



FACULTY OF SCIENCE / MASTER IN DATA SCIENCE & ANALYTICS / DATA 606

## Forecasting S&P 500 Stock Returns using Financial Fundamentals, News Sentiment & Macro Financial Indicators

**February 17<sup>th</sup>, 2026**

1. Alejandro Alvarado - 30301339
2. Aravindh Palaniguru - 30301478
3. Jorge Gutierrez Barajas - 30300374



UNIVERSITY OF  
CALGARY

# Abstract

*This study investigates whether financial fundamentals, news sentiment, and macroeconomic indicators can predict next-day stock returns for firms in the S&P 500 index. Using daily stock returns from November 2024 to October 2025, we implement two modeling frameworks: (i) a sector-wise XGBoost Random Forest regressor to forecast return magnitude, and (ii) a Multinomial Logistic Regression with L1/Elastic Net regularization to classify return direction (Up, Down, Same). Performance is evaluated using time-series cross-validation and out-of-sample validation. Results indicate modest predictive performance, with volatility-related variables emerging as the dominant predictors. The XGBoost model achieves a weighted directional accuracy of 52.46% (vs. 50% random baseline), while the multinomial model reaches 38.17% validation accuracy and 34.94% macro F1-score (vs. 33.3% random baseline), outperforming sector-specific naive classifiers in 9 of 11 sectors. Overall, results indicate that short-term equity return prediction remains challenging, consistent with semi-strong market efficiency. Nevertheless, modest and sector-dependent predictive structure exists.*



## Table of Contents

<b>1. Introduction .....</b>	<b>4</b>
Research Question .....	4
<b>2. Data .....</b>	<b>4</b>
Variables .....	5
Time Alignment and Sample .....	6
<b>3. Methodology .....</b>	<b>6</b>
Data Pre-processing and Feature Engineering.....	6
Train / Validation Split .....	6
Model 1: XGBoost Random Forest Regressor .....	7
Model 2: Multinomial Logistic Regression with Lasso Regularization .....	8
<b>4. Results .....</b>	<b>9</b>
Model 1: XGBoost Regressor Results .....	9
Model 2: Multinomial Logistic Regression.....	10
<b>5. Final Remarks .....</b>	<b>12</b>
Interpretation of Results.....	12
Limitations .....	13
Conclusions .....	13
Future Work .....	14
<b>References.....</b>	<b>15</b>
<b>Appendix.....</b>	<b>16</b>
1. Dictionary Table .....	16
2. Top 10 Variables by Mean Absolute Coefficient (by sector).....	18

# 1. Introduction

Equity markets play a central role in the global financial system, influencing investment decisions, corporate financing, and long-term economic growth. Understanding the dynamics of stock price movements is therefore a fundamental problem in finance and data-driven decision making for individual investors, analysts, and policymakers, as well as institutional investors such as pension funds, whose long-term strategies rely on separating market noise from meaningful signals.

Traditional asset pricing theory emphasizes the role of financial fundamentals such as profitability, leverage, and liquidity, while more recent research highlights the importance of market sentiment and macroeconomic conditions in shaping short-term price dynamics. In modern financial markets, information is disseminated rapidly through news and media coverage, which can influence investor perceptions beyond what is captured by accounting data alone.

The goal of this project is to empirically examine to what extent fundamentals, news sentiment, and macro-financial indicators predict next-day stock returns for S&P 500 firms. Two complementary modeling strategies are employed: an XGBoost Random Forest regression model that predicts the magnitude of next-day returns, and a Multinomial Logistic Regression model with Lasso regularization that predicts the direction of those returns (Up, Down, or Same).

## Research Question

The central research question guiding this study is: Can financial fundamentals, news sentiment, and macroeconomic indicators reliably predict the next-day returns of S&P 500 firms? This question is decomposed into two sub-questions: (i) To what extent can regression models predict the magnitude of next-day equity returns? and (ii) Can a classification model predict the direction of next-day returns better than a naive sector-specific baseline?

The ultimate goal is to better understand the drivers of stock performance (*via* Lasso regularization approach), and to evaluate the extent to which equity market prices could be predicted.

# 2. Data

The dataset used in this project combines firm-level, sentiment-based, and macroeconomic data from multiple sources. Stock price data and quarterly financial statements for S&P 500 companies are obtained using Python's `yfinance` library, which provides daily closing prices as well as balance sheet, income statement, and cash flow information. From these statements, key financial ratios and performance measures are constructed, including profitability, leverage, liquidity, and cash flow metrics.

Daily news sentiment data are sourced from the Bloomberg Terminal through the Bloomberg News Sentiment Index. These sentiment scores are time-aligned with firm-level stock prices and financial data at the company-date level. In addition, macroeconomic indicators such as Gross Domestic Product (GDP) growth, the Industrial Production Index (IPI), and the unemployment rate are obtained from the Federal Reserve Economic Data (FRED) database.

The population of interest consists of all companies listed in the S&P 500 index during the period from November 1st, 2024, to October 31st, 2025. The dataset includes daily observations for stock prices and sentiment, quarterly observations for financial fundamentals, and monthly or quarterly

observations for macroeconomic indicators. All data are used strictly for academic purposes. Figure 1 summarizes the three data sources.

Figure 1: Summary of Data Sources

Source	Dataset	Coverage	Format
Yahoo Finance (yfinance)	Daily prices, quarterly financials, VIX, sector	S&P 500, daily prices of 2024 and 2025	Python library
Bloomberg Terminal	News Sentiment Index (daily)	S&P 500 with Bloomberg coverage	Excel / CSV via VBA
FRED	GDP growth, IPI growth, Unemployment Rate	Historical macro indicators	CSV downloads

## Variables

The full feature set comprises 62 numeric variables spanning five categories, plus industry as a one-hot encoded categorical predictor. The categories are: **(i)** autoregressive price signals — six lagged daily returns (lags 1, 2, 3, 5, 10, 20) and six lagged dividend observations; **(ii)** technical indicators — five Simple Moving Averages (SMA\_5 through SMA\_200), rolling volatility measures, and Average True Range proxies (ATR\_5, ATR\_10, ATR\_20), all computed from prior-day OHLCV<sup>1</sup> data; **(iii)** news sentiment — lagged Bloomberg Sentiment Index values; **(iv)** firm-level fundamentals — nine financial ratios (RoE, RoA, operating margin, debt-to-equity, liquidity ratio, current ratio, free cash flow margin, revenue growth, and OCF-to-assets), two size/value metrics (log market cap and book-to-market), and **(v)** 22 macro-financial series covering Volatility Index, Gross Domestic Product, the Industrial Production Index, and unemployment in multiple seasonal-adjustment and growth-rate forms.

Figure 2: Variable Definitions by Category

Category	Variables	n	Source
<b>Autoregressive</b>	Lagged daily returns: Return_Lag1,2,3,5,10,20 Lagged dividends: Dividends_Lag1,2,3,5,10,20	<b>12</b>	Yahoo Finance
<b>Technical</b>	Simple Moving Averages: SMA_5,10,20,50,200 Rolling Volatility: Volatility_5,10,20 Avg True Range (proxy): ATR_5,10,20	<b>11</b>	Yahoo Finance
<b>Sentiment</b>	Bloomberg Sentiment Index: Sentiment_Lag1,2,3,5,10,20	<b>6</b>	Bloomberg
<b>Financial Ratios</b>	RoE, RoA, Operating Margin, Debt-to-Equity, Liquidity Ratio, Current Ratio, FCF Margin, Revenue Growth (QoQ), OCF-to-Assets	<b>9</b>	Yahoo Finance
<b>Size / Value</b>	Log Market Cap (log_mktcap), Book-to-Market Ratio	<b>2</b>	Yahoo Finance
<b>Macro — VIX</b>	CBOE VIX Index: VIX_Lag1,2,3,5,10,20	<b>6</b>	Yahoo Finance
<b>Macro — GDP</b>	GDP: SA/NSA × QoQ/YoY growth rates (4 series)	<b>4</b>	FRED
<b>Macro — IPI</b>	Industrial Production Index: raw, SA, QoQ, YoY, SA-QoQ, SA-YoY (6 series)	<b>6</b>	FRED
<b>Macro — Unemployment</b>	Unemployment Rate: SA/NSA levels and growth rates (6 series)	<b>6</b>	FRED
<b>Categorical</b>	GICS <sup>2</sup> Sector, GICS Industry (one-hot encoded)	<b>2*</b>	Yahoo Finance
<b>Total numeric features (before VIF screening)</b>		<b>62</b>	

<sup>1</sup> OHLCV stands for Open, High, Low, Close and Volume figures.

<sup>2</sup> Companies are categorized following Global Industry Classification Standard (GICS).

## Time Alignment and Sample

The population of interest consists of all 503 S&P 500 constituents spanning 11 GICS sectors, for the period from November 1, 2024, to October 31, 2025. Given the mixed observation frequencies — daily for prices and sentiment, quarterly for fundamentals, and monthly or quarterly for macro indicators — all variables were lagged and forward-filled to produce a clean panel at the daily frequency. The goal was to create a real-time alike dataset that could portray the information known at each moment (real-time historical dataset).

# 3. Methodology

## Data Pre-processing and Feature Engineering

Prior to modeling, the data underwent a pre-processing procedure. Financial ratios were computed from raw quarterly statements and aligned to their releasing date, then forward-filled to produce daily observations, ensuring no future information leakage. Technical indicators (SMA, rolling volatility, and ATR proxies) were constructed from OHLCV data and shifted by one trading day, so each observation at time  $t$  reflects only information available at  $t-1$ . Bloomberg sentiment scores were included as six lagged features (lags 1 through 20 days) to capture both immediate and persistent sentiment effects. All numeric variables were standardized to zero mean and unit variance within each training fold to facilitate comparison of regularized coefficients across the models.

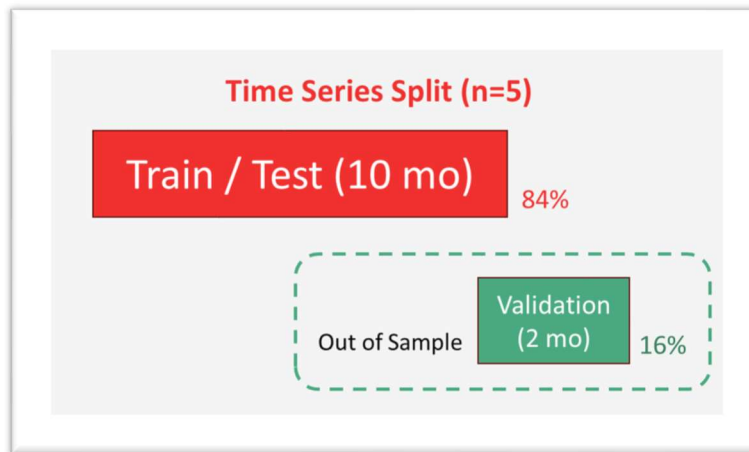
A key pre-processing step was the application of a Variance Inflation Factor (VIF) filter with threshold 5.0 to remove multicollinear features. Starting from 62 candidate features, the VIF procedure produced sector-specific feature sets ranging from 38 to 43 retained features, with an average of approximately 21 features dropped per sector. This step improves the stability and interpretability of the Lasso-penalized model.



## Train / Validation Split

Given the time-series nature of the data, a strict chronological split was imposed. The training and test window (November 2024 – August 2025) accounts for the majority of observations, within which a rolling *TimeSeriesSplit* with five folds was applied for hyperparameter tuning via cross-validation. The final two months (September – October 2025) form the out-of-sample validation set, held completely apart and used solely for final performance evaluation.

Figure 3: Data Preparation



The training window spans November 2024 through August 2025, resulting in 103K observations across all sectors. The out-of-sample validation window covers September–October 2025 (22K total firm-day observations). This strict temporal split prevents look-ahead bias and leakage from your validation dataset.

### Model 1: XGBoost Random Forest Regressor

The first modeling approach uses a sector-wise XGBoost Random Forest Regressor (XGBRFRegressor) to predict next-day stock returns. For each ticker, the target is defined as the one-day-ahead return, and predictors include lagged returns/dividends, technical indicators (SMA, volatility, ATR), lagged sentiment and VIX, firm fundamentals, and macroeconomic variables. To avoid look-ahead bias, the model is trained on November 2024 to August 2025, tuned with time-series cross-validation (up to 5 folds, maximum training window of 126 trading dates), and evaluated on a fully out-of-sample validation period (September to October 2025). Numeric features are median-imputed and filtered with iterative VIF selection (threshold = 5), while ticker and industry are one-hot encoded. Hyperparameters are selected by minimizing cross-validated RMSE, and performance is reported using RMSE, MAE,  $R^2$ , and directional accuracy (sign agreement between predicted and realized returns).

#### Hyperparameter Tuning for Model 1 (XGBoost Random Forest)

For the XGBoost Random Forest regressor (XGBRFRegressor), we tuned the main complexity and regularization controls: number of trees (`n_estimators`), tree depth (`max_depth`), minimum child weight (`min_child_weight`), row subsampling (`subsample`), column subsampling at each split (`colsample_bynode`), L1 penalty (`reg_alpha`), and L2 penalty (`reg_lambda`).

The search grid was:

- `n_estimators` {200, 400},
- `max_depth` {4, 6},
- `min_child_weight` {1, 5},
- `subsample` {0.7, 0.9},

- colsample\_bynode {0.7, 0.9},
- reg\_alpha {0.0, 0.1},
- and reg\_lambda {1.0, 5.0}.

Figure 4: XGBoost Regressor — Hyperparameters Obtained from Tuning

Sector	n_estimators	max_depth	min_child_weight	subsample	colsample_bynode	reg_alpha	reg_lambda
Basic Materials	200	6	1	0.7	0.7	0.1	5
Communication Services	200	6	5	0.7	0.7	0.1	5
Consumer Cyclical	400	6	1	0.7	0.7	0.1	5
Consumer Defensive	400	6	1	0.7	0.7	0.1	5
Energy	200	4	1	0.7	0.7	0.1	5
Financial Services	400	6	1	0.7	0.7	0.1	5
Healthcare	400	6	1	0.7	0.7	0.1	5
Industrials	200	4	1	0.7	0.7	0.1	5
Real Estate	200	6	1	0.7	0.7	0.1	5
Technology	400	6	5	0.7	0.7	0.1	5
Utilities	200	6	1	0.7	0.7	0.1	5

Tuning was performed independently by sector using a chronological training design to avoid look-ahead bias. Within the training window, we used rolling time-series cross-validation (up to 5 folds, maximum training window of 126 trading dates) and selected the best configuration by minimizing cross-validated RMSE (selection\_metric = min\_rmse). This approach preserves temporal order while allowing each sector to choose its own bias-variance tradeoff.

The selected values (Figure 4) show a consistent pattern across sectors. Stronger regularization was preferred in all 11 sectors (reg\_alpha = 0.1, reg\_lambda = 5.0), with subsample = 0.7 and colsample\_bynode = 0.7 also selected in all sectors. For model size, n\_estimators = 200 was selected in 6 sectors and 400 in 5 sectors; max\_depth = 6 in 9 sectors (and 4 in Energy and Industrials); min\_child\_weight = 1 in 9 sectors (and 5 in Communication Services and Technology). With these tuned settings, validation performance reached a sector-level weighted directional accuracy of 52.46% and weighted RMSE of 0.01926.

## Model 2: Multinomial Logistic Regression with Lasso Regularization

The second modeling approach treats return direction as a three-class classification problem (Up, Down, Same) and estimates a Multinomial Logistic Regression with L1 (Lasso) or Elastic Net regularization. Models were estimated independently for each of the 11 GICS sectors covering the full 503-ticker sample. The target variable, Return\_Status, is engineered using a balanced-threshold approach: for each sector, the threshold is selected by scanning the percentiles of absolute returns and choosing the value that best balances the three class shares.

Hyperparameter tuning was performed via grid search over:

- the inverse regularization strength  $C \in \{0.001, 0.01, 0.1, 1.0\}$ ,
- penalty type  $\in \{L1, \text{Elastic Net}\}$ , and
- the Elastic Net mixing ratio  $l1\_ratio \in \{0.2, 0.5, 0.8\}$ .

In this model,  $C$  controls the overall strength of regularization because it is the inverse of the penalty: a lower  $C$  implies stronger shrinkage, which pushes more coefficients toward zero, yielding a simpler and usually lower-variance model; a higher  $C$  weakens shrinkage, allowing more flexibility but increasing overfitting risk. The penalty choice determines the type of shrinkage:  $L1$  induces stronger sparsity and tends to select a smaller subset of predictors, while *elasticnet* combines  $L1$  and  $L2$ , preserving sparsity but improving stability when predictors are correlated. When *penalty='elasticnet'*, *l1\_ratio* determines the mix: values closer to 1 behave more like lasso (more aggressive sparsity), whereas values closer to 0 behave more like ridge (less sparsity, more shared weight across correlated variables). Intuitively, *l1\_ratio* lets you choose between hard variable selection and smoother, more distributed shrinkage.

The selection criterion was the macro-averaged F1-score alone, prioritizing balanced predictive performance across all three return-direction classes. Model performance is reported using overall accuracy and macro-averaged F1-score across train, out-of-fold test, and out-of-sample validation partitions.

## 4. Results

### Model 1: XGBoost Regressor Results

Figure 5 presents sector-level out-of-sample validation performance of the XGBoost regressor across all 503 S&P 500 firms. The overall directional accuracy is 52.46%, which is modestly above the 50% random baseline. Performance varies across sectors: Utilities shows the highest directional accuracy (59.49%), followed by Basic Materials (53.72%) and Healthcare (53.02%), while Consumer Cyclical is the weakest at 49.01% (below baseline).

Figure 5: XGBoost Regressor — Validation Performance Summary

Sector	RMSE (OOS)	Dir. Accuracy (OOS)	vs. 50% Baseline
Utilities	0.014586	59.49%	Above 50% baseline
Basic Materials	0.022145	53.72%	Above 50% baseline
Healthcare	0.020194	53.02%	Above 50% baseline
Financial Services	0.016842	52.71%	Above 50% baseline
Technology	0.026435	52.63%	Above 50% baseline
Real Estate	0.014218	52.51%	Above 50% baseline
Communication Services	0.023532	51.91%	Above 50% baseline
Energy	0.018656	51.16%	Above 50% baseline
Consumer Defensive	0.016719	50.85%	Above 50% baseline

Sector	RMSE (OOS)	Dir. Accuracy (OOS)	vs. 50% Baseline
Industrials	0.018084	50.10%	Above 50% baseline
Consumer Cyclical	0.020424	49.01%	Below 50% baseline
Weighted Mean (all sectors)	0.01926	52.46%	Above 50% baseline

Validation RMSE also differs across sectors, with a weighted mean of 0.01926. The lowest RMSE values are observed in Real Estate (0.014218) and Utilities (0.014586), while Technology records the highest RMSE (0.026435), indicating larger prediction errors in that sector. Overall, 10 of 11 sectors outperform the 50% directional benchmark, suggesting that the model captures useful short-term signal, but the strength of predictability remains clearly sector-dependent.

An additional insight is the cross-sector dispersion in performance. Directional (up and down) accuracy ranges from 49.01% (Consumer Cyclical) to 59.49% (Utilities), a 10.48 percentage-point spread, while validation RMSE ranges from 0.014218 (Real Estate) to 0.026435 (Technology). Importantly, RMSE and directional accuracy are not perfectly aligned: for example, Real Estate has the lowest RMSE but only mid-range directional accuracy (52.51%), whereas Basic Materials shows higher RMSE (0.022145) but stronger directional accuracy (53.72%). This suggests the model may capture return direction more reliably than return magnitude in some sectors, and that predictive strength is heterogeneous rather than uniform across the market.

## Model 2: Multinomial Logistic Regression

Figure 6 reports ticker-weighted aggregate accuracy and macro-averaged F1-score for the Multinomial Logit model across the train, out-of-fold (OOF) test, and out-of-sample (OOS) validation partitions, now estimated on the full 503-ticker sample. The validation accuracy of 0.3817 and validation F1-score of 0.3494 both comfortably exceed the 33.3% theoretical random baseline for a balanced three-class problem, confirming that the combined feature set provides statistically meaningful directional signal at full scale.

The gap between training accuracy (0.4307) and validation accuracy (0.3817) reflects expected overfitting, which the Lasso and Elastic Net regularization partially mitigates. Validation accuracy (0.3817) slightly exceeds out-of-fold test accuracy (0.3719), suggesting the model generalizes robustly to the held-out period.

Figure 6: Multinomial Logit — Aggregate Performance (across 11 sectors)

Metric	Train	Test (OOF)	Validation (OOS)
Overall Accuracy	0.4307	0.3719	0.3817
Macro F1-Score	0.4226	0.3537	0.3494

## Multinomial Logistic Regression — Sector-Level Results

Figure 7 presents validation accuracy and macro F1-score broken down by GICS sector, alongside the sector-specific naive baseline accuracy (the most-frequent-class share in the validation window), the best-tuned regularization penalty, and the inverse regularization strength C. Performance varies considerably across sectors; results reflect the full 503-ticker sample.

At full scale (503 tickers), 9 of 11 sectors beat their sector-specific naive baseline and have a F1 Score above 0.3333. Utilities leads with 42.30% validation accuracy, followed by Technology (41.22%) and Basic Materials (39.09%). Only Financial Services (35.87%) and Real Estate (34.90%) fall below their

respective baseline rates, both sectors having relatively high naive baseline rates due to class concentration in the validation window (36.86% and 36.07% respectively), as well as a low F1 Score.

Figure 7: Multinomial Logit — Sector-Level Validation Performance (OOS)

Sector	n	Val Acc.	Val F1	Baseline <sup>3</sup>	Beats? <sup>4</sup>	Penalty	C	l1_ratio
Utilities	31	0.4230	0.3625	0.3761	Yes	L1	1.0	N/A
Technology	81	0.4122	0.3911	0.3488	Yes	Elastic Net	1.0	0.2
Basic Materials	20	0.3909	0.3640	0.3466	Yes	Elastic Net	1.0	0.2
Comm. Services	25	0.3900	0.3547	0.3664	Yes	Elastic Net	1.0	0.2
Energy	22	0.3822	0.3409	0.3605	Yes	Elastic Net	1.0	0.2
Industrials	72	0.3778	0.3486	0.3387	Yes	Elastic Net	1.0	0.2
Healthcare	60	0.3758	0.3462	0.3417	Yes	Elastic Net	1.0	0.2
Consumer Defensive	37	0.3741	0.3734	0.3692	Yes	Elastic Net	1.0	0.8
Consumer Cyclical	55	0.3698	0.3541	0.3562	Yes	Elastic Net	1.0	0.2
Financial Services	69	0.3587	0.3055	0.3686	No	Elastic Net	1.0	0.2
Real Estate	31	0.3490	0.2888	0.3607	No	Elastic Net	0.01	0.2

Regarding hyperparameter choices, seven of the eleven sectors selected Elastic Net regularization over pure L1 (Lasso), suggesting that combining L1 sparsity with L2 shrinkage provides better out-of-sample generalization in most sectors. The optimal inverse regularization strength C was either 0.1 (stronger regularization) or 1.0 (weaker regularization), with no sector favoring the most extreme values in the grid (0.001 or 0.01), indicating that moderate regularization levels are most appropriate for this dataset.

### Model Interpretability: Lasso-Selected Features

A key advantage of the Multinomial Logit with L1 regularization is its ability to identify the most relevant predictors of return direction through coefficient shrinkage. Figure 8 ranks the top ten variables by their coefficient magnitude, averaged across all sector models. Because all variables were standardized prior to estimation, coefficients are directly comparable across sectors and models.

The ranking shows a clear structure in predictive importance across sectors. The strongest signals are led by market volatility and short-horizon state variables, with VIX\_Lag at the top, followed by SMA\_5 (0.191) and valuation terms such as book\_to\_market (0.168) and log\_mktcap (0.140). This pattern is consistent with the idea that short-run risk conditions and market state capture much of the directional variation needed to classify Up/Down/Same outcomes. Macroeconomic activity also contributes meaningfully through the Industrial Production Index, while balance-sheet structure remains relevant via debt\_to\_equity (0.120). Overall, the top 10 variables indicate that risk, technical dynamics, valuation, and selected macro indicators are the main cross-sector drivers in this multinomial framework. Notably, Bloomberg news sentiment variables do not appear among the top ten most influential predictors when averaged across sectors, suggesting that sentiment information is either rapidly priced in.

<sup>3</sup> Baseline: validation accuracy of a naive majority-class classifier. For each sector, it always predicts the most frequent Return\_Status class (up, down, or same) in that sector's validation sample; it represents the market's "safe bet" benchmark.

<sup>4</sup> Beats? Compares Val. Acc. vs Baseline (majority-class accuracy on validation).

Figure 8: Top 10 Variables by Mean Absolute Coefficient (across sectors)

Rank	Variable	Mean $ \beta $	Category
1	VIX_Lag1	0.248	Macro / Autoregressive
2	SMA_5	0.191	Technical Indicator
3	Book-to-Market	0.168	Financial Fundamental
4	Log Market Cap	0.140	Financial Fundamental
5	IPI Growth (QoQ)	0.129	Macroeconomic
6	VIX_Lag5	0.127	Macro / Autoregressive
7	Debt-to-Equity	0.120	Financial Fundamental
8	Volatility_10	0.111	Technical Indicator
9	VIX_Lag10	0.105	Macro / Autoregressive
10	IPI (Seasonally Adjusted)	0.092	Macroeconomic

The Appendix (section 2) provides the full sector-by-sector breakdown with the top 10 absolute Lasso coefficient magnitude. Across sectors, market-volatility and short-horizon technical signals dominate: VIX\_Lag1 appears as the top variable in most sectors, with other VIX lags, volatility measures, and SMA\_5 frequently entering the top 10. Financial and valuation variables (book\_to\_market, log\_mktcap, debt\_to\_equity) remain consistently relevant, while macro production indicators (IPI) provide additional sector-level explanatory power.

## 5. Final Remarks

### Interpretation of Results

Taken together, the results from both models tell a coherent and consistent story: predicting short-term equity returns from publicly available financial fundamentals, news sentiment, and macroeconomic indicators is genuinely difficult, but the signal is real and sector-dependant. This finding aligns with the semi-strong form of the Efficient Market Hypothesis, which posits that publicly available information should already be reflected in current prices, leaving little room for systematic return predictability. However, the results are not entirely null: the Multinomial Logit model consistently outperforms the random 33.3% baseline at the aggregate level, and beats sector-specific naive baselines in 9 of 11 sectors at full scale (503 tickers).

The dominance of lagged VIX values in the coefficient rankings is economically intuitive: implied volatility captures market-wide uncertainty and risk appetite, both of which influence the cross-sectional distribution of next-day return outcomes. During high-VIX regimes, return dispersion increases and directional outcomes become more predictable in a statistical sense, as extreme movements in either direction become more likely than near-zero returns. This mechanism explains why VIX lags at multiple horizons retain significant and positive Lasso coefficients across sectors.

The feature ranking indicates that financial fundamentals remain economically meaningful, especially valuation and capital-structure variables such as book\_to\_market, log\_mktcap, and debt\_to\_equity, with profitability metrics (roe, roa, op\_margin) also entering the top set in several sectors. This pattern is consistent with cross-sectional asset-pricing intuition: firms with different balance-sheet strength and profitability profiles exhibit different return behavior, and those structural differences can still contribute to short-horizon directional classification when modeled jointly with market-risk and macro signals.

The failure of news sentiment to appear among the top predictors is more nuanced: it does not imply that sentiment is irrelevant, but rather that — averaged across sectors and firms — Bloomberg sentiment signals appear to be incorporated into prices before the next trading day, consistent with rapid information processing in liquid markets.

The marked heterogeneity in model performance warrants comment. The XGBoost regressor delivers stronger pure directional performance, with weighted directional accuracy of 52.46% and 10 of 11 sectors above the 50% baseline, while the multinomial model's validation accuracy ranges from 34.9% to 42.3% and beats its sector-specific majority-class baseline in 9 of 11 sectors. Both approaches rank Utilities and Technology near the top, but XGBoost also clears baseline in Financial Services and Real Estate, where the multinomial model does not.

## Limitations

Several limitations should be noted when interpreting these results. First, the one-year sample period (November 2024 – October 2025) may not span a sufficient range of market regimes, limiting the generalizability of the findings to different economic environments. Second, Bloomberg sentiment coverage may not be uniform across all S&P 500 companies, which may introduce selection bias if firms with higher media coverage differ systematically from those without. Third, the classification of returns into Up, Down, and Same categories is sensitive to the threshold used to define near-zero returns; the balanced-threshold selection procedure partially addresses this, but alternative thresholds could yield different results.



## Conclusions

This study examined whether financial fundamentals, news sentiment, and macroeconomic indicators can predict next-day stock returns for 503 S&P 500 companies across 11 GICS sectors. Using two complementary modeling approaches, an XGBoost Random Forest regressor for return-magnitude prediction and Multinomial Logistic Regression with Lasso/Elastic Net regularization for directional classification, the analysis yields a consistent conclusion: short-term equity return prediction from publicly available information is extremely difficult, but not entirely without signal.

The XGBoost Random Forest regressor achieves a weighted mean directional accuracy of 52.46% across sectors, a modest but consistent improvement over the 50% random baseline, with a weighted validation RMSE of 0.01926. Performance is heterogeneous across sectors: Utilities (59.49%) shows the strongest directional performance, while Consumer Cyclical (49.01%) is the weakest; overall, 10 of 11 sectors remain above the 50% benchmark. The Multinomial Logit model performs meaningfully above the three-class random baseline at full scale, with a ticker-weighted validation accuracy of 38.17% and F1-score of 34.94% against a 33.3% floor. At the sector level, 9 of 11 sectors beat their sector-specific naive baseline, with Utilities (42.30%), Technology (41.22%), and Basic Materials (39.09%) showing the strongest out-of-sample performance. Lasso regularization reveals that VIX-based autoregressive factors

dominate coefficient magnitudes across all sectors, followed by firm-level financial fundamentals, while news sentiment variables do not emerge as top contributors to directional predictability.

These findings have practical implications: combining multiple information layers, market volatility dynamics, firm fundamentals, and macroeconomic conditions, provides incremental forecasting value over any single source, but the magnitude of that advantage remains modest (at best) and sector dependent. Short-term equity returns forecasting remains a genuinely hard problem, consistent with decades of evidence in empirical finance.

## Future Work

Several natural extensions could improve predictive performance and deepen the analysis. Expanding the sample from S&P 500 constituents to all NYSE listed companies would increase cross-sectional statistical power. The current technical indicator set already includes SMA, rolling volatility, and ATR proxies; however, including Relative Strength Index (RSI), Moving Average Convergence Divergence (MACD), Bollinger Band width, and volume-relative measures could add momentum and mean-reversion signals that complement the existing volatility-based predictors. The Put-Call Ratio (PCR) could also be incorporated as an additional market sentiment proxy. Also, predicting medium- to long-term returns at weekly or monthly horizons may reveal stronger fundamental-driven predictability, as the daily noise decreases at longer horizons.

On the modeling side, architectures that jointly model interactions between stocks and sectors, such as Recurrent Neural Networks (RNNs), Temporal Convolutional Networks (TCNs), and Graph Neural Networks (GNNs), represent promising directions for capturing the dynamic relational structure of equities. On the content side, specialization could provide the greatest improvement. Given the nature of each sector, it is difficult to generalize all industries. For instance, in energy sector, incorporating information of futures on the oil prices could meaningfully enhance the accuracy of the forecast.

---

*'To the extent we have been successful, it is because we concentrated on identifying one-foot hurdles that we could step over rather than because we acquired any ability to clear seven-footers.' — Warren Buffett*

---

# References

- Bloomberg L.P. (n.d.). Bloomberg Market Concepts (BMC) [Online training course]. Bloomberg Terminal.
- Federal Reserve Economic Data (FRED). (n.d.). Economic data, statistics, and graphs. Federal Reserve Bank of St. Louis. <https://fred.stlouisfed.org>
- Ross, S. A., Westerfield, R. W., & Jaffe, J. (2017). Corporate Finance: Core Principles and Applications (5th ed.). McGraw-Hill Education.
- Yahoo Finance. (n.d.). Yahoo Finance API and data tools. <https://finance.yahoo.com>

# Appendix

## 1. Dictionary Table

This annex provides a complete enumeration of all variables in the feature set prior to Variance Inflation Factor (VIF) screening. Variables are grouped by category and colour-coded by data source: blue shading indicates Yahoo Finance (yfinance), amber shading indicates Bloomberg Terminal, and green shading indicates the Federal Reserve Economic Data (FRED) database. All technical indicators are computed from prior-day OHLCV data and shifted by one trading day to prevent look-ahead bias. Quarterly and monthly macroeconomic and fundamental variables are aligned to their public release dates and forward-filled to produce daily observations.

**Complete Variable Dictionary — All Features Before VIF Screening**

Variable / Group	Category	Description	Formula / Transformation	Source	Freq.	Type
Return_Lag1–20	Autoregressive	Daily close-to-close log return, lagged 1,2,3,5,10,20 days	$Return_{(t-k)}$	yfinance	Daily	Cont.
Dividends_Lag1–20	Autoregressive	Dividend per share, lagged 1,2,3,5,10,20 days	$Dividends_{(t-k)}$	yfinance	Daily	Cont.
SMA_5/10/20/50/200	Technical	Simple Moving Average of Close over 5,10,20,50,200 days	$MA_k(Close), shift(1)$	yfinance	Daily	Cont.
Volatility_5/10/20	Technical	Rolling std. dev. of returns over 5,10,20 days	$STD_k(Return), shift(1)$	yfinance	Daily	Cont.
ATR_5/10/20	Technical	Avg True Range proxy: rolling mean of (High–Low)	$MA_k(H-L), shift(1)$	yfinance	Daily	Cont.
Sentiment_Lag1–20	Sentiment	Bloomberg News Sentiment Index, lagged 1,2,3,5,10,20 days	$Sentiment_{(t-k)}$	Bloomberg	Daily	Cont.
VIX_Lag1–20	Market Risk	CBOE Volatility Index, lagged 1,2,3,5,10,20 days	$VIX_{(t-k)}$	yfinance	Daily	Cont.
roe	Financial Ratio	Return on Equity	$100 \times net\_income / equity$	yfinance	Quarterly	Cont.
roa	Financial Ratio	Return on Assets	$100 \times net\_income / assets$	yfinance	Quarterly	Cont.
op_margin	Financial Ratio	Operating Margin	$100 \times op\_income / revenue$	yfinance	Quarterly	Cont.
debt_to_equity	Financial Ratio	Debt-to-Equity Ratio	$total\_debt / equity$	yfinance	Quarterly	Cont.
liquidity_ratio	Financial Ratio	Cash-to-Assets Ratio	$cash / total\_assets$	yfinance	Quarterly	Cont.
current_ratio	Financial Ratio	Current Ratio	$curr\_assets / curr\_liab$	yfinance	Quarterly	Cont.
free_cf_margin	Financial Ratio	Free Cash Flow Margin	$100 \times FCF / revenue$	yfinance	Quarterly	Cont.
revenue_growth	Financial Ratio	Quarter-over-Quarter Revenue Growth	$100 \times \Delta revenue / rev\_prev$	yfinance	Quarterly	Cont.

Variable / Group	Category	Description	Formula / Transformation	Source	Freq.	Type
ocf_to_assets	Financial Ratio	Operating Cash Flow to Assets	$op\_cash\_flow/assets$	yfinance	Quarterly	Cont.
log_mktcap	Size / Value	Natural log of market capitalization (lagged)	$\ln(Close_{t-1} \times shares)$	yfinance	Daily	Cont.
book_to_market	Size / Value	Book-to-Market Ratio (lagged equity / market cap)	$equity_{t-1}/mktcap$	yfinance	Daily	Cont.
GDP_SA_PC_QOQ	Macro — GDP	GDP Seasonally Adjusted, % change quarter-over-quarter	raw FRED series	FRED	Quarterly	Cont.
GDP_SA_PC_YOY	Macro — GDP	GDP Seasonally Adjusted, % change year-over-year	raw FRED series	FRED	Quarterly	Cont.
GDP_NSA_PC_QOQ	Macro — GDP	GDP Non-Seasonally Adjusted, % change QoQ	raw FRED series	FRED	Quarterly	Cont.
GDP_NSA_PC_YOY	Macro — GDP	GDP Non-Seasonally Adjusted, % change YoY	raw FRED series	FRED	Quarterly	Cont.
IPI_IPI	Macro — IPI	Industrial Production Index, raw level	raw FRED series	FRED	Quarterly	Cont.
IPI_IPI_YOY	Macro — IPI	IPI year-over-year change	raw FRED series	FRED	Quarterly	Cont.
IPI_IPI_QOQ	Macro — IPI	IPI quarter-over-quarter change	raw FRED series	FRED	Quarterly	Cont.
IPI_IPI_SA	Macro — IPI	IPI seasonally adjusted level	raw FRED series	FRED	Quarterly	Cont.
IPI_IPI_SA_YOY	Macro — IPI	IPI SA year-over-year change	raw FRED series	FRED	Quarterly	Cont.
IPI_IPI_SA_QOQ	Macro — IPI	IPI SA quarter-over-quarter change	raw FRED series	FRED	Quarterly	Cont.
UNEMP_UNRATE	Macro — Unemp.	Unemployment Rate (seasonally adjusted)	raw FRED series	FRED	Monthly	Cont.
UNEMP_UNRATE_PC1	Macro — Unemp.	Unemployment Rate, 1-period % change	raw FRED series	FRED	Monthly	Cont.
UNEMP_UNRATE_PCH	Macro — Unemp.	Unemployment Rate, % change from year ago	raw FRED series	FRED	Monthly	Cont.
UNEMP_UNRATEN_SA	Macro — Unemp.	Unemployment Rate, non-seasonally adjusted	raw FRED series	FRED	Monthly	Cont.
UNEMP_UNRATEN_SA_PC1	Macro — Unemp.	NSA Unemployment Rate, 1-period % change	raw FRED series	FRED	Monthly	Cont.
UNEMP_UNRATEN_SA_PCH	Macro — Unemp.	NSA Unemployment Rate, % change from year ago	raw FRED series	FRED	Monthly	Cont.
sector	Categorical	GICS sector label (used for model stratification)	classification	yfinance	Static	Cat.
industry	Categorical	GICS industry label, one-hot encoded as model predictor	OHE in pipeline	yfinance	Static	Cat.

Colour key: Blue = Yahoo Finance (yfinance); Amber = Bloomberg Terminal; Green = FRED (Federal Reserve Economic Data).

\* Industry is one-hot encoded directly in the sklearn pipeline (ColumnTransformer). Sector is used for model stratification only (not a direct predictor).

(1) All numeric variables are standardized ( $\mu=0$ ,  $\sigma=1$ ) within each training fold before model estimation.

(2) VIF screening (threshold = 5.0) is applied per sector, retaining 38–43 of the 62 numeric features.

## 2. Top 10 Variables by Mean Absolute Coefficient (by sector)

Variable	1	2	3	4	5	6	7	8	9	10
<b>Basic Materials</b>	VIX_Lag1	book_to_m arket	VIX_Lag10	log_mktcap	IPI_IPI_QO Q	IPI_IPI_SA	Return_Lag 20	VIX_Lag20	Volatility_5	debt_to_eq uity
<b>Communication Services</b>	log_mktcap	roe	VIX_Lag1	debt_to_equ ity	SMA_200	book_to_m arket	revenue_gr owth	VIX_Lag10	Volatility_5	IPI_IPI_QO Q
<b>Consumer Cyclical</b>	VIX_Lag1	book_to_m arket	Volatility_ 10	log_mktcap	Volatility_5	VIX_Lag10	IPI_IPI_QO Q	op_margin	Volatility_20	debt_to_eq uity
<b>Consumer Defensive</b>	VIX_Lag1	Volatility_1 0	IPI_IPI_QO Q	log_mktcap	book_to_m arket	liquidity_rat io	Return_Lag 1	Sentiment_ Lag3	Sentiment_L ag20	VIX_Lag5
<b>Energy</b>	book_to_m arket	VIX_Lag1	debt_to_e quity	VIX_Lag20	VIX_Lag5	Volatility_2 0	VIX_Lag10	log_mktcap	Return_Lag2	IPI_IPI_QO Q
<b>Financial Services</b>	VIX_Lag1	log_mktcap	IPI_IPI_QO Q	IPI_IPI_SA	VIX_Lag20	roa	VIX_Lag5	op_margin	Return_Lag5	book_to_m arket
<b>Healthcare</b>	VIX_Lag1	SMA_5	VIX_Lag5	book_to_ma rket	IPI_IPI_QO Q	IPI_IPI_SA	roe	VIX_Lag10	debt_to_equ ity	VIX_Lag20
<b>Industrials</b>	VIX_Lag1	IPI_IPI_QO Q	VIX_Lag5	log_mktcap	current_rati o	IPI_IPI_SA	VIX_Lag10	Volatility_5	Sentiment_L ag20	IPI_IPI_SA_ QOQ
<b>Real Estate</b>	VIX_Lag1	Volatility_1 0	Volatility_ 5	Sentiment_L ag10	Volatility_2 0	VIX_Lag20	Return_Lag 1	VIX_Lag5	Return_Lag5	IPI_IPI_SA
<b>Technology</b>	VIX_Lag1	VIX_Lag5	SMA_5	IPI_IPI_QOQ	VIX_Lag10	current_rati o	IPI_IPI_SA	debt_to_eq uity	VIX_Lag20	Return_Lag 10
<b>Utilities</b>	book_to_m arket	debt_to_eq uity	VIX_Lag5	Sentiment_L ag10	VIX_Lag1	log_mktcap	Volatility_5	IPI_IPI_QO Q	Sentiment_L ag3	IPI_IPI_SA