Final Project- Refining Equities Analysis- Return Prediction and Volatility Modeling

MSDS 492- Analysis of Financial Markets

Northwestern University

2/27/2026

By David Van Dyke

AI used in this report for custom coding, background search, and wording clarification. The work is my own.

Contents

Abstract 3

Introduction 4

Literature Review 5

Summary of Insights from Literature 7

References : 8

Theoretical Framework 9

Data 11

Methodology 13

Feature Construction and Engineering- 13

Engineered Features for Nonlinear Signals: 14

Incorporation of GARCH-Based Volatility Features: 16

Incorporation of News-Based Features: 17

Model Training and Selection: 19

Train/Test Splitting and Temporal Validation: 20

Model Evaluation and Performance Metrics: 21

Investment Strategy Backtesting: 22

Model Validation and Robustness: 23

Results 24

Part 1 – Empirical Asset Market Findings 24

Part 2 – Machine Learning Modeling Results for Predicting Next-Day Return Direction of Valero (VLO) 34

EDA on VLO Machine Learning Dataset 35

Baseline Logistic Regression Model Setup 38

Investment Backtests (Test Period Only) 42

Comparing Modifications to the Logistic Regression Model Setup 43

Comparison of More Complex Machine Learning Models to the Baseline Regularized Logistic Regression 46

Cross-Model Feature Importance 50
Impact of GARCH-Based Volatility Features on Model Performance 52
Impact of OPEC Press-Release Text Features on Model Performance 55
Model Performance with GARCH and OPEC News Feature Augmentation..... 57
Model Results Comparison (Baseline Logistic Regression vs. Advanced CNN with all Features) 59
Conclusions 61
Keywords 63
Bibliography 63

Abstract

This study investigates the daily stock return behavior of three major U.S. petroleum refining companies Phillips 66 (PSX), Valero Energy Corporation (VLO), and Marathon Petroleum Corporation (MPC) over a 15-year period from 2011 through early 2026. The research is conducted in two phases. The first phase provides a comparative analysis of return distributions, volatility patterns, and sensitivities to key external drivers, including the S&P 500 Index, crude oil prices, gold prices, and market volatility indicators (VIX). These diagnostics guide the selection of one firm, VLO, for focused predictive modeling. In the second phase, the study develops a supervised machine learning framework to forecast the binary direction of VLO's next-day returns (up or down). The modeling approach begins with a regularized logistic regression using market-based predictors and is extended to include engineered features that capture nonlinear interactions, volatility regimes, and relative performance metrics. To further enhance predictive power, the model incorporates forward-looking volatility forecasts from GARCH models and binary indicators derived from OPEC press releases, which flag the presence and thematic content of oil market news. Tree-based and neural network models are also evaluated using the same feature set to assess potential gains from nonlinear methods. Model performance is assessed using both statistical classification metrics and a simulated trading strategy to evaluate economic value. The results provide practical insights into the directional predictability of daily stock movements in the refining sector and demonstrate a transparent, data-driven methodology for short-horizon forecasting.

Introduction

The petroleum refining and downstream energy sector operates in a highly dynamic environment shaped by volatile commodity prices, shifting macroeconomic conditions, and evolving investor sentiment. Companies such as Phillips 66 (PSX), Valero Energy Corporation (VLO), and Marathon Petroleum Corporation (MPC) experience daily stock price fluctuations that are particularly sensitive to changes in crude oil prices, equity market performance, and perceived market risk. Despite the importance of short-term price behavior for operational and financial decision-making, accurately forecasting the daily direction of stock returns remains a formidable challenge. This difficulty stems from the complex, often nonlinear interactions between market variables and the rapid assimilation of new information, as posited by the Efficient Market Hypothesis (EMH).

This study addresses the need for a practical and interpretable framework to forecast next-day stock return direction in the refining sector. The research is structured in two phases. The first phase conducts a comparative analysis of the return distributions, volatility patterns, and cross-market sensitivities of PSX, VLO, and MPC over a 15-year period (2011–2026). These diagnostics inform the selection of VLO as the focal company for predictive modeling. This was selected because it has a dataset for the full 15 years in the study which allows for a larger training dataset with the engineered features calculated. The second phase develops a supervised machine learning framework to classify the binary direction (up or down) of VLO's next-day return. The modeling approach begins with a regularized logistic regression using market-based predictors and is extended to include engineered features that capture nonlinear interactions, volatility regimes, and relative performance metrics.

To further enhance the model's predictive power, the study incorporates forward-looking volatility forecasts derived from GARCH models and introduces a novel set of binary indicators based on OPEC press releases. These news-based features flag the presence and thematic content of official OPEC communications such as mentions of "oil," "petroleum," or "market stability" to account for the potential impact of fundamental news on market behavior. The predictive framework is evaluated using both statistical classification metrics and a simulated trading strategy to assess its economic value. By combining traditional financial modeling with volatility-aware and news-sensitive features, this study offers a transparent and actionable methodology for short-horizon forecasting in the energy sector.

Literature Review

Research on stock return prediction using machine learning has expanded significantly in recent years. This section reviews five studies that are closely aligned with the present project's focus: short-term stock movement prediction using machine learning models with financial and commodity-based features. Each study contributes valuable insights into modeling approaches, data selection, and performance outcomes, offering both parallels and contrasts to our methodology.

Machine Learning vs. Logistic Regression for Clean Energy Stock Direction (Sadorsky, 2021)

Sadorsky (2021) conducted a comprehensive analysis of clean energy stock price direction forecasting, comparing random forest models with traditional logistic regression. Using daily price data from clean energy exchange-traded funds and a suite of technical indicators (e.g., moving averages, oscillators), the study found that ensemble methods like random forests significantly outperformed logistic regression, achieving directional prediction accuracies between 85–90% over a 20-day horizon. Logistic regression, while less accurate (~55–60%), still exceeded random guessing, demonstrating its ability to extract predictive signals from technical features. This study aligns with our project's emphasis on directional classification and supports the inclusion of machine learning models alongside logistic regression. However, unlike Sadorsky's focus on technical indicators, our study incorporates fundamental and commodity-market variables such as oil prices and volatility indices, which are particularly relevant to oil refining firms. Additionally, while Sadorsky emphasized predictive accuracy, our evaluation also considers economic outcomes through trading strategy performance. This study establishes a useful benchmark for assessing the added value of non-linear models over logistic regression in energy sector forecasting.

Integrating Crude Oil Prices into Stock Return Prediction (Si, 2020)

Si (2020) explored the relationship between crude oil price fluctuations and stock market performance in the Indian context, focusing on the S&P BSE Oil & Gas index. The study employed 39 financial and accounting ratios from 17 oil and gas companies, along with crude oil price trends, to classify market performance as "GOOD" or "NOT GOOD." A hybrid modeling approach combining binary logistic regression and decision trees was used, with nine statistically significant financial ratios selected through normality and multicollinearity tests. The logistic regression model achieved approximately 75% classification accuracy, highlighting its effectiveness in capturing the influence of financial and commodity-based features. This approach resonates with our methodology, which similarly emphasizes domain-specific feature engineering such as crack spreads and volatility indicators for predicting the daily direction of VLO stock. While Si's study focused

on annual index-level performance, our project targets daily movements of an individual stock, integrating both market and macroeconomic variables. Nonetheless, the findings reinforce the value of incorporating oil price dynamics and sector-specific financial indicators into predictive models.

Macroeconomic Indicators and Bull/Bear Market Classification (Nafalana & Kartikasari, 2023)

Taking a macroeconomic perspective, Nafalana and Kartikasari (2023) examined the classification of stock market regimes—bullish versus bearish—on the Indonesia Stock Exchange (IDX) Composite index. Using the Bry–Boschan algorithm to identify market phases and monthly data from 2003 to 2022, they selected four macroeconomic indicators—inflation, interest rate, exchange rate, and money supply—as predictors in a binary logistic regression model. The model achieved an out-of-sample accuracy of 81.5%, demonstrating the predictive power of macro-financial variables. Although our study operates at a finer temporal resolution (daily) and focuses on a single stock (VLO), this research supports the inclusion of broader economic indicators such as crude oil and gold prices, as well as market indices, in our feature set. The study also underscores the importance of clearly defining prediction targets and labeling data appropriately, principles we apply by constructing a binary next-day return direction variable. While the time scale differs, the high accuracy achieved in this study confirms that logistic regression can effectively model complex financial dynamics when informed by relevant macroeconomic features.

Stock Price Prediction with Logistic vs. Decision Tree in Emerging Markets (Gavirineni et al., 2024)

Gavirineni, Selvakumar, and Sivakumar (2024) investigated stock price prediction in the context of the Bombay Stock Exchange (BSE), comparing logistic regression and decision tree algorithms. Although the abstract provides limited methodological detail, the authors describe analyzing trading patterns and constructing binary classification models using both logistic regression and CART decision trees. The results indicated satisfactory performance from both models, suggesting that hybrid or comparative modeling approaches can yield valuable insights. This study reinforces the continued relevance of logistic regression in stock prediction tasks, even when compared to non-linear classifiers. It also highlights the importance of pattern recognition and the need for rapid, automated predictions in volatile markets—an objective shared by our project. While their focus was on broader market patterns in an emerging economy, our study narrows in on a single U.S.-based refining stock, incorporating sector-specific exogenous variables such as oil prices and volatility indices. The findings support our approach of evaluating logistic regression

alongside more complex models, even as we prioritize interpretability and domain relevance.

Sentiment-Enhanced Stock Direction Prediction with Advanced NLP vs. Logistic Models (Shobayo et al., 2024)

Shobayo et al. (2024) bridged natural language processing (NLP) and financial forecasting by comparing FinBERT, GPT-4, and logistic regression in predicting the Nigerian Stock Exchange All-Share Index. Using news sentiment as input, the study found that a well-tuned logistic regression model outperformed both FinBERT and GPT-4, achieving 81.8% accuracy and a ROC AUC of ~0.90. Despite the sophistication of the NLP models, logistic regression proved more effective, likely due to efficient feature engineering and lower computational complexity. This finding is particularly relevant to our study, which incorporates structured news-based features specifically, binary indicators derived from OPEC press releasesto capture the impact of fundamental events on stock movements. While we do not employ full-text sentiment analysis, the study validates our strategy of using low-dimensional, interpretable news features to enhance model responsiveness. The broader implication is that logistic regression remains a strong benchmark, even in the face of advanced AI models, when paired with thoughtfully engineered inputs.

Summary of Insights from Literature

1. Domain-Specific Features Matter: Incorporating relevant external variables—such as technical indicators (Sadorsky, 2021), commodity prices (Si, 2020), macroeconomic indicators (Nafalana & Kartikasari, 2023), and news sentiment (Shobayo et al., 2024)—consistently improves predictive performance. For an oil-refining firm like VLO, this supports our inclusion of oil prices, market indices, and volatility metrics.

2. Logistic Regression as a Competitive Baseline: Across diverse contexts, logistic regression has demonstrated strong performance, particularly when paired with effective feature engineering. Studies by Gavirineni et al. (2024) and Shobayo et al. (2024) reinforce its value as a benchmark model.

3. Classification over Regression: All reviewed studies framed their objectives as classification tasks (e.g., up/down or bull/bear), using metrics like accuracy and AUC rather than price-level errors. Our study follows this paradigm, while also evaluating economic outcomes such as trading strategy returns.

4. Contextual Adaptability: Despite differences in geography, asset class, and time scale, the consistent success of logistic regression suggests its generalizability when features are carefully selected. While more complex models may offer incremental gains, logistic regression often delivers competitive results with greater interpretability.

These insights provide a strong foundation for our study, which builds on this literature by applying a feature-rich, interpretable modeling framework to the directional prediction of VLO stock returns. By integrating engineered features from financial, commodity, volatility, and news domains, we aim to test and extend the findings of prior research in the context of a single energy-sector equity.

References :

Sadorsky, P. (2021). A random forests approach to predicting clean energy stock prices. Journal of Risk and Financial Management, 14(2), Article 48. DOI: 10.3390/jrfm14020048.

https://www.mdpi.com/1911-8074/14/2/48

Si, R. K. (2020). Relationship between crude oil price and S&P BSE stock index – an integration of binary decision tree and logistic regression approach. Elixir International Journal (Statistics), 141, 54271-54279.

https://www.elixirpublishers.com/articles/1672815351_202004004.pdf

Nafalana, R. D., & Kartikasari, M. D. (2023). Predicting stock markets using binary logistic regression based on Bry–Boschan algorithm. Jurnal Varian, 6(2), 127-136. DOI: 10.30812/varian.v6i2.2385.

https://journal.universitasbumigora.ac.id/Varian/article/view/2385

Gavirineni, S., Selvakumar, I., & Sivakumar, T. K. (2024). Stock price prediction using logistic regression and decision tree. AIP Conference Proceedings, 3075(1), 020225. DOI: 10.1063/5.0217082.

https://pubs.aip.org/aip/acp/article-abstract/3075/1/020225/3305158/Stock-price-prediction-using-logistic-regression?redirectedFrom=PDF

Shobayo, O., Adeyemi-Longe, S., Popoola, O., & Ogunleye, B. (2024). Innovative sentiment analysis and prediction of stock price using FinBERT, GPT-4 and logistic regression: A data-

driven approach. Big Data and Cognitive Computing, 8(11), 143. DOI: 10.3390/bdcc8110143.

https://www.mdpi.com/2504-2289/8/11/143

Theoretical Framework

This study is grounded in several interrelated theories from financial economics and time-series modeling that collectively support the plausibility of short-horizon predictability in equity returns, particularly within the petroleum refining sector. The framework integrates concepts from the Efficient Market Hypothesis (EMH), volatility modeling, cross-market transmission theory, and predictive modeling paradigms.

Market Efficiency and Short-Term Predictability

The EMH posits that asset prices fully reflect all available information, rendering consistent outperformance of the market through prediction infeasible (Fama, 1970). However, empirical evidence has demonstrated that financial markets, while broadly efficient, exhibit short-term anomalies such as autocorrelation, momentum, and mean reversion (Lo & MacKinlay, 1988; Jegadeesh & Titman, 1993). These deviations are particularly pronounced in sectors exposed to rapid information flow and exogenous shocks—characteristics inherent to the petroleum refining industry. Firms such as Valero Energy Corporation (VLO) operate in a domain where equity prices are highly sensitive to fluctuations in crude oil prices, macroeconomic indicators, and investor sentiment. These dynamics create conditions under which short-term return direction may be partially predictable, even if long-term price levels remain stochastic.

Volatility Clustering and GARCH Theory

A central feature of financial time series is volatility clustering—the tendency for large price changes to be followed by further large changes, and small changes by small ones (Mandelbrot, 1963; Engle, 1982). Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models formalize this behavior by modeling conditional variance as a function of past squared returns and past variances. Although this study does not estimate full GARCH models for return prediction, it leverages GARCH-based volatility forecasts as exogenous features. These forecasts serve as forward-looking indicators of market uncertainty, enabling the predictive model to adjust its expectations based on anticipated volatility regimes. This approach aligns with the theoretical expectation that volatility itself contains information about future return distributions and investor behavior (Bollerslev, 1986).

Cross-Market Spillovers and Transmission Effects

The petroleum refining sector operates at the intersection of multiple markets, including equities, commodities, and macroeconomic indicators. The theory of cross-market spillovers suggests that shocks in one market—such as a sudden change in crude oil prices or a shift in monetary policy—can propagate to others, influencing asset prices through correlated risk factors and investor sentiment (King & Wadhwani, 1990; Forbes & Rigobon, 2002). In this context, the inclusion of features such as S&P 500 returns, WTI crude oil prices, gold prices, and the VIX index is theoretically justified. These variables capture systematic risk, input cost dynamics, and market-wide uncertainty, all of which are known to influence the valuation of refining firms. Moreover, the use of interaction terms and regime indicators (e.g., "Equity Down & Oil Down" flags, high-volatility dummies) reflects the non-linear and state-dependent nature of these transmission mechanisms.

Predictive Modeling and Feature Engineering

From a methodological standpoint, the study draws on the theory of supervised learning, wherein models are trained to map input features to a target variable—in this case, the binary direction of next-day returns. Logistic regression serves as the primary modeling framework due to its interpretability and established performance in financial classification tasks (Ohlson, 1980; Lo, Mamaysky, & Wang, 2000). However, recognizing the potential for non-linear relationships and complex interactions, the study also evaluates tree-based and neural network models. The inclusion of engineered features—such as lagged returns, momentum indicators, excess returns, and news-based binary flags derived from OPEC press releases—is grounded in the principle that domain knowledge can enhance model performance by capturing economically meaningful patterns (Gu et al., 2020). These features are constructed to reflect both technical and fundamental drivers of stock price movements, enabling the models to learn from a richer representation of market dynamics.

Summary

In sum, the theoretical framework supports the hypothesis that short-term directional predictability in refining-sector equities is feasible when models are informed by a comprehensive set of market, commodity, volatility, and news-based features. While the EMH cautions against naïve forecasting, the presence of volatility clustering, cross-market dependencies, and behavioral biases provides a rationale for exploring predictive modeling approaches. By integrating insights from financial theory and machine learning, this study aims to contribute to the growing literature on interpretable, data-driven forecasting in high-noise financial environments.

Data

This study utilizes daily financial market data over approximately a 15-year period (early 2011 through early 2026), obtained from Yahoo Finance. The data covers seven key tickers that capture the performance of major U.S. refining companies and relevant market indicators. These include three refining sector equities Marathon Petroleum Corporation (MPC), Phillips 66 (PSX), and Valero Energy Corporation (VLO), along with broad market and commodity indicators: the S&P 500 Index (^GSPC) as a proxy for overall U.S. equity market conditions; WTI Crude Oil futures (CL=F) and Gold futures (GC=F) to represent key commodity price influences on refining margins and risk sentiment, respectively; and the CBOE Volatility Index (^VIX) to reflect market-implied volatility (often dubbed the "fear index"). Each of these tickers plays a distinct role in capturing the fundamental and market forces that may drive daily movements in refiner stock prices, ensuring a comprehensive set of input features for analysis.

All price series were sourced at a daily frequency using adjusted closing prices, which properly account for corporate actions like dividends and stock splits. For instruments that do not have adjustments, the "Adjusted Close" was set equal to the regular close price to maintain consistency across series. This includes the commodity futures, VIX and S&P 500 index. The data was pulled from ~02/2011through ~02/2026, providing roughly 15 years of history for each series.

(Note: PSX and MPC have shorter trading histories, as they were spun off in 2012 and 2011 respectively, so their data begins on their inception dates and covers slightly less than 15 years. All other series—VLO, ^GSPC, CL=F, GC=F, ^VIX—span the full period, ensuring the overall analysis window extends to early 2026.)

For each trading day in the dataset, we computed daily log returns for all price series. This was also done for the VIX even though it is not a tradable asset or security like the others. This calculation gives a machine learning variable for the VIX price changes. Using log returns standardizes percentage changes across assets and stabilizes variance, making the features more comparable across different price levels. We then aligned and merged the data into a single time-series dataset indexed by trading date, including all chosen tickers. Because not all tickers trade on exactly the same days (particularly around market holidays), we merged on the intersection of trading days, thereby excluding dates where any series was missing. This effectively means the modeling dataset starts in mid-2012, once all three refinery stocks were actively trading. The merged dataset contains a rich set of variables for each trading day t, all of which are known by the end of day t. These include each asset's log return on day t, as well as additional features described below.

The prediction target for our models is the direction of VLO's next-day return. Specifically, for each day t, we define yt = 1 if VLO's return on day t+1 was positive, and yt = 0 if the next day's return was zero or negative. This binary variable (Up vs. Not Up) is constructed by shifting VLO's own daily return series by one day and thresholding it at zero, a common approach in directional return forecasting. We also calculated and stored VLO's actual next-day return (e.g. as vlo_next_ret) and a categorical label (vlo_next_dir indicating "positive" or "negative" next-day movement) for analysis purposes, but these were excluded from the feature set to prevent any lookahead bias or target leakage. The final dataset omits the last trading day (which lacks a t+1 outcome) and any initial days required for lagged calculations, resulting in a complete and clean matrix of features and targets ready for model training.

By construction, this dataset ensures that on each day's record, all feature values are information that would have been available on that day, and the target refers to the following day's outcome. This careful alignment and target definition uphold a strict chronological order in our analysis, eliminating forward-looking biases. In total, the data preparation yields a rich daily dataset for VLO spanning from 2012 through the start of 2026, with several thousand observations of aligned market features and an associated next-day return direction for VLO.

To incorporate the influence of fundamental news events into the predictive framework, the dataset was augmented with a set of binary indicators derived from OPEC press releases. These features were constructed by collecting official OPEC communications (including press release dates, titles, and summaries) and aligning them with the nearest subsequent trading day to simulate the market's first opportunity to react. Four binary variables were created: (1) a general flag indicating the presence of any OPEC press release on a given day, and three content-specific flags capturing whether the release mentioned the terms "oil," "petroleum," or included a "market summary." These indicators were designed to capture the potential impact of oil market news on investor sentiment and refining sector equities, particularly Valero (VLO), by flagging days when new information from OPEC could influence price behavior.

The Organization of the Petroleum Exporting Countries (OPEC) is an intergovernmental organization founded in 1960 to coordinate and unify petroleum policies among its member countries, primarily oil-producing nations. OPEC plays a significant role in the global oil market by collectively managing oil production levels to influence oil prices and ensure market stability. By adjusting output quotas, OPEC can tighten or loosen global oil supply, which directly impacts crude oil prices. For example, production cuts typically lead to higher prices by reducing supply, while increased output can lower prices. As a result,

OPEC's decisions, often communicated through official press releases, are closely monitored by market participants, policymakers, and analysts for their potential to shift market dynamics and investor sentiment.

Methodology

The analysis is structured in two main phases. Phase 1 consists of an exploratory analysis of historical data for the three refining companies (PSX, VLO, MPC) to characterize their return distributions, volatility patterns, and correlations with external market variables. This phase helps identify which stock is most amenable to prediction and which features are potentially informative. Phase 2 then builds a predictive modeling framework focused on the chosen stock (VLO), aiming to forecast its next-day stock movement (up or down) using a suite of machine learning classifiers. Again, VLO was selected because its stock return data covers the full 15 years of the study and this allowed for a larger training and test data set. Initially, we develop a baseline logistic regression model using a core set of market-based predictors (returns and price changes of equities, commodities, and indices). We then expand this model by incorporating additional variables, specifically GARCH-based volatility forecasts and text-based news indicators, to test if these augment the model's performance. Finally, we also train non-linear models (a tree-based ensemble and a neural network) on the same dataset to evaluate whether they can offer improvements over logistic regression. All models are developed and evaluated under identical data splits and performance metrics for a fair comparison, and their results are reported in parallel in the Results section.

Feature Construction and Engineering-

The base feature set for the predictive models consists of standard financial variables that capture recent market movements and conditions, all measured up to and including day t (to predict day t+1). Key features include:

Daily log returns of each asset on day t – for VLO, PSX, MPC, ^GSPC, CL=F, GC=F, and ^VIX. These are the primary predictors, reflecting the day's performance across equities, commodities, and volatility markets.

Lagged returns (1-day lag) for each asset (often notated as ret t-1 or ret2_* in the dataset) – representing momentum or mean-reversion effects from the previous trading day.

Price level indicators and technicals: For selected series, the raw price or level (e.g., the level of VIX or the S&P 500 on day t) and technical measures such as High–Low trading range and Open–Close change for the day. These capture intraday volatility and market

sentiment (for instance, a wide high–low range may indicate market uncertainty or news impact on that day).

Trading volume for the equity tickers (including VLO's own volume, if available), as a proxy for liquidity and investor attention, which can precede price moves.

All features are carefully constructed to avoid any lookahead bias. For example, returns are based on same-day price changes, and any indicator that inherently looks forward (like VLO's next-day return or direction) is excluded from the feature set during training. After creating these base features, we perform an inner join on the date index to ensure a unified dataset: each row represents one trading day where all feature values are present. This means that the early days of VLO's history before PSX's existence (pre-2012) are omitted in the model training dataset. The result is a time-aligned feature matrix X and target vector y with no missing values, ready for modeling.

Engineered Features for Nonlinear Signals:

To enhance the predictive power of the machine learning model, a comprehensive set of engineered features was developed. These features were designed to capture market dynamics, asset-specific behavior, and macroeconomic signals that may influence the short-term direction of Valero Energy (VLO) stock returns. The feature engineering process was applied after assembling the base dataset and was structured to ensure all features were derived using only information available up to each prediction date, thereby avoiding lookahead bias.

The engineered features fall into several key categories:

Cross-Market Interaction Flags

Binary indicators were created to capture joint movements across major asset classes. For example, the EquityDown_OilDown flag identifies days when both the S&P 500 and WTI crude oil declined, signaling potential broad-based risk-off sentiment. Similarly, the RiskOff indicator flags days when gold prices rose, equities fell, and market volatility (VIX) increased simultaneously—conditions often associated with heightened investor caution.

Volatility Regime Indicators

To account for changing market conditions, features such as HighVIX were introduced. This flag identifies periods when the VIX index is above its one-year moving average, indicating elevated market uncertainty. Additionally, rolling realized volatility measures for VLO (over 5, 10, and 20-day windows) were computed to quantify short-term risk. The ratio of short-term to long-term volatility was also included to detect volatility spikes.

Commodity Shock Flags

Features like OilShock and GoldShock were designed to detect unusually large price movements in WTI crude oil and gold, respectively. These binary indicators flag days when the absolute return exceeds two standard deviations of the recent 60-day return distribution, capturing potential market disruptions or significant news events.

Relative and Excess Return Metrics

To assess VLO's performance in context, features were created to measure its return relative to benchmarks. These include VLO's excess return over the S&P 500 (VLO_minus_SPX) and over a peer average of other refiners (VLO_minus_Peers). These metrics help the model identify whether VLO is outperforming or underperforming its sector or the broader market.

Refining Margin Proxies

Features such as Oil_minus_Refiners and Oil_minus_SPX serve as proxies for refining margins by comparing oil price movements to those of refiners and the broader market. These indicators aim to capture the economic environment for downstream energy companies like VLO.

Momentum and Mean Reversion Signals

Short-term momentum indicators were constructed by summing VLO's returns over the past 3 and 5 days. A mean reversion signal was also included by computing the negative of the 5-day average return, based on the hypothesis that extreme short-term gains or losses may reverse in the near term.

Standardized Return Z-Scores

To normalize recent return behavior, rolling z-scores were calculated for VLO, the S&P 500, and oil. These features measure how unusual a given day's return is relative to its recent 20-day history, helping the model detect outlier behavior or regime shifts.

All engineered features were constructed using robust, transparent logic and were carefully validated to ensure correctness. Missing values introduced by rolling calculations were removed to maintain data integrity. The inclusion of these features significantly enriched the dataset, allowing the model to capture complex interactions between market signals, asset-specific behavior, and macroeconomic conditions. This comprehensive feature set was critical to improving the model's ability to forecast next-day return direction and to support the development of effective trading strategies.

Incorporation of GARCH-Based Volatility Features:

After testing the base features and model performance, models were then setup with GARCH based features for comparison. This was done to see if volatility forecasting improves the machine learning modeling performance

In order to explicitly include forward-looking volatility information, we integrate features from a GARCH (Generalized Autoregressive Conditional Heteroskedasticity) model into our predictors. Financial returns often exhibit volatility clustering (periods of high volatility followed by high volatility, etc.), and GARCH models are a standard econometric tool to capture time-varying volatility. Here, we use a univariate GARCH(1,1) model on VLO's own returns to generate one-day-ahead volatility forecasts for VLO. The GARCH model is specified such that VLO's daily return rt (in percentage form for stability) is represented with a conditional variance that evolves as:

$$\sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2,$$

with epsilon_{t} being the return shock on day t (residual) and sigma_{t}^2 its conditional variance. The model assumes epsilon_{t} follows a heavy-tailed distribution (e.g., Student-t) to account for extreme moves. The GARCH model is re-estimated regularly using a rolling window of the past ~1 year (≈ 252 trading days), updating its parameters (ω, α, β) periodically (e.g., every week) to produce an out-of-sample forecast of volatility. This rolling estimation ensures that at each day t, the GARCH model's next-day volatility forecast uses only information up to time t, preserving the integrity of out-of-sample prediction.

From the GARCH model's output, we derive four features that quantify the predicted volatility environment for VLO on day t+1, and include them in our machine learning dataset (with indexing such that these features are aligned with day t so that they can be used to predict the day t+1 outcome):

garch_var_1d: The one-day-ahead forecast of VLO's return variance (σt+12) from the GARCH(1,1) model. This represents the model's expectation of return variability for the next trading day, in variance terms.

garch_vol_1d: The one-day-ahead forecast of volatility (σt+1) i.e. the square root of the forecasted variance. This is easier to interpret (in percent daily move terms) and can capture the anticipated risk on the next day.

garch_vol_z: A volatility regime Z-score for the next day's forecasted volatility. We compute this by taking the forecasted σt+1 and standardizing it based on its trailing 252-trading-day mean and standard deviation (effectively comparing tomorrow's expected volatility to the past year's average). A high positive value of garch_vol_z means the GARCH model expects an unusually high volatility day relative to recent history (potentially a turbulent market regime), whereas a negative value indicates an unusually calm expected day.

garch_vol_chg: The day-over-day change in the forecasted volatility, calculated as the difference between the forecasted σt+1 and the forecast made for σt (the prior day's forecast for today's volatility). This measures whether volatility is expected to rise or fall from one day to the next, which might influence the likelihood of an up or down move.

By including these GARCH-based features, we provide the model with a forward-looking measure of market risk. This is information that other predictors (like recent returns or even the VIX) only partly capture, since those are backward-looking or contemporaneous. The GARCH features allow the model to condition its predictions on whether the upcoming day is expected to be in a high-volatility state (which could portend sharper movements or trend reversals) or a low-volatility state (where mean reversion might prevail). In essence, the model can learn relationships such as "if predicted volatility is very high, the likelihood of a large downward move might increase," thereby potentially improving classification performance. Importantly, these features are generated in a way that maintains a strictly out-of-sample regime—the GARCH model's predictive parameters are never learned on data that include the day being predicted.

The goal is to see if the best models show improved performance by incorporating the volatility features.

Incorporation of News-Based Features:

To account for important exogenous information shocks, we incorporate a set of binary news indicator features related to OPEC's press releases. Major announcements from OPEC (the Organization of the Petroleum Exporting Countries) can influence global oil prices and market sentiment, which in turn could affect VLO's stock. Instead of performing complex textual analysis or sentiment modeling, we opted for a simple but effective approach: using indicators for the presence of certain types of OPEC news on or just before a given trading day. We obtained a chronologically ordered list of OPEC press releases, including their dates, titles, and summary texts, and aligned them with our trading dates via a nearest-date mapping (assigning each press release to the closest trading day on or after its release, to simulate the market's first opportunity to react). This way, if OPEC released a

statement on a weekend or holiday, it would be associated with the next trading day, ensuring we do not "look ahead" to information that wasn't available to market participants.

From the press release corpus, we designed four daily binary features that capture the occurrence of potentially impactful news themes:

PR_has_release: Indicates (1/0) whether at least one OPEC press release was issued on or immediately before day t (mapped to day t). This simply flags the arrival of new information from OPEC, regardless of content.

PR_has_oil: Indicates (1/0) whether the word "oil" appears in the title or summary of an OPEC release on day t. This flag suggests the press release contained content explicitly about oil markets (e.g., production quotas, oil price outlooks).

PR_has_petroleum: Indicates (1/0) whether the word "petroleum" appears in the title or summary on day t. This often signifies more formal or technical communications using the term "petroleum," potentially pointing to discussions of industry-specific topics or detailed market conditions.

PR_has_market_stability: Indicates (1/0) whether the phrase "market stability" appears in the press release title/summary on day t. OPEC frequently publishes "market stability" updates in its press releases; presence of this phrase may signal a broader overview of market conditions or outlook, which could carry information content affecting investor sentiment.

These four features provide the model with awareness of days that are likely to be news-sensitive. They effectively allow the classifier to adjust predictions on days when OPEC's communications (and by extension, possibly other macro-news about oil markets) are influencing trader expectations. The binary design of the features keeps them simple and interpretable, avoiding overfitting that might arise from high-dimensional textual data in a relatively limited sample of press releases. We expect these features to improve model performance by flagging "event" days and potentially capturing the impact of fundamental news that isn't reflected fully in the price-based technical features.

After constructing all engineered features (technical flags, GARCH outputs, and news indicators), we merge them into our main dataset keyed by date. Any day without a press release simply has 0 values for those news features, and days with releases have the corresponding indicators set to 1. Similarly, GARCH features are merged for all dates from the start of the rolling estimation window onward; they are NaN (and then dropped) for the initial period where a 1-year history wasn't available. Through careful merging and dropping of any residual missing values, we obtain a finalized feature matrix that includes all the

above-described variables for each day, alongside the target. The resulting dataset (after excluding the final day of the sample for which no next-day target is available) is then partitioned for model training and evaluation as described below.

Model Training and Selection:

For the primary prediction model, we selected a logistic regression classifier due to its transparency, interpretability, and established performance in financial classification tasks. The model was trained using scikit-learn with both L1 (Lasso) and L2 (Ridge) regularization to mitigate overfitting and facilitate feature selection. A comprehensive hyperparameter grid search was conducted over a range of regularization strengths and penalty types, with time-series-aware cross-validation employed to preserve temporal integrity. The optimal configuration was selected based on the highest average Area Under the Receiver Operating Characteristic Curve (ROC-AUC), a metric chosen for its robustness to class imbalance and its ability to evaluate ranking performance.

To evaluate whether targeted methodological refinements could improve model performance, several modifications were applied to the baseline logistic regression framework. These included filtering highly collinear predictors by removing variables with pairwise correlations exceeding 0.85, replacing randomized cross-validation with time-series-aware cross-validation to preserve temporal ordering, and optimizing model selection using balanced accuracy in addition to ROC–AUC. Balanced accuracy was used to weight up and down return predictions symmetrically by averaging class-specific recall, reducing the influence of class imbalance. Each modification was evaluated independently to assess its effect on classification behavior while maintaining a consistent modeling and evaluation framework.

To further benchmark the logistic regression model, we implemented several additional classifiers using the same feature set and training data. These included:

- A decision tree classifier, which captures hierarchical decision rules and non-linear interactions. Hyperparameters such as maximum depth and minimum samples per leaf were tuned using randomized grid search with cross-validation.
- A gradient boosted decision tree model, which builds an ensemble of weak learners in a sequential manner to minimize prediction error. Regularization parameters, learning rate, and tree complexity were optimized through randomized search tailored to the boosting framework.
- A feed-forward neural network (multi-layer perceptron) with different levels hidden layers tested. Input features were standardized, and hyperparameters including the

number of hidden units and regularization strength were tuned via randomized grid search.

- Two convolutional neural network (CNN) architectures designed to capture temporal and cross-feature dependencies:
    - The first CNN model applied convolutional filters over a fixed-width window of features for each individual day, enabling the model to learn spatial patterns across engineered predictors.
    - The second CNN model extended this approach by incorporating a 10-day look-back horizon, allowing the network to extract temporal dynamics and short-term trends across multiple days. Both CNN models were trained with architecture-specific regularization strategies and hyperparameter tuning via randomized grid search.
- In addition to fully supervised models, a zero-shot deep learning baseline was included using the pretrained IBM Granite TinyTimeMixer (TTM) time-series foundation model. This approach did not train on VLO labels directly; instead, it generated forecasts of next-day returns using a fixed context window of 512 observations and converted the sign of the forecast into a directional prediction.

All models were trained using consistent train/test splits to ensure fair comparison and evaluated on the same out-of-sample test set. Performance metrics, including ROC-AUC and directional accuracy, were used to assess each model's predictive capability. This multi-model evaluation framework enabled a rigorous assessment of whether more complex, non-linear models could offer meaningful improvements over the baseline logistic regression, or whether the simpler model remained sufficient given the structure and quality of the engineered features.

Train/Test Splitting and Temporal Validation:

In developing and testing the models, we respected the time-series nature of the data to emulate real-world forecasting. We partitioned the data into a training period spanning from the start of the dataset (mid-2012) through December 31, 2022, and reserved January 1, 2023 onward as the out-of-sample test set (approximately three years of data for testing). All model fitting, feature engineering (such as computing rolling statistics or GARCH forecasts), and hyperparameter tuning were performed using only the training data (~2012–2022). The test set was completely held out until the final evaluation, thereby simulating how the models would perform on "future" unseen data in an actual live forecasting scenario. This chronological split ensures that our performance metrics reflect true out-of-sample predictive power and not merely in-sample or retrospective fit.

Before model training, we applied feature scaling to standardized continuous features. Using the training set, we computed mean and standard deviation for each continuous feature (e.g., returns, volatility measures, volumes) and normalized these features to have zero mean and unit variance. This scaling was then applied to the test set using the training-set parameters, so no information from the test period leaked into the training process. Categorical features (the binary flags like our news and regime indicators) were left in their 0/1 form. We also carefully removed any potential leakages: for instance, the columns containing VLO's actual next-day return or direction (which we had added for analysis convenience) were dropped from the feature matrix before training. The dataset was then ready for modeling, with feature vectors X and outcome y.

Model Evaluation and Performance Metrics:

We evaluate model performance using a variety of metrics to capture both classification accuracy and financial significance of the predictions. The primary metrics and evaluation techniques include-

Directional Accuracy (Hit Rate): The percentage of days in the test set for which the model correctly predicted the up/down direction of VLO's return. This is compared against the null hypothesis of 50% (random guessing) to see if the model adds value in predicting direction.

Confusion Matrix & Derived Stats: We analyze the confusion matrix (True Ups, True Downs, False Ups, False Downs), computing metrics such as Precision (for "Up" predictions), Recall/Sensitivity (true positive rate for up-days), and the F1-score for up-day predictions. Given the application, we pay particular attention to False Positives vs False Negatives: a False Positive means the model predicted an up-day that turned out down, which could lead to a losing trade, while a False Negative means missing a profitable opportunity. To contextualize these, we calculate the average actual VLO return for each cell of the confusion matrix (e.g., mean return on days the model predicted up and it was actually up, etc.). This provides economic insight into what different types of errors cost in terms of missed gains or incurred losses.

ROC Curve and AUC: We consider the continuous output of the models (predicted probability from logistic and NN, vote proportion from Random Forest) and plot the Receiver Operating Characteristic (ROC) curve for the test set, summarizing it via the Area Under the Curve (AUC) metric. The ROC/AUC reflects the model's ability to rank positive vs. negative days independent of a specific threshold. An AUC significantly above 0.5 indicates genuine predictive skill.

Feature Importance and Coefficients: For the logistic regression, we examine the learned coefficients (after scaling) to identify which variables had the strongest influence on the prediction. For the Random Forest, we look at the usual feature importance scores. This allows us to interpret the models and validate that the important predictors make sense (e.g., we might expect that VIX changes, oil returns, or the GARCH volatility forecast have significant impact on the next-day direction).

All evaluation outputs (confusion matrices, ROC and precision-recall curves, classification reports, and coefficient plots) are compiled into a multi-page PDF report for documentation and review. This ensures transparency and reproducibility in model assessment and provides a foundation for comparing alternative models under consistent evaluation criteria.

Investment Strategy Backtesting:

Beyond statistical metrics, we evaluate the economic value of the model's predictions through a simple backtesting exercise. We design a set of rule-based trading strategies that act on the model's daily forecasts for VLO's return direction, and we simulate these strategies on the test period (2023–2026):

Daily Long/Short Strategy: Each day, take a long position in VLO at the market close if the model predicts an "Up" day for tomorrow, or take a short position if the model predicts a "Down" (or no gain) day. Close the position at the next market close, then repeat. This strategy profits from correct directional calls in either direction.

Conservative Long-Only Strategy: Take a long position in VLO at the close only if an up-day is predicted; if a down-day is predicted, move to cash (i.e., take no position) for the next day. This strategy avoids shorting, which may be preferable for certain investors and isolates the value of predicting upward movements.

Aggressive Leveraged Strategy: Similar to the long-only strategy, but when an up-day is predicted, take a leveraged long position (e.g., using a 3× leverage factor) in order to potentially amplify returns. If no up-day is predicted, stay in cash. This tests if the model's confidence in upward moves is high enough to justify leverage.

Compare the performance of these model-driven strategies against two benchmarks:

A passive Buy-and-Hold of VLO and the S&P 500 index over the test period (which gives the baseline return of simply holding the stock).

Each strategy's cumulative returns are calculated over the test period, and we compute annualized performance metrics such as total return and average return per year. To make the back test realistic, we incorporate trading frictions and costs: we assume transaction costs of 0.1% per round-trip trade (approximately 0.05% each buy or sell), a short borrowing cost of about 0.3% per year for short positions, and for the leveraged strategy, a financing cost of ~15% per annum on borrowed funds. We also consider the impact of taxes in a simplified way (e.g., higher short-term tax rates on profits from positions closed within a year). These assumptions are intentionally conservative to avoid overstating the strategy performance.

The back test assumes a starting $100,000 is invested in the strategies at the beginning of 2023. The total capital results over time are plotted through early 2026 using each of the investment strategies. This visually shows how each investment performed and if the model improved on the buy and hold strategy.

By examining the risk-adjusted returns of the model-driven strategies versus the benchmarks, we can evaluate whether the predictive signals translate into practical investment value after accounting for real-world constraints. A model with modest predictive ability might still generate appreciable economic gains (especially the long-short strategy, if it avoids down days), but high transaction costs or other frictions could erode these gains. The backtest thus serves as a critical validation of the model's usefulness. if the best model cannot outperform a buy-and-hold or random strategy after costs, its real-world utility is questionable.

Model Validation and Robustness:

To ensure the reliability of our findings, we took several steps to validate the models:

All data preprocessing (lag calculations, feature scaling, etc.) was derived from the training set alone, and then applied to the test set, to avoid any information leaking from test to train.

We used cross-validation within the training period for model tuning, and we also tested the stability of model coefficients (for logistic regression) across different folds to see if certain predictors consistently had influence.

We saved the final model parameters and scaler, enabling out-of-time validation on new data as it becomes available, and ensuring the study could be replicated or updated in the future.

Through these methods, our aim was to present a fully documented and rigorous Data and Methodology section. The steps above comprehensively describe how we collected and preprocessed data, what features (including novel volatility and news-based indicators) were engineered, how we trained and tuned multiple predictive models, and how we planned to evaluate their performance both statistically and in practical trading terms. This robust setup sets the stage for the Results section that follows, where we will compare the logistic regression, tree-based, and neural network models and discuss their predictive performance using the outlined metrics and backtesting results.

This study is a thorough test if the random nature of daily returns can be predicted using modern machine learning tools and volatility prediction findings from Taylor(2005). The EMH implies that returns are a random walk and cannot be predicted. Since this is directly going against significant economic theory and history, it is a difficult task. Even Taylor(2005) did not find significant evidence of being able to predict daily returns and then focused on the volatility clustering and prediction. Any positive performance improvement over random guessing would be a significant finding for our study.

Results

This Results section is organized into two parts. The first part presents empirical market findings for the assets included in the study. The second part reports the machine learning modeling results for predicting the next-day return direction of Valero Energy Corporation (VLO).

Part 1 – Empirical Asset Market Findings

This section presents the empirical financial market results for the assets examined in the study. The analysis covers crude oil (CL=F), gold (GC=F), refining equities—Marathon Petroleum Corporation (MPC), Phillips 66 (PSX), and Valero Energy Corporation (VLO)—the S&P 500 index (^GSPC), and the CBOE Volatility Index (^VIX). Results are evaluated across multiple dimensions, including price dynamics, return distributions, serial dependence, volatility behavior, seasonality, and cross-asset dependence.

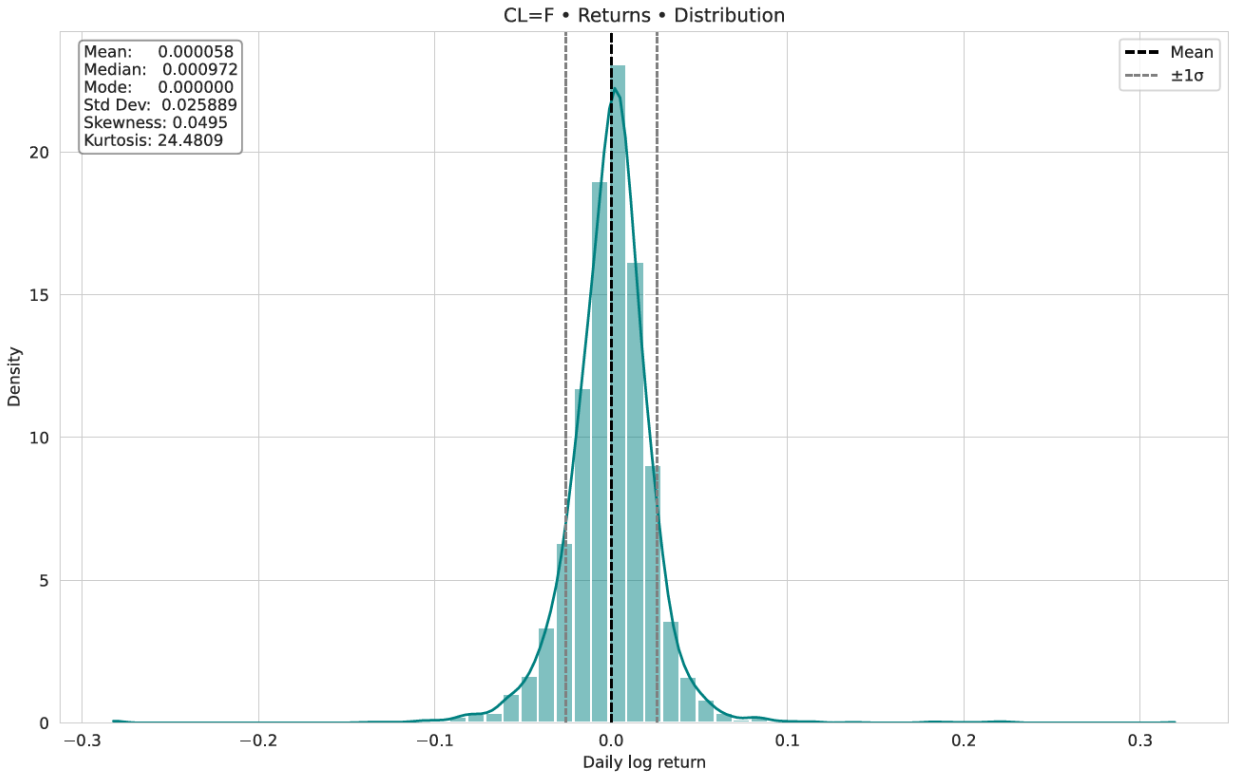
Across all assets analyzed, the results are consistent with the well-documented stylized facts of financial markets described in Taylor (2005):

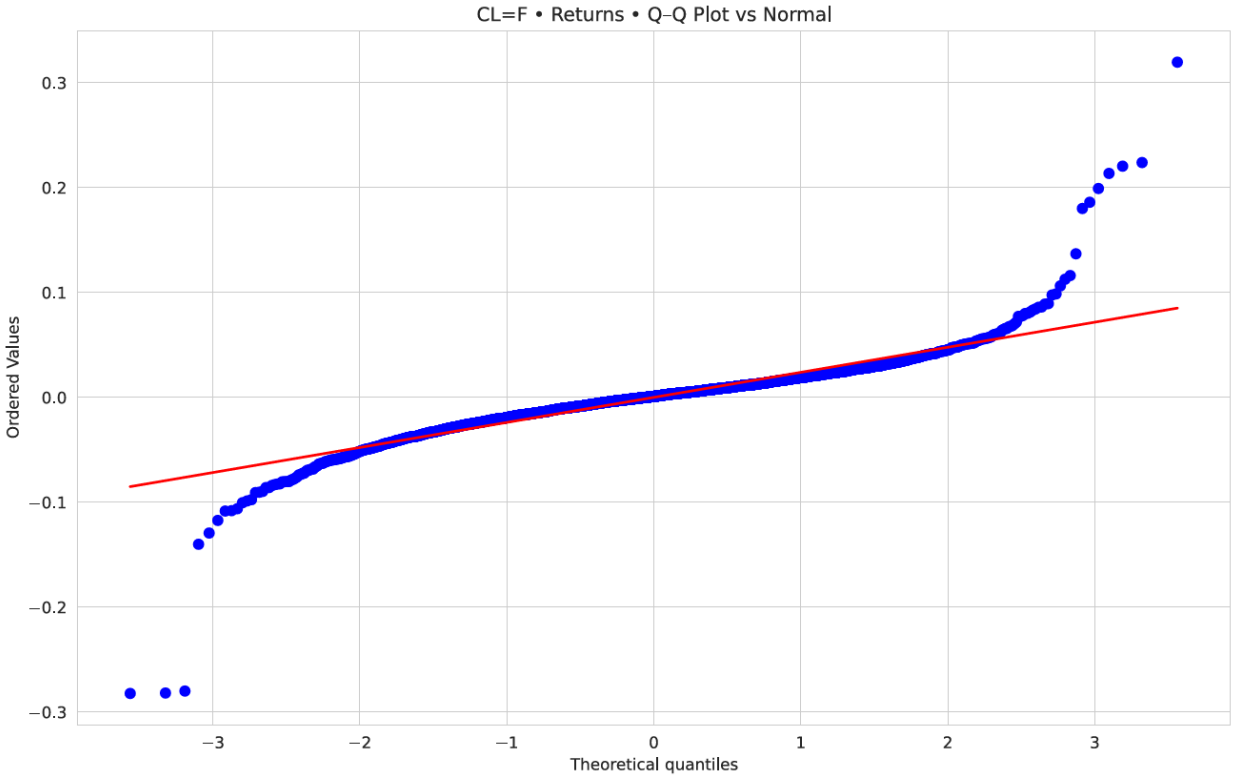
- Return distributions are approximately symmetric but exhibit significantly fatter tails than the normal distribution.

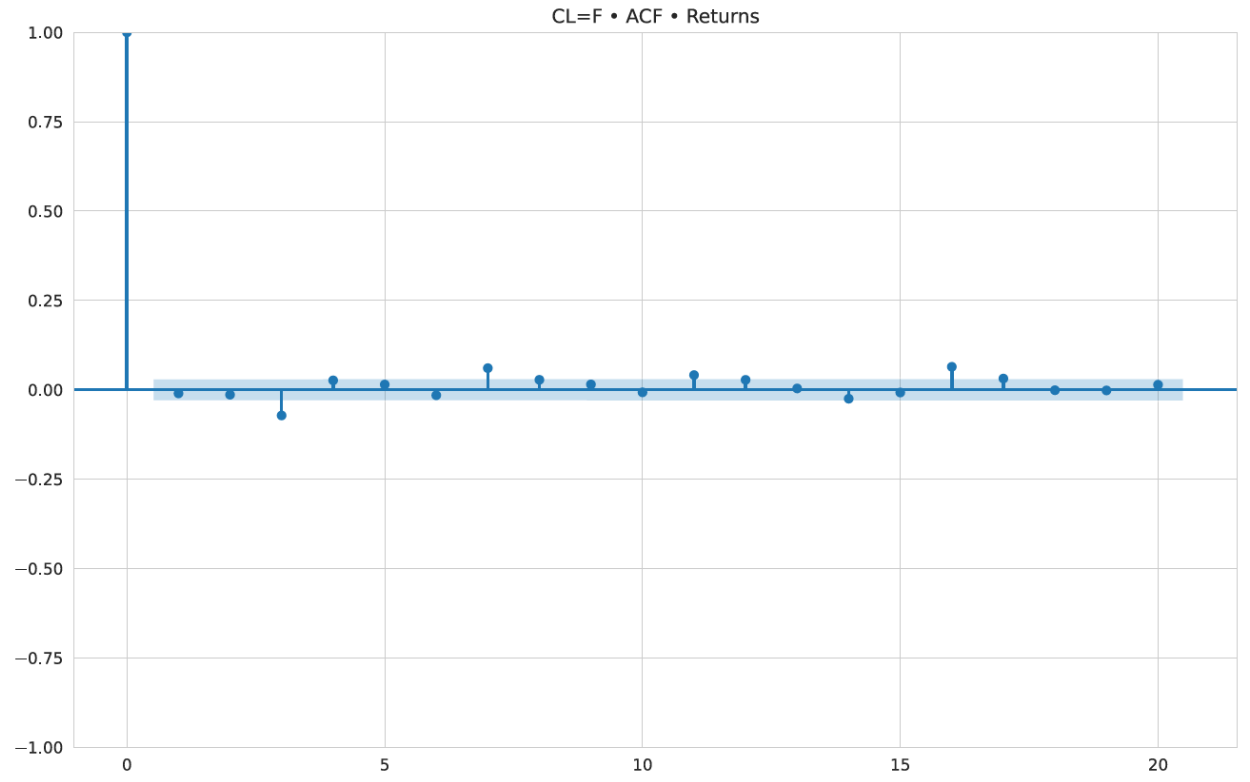
- Autocorrelations of raw returns are close to zero, indicating little to no linear dependence.
- Autocorrelations of absolute and squared returns are positive and persistent across many lags, reflecting volatility clustering and nonlinear dependence.

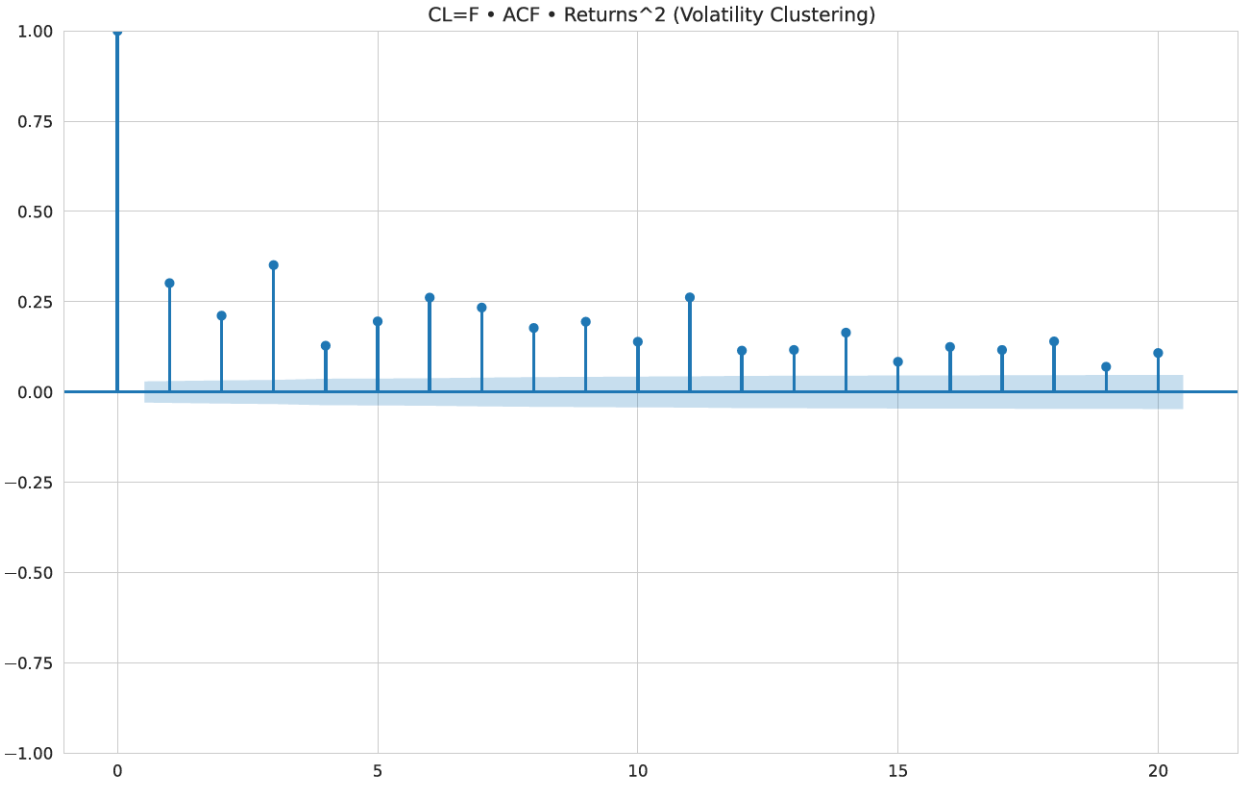
The *Refining Stock Report.pdf* contains the full set of diagnostic plots for each asset. To illustrate these stylized facts, results for crude oil are shown below as a representative example; the remaining assets display similar qualitative patterns.

Crude Oil (CL=F)- Example plots showing stylized facts for financial markets


CL=F • Returns • Distribution
Mean: 0.000058
Median: 0.000972
Mode: 0.000000
Std Dev: 0.025889
Skewness: 0.0495
Kurtosis: 24.4809
Mean
±1σ
Density
Daily log return


CL=F • Returns • Q–Q Plot vs Normal
Ordered Values
Theoretical quantiles


CL=F • ACF • Returns
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
5
10
15
20


CL=F • ACF • Returns^2 (Volatility Clustering)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
5
10
15
20

Review of Gold (GC=F) market results

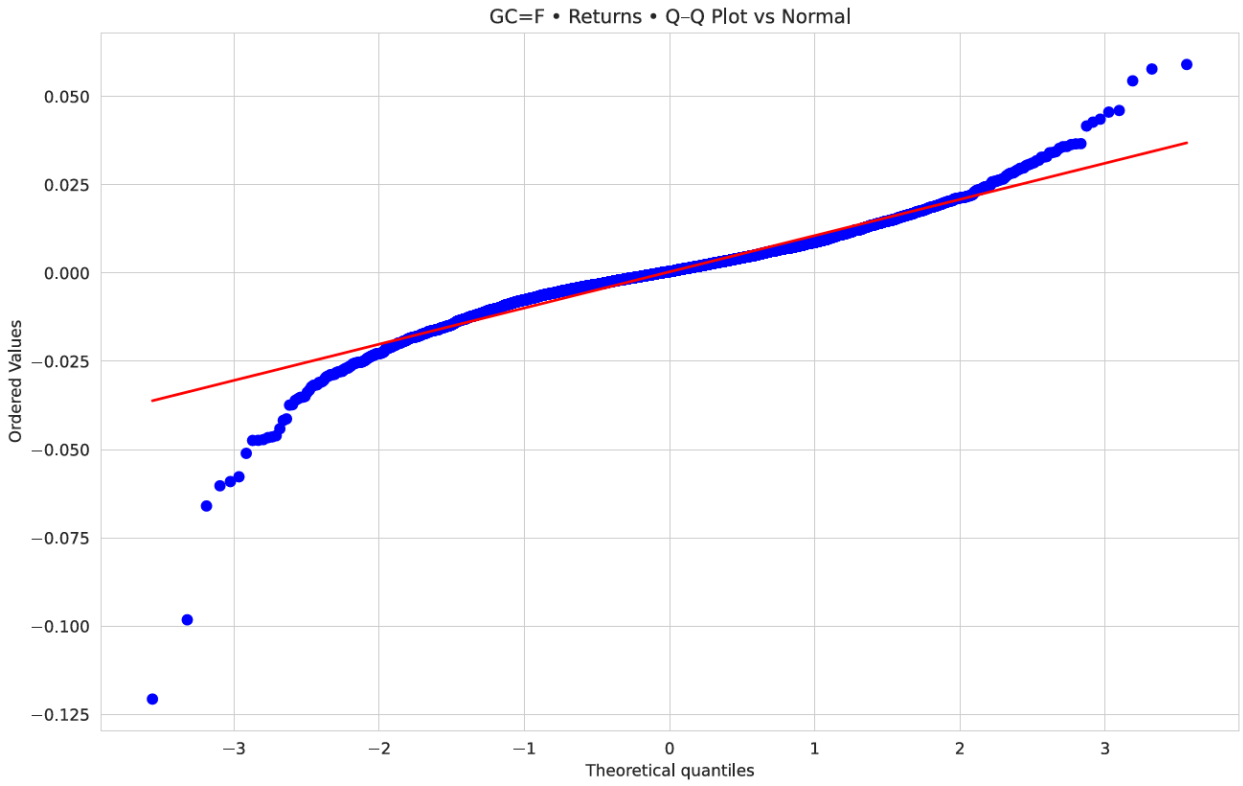
Next, we examine how the gold market results compare with those of the other assets analyzed. Gold returns exhibit noticeably lower volatility relative to crude oil, while still adhering to the standard stylized facts of financial markets, including weak serial correlation in returns and persistence in volatility.

Cross-asset analysis indicates that gold returns are weakly correlated with both crude oil and refining equities. This low correlation is stable across different market regimes, reinforcing gold's role as a diversification asset rather than a primary explanatory driver of refining equity price movements.

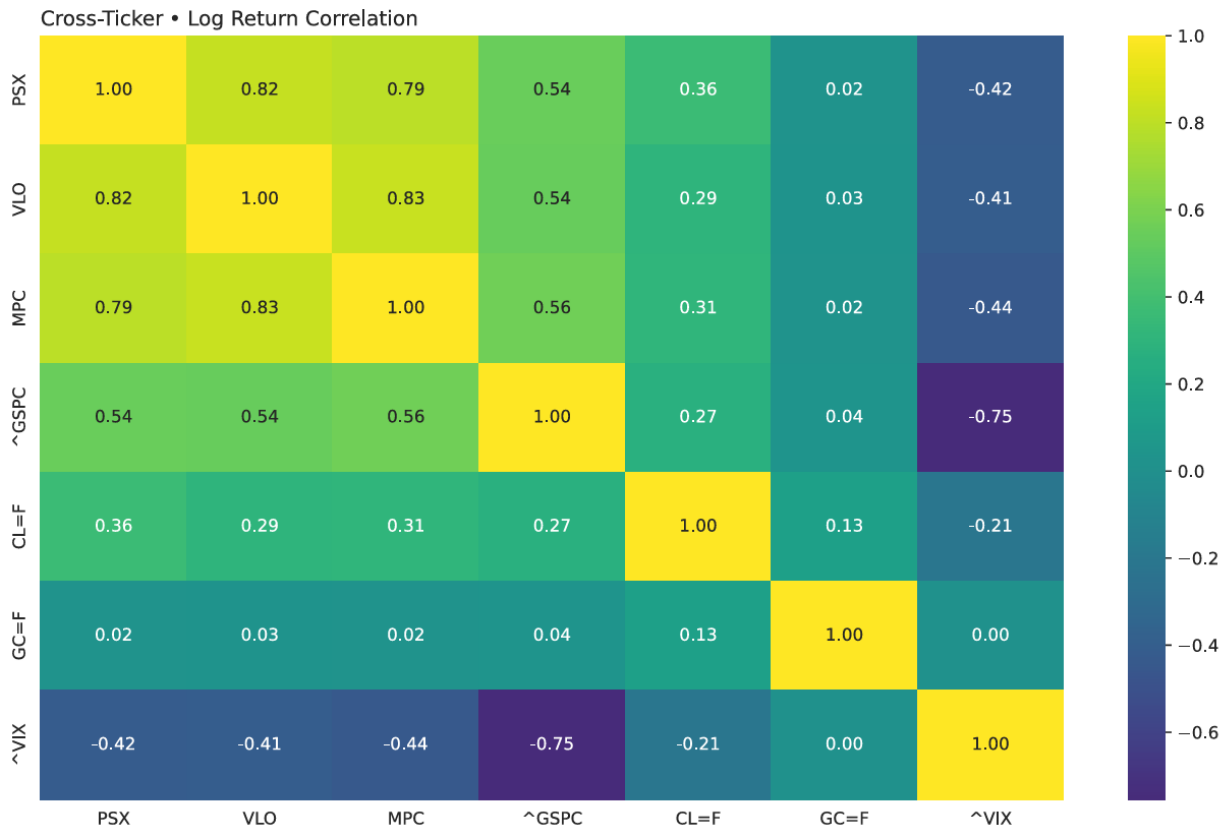
Within the context of refining and energy-focused analysis, gold therefore functions primarily as a portfolio diversifier and macro-risk hedge. Its price dynamics have limited direct relevance to refinery margins or input-cost fundamentals, and its inclusion in the feature set is best interpreted as capturing broader risk-aversion or flight-to-safety behavior rather than sector-specific economic effects.

The Q–Q plot further illustrates that gold exhibits substantially lower volatility relative to crude oil. Extreme gold return observations are generally bounded within approximately ±10%, whereas crude oil returns display much heavier tails, with occasional extreme movements exceeding ±20%. This contrast highlights gold's comparatively stable return

distribution and reinforces its role as a lower-volatility asset within the broader market context.



The cross-ticker return correlation matrix below further demonstrates that gold returns exhibit low correlation with the other assets in the study. Specifically, gold shows weak contemporaneous correlation with crude oil, refining equities, and the broader equity market, reinforcing its distinct behavior relative to energy-related assets. The cross correlation of the other asset returns are discussed in the following sections.



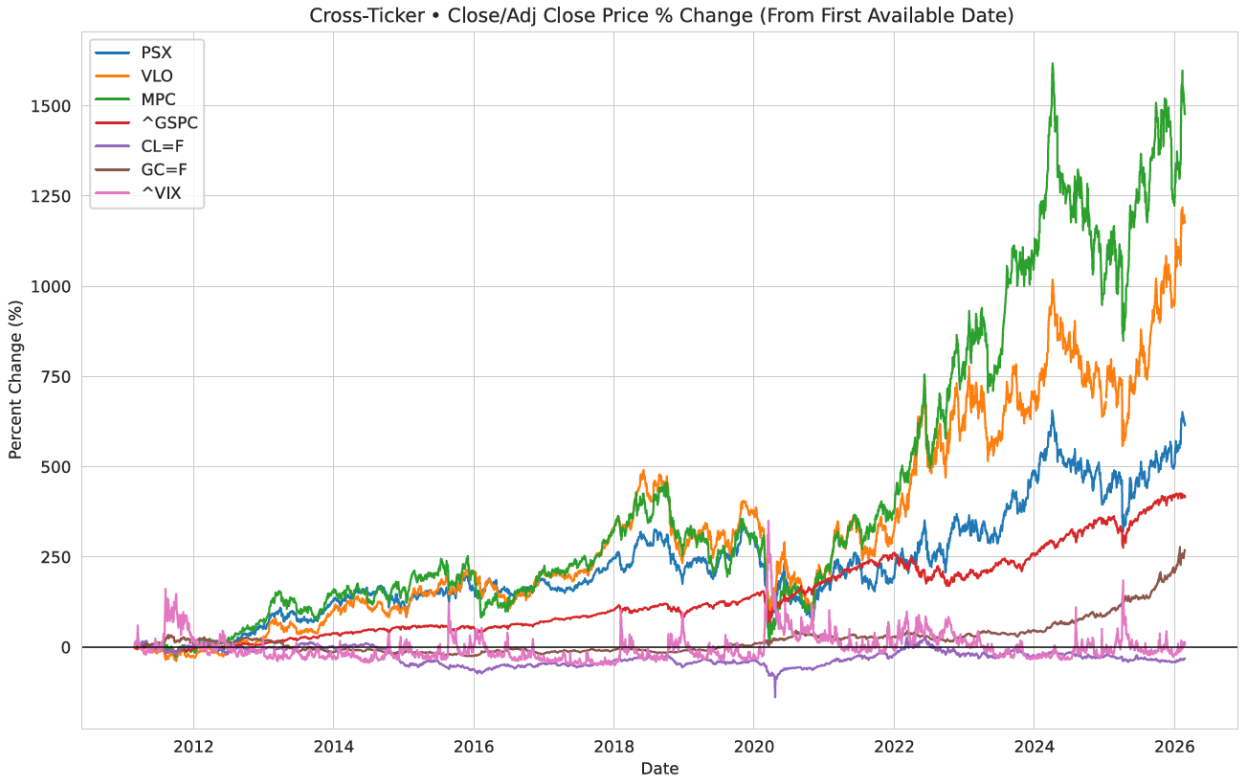
Refining Equities (MPC, PSX, VLO) market results

The refining equities MPC, PSX, and VLO exhibit broadly similar return distributions and volatility profiles compared to each other. Overall returns are positive over the sample period. The volatility levels are lower than those observed in crude oil but higher than those of the broader equity market.

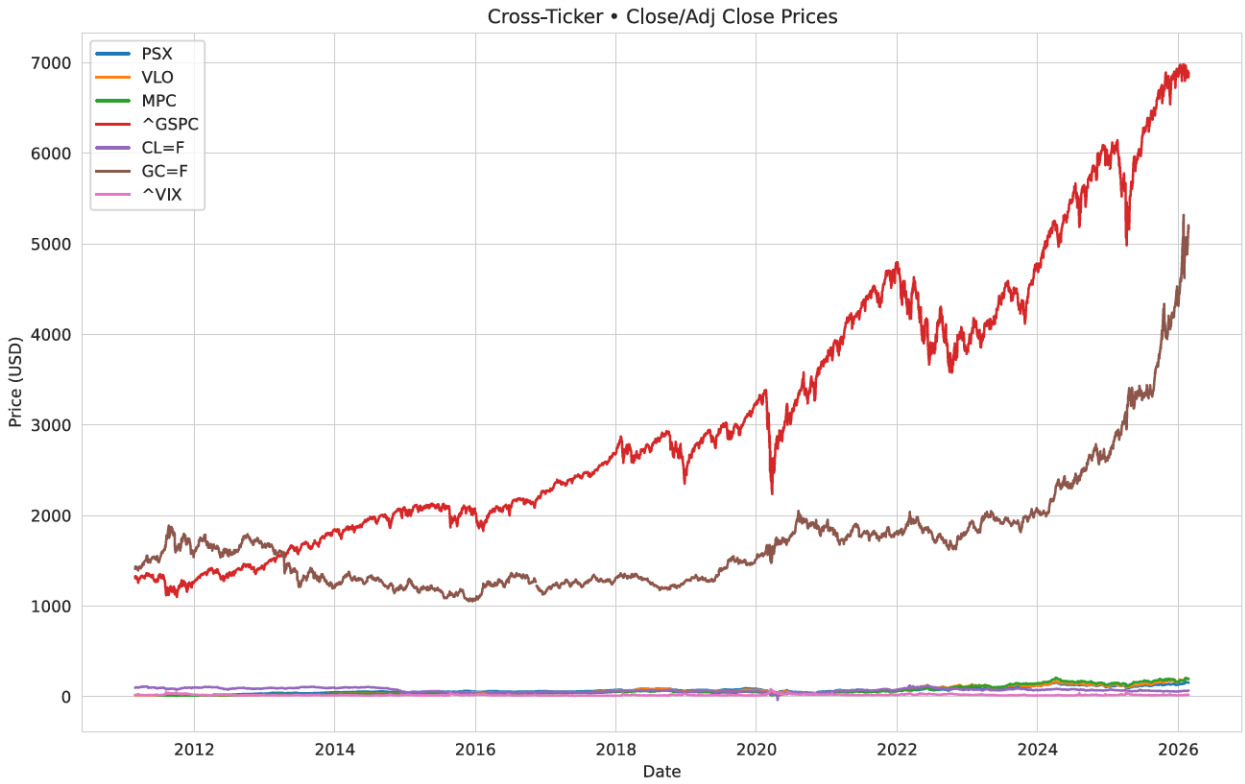
Cross-correlation of returns heat map reveals strong co-movement among refining equities with correlations around 0.80, indicating that these securities largely trade as a cohesive sector. Correlations with the S&P 500 are moderate and positive at around 0.55, reflecting systematic equity market exposure. As with other assets examined, volatility clustering is evident, though less extreme than in crude oil.

These results imply that sector-level and macroeconomic factors dominate firm-specific effects in refining equity performance. Consequently, incremental informational gains from specific equity timing strategies appear limited relative to sector-based or macro-driven approaches.

The overall percentage price change for the full sample period is shown below. This comparison indicates that refining equities outperformed the other analyzed assets over the study horizon, reflecting strong long-term appreciation relative to crude oil, gold, and the broader market benchmarks.



The overall price trends for all assets over the sample period are shown below. Because the assets differ substantially in price levels and volatility, some movements are difficult to discern when plotted on a common scale. As a result, the percentage price change view presented above provides a clearer and more comparable representation of how asset values evolved over time.



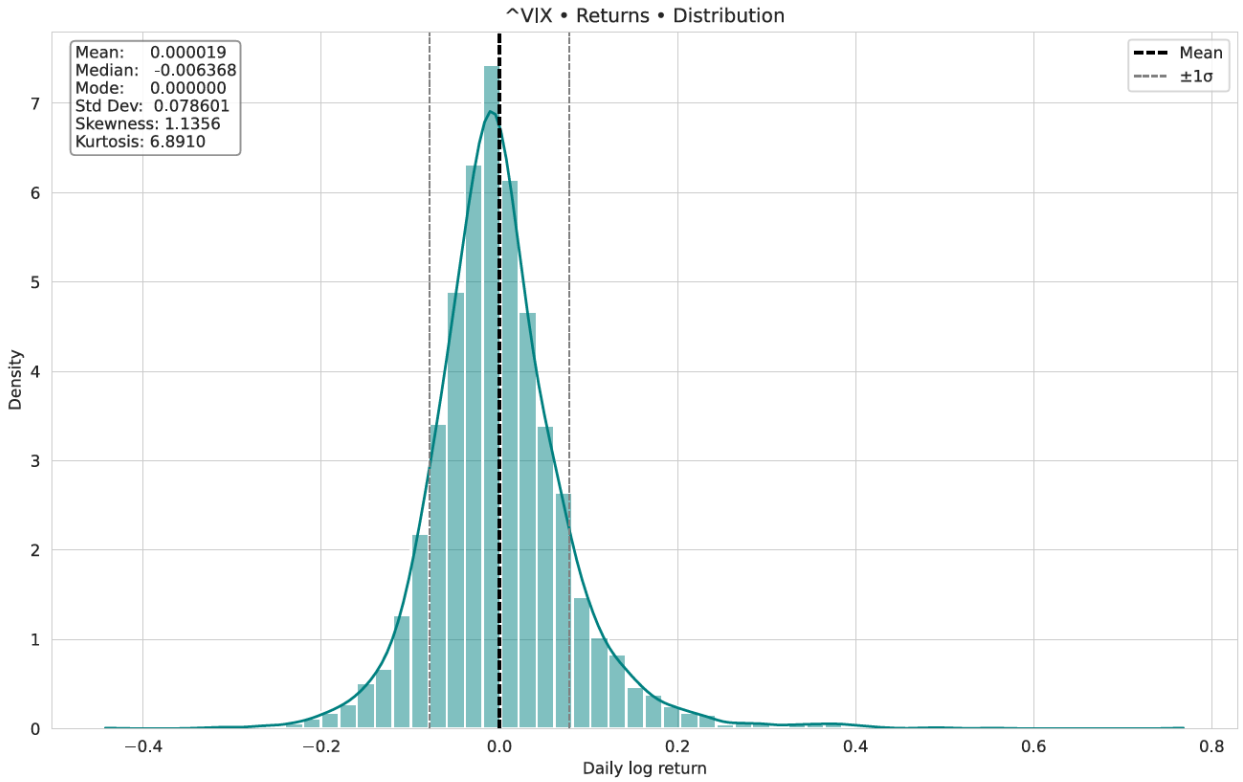
Overall Market and Volatility Benchmarks (^GSPC and ^VIX)

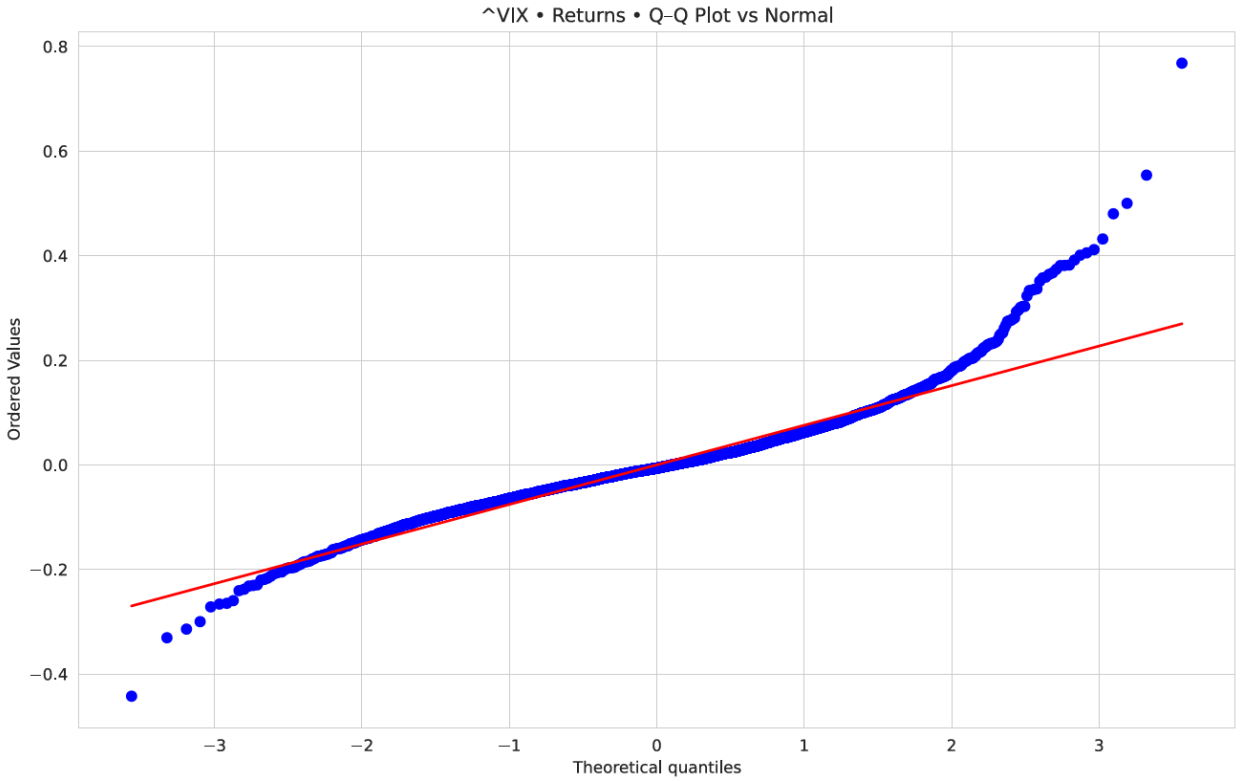
The S&P 500 exhibits the lowest return volatility among the analyzed assets, with weak serial correlation in returns and persistent volatility clustering in squared returns. These characteristics are consistent with well-documented stylized facts of broad equity markets.

The VIX displays markedly different behavior, characterized by extreme volatility, strong persistence, and pronounced non-normality. Return correlations between the VIX and equity assets, including refining equities, are strongly negative, confirming the VIX's role as a market-wide risk indicator.

For refining equity analysis, these findings underscore the importance of market-level risk conditions. Periods of elevated volatility, as captured by the VIX, are systematically associated with adverse equity performance in the sector.

Plots showing the positive skew of the ^VIX. Large increases are more likely than large decreases.


^VIX • Returns • Distribution
Mean: 0.000019
Median: -0.006368
Mode: 0.000000
Std Dev: 0.078601
Skewness: 1.1356
Kurtosis: 6.8910
Mean
±1σ
Density
Daily log return


^VIX • Returns • Q–Q Plot vs Normal
Ordered Values
Theoretical quantiles

Summary of Overall Market Results

Cross-asset correlation analysis highlights several key structural relationships. Refining equities are highly correlated with one another and moderately correlated with the S&P 500, indicating strong sector and market influences. Crude oil exhibits modest positive correlation with refining equities, consistent with its role as a key input cost rather than a direct price proxy. Gold remains largely uncorrelated with the other assets, while the VIX shows strong negative correlation with equities and refining stocks.

Overall, the empirical evidence indicates weak return predictability across all analyzed assets, pervasive volatility persistence, and dominant sector-level and macroeconomic effects in refining equities. These findings support the use of volatility-aware risk frameworks, sector-based exposure analysis, and macro-risk indicators when evaluating refining stocks and related market dynamics.

Part 2 – Machine Learning Modeling Results for Predicting Next-Day Return Direction of Valero (VLO)

This section presents the machine learning modeling results for predicting the next-day return direction of Valero Energy Corporation (VLO). A wide range of models and engineered features were developed and evaluated in the analysis. To maintain clarity and focus, one representative regularized logistic regression model is reviewed in detail to illustrate the modeling setup, evaluation process, and interpretation of results. The remaining models were implemented using the same general framework, and their outcomes are summarized through comparative performance metrics. In addition, exploratory data analysis (EDA) on VLO's return direction is presented prior to modeling to provide context for the classification task.

It is important to emphasize that next-day stock returns are close to random, as predicted by the Efficient Market Hypothesis (EMH). As a result, even modest predictive power above random guessing can be economically meaningful. Classification accuracy in the range of 50–55% may still generate valuable trading signals under certain conditions, whereas accuracy levels near 60% are generally unrealistic for daily equity return prediction given the inherent noise and efficiency of financial markets.

The analysis first evaluates several variations of the baseline logistic regression model to assess whether targeted methodological adjustments improve performance. These modifications include:

- Removing features with pairwise collinearity above 0.85 to reduce redundancy and instability in coefficient estimates.
- Replacing randomized cross-validation with time-series-aware cross-validation to better respect the temporal structure of the data.
- Evaluating model performance using the balanced accuracy metric to account for asymmetries between up and down return classifications.

Next, more complex machine learning models are compared against the baseline regularized logistic regression, including:

- Decision Tree Classifier
- Gradient Boosted Trees Classifier
- Random Forest Classifier
- Deep neural network model
- Convolutional neural network (CNN) model applied across the feature dimension
- Convolutional neural network model incorporating a 10-day lookback window
- IBM TinyTimeMixer (TTM) time-series foundation model

The study then evaluates whether augmenting the baseline logistic regression with additional feature sets improves predictive performance. Specifically, the following features are introduced:

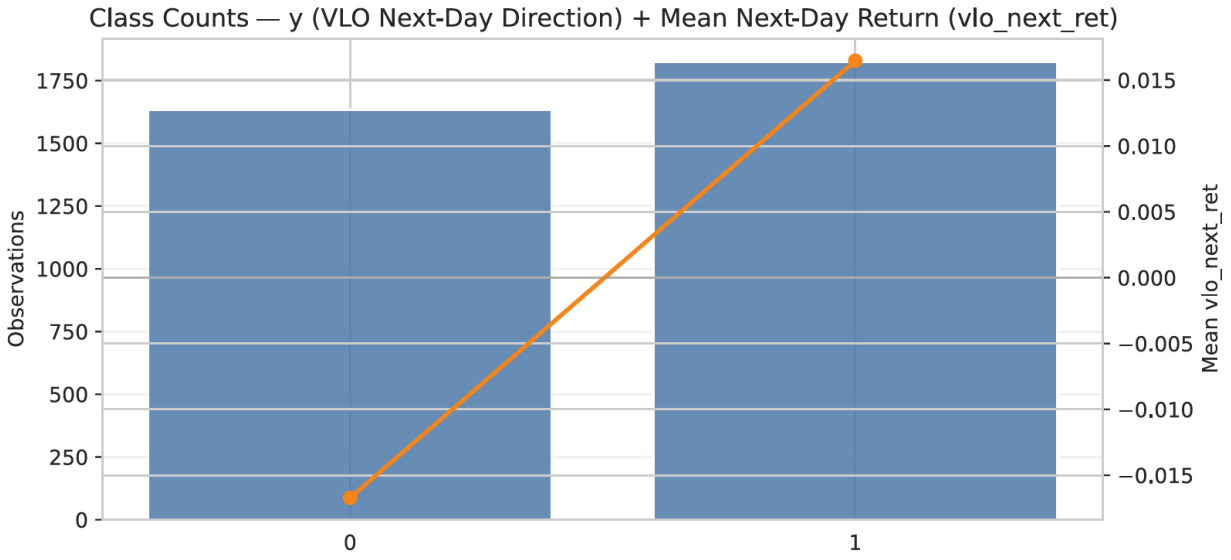
- GARCH(1,1)-based volatility forecasting features
- OPEC press-release-based news indicators

Finally, the results compare the baseline logistic regression model with the best-performing model using the full combined feature set. This progression allows for a structured assessment of how methodological refinements, nonlinear modeling approaches, and enriched feature engineering affect both statistical classification performance and economic outcomes.

EDA on VLO Machine Learning Dataset

This section provides exploratory data analysis (EDA) to establish context for the machine learning modeling that follows. The goal is to summarize key characteristics of the VLO return direction dataset and highlight baseline patterns relevant to the classification task.

The table and chart below present the distribution of positive and negative next-day returns in the machine learning dataset, along with the mean return for each group. This summary illustrates the class balance and the magnitude of returns associated with up and down days, providing a reference point for interpreting subsequent model performance.



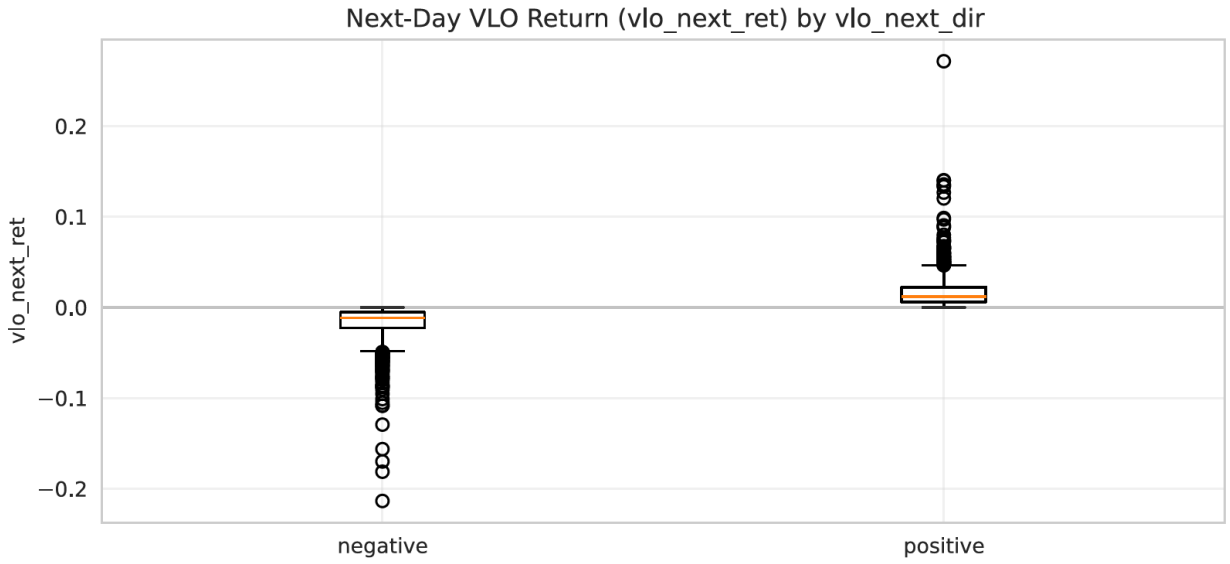
metric	negative	positive
n	1634.0	1825.0
mean	-0.01671975	0.01647241
median	-0.01155561	0.01214429
std	0.01790377	0.01676052

The higher frequency of positive returns is consistent with VLO's substantial long-term price appreciation over the sample period. This distribution also provides a basic validation check, confirming that the return direction classification was implemented correctly in the data preparation process. For simplicity in the binary classification framework, days with exactly zero returns were grouped with negative returns, resulting in a two-class target of positive versus non-positive next-day outcomes.

If the model were to naïvely predict a positive return on every day, it would achieve an accuracy of approximately 52.8%, reflecting the underlying class imbalance in the dataset. During initial experimentation, some model configurations effectively converged toward this trivial strategy when optimized solely on raw accuracy. However, this behavior is not informative, as it fails to demonstrate an ability to distinguish between up and down market conditions.

Because the objective of the modeling exercise is to generate signals that are useful in both rising and declining markets, accuracy alone is an insufficient evaluation metric. As a result, the ROC–AUC metric was adopted as the primary performance measure, as it evaluates the model's ability to discriminate between positive and negative outcomes across all possible classification thresholds rather than rewarding majority-class predictions.

Box plot of the two classifications



The following chart displays the features with the highest correlation to VLO's next-day return. All correlations are close to zero, underscoring the weak linear relationship between individual predictors and next-day price direction and reinforcing the near-random nature of daily equity returns.

feature	corr_y	abs_corr_y
vlo_next_ret	0.691619	0.691619
HL_VLO	-0.044332	0.044332
ret_MPC	-0.042287	0.042287
ret2_CL=F	0.042253	0.042253
RefinerPeerAvg	-0.033644	0.033644
ret2_^VIX	-0.033354	0.033354
ret_VLO	-0.031618	0.031618

The variable *vlo_next_ret* was not included as an input feature in the machine learning dataset. It was retained solely for exploratory data analysis to facilitate diagnostic review and was removed prior to all model training to prevent any target leakage. The correlation table further highlights the near-random nature of next-day returns: none of the candidate predictors exhibit a correlation with the target exceeding 0.05.

The variables with the highest observed correlations to VLO's next-day return though still very small in magnitude were:

- Previous-day high–low price spread
- Previous-day return of Marathon Petroleum Corporation (MPC)
- Two-day lagged return of crude oil
- Prior-day average return of the refinery peer group
- Two-day lagged return of the VIX

These correlations are only marginally larger than that of VLO's own previous-day return, which, consistent with established stylized facts of financial markets, is known to have little predictive power for future returns. The remaining approximately 70 engineered features exhibit even weaker linear relationships with the target variable.

Taken together, these results indicate that all models begin from a highly challenging baseline in which no individual predictor provides meaningful standalone predictive power. In effect, the modeling task resembles extracting weak structure from an environment that is close to white noise, underscoring the inherent difficulty of short-horizon equity return prediction.

Baseline Logistic Regression Model Setup

The baseline regularized logistic regression model was trained and evaluated using the following data partitions and configuration:

- Training period: February 2011 through December 31, 2022
- Test period: January 1, 2023 through February 2026
- Training sample size: 2,672 observations
- Test sample size: 787 observations

Feature set:

- 72 scaled continuous features, including returns, volatility measures, and engineered numeric predictors, standardized using a *StandardScaler*
- 5 unscaled passthrough numeric features consisting of binary (0/1) market-regime and classification indicators derived from engineered market condition flags

Model specification:

- Logistic Regression classifier with L1 and L2 regularization

Model selection and validation:

- Hyperparameter tuning performed using *GridSearchCV* over:
    - Regularization penalty ∈ {L1, L2}
    - Inverse regularization strength $C \in \{0.001, 0.01, 0.1, 1, 10, 100\}$
- Model refit criterion: ROC–AUC, selected to emphasize discrimination rather than majority-class accuracy

Leakage prevention:

- The variables *vlo_next_ret* and *vlo_next_dir* were explicitly removed from the feature matrix prior to model training to prevent target leakage

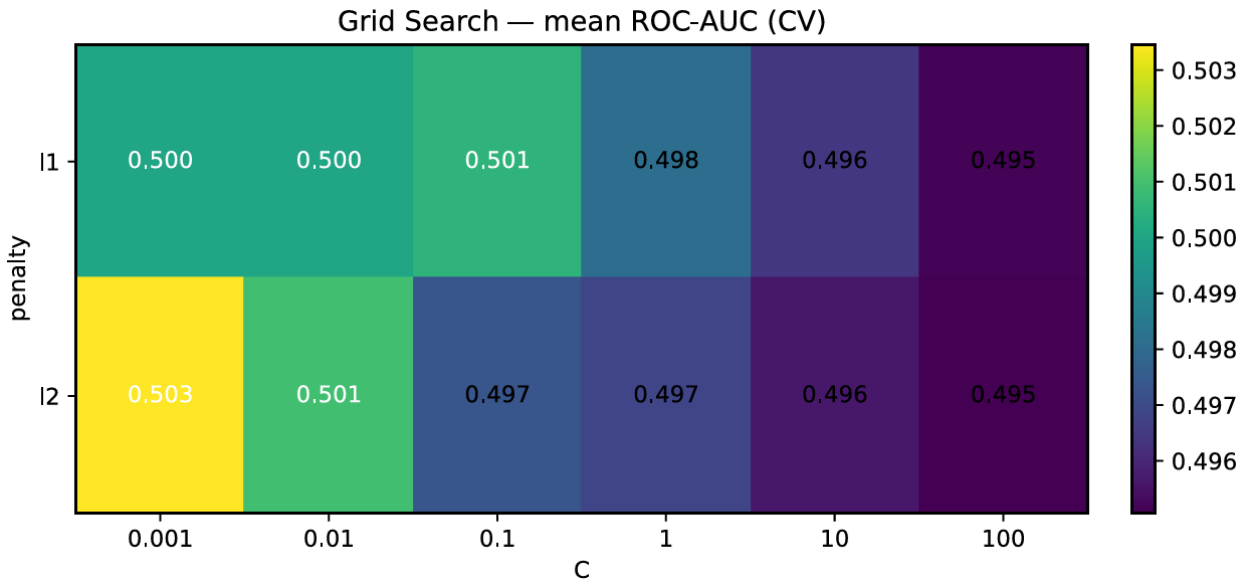
Selected model parameters:

- L2 regularization
- $C = 0.001$(strong regularization)

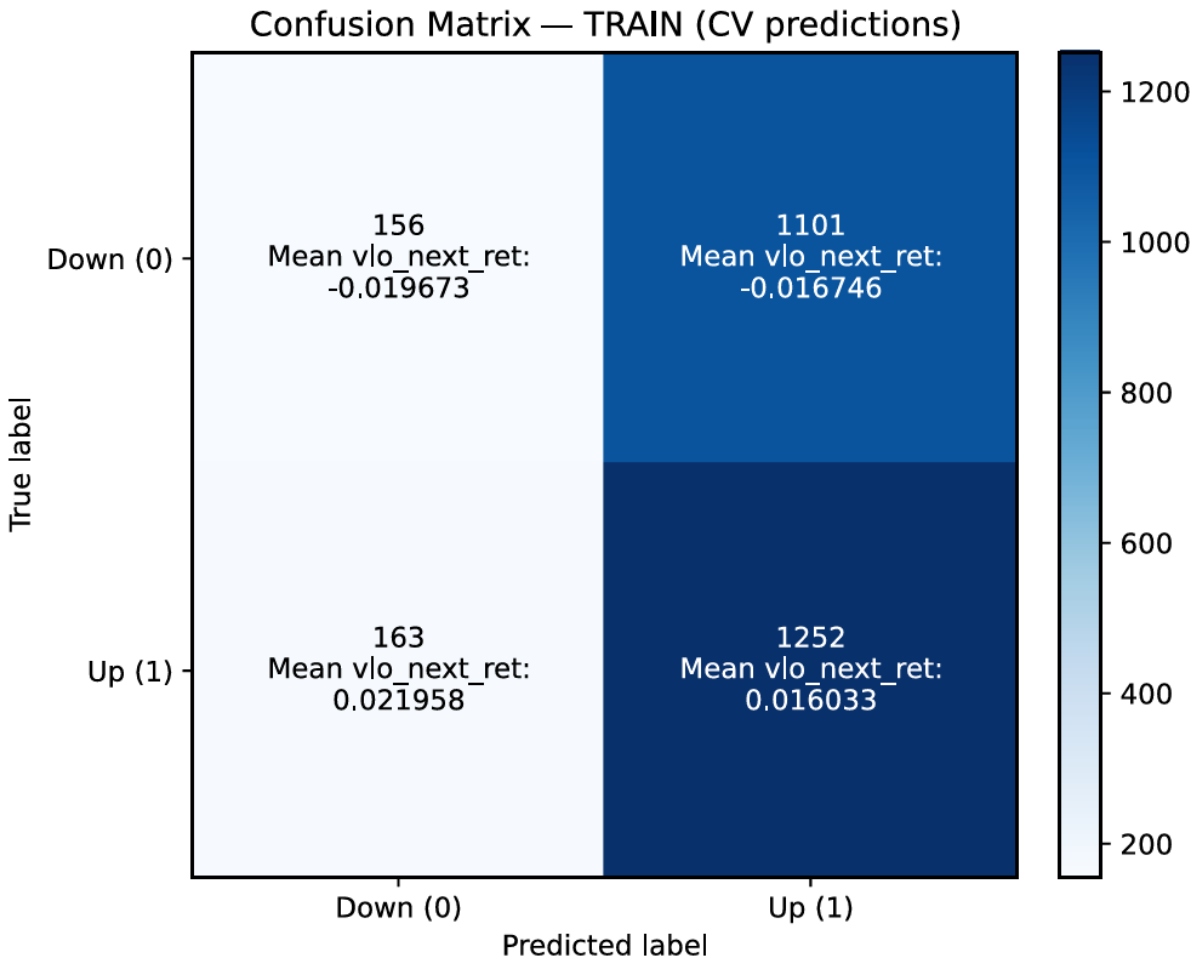
Classification Predictive Performance

Cross-Validation Results (Training Set, Pooled CV Predictions)

- Mean cross-validated ROC–AUC: 0.5035
- Standard deviation of ROC–AUC: 0.0298
- Pooled training ROC–AUC: approximately 0.502



Training confusion matrix with mean next-day returns:

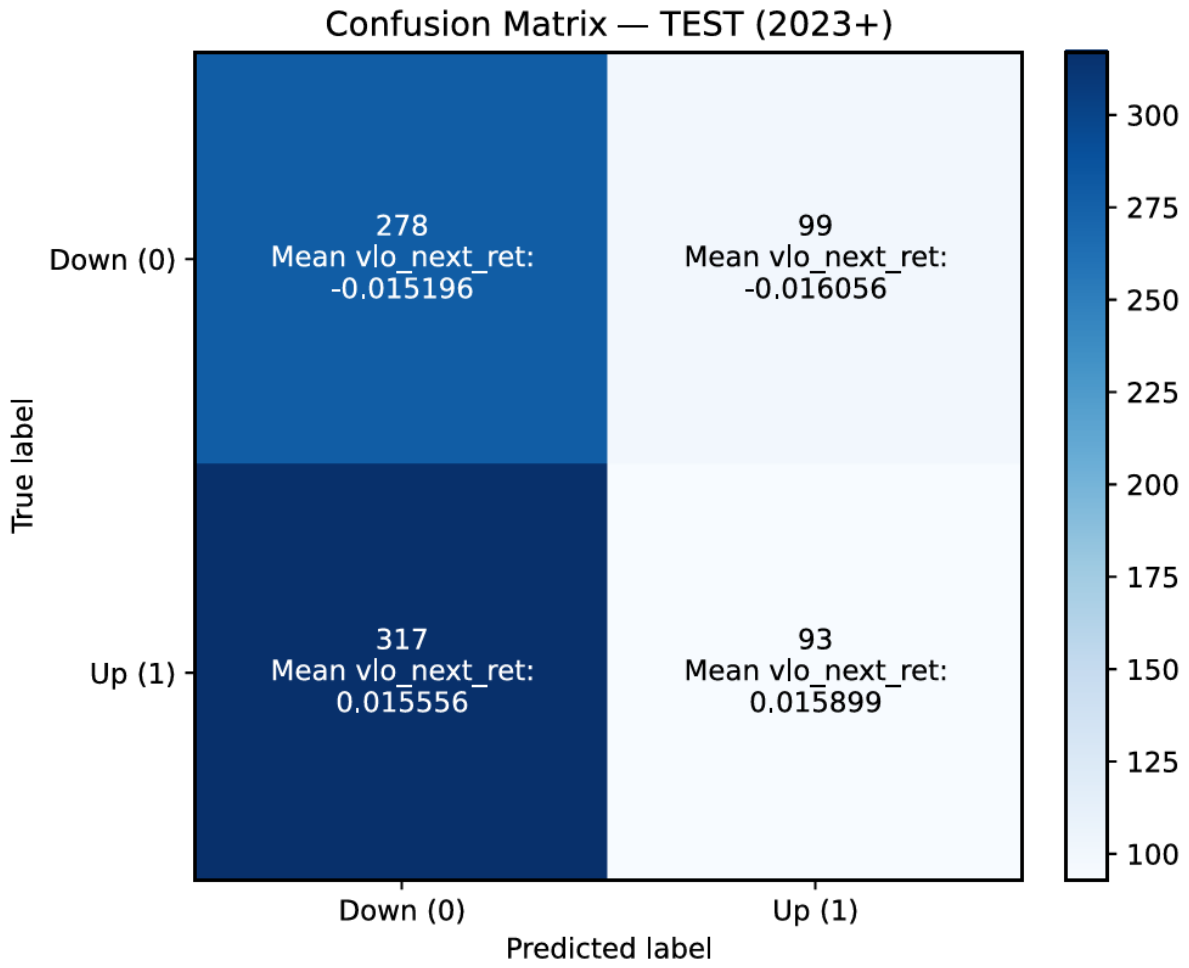


While the mean returns align in sign (Up cells positive, Down cells negative), classification discrimination is still near and almost random guesses based on AUC.

Out-of-Sample (TEST, 2023+ data)

- TEST ROC-AUC: 0.5055
- TEST accuracy: 0.4714

Testing confusion matrix with mean next-day returns per cell:



TEST classification report:

- Class 0 (Down): precision 0.467, recall 0.737
- Class 1 (Up): precision 0.484, recall 0.227
- Overall accuracy: 0.471

This base model leans heavily toward predicting “Down” (high recall for Down, very low recall for Up), while overall performance remains weak with a 47% accuracy.

Investment Backtests (Test Period Only)

Investment performance was evaluated using out-of-sample backtests conducted exclusively over the test period beginning January 1, 2023. All reported results are based on after-tax equity curves to better reflect realistic investor outcomes.

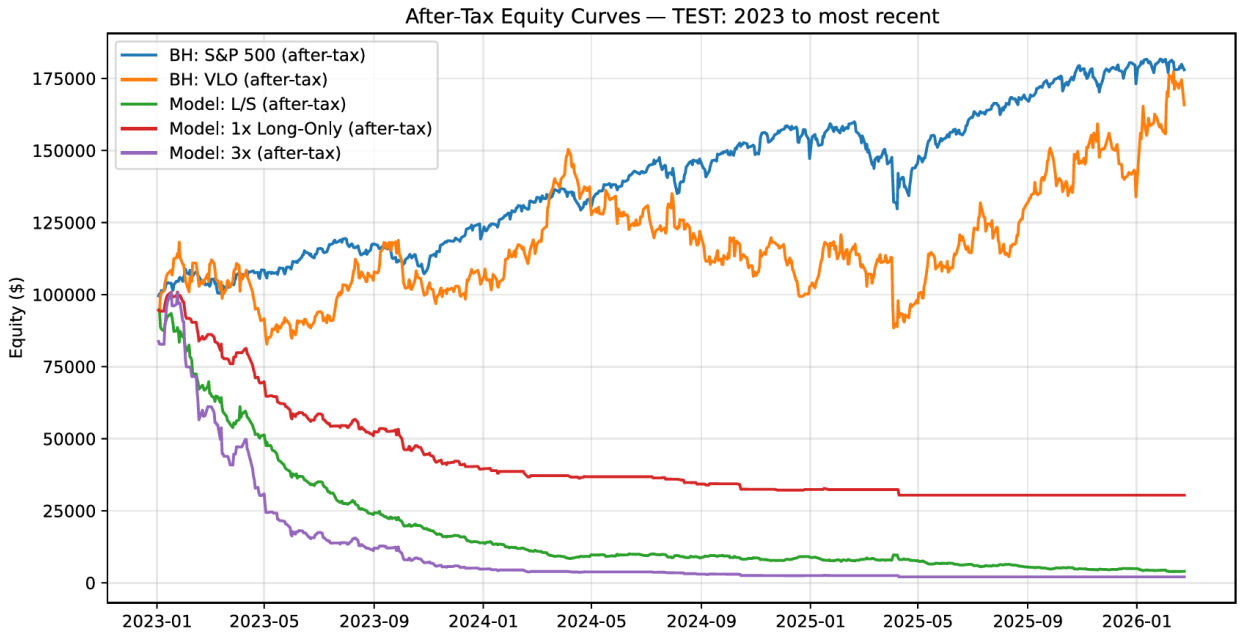
The backtesting framework incorporates several real-world trading frictions and financing assumptions, including:

- Transaction costs: 2 basis points per side per trade
- Tax treatment: long-term capital gains proxy applied to buy-and-hold strategies; short-term rates applied to active strategies, with loss carryforward
- Short borrow cost: 0.30% annualized rate applied on days with short exposure in long/short strategies
- Leverage costs: daily financing and expense drag applied on days with 3× leveraged long exposure

The following strategies were evaluated and compared:

- Buy-and-Hold S&P 500, after tax
- Buy-and-Hold VLO, after tax
- Model-driven Long/Short VLO strategy (±1 exposure)
- Model-driven 1× Long-Only VLO strategy, invested only on predicted up-days and otherwise held in cash
- Model-driven 3× Long-Only VLO strategy, invested only on predicted up-days and otherwise held in cash

As expected, given the weak predictive performance of the baseline model, the investment backtests also perform poorly. All active strategies generate negative risk-adjusted returns over the test period, with Sharpe ratios remaining negative, reflecting insufficient return generation relative to volatility and trading costs.



Annual After-Tax Performance (2023+), incl. Average

	BH: S&P 500 (after-tax	BH: VLO (after-tax)	Model: L/S (after-tax)	Model: 1x Long-Only (after	Model: 3x (after-tax)
2023	19.78%	9.50%	-85.34%	-58.31%	-94.32%
2024	19.21%	-3.45%	-36.14%	-18.26%	-47.59%
2025	13.32%	29.79%	-43.61%	-6.09%	-18.43%
2026	-0.32%	16.09%	-18.87%	0.00%	0.00%
Average	13.00%	12.98%	-45.99%	-20.66%	-40.09%

Sharpe Ratios (Annualized, Daily Returns) — TEST (2023+ only)

	Sharpe
BH: S&P 500	1.311
BH: VLO	0.682
Model: L/S	-2.936
Model: 1x Long-Only	-2.691
Model: 3x	-2.771

Comparing Modifications to the Logistic Regression Model Setup

To assess whether targeted methodological adjustments could improve predictive performance, several modifications were applied to the baseline logistic regression model. Each change was evaluated independently to isolate its effect on classification quality and economic outcomes.

First, features exhibiting high pairwise collinearity were removed. Specifically, variables with correlation coefficients exceeding 0.85 were excluded from the feature set. High collinearity can destabilize coefficient estimates in linear models and complicate interpretation. While regularization can mitigate some of these effects, it does not always fully resolve coefficient redundancy, particularly in the presence of many correlated predictors.

Second, model evaluation was repeated using time-series-aware cross-validation in place of randomized cross-validation. Although each observation represents a single day's feature set used to predict the following day's return direction, rather than an explicit sequence forecast, financial time series exhibit regime persistence and temporal dependence. Time-series cross-validation better reflects real-world forecasting conditions by preserving chronological ordering and preventing information leakage across regimes.

Third, model performance was evaluated using **balanced accuracy** as an alternative optimization metric. Balanced accuracy is designed to give equal weight to both outcome classes and is defined as the average of recall computed separately for each class:

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\text{Recall}_{\text{Down}} + \text{Recall}_{\text{Up}}\right)$$

where:

- $\text{Recall}_{\text{Down}} = \frac{TN}{TN+FP}$
- $\text{Recall}_{\text{Up}} = \frac{TP}{TP+FN}$

This metric treats up and down return predictions symmetrically, regardless of class imbalance, and discourages models from achieving deceptively high performance by favoring the majority class. Balanced accuracy therefore provides a more informative assessment of whether a model can meaningfully discriminate between positive and negative return outcomes.

Comparison table

Model variant	Key modification	Best params	TEST ROC-AUC	TEST accuracy	Long Short Sharpe	Net effect vs base

Base logistic (earlier)	ROC-AUC refit, strong L2	L2, C=0.001	~0.506	~0.471	−2.94	Baseline
Collinearity filtered	Drop corr>0.85 (dropped 26 feats)	L1, C=0.1	~0.516	~0.497	−2.35	Slight metric gain but worse economics
Time series cross validation	TimeSeriesSplit out-of-fold (OOF) predictions	L2, C=0.01	~0.509	~0.479	−1.55	Best trade-off overall
Balanced accuracy refit	New optimization target for model	L1, C=100	~0.517	~0.483	−1.30	Best classificati on balance

Across all evaluated variants, the three model modifications produce modest improvements in performance relative to the baseline logistic regression. In particular, each modification yields incremental gains in ROC–AUC and classification accuracy, indicating slight improvements in discriminatory ability.

Despite these improvements, overall test-set accuracy remains below 50% for all model variants. More importantly, none of the modified specifications are able to translate the observed gains into economically profitable trading strategies once transaction costs, taxes, and financing assumptions are applied.

Among the evaluated approaches, the balanced-accuracy-optimized model performs best when the objective is classification symmetry between up and down outcomes. However, this improvement in classification fairness does not translate into superior trading performance. The collinearity-filtered model enhances interpretability by reducing feature redundancy, but this simplification comes at the expense of weaker economic outcomes in backtesting.

While the use of time-series-aware cross validation shows the most consistent improvement among the logistic regression variants, preliminary testing indicated that similar benefits did not extend to the more complex machine learning models. As a result, randomized cross validation was retained for the remaining models to maintain consistency and simplicity in the broader comparative analysis.

Comparison of More Complex Machine Learning Models to the Baseline Regularized Logistic Regression

In addition to the baseline regularized logistic regression, a range of more complex machine learning models were evaluated to assess whether nonlinear methods could extract additional predictive signal from the engineered feature set. The models tested include:

- Decision Tree Classifier
- Gradient Boosted Trees Classifier
- Random Forest Classifier
- Deep neural network model
- Convolutional neural network (CNN) model applied across the feature dimension
- Convolutional neural network model incorporating a 10-day lookback window
- IBM TinyTimeMixer (TTM) time-series foundation model

This section provides an overall performance comparison across these more advanced models. Each model was configured using either grid search or randomized search to identify appropriate hyperparameters, and model evaluation followed the same general framework used for the baseline logistic regression, including consistent data splits and performance metrics.

Model result table

The following model results table summarizes the comparative performance across all approaches. Overall, the more complex models deliver only modest improvements relative to the baseline, and none produce a substantial or consistent enhancement in predictive accuracy or ranking performance. As a result, increased model complexity does not materially improve next-day return direction forecasts in this setting. Detailed results and diagnostics for each model are discussed in the subsections that follow.

Rank	Model	CV ROC-AUC	TEST ROC-AUC	Δ TEST AUC	TEST Acc	Δ TEST Acc	Avg 1× Long-Only trading	Comment
1	CNN (TIME, lookback=10)	**0.555**	0.5172	0.0117	0.4769	0.0055	-5.45%	Strongest CV signal but weaker test result
2	Random Forest	0.5415	**0.5423**	0.0368	0.4956	0.0242	-7.73%	Best out-of-sample ranking; poor test accuracy
3	CNN (Feature Conv)	0.5396	0.5074	0.0019	**0.5248**	**0.0534**	-30.39%	Best accuracy/F1; weak ranking; poor economics
—	**LogReg (Baseline)**	**0.5035**	**0.5055**	**0**	**0.4714**	**0**	**-20.66%**	**Linear baseline; near-random signal**
4	Gradient Boosted Trees	0.5363	0.5044	-0.0011	0.4905	0.0191	-15.26%	CV strength doesn't translate to test result
5	Neural Net (MLP)	0.5361	0.4999	-0.0056	0.4828	0.0114	-12.82%	Mild nonlinearity; limited generalization
6	TTM (zero-shot)	0.4764	0.4853	-0.0202	0.4943	0.0229	*(ignored)*	Below baseline signal; econ test didn't calculate correctly
7	Decision Tree	0.503	0.4785	-0.027	0.4663	-0.0051	**-3.34%**	Weak signal but least-bad economics. Likely luck

The economic backtest results for the IBM TinyTimeMixer (TTM) model were not computed correctly and are therefore excluded from the comparison table.

Baseline Predictive Performance Summary

Cross-Validation Results

Ranking models by cross-validated ROC–AUC reveals a clear distinction between linear and nonlinear approaches. Overall, nonlinear models demonstrate a greater ability to extract signal from the engineered feature set during in-sample evaluation:

- The CNN with temporal convolution (TIME, lookback = 10) achieved the highest cross-validated ROC–AUC, indicating the strongest in-sample signal extraction from short-term temporal structure.
- Random Forest, CNN Feature Convolution, Gradient Boosted Trees, and MLP models formed a second tier, with similar cross-validated ROC–AUC values that were all meaningfully above the baseline.
- The baseline logistic regression and the single decision tree produced cross-validated ROC–AUC values close to 0.50, consistent with weak linear separability in the data.
- The TTM zero-shot model underperformed in cross-validation, suggesting limited alignment between its pretrained forecasting objective and the specific next-day classification task considered here.

Taken together, the cross-validation results suggest that nonlinear models are able to extract modest predictive structure from the feature set. However, the magnitude of this signal remains small, even for the strongest in-sample performers.

Out-of-Sample (Test) Results

Out-of-sample test performance presents a more nuanced and practically relevant picture:

- The Random Forest model achieved the highest test-set ROC–AUC, substantially outperforming the baseline and all other models. This result indicates the strongest generalization ability in terms of ranking positive versus negative return outcomes.
- The CNN TIME10 model retained a positive improvement over the baseline in test ROC–AUC but exhibited noticeable degradation relative to its cross-validated performance, suggesting some degree of overfitting or sensitivity to the chosen lookback window.
- The CNN Feature Convolution model marginally exceeded the baseline in test ROC–AUC but primarily distinguished itself through higher classification accuracy rather than improved ranking power.
- Gradient Boosted Trees and MLP models failed to translate their cross-validation strength into improved test-set ROC–AUC, ultimately performing slightly below the baseline.
- The Decision Tree and TTM zero-shot models performed worst in out-of-sample ROC–AUC.

Classification Accuracy vs. Ranking Performance

A key observation across models is the divergence between classification accuracy and ranking performance as measured by ROC–AUC. These two metrics capture different aspects of model behavior and do not always move together:

- The CNN Feature Convolution model achieved the highest test-set accuracy, yet only marginally improved ROC–AUC relative to the baseline, indicating limited gains in probabilistic ranking despite improved threshold-based classification.
- In contrast, the Random Forest model produced the highest test-set ROC–AUC, reflecting superior ranking of positive versus negative outcomes, but exhibited only moderate classification accuracy.

This divergence reflects differences in decision-threshold behavior and class prediction asymmetry. Models optimized for accuracy may favor a particular class or threshold configuration, while models with stronger ranking performance may not translate that advantage into higher accuracy at a fixed cutoff.

Economic Performance

Economic backtests were conducted using consistent transaction cost, tax, and leverage assumptions across all models. Several important findings emerge from this analysis:

- All supervised machine learning models generated negative average after-tax returns across long/short, 1× long-only, and 3× leveraged implementations over the test period.
- The baseline logistic regression was among the weakest performers economically, particularly in leveraged strategies.
- Notably, the single decision tree, despite weak predictive metrics, produced the least negative economic outcomes, especially for the 1× and 3× long-only strategies. This result is most likely attributable to favorable signal timing rather than robust or persistent predictive power.
- The Random Forest and CNN TIME10 models demonstrated improved economic performance relative to the baseline but remained meaningfully negative after costs.
- The CNN Feature Convolution model, despite strong classification accuracy, delivered the poorest economic performance, highlighting the risks associated with overtrading and poorly calibrated signals.

Overall, these findings indicate that incremental improvements in predictive metrics do not reliably translate into profitable trading performance under realistic transaction cost, tax, and financing assumptions. This disconnect underscores the difficulty of converting weak short-horizon predictive signals into economically viable trading strategies.

Overall Findings from the More Complex ML Models

The evaluation of more advanced machine learning models yields several key takeaways:

Best predictive performance (ranking): The Random Forest model demonstrates the strongest out-of-sample ranking ability, as measured by test-set ROC–AUC, indicating the most robust generalization across models.

Strongest learned signal (cross-validation): The CNN TIME10 model achieves the highest cross-validated performance, suggesting it is most effective at extracting short-term temporal structure in-sample.

Best classification accuracy: The CNN Feature Convolution model delivers the highest test-set accuracy, reflecting improved threshold-based classification performance.

Least negative economic outcomes: The single Decision Tree, despite weak predictive metrics, produces the least adverse economic results, likely due to favorable timing rather than persistent predictive strength.

Overall conclusion: The engineered feature set contains weak but non-zero predictive signal that can be partially exploited by nonlinear models; however, this signal is insufficient to overcome trading frictions, costs, and noise in its current form.

Taken together, these findings suggest that meaningful performance gains are more likely to arise from improved feature engineering, alternative target formulations, or refined decision rules, rather than from additional increases in model complexity alone.

Cross-Model Feature Importance

To identify the most economically and statistically meaningful predictors of next-day VLO price direction, feature importance outputs were examined across all modeling classes, including the baseline logistic regression, tree-based models (decision tree, gradient boosting, random forest), neural networks (MLP), convolutional neural networks (feature-wise and time-wise CNNs), and a zero-shot time-series foundation model (TTM). Rather than relying on a single importance metric, this analysis emphasizes consistency of signal across heterogeneous model families, which provides greater robustness than any individual model's attribution.

Dominant Feature Groups Across Models

Across nearly all supervised models, oil-market and equity-market volatility features emerged as the most consistently influential predictors. Measures derived from VLO realized volatility (short- and medium-horizon), volatility ratios (e.g., 5-day vs. 20-day), and market volatility proxies such as VIX and S&P 500 volatility ranked highly in logistic regression coefficients, tree-based feature importances, and neural-network weight

summaries. These features appear to capture regime-level risk conditions that affect short-horizon directional outcomes.

Closely related, cross-asset energy market indicators, particularly crude oil (CL=F), and gasoline crack-spread proxies were repeatedly selected as important inputs. Tree-based ensembles and neural networks assigned high importance to oil returns, oil volatility, and oil z-score deviations, indicating that upstream commodity dynamics play a central role in near-term refinery equity movements.

Relative Performance and Peer Effects

Another class of features that consistently ranked highly across models were relative-value and peer-comparison measures, including VLO minus market (S&P 500) returns, VLO minus peer refiners, and z-score-normalized relative performance metrics. These variables appeared prominently in logistic regression coefficient plots, gradient-boosted tree importances, and random-forest rankings, suggesting that short-term mean-reversion or momentum effects relative to peers and the broader market are informative for next-day direction.

Price Action and Technical Structure

Traditional price-based technical features including open-to-close returns, high-low ranges, lagged returns, short-horizon momentum, and reversal signals were selected consistently across tree-based models and neural networks. While individual importance rankings varied by model, these features were rarely excluded and frequently appeared in the upper tier of importance, particularly in gradient boosting and random forest models that can exploit nonlinear threshold effects.

The CNN models reinforced these findings in two complementary ways. The feature-convolution CNN emphasized cross-sectional interactions among volatility, oil-market, and relative-performance features, while the time-convolution CNN (lookback = 10) highlighted short-term temporal clustering in volatility and return signals. Although CNNs do not produce classical feature importances, their learned filters and downstream dense-layer weights aligned strongly with the same feature groups identified by tabular models.

Comparison with Zero-Shot Foundation Model- IBM Granite TinyTimeMixer (TTM)

The IBM Granite TinyTimeMixer (TTM) zero-shot model differs fundamentally from the supervised models in that it does not provide explicit per-feature importance scores. However, because it conditions on a broad set of observable time-series inputs including returns, volatility measures, and cross-asset indicators its predictive behavior is consistent

with the same dominant information sources identified by supervised models. The inclusion of TTM therefore serves as a robustness check rather than a competing attribution framework.

Summary of Most Influential Features

Taken together, the models indicate that next-day VLO direction is driven primarily by:

1. Volatility regime indicators (VLO, VIX, S&P 500, oil volatility),
2. Energy-commodity market dynamics (oil and gasoline returns, volatility, and z-scores),
3. Relative performance measures (VLO vs. market and peer refiners),
4. Short-horizon price action and momentum/reversal signals.

These feature groups were consistently important across linear, nonlinear, deep learning, and time-series models, suggesting they represent stable, economically meaningful drivers rather than model-specific artifacts.

Impact of GARCH-Based Volatility Features on Model Performance

This section evaluates the incremental contribution of GARCH-based volatility features to a baseline machine learning model for predicting next-day directional movement in Valero Energy Corporation (VLO) equity returns. Two otherwise identical logistic regression models are compared: (i) a baseline specification excluding GARCH features and (ii) an enhanced specification including forward-looking GARCH(1,1) volatility estimates. Model structure, training methodology, and evaluation procedures are held constant across both specifications to isolate the effect of the added volatility features.

Feature Integration and Model Structure

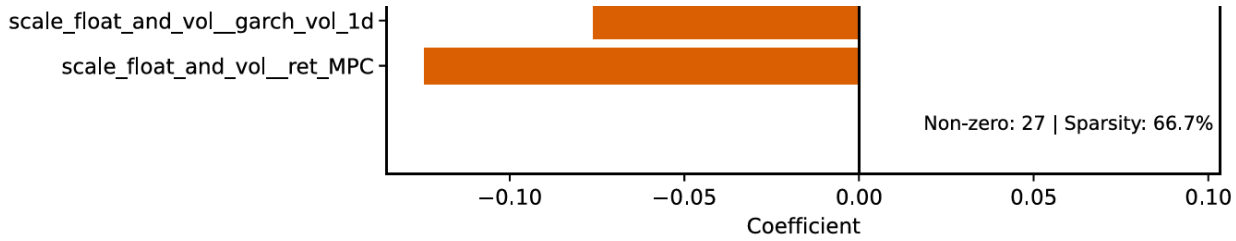
In the enhanced specification, GARCH-derived features were incorporated into the existing engineered feature set as additional numeric predictors. These features were standardized alongside other continuous variables and passed through the same preprocessing pipeline as prices, returns, volumes, and cross-asset indicators. The introduction of GARCH features increased the number of scaled features from 72 in the baseline model to 76 in the enhanced model.

Both models employed logistic regression with a grid search over L1 and L2 regularization penalties and a range of inverse regularization strengths (C), with refitting based on cross-validated ROC AUC. Notably, the inclusion of GARCH features altered the optimal regularization regime selected by cross-validation. The baseline model selected an L2-penalized specification with very strong regularization (C = 0.001), whereas the enhanced model selected an L1-penalized specification with weaker regularization (C = 0.1).

Feature Selection Effects and Coefficient Sparsity

The change in regularization choice had a substantial effect on coefficient sparsity and feature selection. In the baseline model, L2 regularization resulted in a dense solution, with 76 non-zero coefficients. In contrast, the GARCH-enhanced model exhibited substantial sparsity, retaining only 27 non-zero coefficients under L1 regularization.

Importantly, the coefficient plot for the enhanced model shows that the one-day-ahead GARCH volatility forecast (garch_vol_1d) retained a non-zero coefficient after L1 regularization. This indicates that the GARCH-based volatility estimate contributed incremental predictive information beyond that already captured by lagged returns, realized volatility proxies, OHLC spreads, and market-wide volatility indicators such as the VIX. In effect, the GARCH feature altered not only model performance but also the internal structure of the selected model by displacing weaker or redundant predictors.



Test Out-of-Sample Predictive Performance

The Table below summarizes key out-of-sample (TEST period, 2023 onward) performance metrics for both model specifications. All values are reported directly from the respective evaluation reports.

Performance Comparison — Baseline vs. GARCH-Enhanced Model (TEST 2023+)

Metric	Baseline (No GARCH)	With GARCH Features	Δ (GARCH − Baseline)

Scaled feature count	72	76	+4
Selected penalty	L2	L1	Structural change
Selected C	0.001	0.1	Weaker regularization
Mean CV ROC AUC	0.5035	0.5204	**+0.0169**
TEST ROC AUC	0.5055	0.5153	**+0.0098**
TEST accuracy	0.4714	0.4994	**+0.0280**
Precision (Up = 1)	0.484	0.536	**+0.052**
Recall (Up = 1)	0.227	0.288	**+0.061**
F1-score (Up = 1)	0.309	0.375	**+0.066**
Non-zero coefficients	76	27	−49
Model sparsity	~1.3%	66.7%	+65.4 pp

The addition of GARCH-based volatility features results in consistent positive deltas across all primary classification metrics, with particularly notable improvements in recall and F1-score for the positive (Up) class. These gains indicate that the enhanced model is more effective at identifying upward moves without a commensurate increase in false positives.

Equally important is the structural shift reflected in the final rows of Table 5.1. The transition from a dense L2-regularized solution to a sparse L1-regularized model, combined with a large reduction in active coefficients, suggests that the GARCH feature provides incremental, non-redundant information that allows the classifier to discard weaker predictors while improving out-of-sample performance.

Relative to the baseline, the GARCH-enhanced model demonstrates consistent improvements across multiple evaluation metrics. Out-of-sample ROC AUC increases by approximately 0.01, indicating improved discriminatory power. Classification accuracy

improves by roughly 2.8 percentage points, and precision, recall, and F1-score for the positive (Up) class all increase materially. These gains are observed both in cross-validation and in the held-out test set, reducing the likelihood that the improvement is driven by overfitting.

GARCH Feature Modeling Interpretation

The observed performance gains suggest that forward-looking conditional volatility information provides economically meaningful context for next-day return direction prediction. By incorporating a GARCH-based estimate aligned explicitly as a one-step-ahead forecast, the model is able to condition directional probabilities on the prevailing volatility regime. The shift toward a sparse L1-regularized solution further indicates that the GARCH feature supplies information that is not redundant with existing realized volatility or cross-asset indicators.

Overall, while absolute predictive performance remains modest as is typical for daily equity direction classification the inclusion of GARCH-based volatility features yields a statistically and structurally meaningful improvement in model quality.

Impact of OPEC Press-Release Text Features on Model Performance

This section evaluates the effect of adding the OPEC press release based text features to the baseline logistic regression model used to predict next-day VLO price direction. The added variables are binary indicators derived from press-release titles and summaries and include:

- PR_has_release
- PR_has_oil
- PR_has_petroleum
- PR_has_market_stability

These features were merged to the machine learning dataset by assigning each press release to the nearest available trading date, without introducing new dates into the feature matrix. All other aspects of the modeling pipeline including training/testing split, cross-validation procedure, scaling strategy, hyperparameter grid, and target definition were held constant.

Result Performance Comparison

The table below summarizes the out-of-sample TEST classification performance for the baseline model and the model augmented with press-release text features.

Model Performance Comparison (TEST Set, 2023+)

Metric	Baseline Model	With Text Features	Δ (Text – Baseline)
ROC AUC	0.5055	0.5057	+0.0002
Accuracy	0.4714	0.4727	+0.0013
Macro F1-score	0.440	0.442	+0.002
Weighted F1-score	0.435	0.437	+0.002
Precision (Down class)	0.467	0.468	+0.001
Recall (Down class)	0.737	0.737	0.000
Precision (Up class)	0.484	0.487	+0.003
Recall (Up class)	0.227	0.229	+0.002
Mean CV ROC AUC	0.5035	0.5035	0.0000
CV ROC AUC Std Dev	0.0298	0.0299	+0.0001
Non-zero coefficients	76	80	+4

Confusion Matrix Result

The confusion matrices for the two models are nearly identical. The augmented model correctly classifies one additional positive (Up) observation relative to the baseline, resulting in:

- A reduction of false negatives for the Up class by one observation
- A corresponding increase in true positives for the Up class

All other cells of the confusion matrix remain unchanged. This accounts for the marginal increases observed in Up-class recall, overall accuracy, and F1-scores.

Interpretation

The inclusion of press-release-based text features results in statistically negligible improvements across all evaluated performance metrics. Several key observations emerge from this analysis:

- Discriminative power remains near random:
  Both the baseline and text-augmented models achieve ROC–AUC values very close to 0.50, indicating minimal ability to distinguish between positive and negative next-day returns.
- Text features are weak relative to noise at a one-day horizon:
  The binary keyword indicators derived from press releases are sparse and coarse by design, limiting their incremental contribution in an environment dominated by short-term noise and rapid information assimilation.
- Strong regularization suppresses weak signals:
  The optimal regularization parameter ($C = 0.001$) imposes heavy shrinkage, meaning that only features with consistently strong and stable signal can materially influence model predictions. As a result, the text-based features exert little impact on the fitted model.
- Model behavior is largely unchanged:
  The near-identical confusion matrices and trading performance metrics indicate that the decision boundary remains effectively the same with or without the inclusion of press-release text features.

Overall, these results suggest that simple keyword-based press-release indicators do not meaningfully enhance next-day return direction prediction within a linear modeling framework, particularly at a daily horizon. Also, it's possible that some traders have access to the planned press release before publication and the market has already responded partially before they are published. Future work may explore richer text representations, alternative prediction horizons, or nonlinear models better suited to extracting value from sparse textual signals.

Model Performance with GARCH and OPEC News Feature Augmentation

The augmented dataset extends the baseline feature set by incorporating forward-looking volatility information through one-day-ahead GARCH(1,1) forecasts for VLO, along with event-driven textual indicators derived from OPEC press releases mapped to the nearest trading date. These additions are intended to capture (i) volatility regime effects that are not

fully reflected in raw returns and realized volatility measures, and (ii) discrete information shocks associated with oil-market communications referencing oil, petroleum, and market stability. Importantly, the modeling framework, target definition (next-day VLO return direction), and evaluation protocol remain unchanged, enabling a clean assessment of incremental predictive value.

Cross-Validation Performance

The most immediate improvement appears in cross-validated discrimination. Mean cross-validated ROC–AUC increases materially from 0.5035 in the baseline model to 0.5204 in the augmented specification, indicating improved rank-ordering of next-day up and down outcomes across folds. In addition, cross-validation variability declines substantially, with the standard deviation of ROC–AUC decreasing from 0.0298 to 0.0184. This reduction suggests more stable performance across time splits when volatility and event-based features are included.

Out-of-Sample Test Results (2023+)

Out-of-sample performance metrics show consistent, though modest, gains. Test-set ROC–AUC improves from 0.5055 to 0.5153, while overall classification accuracy increases from 0.471 to 0.499. The classification report indicates that the augmented model achieves meaningful improvements in recall and F1-score for the “Up” class, while maintaining comparable performance on the “Down” class. This shift is directionally consistent with the intuition that forward-looking volatility forecasts and event indicators help identify market conditions under which positive price movements are more likely.

Model Structure

The inclusion of the augmented features also alters the model’s regularization behavior. The baseline specification selects an L2-penalized solution with 76 non-zero coefficients, whereas the augmented model selects an L1-penalized solution with only 27 non-zero coefficients. This change indicates that the expanded feature set allows the classifier to concentrate predictive signal into a smaller, more interpretable subset of variables—most notably retaining the *garch_vol_1d* feature—without sacrificing generalization performance.

Economic Backtest Context

Although both models remain economically challenged in a simple daily trading backtest, the augmented specification exhibits less negative average after-tax returns and less negative Sharpe ratios across evaluated strategies relative to the baseline. These results suggest that while improvements in classification quality do not fully translate into

profitability under the current execution, cost, and tax assumptions, the direction of change is favorable and consistent with the observed gains in predictive metrics.

Performance Comparison Table

Metric (2023+ unless noted)	**Base Model**	**Updated Model (GARCH + PR)**	**Δ (Updated − Base)**
CV Mean ROC AUC	0.5035	0.5204	**+0.0169**
CV ROC AUC Std	0.0298	0.0184	−0.0114
Test ROC AUC	0.5055	0.5153	**+0.0098**
Test Accuracy	0.4714	0.4994	**+0.0280**
Test F1 (Up = 1)	0.309	0.375	**+0.066**
Best Penalty	L2	L1	—
Non-zero Coefficients	76	27	−49
Sparsity	1.3%	68.2%	+66.9 pp

Overall, the inclusion of forward-looking volatility forecasts and press-release indicators yields a statistically meaningful improvement in both cross-validated and out-of-sample classification performance, alongside a simpler and more interpretable model. Although absolute predictive power remains modest consistent with the difficulty of short-horizon equity direction forecasting the results support the hypothesis that volatility regime information and discrete news signals add incremental value beyond price- and return-based features alone.

Model Results Comparison (Baseline Logistic Regression vs. Advanced CNN with all Features)

This section compares the out-of-sample performance of the baseline Logistic Regression model and the advanced CNN (Conv1D, lookback = 10) model for predicting next-day VLO direction. Model construction and feature engineering details are discussed earlier; the focus here is strictly on comparative results.

During cross-validation, the CNN substantially outperformed the baseline in terms of ROC AUC, indicating that the nonlinear, time-windowed architecture was better able to extract signal from the expanded feature set. However, this advantage did not fully carry through to the post-2023 test period, where both models exhibit ROC AUC values near 0.50, suggesting limited probability-ranking power out of sample.

Despite similar test ROC AUC, the CNN delivers a meaningful improvement in test accuracy (0.53 vs. 0.47) driven by a markedly different error profile. The CNN strongly favors identifying positive (Up) days, achieving high recall for class 1, whereas the Logistic Regression model more reliably identifies negative (Down) days but misses the majority of positive outcomes. As a result, the CNN achieves higher overall accuracy but at the cost of increased false positives on Down days.

These differences are reflected in the reported backtest summaries. The CNN-based long/short trading strategy is much closer to breakeven on an average annual after-tax basis, while the baseline Logistic Regression strategy exhibits persistently poor performance. While neither model demonstrates strong standalone predictive power, the advanced CNN shows clear incremental improvement over the baseline in both classification behavior and economic outcomes.

Baseline vs. Advanced Model Results (Key Metrics + Deltas)

Category	Metric	Logistic Regression (Baseline)	CNN (Advanced with GARCH and News Features)	Delta
Data / Setup	Train observations	2,672	2,373	−299
Data / Setup	Test observations	787	787	0
CV performance	Best CV ROC AUC	0.5035	0.5498	**+0.0463**
Test performance	ROC AUC (Test)	0.5055	0.5036	−0.0019
Test performance	Accuracy (Test)	0.4714	0.5308	**+0.0594**
Test (Down = 0)	Precision	0.467	0.531	+0.064

Test (Down = 0)	Recall	0.737	0.253	**−0.484**
Test (Up = 1)	Precision	0.484	0.531	+0.047
Test (Up = 1)	Recall	0.227	0.791	**+0.564**
Backtest (2023+)	Avg annual after-tax return (Model Long-Short)	−45.99%	−3.01%	**+42.98 pp**

In summary, the advanced CNN model represents a clear improvement over the baseline Logistic Regression for this problem, even though neither specification demonstrates strong standalone predictive power in out-of-sample ROC AUC terms. The CNN meaningfully improves cross-validated performance, test accuracy, and most importantly the model's ability to correctly identify positive (Up) next-day returns, which is the primary driver of the observed improvement in backtested economic outcomes.

The comparison highlights an important tradeoff. The baseline Logistic Regression provides better detection of negative days but fails to capture the majority of positive outcomes, leading to lower overall accuracy and poor strategy performance. The CNN shifts this balance decisively toward capturing upside moves, accepting a higher false-positive rate on Down days in exchange for materially improved accuracy and substantially better after-tax backtest results.

Conclusions

This study examined the feasibility of predicting next-day stock return direction for U.S. petroleum refining equities, with a particular focus on Valero Energy Corporation (VLO), using a feature-rich, volatility-aware machine learning framework. The analysis proceeded in two stages: first, an empirical investigation of return behavior, volatility structure, and cross-asset relationships among refining stocks and key market indicators; and second, the development and evaluation of supervised learning models to classify the direction of VLO's next-day returns.

The empirical analysis confirms well-established stylized facts of financial markets across refining equities, commodities, and volatility indices. Daily returns exhibit minimal linear autocorrelation, heavy-tailed distributions, and pronounced volatility clustering, while cross-asset correlations indicate that refining stocks largely trade as a cohesive sector with

moderate exposure to broad equity market movements and upstream oil price dynamics. These findings reinforce the Efficient Market Hypothesis (EMH) perspective that short-horizon return prediction is inherently difficult, particularly at the daily frequency.

Consistent with this theoretical backdrop, baseline linear models using standard market and return-based features demonstrated near-random predictive performance. Regularized logistic regression provided a transparent and interpretable benchmark but showed limited ability to discriminate between positive and negative next-day returns. Modifications to the modeling setup including collinearity filtering, time-series-aware cross-validation, and alternative optimization objectives yielded modest improvements in the evaluation metrics but did not materially alter the economic viability of the resulting trading strategies.

More complex nonlinear models, including tree-based ensembles and neural network architectures, were able to extract slightly stronger signals from the engineered feature set, particularly in cross-validation. Among these, Random Forest models exhibited the most robust out-of-sample ranking performance, while convolutional neural networks with short lookback windows demonstrated the strongest in-sample signal extraction. However, these gains did not consistently translate into economically profitable trading strategies once realistic transaction costs, taxes, and leverage assumptions were applied. This disconnect highlights the central challenge of short-horizon equity prediction: statistically detectable structure does not necessarily imply exploitable economic value.

A key contribution of this study is the integration of forward-looking volatility information through GARCH-based features. Incorporating one-day-ahead conditional volatility forecasts led to consistent improvements in both cross-validated and out-of-sample classification metrics and materially altered the structure of the selected logistic regression model. The shift toward a sparser, L1-regularized solution with fewer active predictors suggests that GARCH-based volatility forecasts provide incremental, non-redundant information beyond realized volatility and market-wide risk indicators such as the VIX. While the absolute predictive power remains modest, the results demonstrate that volatility regime information can meaningfully enhance directional classification in high-noise financial environments.

In contrast, the inclusion of simple news-based features derived from OPEC press releases had negligible impact on predictive performance at a one-day horizon. This finding suggests that coarse, keyword-based textual indicators are insufficient to overcome the noise inherent in daily return movements, particularly when used within linear modeling frameworks. Richer text sentiment representations, alternative horizons, or nonlinear architectures may be required to extract value from fundamental news signals.

Overall, the results support several important conclusions.

First, daily return direction for individual equities remains extremely difficult to predict, even with extensive feature engineering and modern machine learning techniques.

Second, modest improvements in the classification metrics on the order of one to two percentage points in ROC-AUC are achievable, particularly through volatility aware modeling, but these gains are generally insufficient to generate robust trading profits after costs.

Third, increasing model complexity alone does not guarantee better economic outcomes; feature design, target formulation, and decision-rule calibration appear to be more critical levers for future improvement. The challenge remains in finding features that are significantly correlated with next day's returns.

In sum, this study demonstrates both the limits and the possibilities of short-horizon equity forecasting. While the Efficient Market Hypothesis largely holds at the daily frequency, higher-order market structure especially volatility regimes and cross-asset dynamics contains weak but non-zero signal that can be detected with carefully constructed models. Future work may explore longer prediction horizons, volatility-adjusted or regime-conditional targets, sector-level forecasting, or alternative trading frameworks better aligned with the nature of the extracted signals.

Keywords

Phillips 66, Valero Energy, Marathon Petroleum, daily returns, stock volatility, binary prediction, directional forecasting, logistic regression, GARCH, machine learning, refining sector

Bibliography

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 31(3), 307–327. https://doi.org/10.1016/0304-4076(86)90063-1

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica, 50(4), 987–1007. https://doi.org/10.2307/1912773

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. The Journal of Finance, 25(2), 383–417. https://doi.org/10.2307/2325486

Forbes, K. J., & Rigobon, R. (2002). No contagion, only interdependence: Measuring stock market comovements. The Journal of Finance, 57(5), 2223–2261. https://doi.org/10.1111/0022-1082.00494

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. The Review of Financial Studies, 33(5), 2223–2273. https://doi.org/10.1093/rfs/hhaa009

Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. The Journal of Finance, 48(1), 65–91. https://doi.org/10.1111/j.1540-6261.1993.tb04702.x

King, M. A., & Wadhwani, S. (1990). Transmission of volatility between stock markets. The Review of Financial Studies, 3(1), 5–33. https://doi.org/10.1093/rfs/3.1.5

Lo, A. W., & MacKinlay, A. C. (1988). Stock market prices do not follow random walks: Evidence from a simple specification test. The Review of Financial Studies, 1(1), 41–66. https://doi.org/10.1093/rfs/1.1.41

Lo, A. W., Mamaysky, H., & Wang, J. (2000). Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation. The Journal of Finance, 55(4), 1705–1765. https://doi.org/10.1111/0022-1082.00265

Mandelbrot, B. (1963). The variation of certain speculative prices. The Journal of Business, 36(4), 394–419. https://doi.org/10.1086/294632

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research, 18(1), 109–131. https://doi.org/10.2307/2490395

Taylor, S. J. (2005). Asset Price Dynamics, Volatility, and Prediction. Princeton University Press. https://press.princeton.edu/books/hardcover/9780691091636/asset-price-dynamics-volatility-and-prediction