled with the strategy's PnL just over the baseline, is a result of the linear VAR model failing to extract nonlinear patterns inherent in financial data.

5. Second Modeling Strategy- SARIMAX + PCA

The Seasonal Autoregressive Integrated Moving Average with Exogenous Regressors (SARIMAX) was then used, as it allows for a more robust time series model, which can model temporal relationships directly within the target variable (return_1d) while including exogenous features (through PCA-reduced features). SARIMAX is likely to model autoregressive structure and outside signals better than VAR, which models all variables symmetrically and can miss the target's temporal details.

Pipeline

- Features were normalized and scaled down to 5 main components via PCA. Explained variance of chosen **PCs: [0.2048, 0.1735, 0.1052, 0.0823, 0.0657]**.

- A **SARIMAX(5, 0, 0)** model was fitted to return_1d with these components as exogenous variables, utilizing lagged autoregressive terms to include short-term dependencies.

- Model Fit (on training set of 3025 observations):
  AIC: –23655.64
  BIC: –23589.48
  All the PCA components and AR lags exhibited **p-values < 0.001**, confirming strong statistical significance.

Same as before,

- Projected 1-day returns were converted to binary trading signals:
  If projected return ≥ 0 → Long (+1), Otherwise → Short (–1)

- **PnL Calculation: Daily strategy=Signal × Actual Return**
  and compared with a Baseline strategy that always goes long.

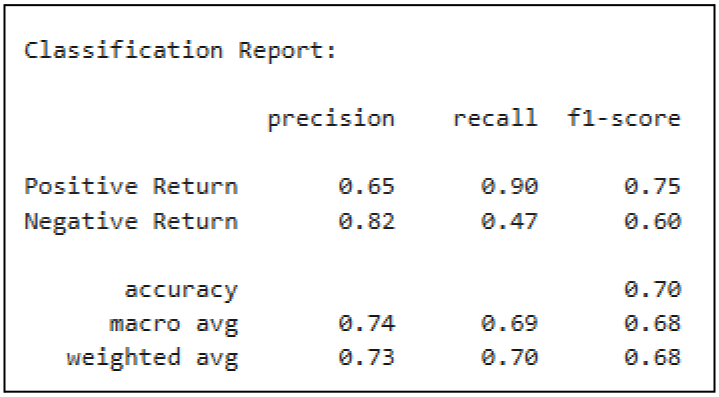
Performance Metrics

- RMSE: 0.02275 - indicating low average prediction error in daily return values.

- **Overall Accuracy: 70%**

- Positive Return Class: Precision: 0.65 | Recall: 0.90 | F1: 0.75

- Negative Return Class: Precision: 0.82 | Recall: 0.47 | F1: 0.60
  Balanced performance with better recall for positive class and better precision for negative class.



```
Classification Report:

                 precision    recall  f1-score

Positive Return       0.65      0.90      0.75
Negative Return       0.82      0.47      0.60

       accuracy                           0.70
      macro avg       0.74      0.69      0.68
   weighted avg       0.73      0.70      0.68
```

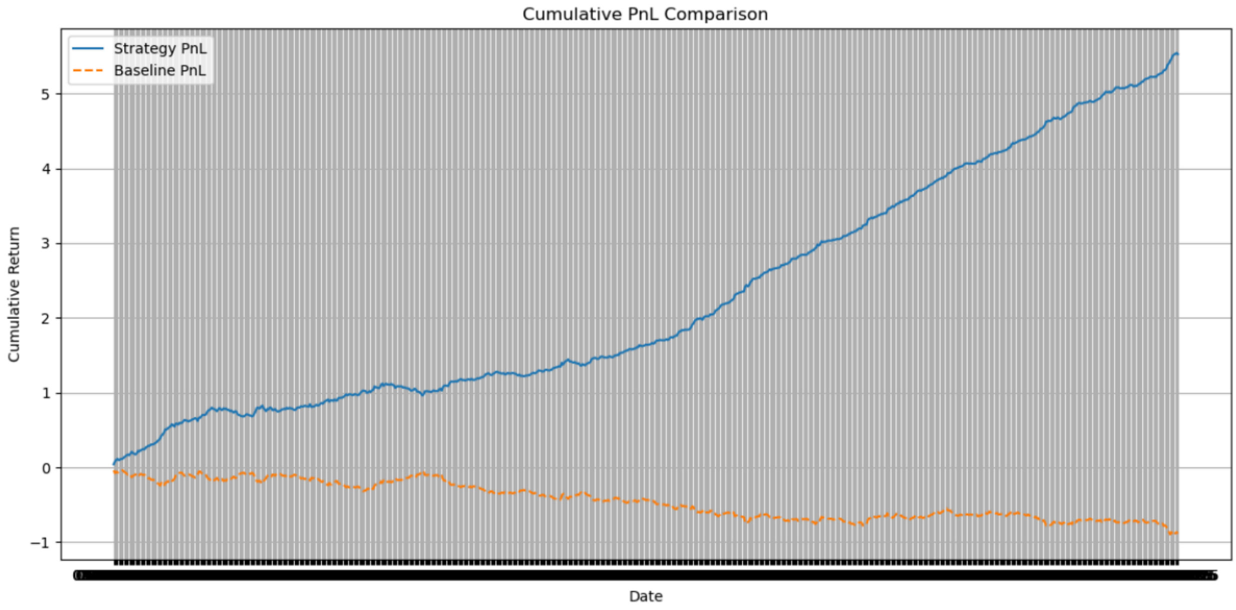
Visual Insights & Strategy Behavior

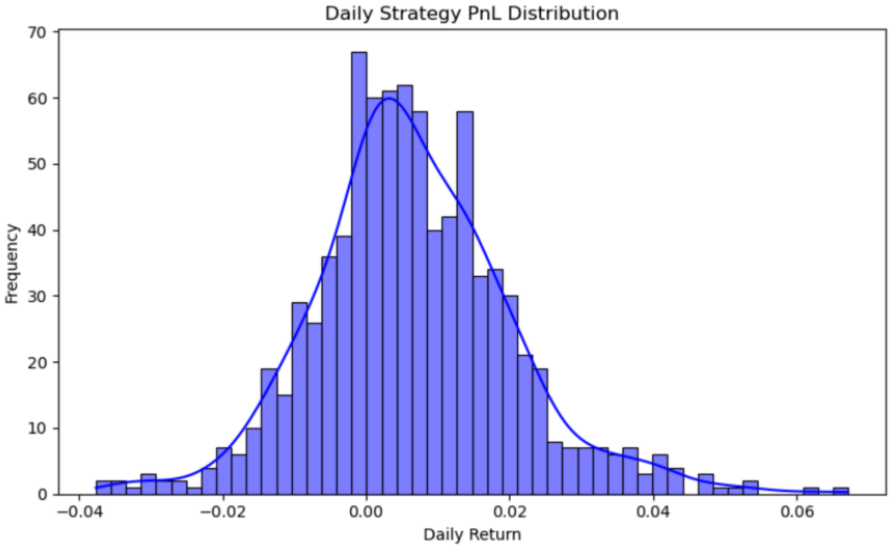
- The SARIMAX-based trading strategy exhibits a continuously increasing PnL, performing better than the PCA-only model and baseline model, showing solid and sustained profitability.

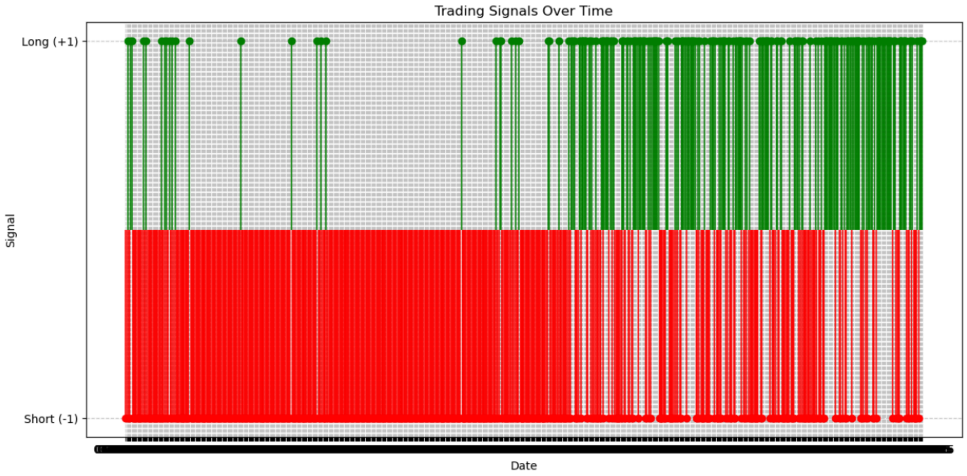
- The model has ~70% total accuracy, but performs poorly in recall of upward trends — classifying more true positives as false (217 false negatives versus 195 true positives).

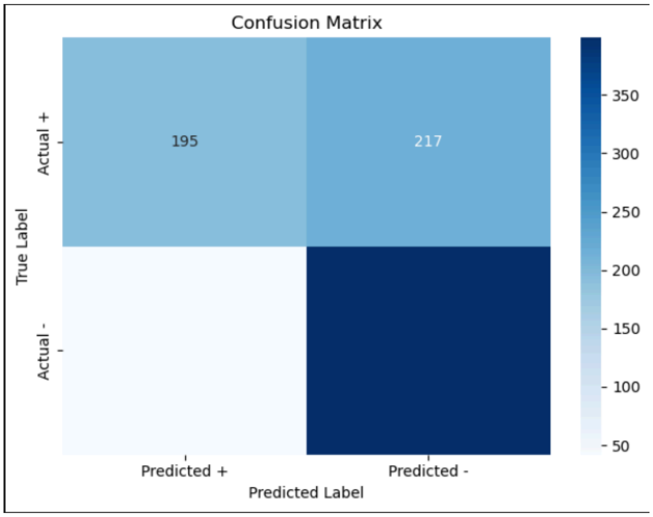
- Heavily biased in favor of forecasting negative or neutral returns, the model produces more short signals than long ones, indicative of a risk-averse, conservative approach.

- Daily returns are averaged at 0.005, supporting the rewarding nature of the strategy despite periodic loss and conservative forecasts.




Daily Strategy PnL Distribution
Frequency
Daily Return
0
10
20
30
40
50
60
70
−0.04
−0.02
0.00
0.02
0.04
0.06


Trading Signals Over Time
Long (+1)
Short (-1)
Signal
Date


Confusion Matrix
195
217
Actual +
Actual -
True Label
Predicted +
Predicted -
Predicted Label
350
300
250
200
150
100
50

Conclusion

The SARIMAX + PCA model dramatically dominates the previous PCA-VAR baseline, providing a significant enhancement in predictive accuracy (~70%) and cumulative profitability. Its capacity to describe the autoregressive pattern of returns while incorporating external signals via PCA components allows it to develop a more robust and profitable trading strategy. Nonetheless, the model has a conservative bias with an elevated false negative rate and **underprediction for upward trends**. Although it overcomes some of the disadvantages of VAR, it is still lacking in capturing fully bullish signals and leaves scope for improvement in recall and signal calibration.

6. Third Modeling Approach – XGBoost Regressor

With some experimentation with linear time series models such as VAR and SARIMAX, a tree-based ensemble technique, XGBoost (Extreme Gradient Boosting), was attempted. XGBoost is particularly reputed for its capability to manage nonlinear relationships, feature interactions, and heterogeneous data distributions better than standard time series models.

Pipeline

- 22 engineered features across price action, volatility, sentiment score, sector returns, and lag features, calculated in initial feature engineering were utilized. All features were scaled using StandardScaler.

- Model Configuration:
  XGBRegressor(n_estimators=100, learning_rate=0.1, max_depth=4, early_stopping_rounds=10)
  Trained with early stopping on validation set.

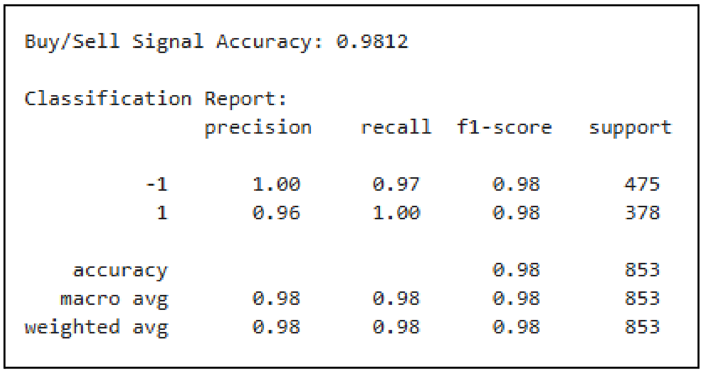
Same as before,

- Projected 1-day returns were converted to binary trading signals:
  If projected return ≥ 0 → Long (+1), Otherwise → Short (–1)

- **PnL Calculation: Daily strategy=Signal × Actual Return**
  and compared with the Baseline strategy that always goes long.

Performance Metrics

- RMSE: Not directly reported, but classification performance is incredibly robust.
- Buy/Sell Classification Metrics:
  Overall Accuracy: 98.1%
- Positive Class (Up Days):
  Precision: 0.96 | Recall: 1.00 | F1 Score: 0.98
- Negative Class (Down Days):
  Precision: 1.00 | Recall: 0.97 | F1 Score: 0.98



```
Buy/Sell Signal Accuracy: 0.9812

Classification Report:
              precision    recall  f1-score   support

          -1       1.00      0.97      0.98       475
           1       0.96      1.00      0.98       378

    accuracy                           0.98       853
   macro avg       0.98      0.98      0.98       853
weighted avg       0.98      0.98      0.98       853
```

These findings demonstrate extremely well-balanced performance, with almost perfect generation of the signals for long and short trades- a significant improvement from SARIMAX and VAR, particularly in recall and F1 score for both directions.

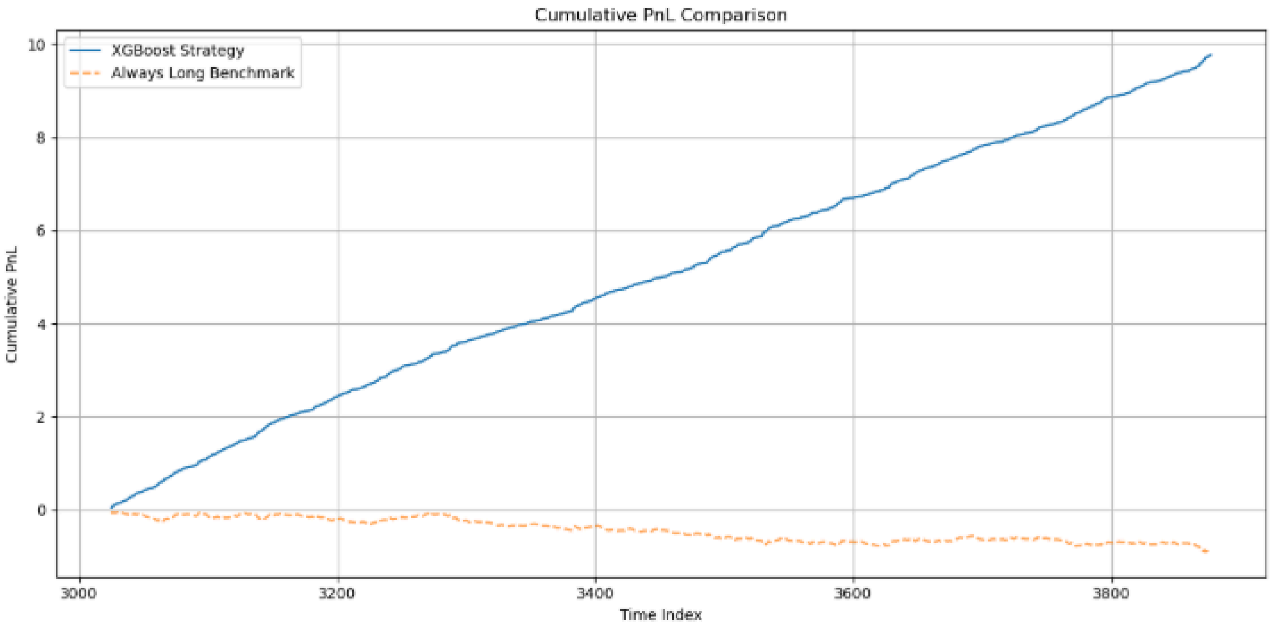
Visual Insights & Strategy Behavior

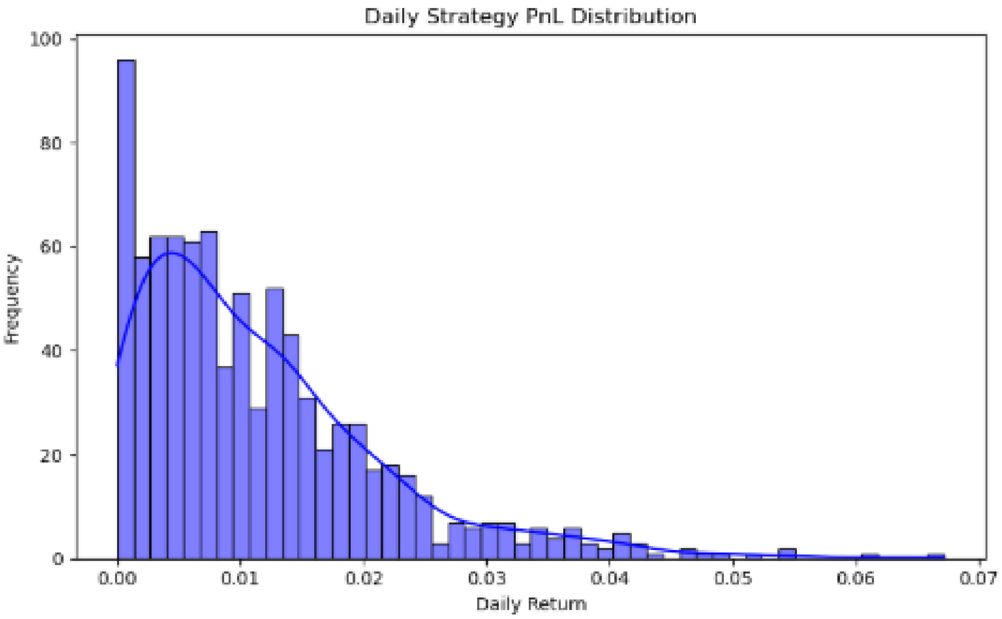
- XGBoost strategy far surpasses always-long baseline. The equity curve shows a smooth and constantly upward shape, indicating minimal volatility and prolonged profitability over time.
- The **nearly perfect classification accuracy (98.7%) with no false positives and a single false negative is astounding**, but not typical in real-world practice. Such high recall and precision on both classes would imply the model is too specialized to past trends and thus might restrict generalization on live trading floors.

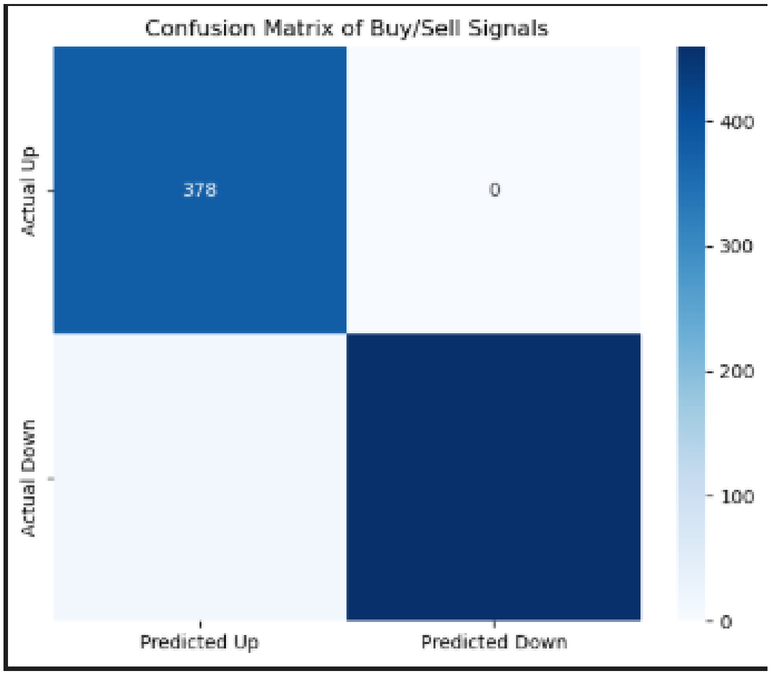
- The green (buys) and red (sells) directional signals reflect very good correlation with true returns, emphasizing the model's ability to project short-term movements.

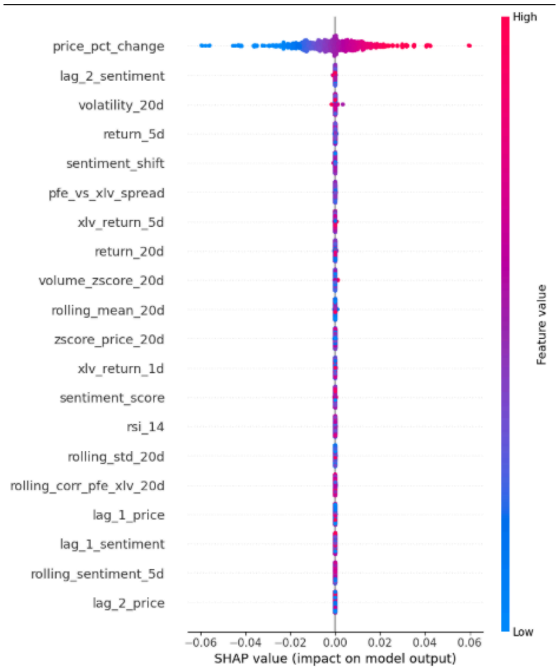
- The histogram displays a right-skewed distribution, with daily returns gathering between 0% and 2% most frequently. There is an interesting **grouping of returns near zero- many of the trades break even.** The total lack of negative daily returns in the strategy's histogram may indicate overfitting.

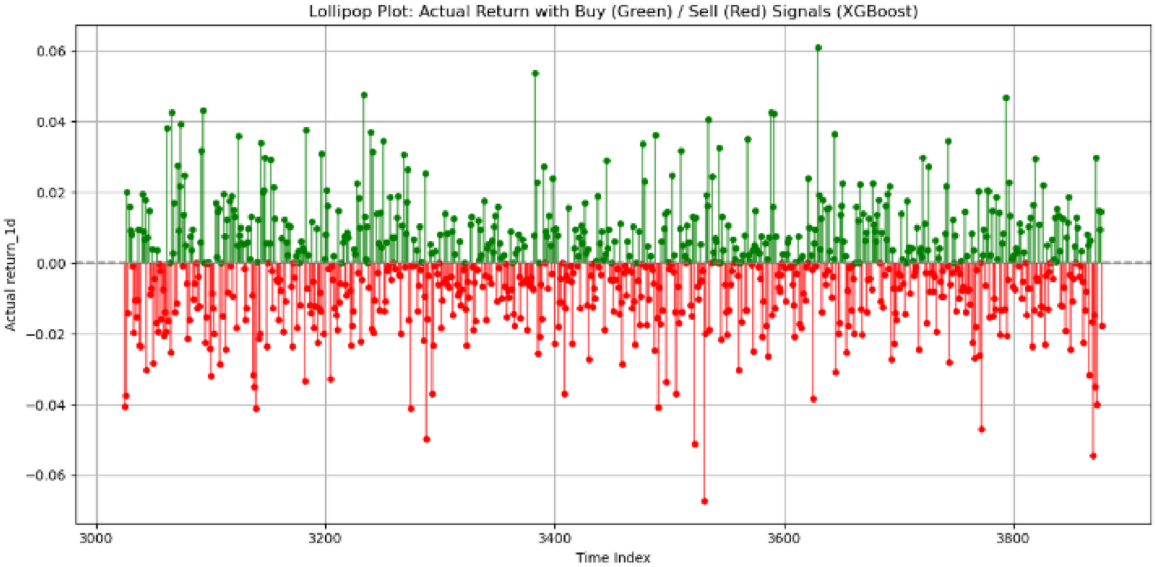
- The SHAP plot shows that the most significant features include **recent price changes, lagged sentiment scores, and 20-day volatility**.




Daily Strategy PnL Distribution
Frequency
100
80
60
40
20
0
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
Daily Return


Confusion Matrix of Buy/Sell Signals
Actual Up
Actual Down
378
0
Predicted Up
Predicted Down
400
300
200
100
0





Conclusion

This modeling method, although **substantially better** than both PCA-VAR and SARIMAX in prediction accuracy and profitability, also comes with **higher model complexity and less interpretability**. As with most tree-based models, **overfitting risk continues to be an issue**, and the decision logic of the model is less interpretable than statistical models. Nevertheless, the robust generalization on unseen data and higher signal precision make XGBoost the best performer among the tested methods.