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

MSDS 492- Analysis of Financial Markets

Northwestern University

2/13/2026

By David Van Dyke

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

Contents

Abstract

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 summary" 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 years (≈ 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_summary: Indicates (1/0) whether the phrase "market summary" appears in the press release title/summary on day t. OPEC frequently publishes "Market Summary" 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 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.

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

Conclusions

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