# Using Volatility as an Event Based Indicator to Predict Short and Medium Term Stock Trends

Garrett John Henry Pittman

University of Nottingham

September 2025

*A Dissertation presented in part consideration for the degree of "MSc Business Analytics"*

# Abstract

Volatility is a ubiquitous signal in equity markets. On one hand it is responsible for market inefficiency, but if harnessed appropriately can be an incredibly strong driver of return. Most studies regarding the equity market either attempt, futilely, to predict volatility itself or simply treat it as one of many regressor in larger return models. This dissertation takes a different route: it treats **volatility shocks as event triggers** and asks whether machine-learning models can forecast the **direction** and **magnitude** of post-shock equity moves over short and medium horizons using firm-level and macro time-series features. General volatility is nearly impossible to predict, but using volatile events as a proxy to predict whether a company will rebound, revert to its mean, or continue on the volatile spike's trend is plausible. Using a two-year panel of S&P 500 constituents (2023–2024) and term-structure data (10y–2y Treasury spread), we build leakage-safe features, enforce strict chronological splits, and compare multiple learners, including a constant-mean benchmark, Random Forest, CatBoost, an LSTM sequence model, and gradient-boosted trees (XGBoost).

Two headline results emerge. First, **predictability rises with horizon**: models deliver **meaningful medium-horizon skill**, with **directional accuracy (DA) up to ~70%** at 20 trading days on the validation window, while very short horizons remain near noise. Second, predictability is **regime-sensitive**. Out-of-sample results on a later test window, coinciding with the **yield-curve's shift from inversion toward de-inversion/steepening** and a higher share of **large shocks**, settle in the **mid-50% DA** range, indicating attenuation rather than disappearance of signal. A polarity analysis further reveals that **negative shocks are easier to forecast**, with **DA approaching ~80%** at the 20 day forecast horizon, consistent with post-shock mean-reversion and liquidity effects.

Methodologically, parsimonious lagging of key signals (t−1, t−5, t−10, t−20) allows tree ensembles to capture local trend and regime interactions without over-learning non-event days; in contrast, the LSTM tends to regress toward zero in abnormal post-shock states. We evaluate models on **RMSE** (magnitude fidelity) and **DA** (decision relevance) and emphasize that, for sign-based strategies, **DA is the economically salient metric**, with selective execution offering a practical path to application.

The contribution is twofold: (i) it **joins** two streams usually studied apart, **return prediction** and **volatility prediction**, by using volatility **as an event signal** to forecast what happens **after** shocks; and (ii) it documents **when** and **under which regimes** such predictions are most credible, offering a clear agenda for regime-aware, polarity-aware extensions.

**Keywords:** event study; directional accuracy; yield curve; XGBoost; volatility shocks; regime shift; stock market; machine learning; business analytics.

# Table of Contents

**Appendices**

A. Engineered feature computations (*avg_return, avg_abs_return, avg_raw_return, rolling_market_vol_10d, return_std, RSI*)

B. Feature correlation heatmap (final features)

C. Tuned hyperparameters used in the final XGBoost model

D. SHAP feature-importance plots

E. Permutation-importance output for 20-day target

# Introduction

Financial markets are noisy, non-linear, and non-stationary, yet investors and researchers repeatedly ask a simple question: *what happens next after a large move in the markets?* While prior work shows that machine learning (ML) can extract weak but exploitable signals from returns (e.g., Rossi, 2018; Lehrer et al., 2020; Tiwari et al., 2025), most studies either attempt to predict volatility or include it as a continuous regressor in a more broad, stock prediction model. This dissertation takes a different perspective: it uses volatility shocks as event-based indicators and investigates whether ML can forecast the direction of equity returns after such shocks over practical horizons.

**Research question:** *Can machine-learning models predict short- and medium-term stock price behavior following extreme volatility shocks, using firm-level and market-wide time-series features?*

**Why this matters and what is new about this study**

Predicting returns and predicting volatility are often pursued independently. In this study, we join them by treating volatility as a trigger, not a regressor, and asking whether post-shock direction is predictable when models are conditioned on the event. This framing clarifies mechanism (what follows a shock) and allows us to study polarity asymmetries (positive vs negative shocks) and macro regimes in a disciplined way. It also addresses a gap in the literature that tends to emphasize "meaningful" volatility while filtering out transitory spikes; here we study all shocks that clear an objective threshold, acknowledging that even so-called "meaningless volatility" can move prices and have an effect on the investors, the market, and individual companies (Mansilla-Lopez et al., 2025).

**Data and event construction**

We assemble a daily panel for S&P 500 constituents over 2023–2024 from S&P Capital IQ (prices, volume, market cap, selected fundamentals) and pair it with the 10y–2y Treasury spread (T10Y2Y) as a macro proxy. Shocks are defined at the firm level as days where the return's z-score exceeds |2| relative to each stock's global two-year baseline; contiguous shock streaks are de-duplicated by keeping only the last day as the shock indicator. For each shock, we form targets as forward returns over 1, 5, and 20 trading days. The sample is split chronologically into Train (2023/02/02 to 2024/06/28), Validation (2024/07/01 to 2024/08/30), and Test (2024/09/03 to 2024/12/30) data.

**Methods in brief**

We compare a constant-mean baseline, Random Forest, CatBoost, an LSTM sequence model (using a 20-day pre-shock window), and XGBoost. Because tree models are not sequence-aware, we inject short-history information via parsimonious lags (t−1, −5, −10, −20) of key signals (returns, prices, volume, RSI; market-stress summaries; macro level/change; size/fundamentals). All features are built as-of and leakage-safe. Hyperparameters are tuned by randomized search with group-aware folds by entity where appropriate. We evaluate using RMSE (magnitude) and Directional Accuracy (DA) (sign), noting that DA maps directly to decision utility in sign-tilt strategies, while RMSE documents calibration, tail errors, and pure accuracy of the models.

**Headline findings**

Predictability in this setting is horizon-dependent and regime-sensitive. On the validation window, models achieve medium-horizon skill, DA up to ~70% at 20 days, with modest (~60%) gains at 5 days and little signal at 1 day, demonstrating that event-conditioned direction is learnable. On a later test window, coinciding with the yield-curve's move from prolonged inversion toward de-inversion/steepening and a higher share of positive shocks, DA attenuates into the high-50s, indicating that skill persists but depends on the macro mix and shock composition. The model was regime aware, but was only trained on negative yield spread, bearish regime (this limitation of data and model is extrapolated in the discussion chapter). A polarity split shows negative shocks are easier to predict (DA approaching ~80% at 20 days), consistent with mean-reversion and liquidity mechanisms.

**Why the LSTM trails trees here**

In this event-study design the signal of interest is sparse relative to abundant normal days; the LSTM's 20-day encoder tends to learn level dynamics and regress toward zero when faced with the abnormal post-shock state. More simply put, the LSTM learns the 20 days prior to the shock well, but these days tend to be 'normal'. When the LSTM is then tasked with predicting what will happen after the shock, when returns tend to be 'non-normal' the model tends to give a 'normal' prediction which is usually around 0. In contrast, XGBoost, fed compact lags and macro/context features, discovers stable, non-linear decision boundaries that translate into stronger DA at further time horizons. This does not contradict the broader literature's success of deep models; it underscores data/feature-regime fit. Most of the literature's models are predicting normal days, with abnormal days as the exception. This study is quite the opposite.

**Implications**

For practice, DA is the salient metric; selective execution (trading only when $|\hat{y}|$ is large) and polarity-aware models are natural ways to sharpen performance. For research, the results point to regime-aware learning (conditioning on curve states), richer macro and text/sentiment features, and longer multi-regime histories to test stability. Using S&P 500 (2023–2024), we demonstrate that volatility shocks can be productively routed into medium-horizon sign forecasts, and we show when that edge is strongest.

**Roadmap**

Section 2 reviews related work on ML for returns and on volatility as a signal. Section 3 details the data, event construction, and features. Section 4 presents the modeling approach and evaluation design. Section 5 reports results across horizons, polarity, and regimes. Section 6 discusses implications, limitations, and future work. Section 7 concludes.

# Literature Review

Predicting stock market returns is a cornerstone pursuit in financial research, made increasingly feasible by recent advances in machine learning techniques. Volatility has long been recognized as a leading indicator of stock market movements, with substantial literature demonstrating its predictive potential (Lehrer et. all 2020, Hung 2015, Rossi 2018). However, much of the existing research either focuses on forecasting volatility itself or incorporates volatility as a continuous regressor within broader return prediction models (Mansilla-Lopez et. all 2025). In Mansilla-Lopez's comprehensive review of literature regarding volatility in the market, almost every one of the most cited research papers regarding volatility uses volatility strictly as a predictor for stock prices or as the output variable. Our goal is to expand the research to, instead of using volatility as one of many input variables, see if volatility can act as a leading indicator to more accurately predict the price movements of stocks after price shocks. According to Mansilla-Lopez's review, volatility, when used in conjunction with other company specific and macroeconomic variables, is typically the most influential variable when predicting stock price. It follows then, that if we pull volatility out of the equation and instead use it as an event-based signal, we will find what other factors have the largest influence on stock market prices during volatile swings in the market and we will find if volatility has more useful predictive power when it is used as a signal rather than an input variable. The current literature lacks nuance, often implying the self-evident point that volatile markets lead to price changes, without offering meaningful insight into what happens next; our research addresses this by using volatility as a diagnostic signal to forecast post-shock behavior, rather than conflating it with the outcome.

The following literature review explores how existing research on stock return prediction using machine learning, alongside the predictive role of volatility, informs and validates our investigation into using **volatility shocks as event-based signals** to forecast subsequent price behavior in equities. By bridging these two domains, this study positions itself within a growing yet underexplored niche in financial forecasting, focusing on whether volatility events, when paired with additional equity and market variables, can lead to increased returns and above-standard predictability.

To create a comprehensive review of the well-researched subjects, the literature review will proceed in three sections. First, it will analyze literature on stock market prediction using machine learning, with an emphasis on methodology and technique found in previous research that will inform our methodology and model. This section will identify the most recent and advanced methods used by researchers to predict returns, highlighting which models, data sources, engineered variables, and preprocessing steps contribute most to predictive accuracy.

Second, the review will examine volatility, moving from foundational definitions to comprehensive reviews illustrating its role as a predictive indicator. The literature is ripe with

two main focuses on volatility. The first is predicting volatility, where research attempt various strategies to predict when volatility within the market will occur, and the magnitude of its effect. The second uses volatility as a tool within the market, typically to predict equity or market returns. This section will both demonstrate that volatility is a powerful tool for capturing equity price and explain the key factors found that help define volatility's function and use within the markets.

Finally, the review will explore the intersection between these topics and expose gaps in the literature. It will argue that while prior research has largely focused on predicting volatility or incorporating it into return models, there remains a gap in leveraging volatility shocks as event indicators to predict post-event equity behavior. Volatility still remains largely unpredictable, but are the events that follow volatility better suited to machine learning's predictive capability? This research aims not only to test whether mean-reverting behavior is likely to follow volatility shocks but also to explore broader patterns of price behavior, developing a **novel, well-informed approach to using volatility as a signal within equity return prediction models**.

## 1. Predicting Stock Market Returns Using Machine Learning

Machine learning has become an increasingly important tool in financial forecasting, offering methods capable of handling the noisy, non-linear, and non-stationary nature of financial time series data. Tiwari et al. (2025), Rossi (2018), Pare and Natarajan (2025), and Lehrer (2020) all highlight that machine learning methods now represent the most advanced and effective approach for forecasting equity returns. These techniques, particularly deep learning and ensemble models, consistently outperform traditional statistical models by capturing higher-order interactions between features that would otherwise be lost in linear frameworks. Tiwari et al. (2025) further argue that machine learning's flexibility also allows researchers to embed behavioral finance concepts, such as investor biases, while preserving predictive accuracy and enhancing interpretability. Taken together, these studies form a strong consensus in the modern literature that machine learning is the most viable approach for forecasting stock returns.

A central challenge in machine learning-based return prediction is the process of feature selection. Patel and Natarajan (2025) provide a comprehensive review of technical indicators that have consistently demonstrated value across datasets, including the Simple Moving Average (SMA), Exponential Moving Average (EMA), Moving Average Convergence Divergence (MACD), Relative Strength Index (RSI), Bollinger Bands, momentum indicators, and volume-based metrics. Rossi (2018) complements this by listing several macroeconomic and firm-level features that have historically shown predictive power, such as the default spread, inflation rate, industrial production growth, dividend yield, book-to-market ratio, and return on equity. More recently, volatility has emerged as one of the most influential variables in forecasting stock prices. Lopez-Gil et al. (2024) found that volatility was the single most important predictor in their models, while Campisi (2024) further reinforces this by showing that

its inclusion significantly enhances model accuracy. Additionally, Lopez-Gil et al. demonstrate that deep learning models, particularly neural networks, outperform traditional models when volatility is used as a core feature. This research will therefore incorporate the features identified in the literature (wherever available and possible) into both neural network and ensemble models to evaluate relative performance.

With respect to methodology, several studies provide useful guidance on how best to structure and evaluate time series prediction models. Rossi (2018) introduces the use of expanding window training and testing, a strategy that acknowledges the temporal dynamics and data leakage risks inherent in financial forecasting. This approach will be used in the current study to mirror real-world predictive scenarios. Furthermore, Hung (2015) empirically tests different forecast horizons and concludes that a 21-day prediction window provides the most robust performance. Based on this, our study will explore forecasting horizons ranging from short-term (e.g., 5-day) up to 20-day windows to assess how predictive accuracy varies over time.

## 2. Volatility in Financial Markets

Volatility, broadly defined as the degree of variation in the price of a financial instrument over time, plays a central role in both risk assessment and return forecasting. In the literature, multiple forms of volatility are identified, each contributing differently to price behavior. *Historical volatility* refers to the realized standard deviation of past returns and is calculated directly from observed price changes over time (Berger et al., 2009). *Implied volatility*, by contrast, is a forward-looking measure derived from options pricing, representing the market's expectations of future risk (Rhoads, 2011). *Transitory volatility* captures short-term fluctuations in asset prices that arise from temporary shocks or noise rather than underlying fundamentals (Lo and MacKinlay, 1988). *Cyclical volatility* refers to volatility patterns driven by macroeconomic cycles, typically rising during recessions and falling during expansions (Sill, 1993). Lastly, *geopolitical volatility* captures market turbulence stemming from exogenous political events, including wars, elections, and international conflicts, which are particularly challenging to predict due to their sudden and non-economic nature (Mansilla-Lopez et al., 2025).

While the existing literature focuses primarily on what might be considered **"***meaningful volatility"*, volatility driven by macroeconomic fundamentals, systemic crises, or political shocks that is mainly identified as significant geopolitical volatility (Mansilla-Lopez et al., 2025), it tends to overlook more ambiguous or short-lived volatility events. Although this study does not explicitly differentiate between these volatility types in its model architecture, it takes a comprehensive and objective approach by treating all volatility spikes as potential information signals, regardless of origin. The term "meaningless volatility" is introduced here not as a classification for modeling, but to acknowledge that current literature underemphasizes volatility events not tied to clearly identifiable economic or political catalysts. One function of this

research is to broaden the literature by putting an emphasis on the importance of all types of volatility, not just significant ‘meaningful’ volatility.

This broader framing is motivated by foundational work from Sill (1993), who examined volatility from a macroeconomic lens and noted the limitations of forecasting it. Sill showed that even when using variables such as inflation or money supply growth, only about 2% of future volatility could be explained, rendering direct volatility prediction highly uncertain. His work laid the groundwork for understanding volatility's relationship with broader economic behavior, noting its heightened presence during recessions and its influence on consumption, savings, and investment.

Building on this foundation, modern literature has improved dramatically in its ability to model and use volatility in predictive contexts. Mansilla-Lopez et al. (2025) offer a comprehensive review of how volatility is measured and forecasted using machine learning. Their study categorizes fifteen drivers of volatility across six domains; news, politics, irrationality, health, economics, and war, and distinguishes between historical and implied volatility as key constructs. They also emphasize the value of high-frequency data in improving the precision of volatility measurement (Berger et al., 2009). While these advances demonstrate the evolution of volatility modeling since Sill’s early findings, significant limitations persist. In particular, certain volatility-inducing events, especially geopolitical shocks, remain inherently unpredictable. Therefore, this research chooses not to predict volatility itself or use volatility as an input variable, but rather to use volatility spikes as event-based signals, signaling a shift in market regime or uncertainty that can be studied for post-event price behavior.

The historical context provided by Mansilla-Lopez et al. (2025) underscores the persistence and influence of volatility throughout financial history, from the Dutch Tulip Mania to the COVID-19 pandemic. Events like the Great Depression or the 2008 financial crisis generated prolonged periods of elevated volatility that often distorted price discovery and model stability (Hayes, 2022; Bekiros and Georgoutsos, 2008). Importantly, models built during these periods may overfit due to extreme and persistent market swings. As a result, this research may consider excluding recessionary windows or isolating them as a separate case, to avoid a situation where every asset, in every window, is flagged as volatile, diluting signal and inflating noise.

Finally, while the unpredictability of volatility events continues to limit the effectiveness of direct forecasting, its behavioral and statistical impact on subsequent price movements offers a promising direction. Rather than treating volatility as just another input variable in return prediction models, this study leverages it as a structural trigger to examine what follows after such events. This approach preserves the value of volatility in predictive modeling while sidestepping the often intractable problem of predicting its occurrence.

## 3. The Intersection – Using Volatility Shocks to Predict Returns in a New Way

While a substantial body of research leverages volatility within predictive frameworks, the predominant approach has been to use volatility as a continuous regressor within broader return models. However, forecasting volatility itself remains highly challenging due to its complex and often non-linear dynamics. An alternative and underexplored approach is to use volatility shocks as post hoc signals to forecast subsequent market behavior, including potential mean reversion, by leveraging machine learning's capacity to detect non-linear patterns during high-variance periods.

Ferreira and Medeiros (2022) investigate the relationship between market returns and volatility measures in a high-frequency environment using machine learning methods, finding that the CBOE Volatility Index (VIX) serves as a strong predictor of intraday market returns, particularly when implemented through Long-Short-Term Memory (LSTM) neural networks and Random Forest models. Their findings demonstrate the feasibility and effectiveness of using volatility indicators within machine learning frameworks to improve predictive performance in short-term forecasting environments, yet their work remains within the conventional approach of incorporating volatility as a regressor (Ferreira and Medeiros, 2022).

Similarly, Campisi et al. (2024) explore the predictive power of volatility indices for forecasting the direction of the US stock market, highlighting the effectiveness of volatility as a key variable within machine learning models. Their findings reinforce the view that volatility contains significant information relevant to return forecasting and validate its inclusion within predictive models for financial markets. Additionally, Wang (2024) demonstrates that machine learning models can effectively predict stock prices in high-volatility scenarios, further emphasizing the relevance of incorporating volatility data in forecasting frameworks.

However, the prevailing literature typically focuses on modelling of aforementioned "meaningful" volatility, defined as persistent and predictive of future risk premia, while intentionally filtering out short-lived volatility spikes considered as noise. Gallo (2025) critiques this narrow focus by proposing a framework for identifying "meaningful volatility events" in the context of monetary policy announcements, arguing that it is insufficient for price movements to be large; they must also be predictive of future returns or volatility. This perspective underscores a significant gap in the literature: the potential predictive value embedded within transitory, short-term volatility spikes that are often discarded in traditional asset pricing frameworks (Gallo, 2025).

This study directly addresses this gap by hypothesizing that transitory volatility shocks, often perceived as "meaningless," may, in fact, provide strong signals for subsequent mean-reverting price behavior and profitable correction predictions. Rather than predicting volatility events themselves, this research leverages volatility spikes as post hoc event-based signals to predict stock price movements, allowing machine learning models to capture the non-linear relationships that characterize post-volatility dynamics. By training machine learning models on the behavior

of equities preceding volatility spikes and the respective prices post shock, this research seeks to determine whether these short-lived volatility events can reliably predict market corrections, providing practical value for dynamic rebalancing and risk management strategies.

The practical relevance of this approach is further supported by Chun et al. (2025), who demonstrate that accurate forecasts of market behavior following volatility spikes can significantly enhance portfolio construction and inform dynamic investment strategies. Their findings indicate that while volatility is highly persistent and long-term forecasts are valuable, short-term forecasts remain useful for capital protection and tactical rebalancing, aligning closely with the aims of this study (Chun et al., 2025). Drawing on insights from Andersen et al. (2006), which show that realized volatility forecasts yield economic value across different horizons, this research evaluates post-volatility corrections over multiple timeframes, including 1-day, 3-day, and 5-day windows, to align with real-world investment practices.

Collectively, these studies establish a foundation that underscores the feasibility and importance of using volatility within predictive models while highlighting a key research gap in leveraging volatility shocks as event-based predictors. By addressing this gap, this research aims to contribute a novel perspective to financial forecasting, demonstrating how transitory volatility events, when paired with machine learning methods, can yield actionable insights for predicting price corrections in equity markets.

## 4. Gaps in Contemporary Research

Most studies in financial forecasting either attempt to predict volatility or incorporate it as a continuous input within return prediction models. This research takes a different approach, using volatility shocks as event-based signals to trigger predictive modelling of subsequent price behavior. By treating volatility as an event indicator rather than a regressor, we hypothesize that transitory volatility, often dismissed as “noise,” may in fact provide opportunities for predictive modelling, particularly in capturing mean-reverting behaviors in equity prices following high-volatility events. This methodological shift addresses a gap in the literature while potentially providing a more tractable and actionable forecasting framework for financial markets. While exogenous volatility events caused by factors like natural disasters, recession, and government intervention are impossible to predict, the reaction to these events may well be predictable.

Machine learning has demonstrated effectiveness in financial prediction, yet challenges remain due to the noisy, non-stationary nature of market data. Volatility is a powerful but underutilized signal when framed within a structured, event-based predictive framework, and using volatility spikes as event triggers rather than as a regressor offers a new lens for market forecasting. By testing whether transitory volatility shocks are followed by predictable price movements, this research contributes to a deeper understanding of market dynamics while offering practical tools for investors and institutions.

Importantly, the implications of this research extend beyond academic contribution. If volatility events can be leveraged to predict subsequent price corrections with high certainty, it has the potential to dampen the negative effects of volatility in financial markets. For institutional investors, hedge funds, and portfolio managers, reliable models that predict post-volatility price behavior can inform dynamic rebalancing strategies and enhance capital protection, allowing them to navigate high-volatility periods with greater confidence and efficiency.

Moreover, from a market stability perspective, improved prediction of post-volatility corrections can contribute to faster price discovery and smoother market adjustments. By enabling investors to respond swiftly and accurately to volatility shocks, such predictive models can facilitate the absorption of these shocks, reduce overreactions and mitigate prolonged periods of mispricing. In essence, while volatility itself may be unavoidable, the capacity to predict market responses to volatility can help contain and correct its destabilizing effects, bringing financial markets closer to stability while maintaining their fundamental role in reflecting economic realities.

In summary, this research aims to bridge a critical gap in the literature by exploring the predictive power of volatility shocks for subsequent price behavior, using advanced machine learning methods to capture non-linear relationships in post-volatility periods. By focusing on the predictive utility of transitory volatility events, this approach offers a practical, impactful strategy for investors and institutions seeking to manage risk and enhance performance in the face of market turbulence. This provides a strong foundation for the methodology that follows, detailing how this research will empirically test the viability of this event-based predictive framework.

# Methodology

## 1. Introduction

The purpose of these methods are to investigate whether **machine learning models can predict short- and medium-term stock price behavior following extreme volatility shocks**, using firm-level and market-wide time-series features. The empirical focus is deliberately narrow: forecasts are made **conditional on shock days using these shock days as event indicators**, while stock movement is predicted and performance is evaluated over **one, five, and twenty trading days** after the shock. These horizons align with realistic decision cycles: next-day execution, weekly positioning, and a one-month tilt. These horizons allow examination of any post-shock signal that is forecasted to change prices with time. Though previous research has neglected to use volatility as an event indicator, it has proven that machine learning can outperform naive models in daily, weekly, and monthly windows (Campisi 2024, Wang 2024, Ferreira 2022).

Our methodological design has four objectives. First, we **quantify the predictive content at 1/5/20 days** after shocks, balancing statistical fidelity with practical actionability. Second, we **contrast firm-specific features** (prices, returns, volume, volatility, and fundamentals) **with macro conditions** (term-structure indicators) to assess whether post-shock behavior is regime-dependent. Third, we **probe polarity asymmetry**, recognizing that post-shock paths after positive and negative events can differ due to behavioral and microstructure effects. Finally, we **evaluate forecasts on both Root Mean Squared Error (RMSE)**, a magnitude-sensitive loss which tells us how close our model is to predicting the real price change outcome, and **Directional Accuracy (DA)**, the probability of predicting the direction of the movement.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \quad \text{and} \quad \mathrm{DA} = \Pr\left[\operatorname{sign}(\hat{y}) = \operatorname{sign}(y)\right]$$

*Equation 1: RMSE and Directional Accuracy*

RMSE penalizes large mistakes disproportionately and returns to the target's units after the square root and acts as a marker for how far off our regression prediction is from the actual price change (Ferreira & Medeiros 2024). DA is a bit more simplistic and is a classifier by nature (positive or negative), but offers a direct link to business interpretation for long/short tilts (Campisi et al., 2024; Rossi, 2018; López Gil et al., 2024).

We adopt a **two-year window (2023–2024)** for the daily S&P 500 companies' data, augmented with macro term-structure data. This interval is long enough to deliver statistical power (many firms × many days) yet short enough to avoid conflating markedly different regimes (e.g., the COVID crisis), which would dilute relevance for today's macro environment which we are

actively trying to predict. Additionally, the further away from today the data is, the less relevant it gets to predicting what is happening today, and many leading projects in this field use a similar dataset size (Campisi 2024). By conditioning on shock days, using **leakage-safe** features that encode short histories and macro state, and enforcing **chronological** evaluation, the methodology directly addresses the research question and allows us to interpret Machine Learning models' ability to predict stock trends after volatile events.

## 2. Data Sourcing and Description

The sample consists of daily observations for S&P 500 constituents across **2023–2024**, drawn from S&P Capital IQ (firm-level markets and fundamentals) and standard macro sources (the 10y–2y Treasury yield spread, "T10Y2Y"). Firm identifiers include *Entity_ID, Company_Name,* and *Ticker*. Daily market variables include *Close, Open, Volume*, and *Market_Cap*. Quarterly fundamentals, e.g., *Book_value_equity* and *Net_income*, are merged to the daily panel using an as-of carry (last known report applied forward until updated). A static credit metric (*Credit_Rating_Global*) is mapped to an ordinal rank 0–20 plus a *was_not_rated* indicator for a few companies whose credit rating is not publicly available.

**Prices (open and close) and daily returns** capture realized drift and reversal forces around the event and remain the primary carriers of short-horizon information. **Volume** proxies attention, information arrival, and liquidity conditions; abnormal volume often coincides with continuation or reversal pressure. **Volatility and market-wide stress** provide common-factor context; when aggregate turbulence is high, post-shock paths may differ systematically. **Fundamentals and risk** (valuation via Book-to-Market, profitability via Net Income, and a credit-risk ordering) reflect balance-sheet strength and valuation posture that can moderate the probability of correction vs. continuation. **Macro term structure** (T10Y2Y) encodes regime (growth/inflation/monetary stance) and changes materially within our sample (including a yield-curve flip), plausibly shifting post-shock dynamics. Including a macro based factor like this proved to be incredibly important to identify current market regimes and add context to the mostly entity based focus of the feature list (Welch & Goyal 2008). **Size (Market Cap)** maps to liquidity, coverage, and investor base, each of which can influence the resilience of deviations from fundamentals.

## 3. Data Preprocessing and Feature Engineering

### 3a. Leakage control

All transformations are forward-looking safe: entity-level lags and rolling statistics use only past observations; quarterly fundamentals are merged as-of the last report; macro series are forward-filled where appropriate but never peek ahead. Features are generated using only in-sample data. Any normalization occurs within the train/validation/test data sets, so no future

information is used to inform previous information. Models are only trained using train data and evaluated using validation data, until the final model is deployed. These safeguards prevent information leakage across time from the data processing stage through the model testing stage.

### 3b. Rationale for lagging

Tree ensembles such as XGBoost, CatBoost, and Random Forest are not sequence-aware. To inject short-history dynamics, we lag key signals at **t−1, t−5, t−10, and t−20**, spanning very-recent, weekly, bi-weekly, and monthly scales. This parsimonious lag scheme introduces trend and short-term autocorrelation without exploding dimensionality or letting the model "overlearn" the structure of ordinary non-event days. In earlier neural-network exploration, we attached the previous 20 days to each shock so the model could learn temporal patterns end-to-end; empirically, the LSTM tended to predict values close to zero, suggesting that it learned the typical pre-shock state well while struggling to extrapolate to the post-shock distribution due to the model learning regular patterns, but predicting largely irregular samples. This practical result strengthened the case for parsimonious lagging with the tree models used. Patton successfully used a similar methodological approach in his study of the effects of volatility events, creating lagged variables and using time series with a limited backwards looking window to inform his model (Patton, 2013).

### 3c. Feature families

Engineering focuses on (i) **returns and prices** (daily returns, lagged returns/prices, moving averages), (ii) **volume and liquidity** (lagged and smoothed volume), (iii) **market-wide stress** (cross-sectional average absolute return and its rolling volatility), (iv) **momentum/oscillators** (e.g., RSI-14 and simple price momentum), (v) **macro term structure** (T10Y2Y level and a 20-day change), and (vi) **fundamentals and risk** (credit-rating rank, Book-to-Market proxy, Net Income). After diagnostic checks (including permutation importance), correlation checks, and SHAP contribution analysis, a compact set of ~20 variables was retained for modeling, while many purely static descriptors were de-emphasized after they were found to be unimportant when fed into the machine learning structure. The final feature correlation heatmap can be found in the appendix B, while the permutation and SHAP analysis are further expanded on in the Section 6 of the Results chapter and can also be found in Appendix D and E.

A structured **Feature Dictionary** appears below, listing the definition of each feature used in final modelling. Supplemental granular definitions of all generated features are given in the Appendix A for additional clarity and reproducibility.

| Feature | Definition |
|---|---|
| **Lagged features: close, return, volume** | Features lagged by 1, 5, 10, and 20 days |
| **close** | Closing price on event day |
| **volume** | Number of shares traded on event day |
| **daily_return** | Open-to-close return on event day |
| **avg_return** | Rolling mean of daily returns over the training period |
| **return_std** | Rolling standard deviation of returns over training period |
| **abs_return_stock** | Absolute return of stock on event day |
| **avg_abs_return** | Average absolute return of all stocks on event day |
| **avg_raw_return** | Average raw return of all stocks on event day |
| **rolling_market_vol_10d** | Rolling std of returns for all stocks for 10 days before event |
| **rsi_14** | Relative Strength Index for all stocks 14 days before event |
| **T10Y2Y** | Level of the 10y–2y Treasury yield spread on event day |
| **T10Y2Y_t-20** | 20- day lagged T10Y2Y |
| **book_value_equity** | Latest reported Book Value of Equity |
| **market_cap** | Market Capitalization on event day |
| **net_income** | Latest reported Net Income |
| **month** | Calendar month extracted from the date (1–12) |

*Table 1: Final Feature List with Definitions*

## 4. Defining the Target Variable

### 4a. Shock identification

We compute daily returns and per-stock z-scores relative to a global two-year baseline (mean and standard deviation for each stock over 2023–2024). A volatility shock is a day with $|z_t|>2$, or more simply, stocks that experienced a daily return at least 2 standard deviations above their average daily movement over the past 2 years. Because shocks often cluster, we deduplicate streaks by retaining the last day of each contiguous run, thus avoiding multiple counts of the


*Figure 1: Net Shock Score Over Time*

*This figure shows the amount of volatility shocks over the data period, as well as the directionality and magnitude of these shocks*

same evolving episode. A global baseline ,rather than a rolling one, stabilizes the reference distribution and reduces spurious “trend-driven” flags during persistent drifts (Patton, 2013).

**4b. Prediction targets**

For each shock at time t, we predict forward returns over the next 1, 5, and 20 trading days:

$$y_t^{(1)} = \frac{C_{t+1} - C_t}{C_t}, \qquad y_t^{(5)} = \frac{C_{t+5} - C_t}{C_t}, \qquad y_t^{(20)} = \frac{C_{t+20} - C_t}{C_t}$$

*Equation 2: Definition of Target Variables*

*These targets are calculated as percentage point change over a 1 trading day, 5 trading day, and 20 trading day horizon*

We analyze performance **overall** and, where informative, **by shock polarity** (positive vs. negative), recognizing that asymmetries can arise from investor behavior, financing constraints, and limits to arbitrage (Construction and validation of an overconfidence scale in investment decisions, 2025).

**4c. Brief Note on Targets by Horizon**

Across horizons, the return distributions widen as you look further out: the 1-day target is tightly clustered around zero with some large outliers, the 5-day target is broader with modest tails, and the 20-day target is the widest with noticeably fatter tails (bigger swings and clearer mean-reversion/continuation patterns). Practically, the shape of the return distributions shape our result analysis (and will be referenced again later in that section.

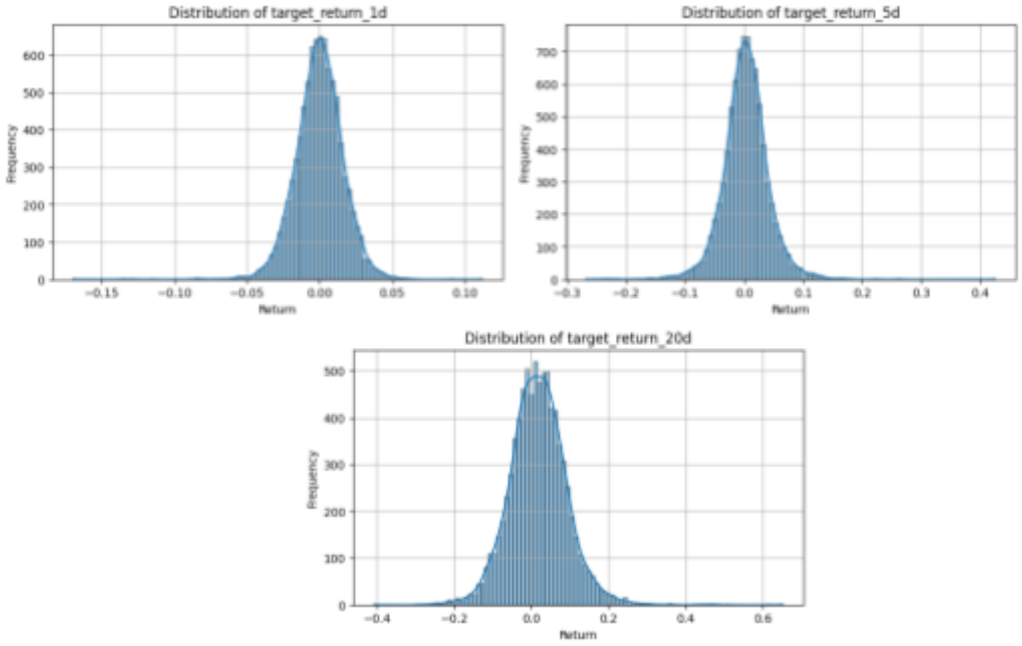



*Figure 2: Target Horizons*

*Figure 2 displays the distribution of returns observed in the three target horizons. Note the differences, and how the distribution widens as the horizons get further away from the event day.*

## 5. Modelling Approach

### 5a. Temporal design

To maintain out-of-sample realism, training and evaluation are strictly chronological and no training data interacts with testing data. After event filtering and windowing, the realized ranges in the final panel were:

| Split | Date range | Shocks | Firms | Pos | Neg |
|---|---|---|---|---|---|
| Train | **2023-02-02 → 2024-06-28** | 5929 | 435 | 2889 | 3040 |
| Validation | **2024-07-01 → 2024-08-30** | 1003 | 399 | 488 | 515 |
| Test | **2024-09-03 → 2024-12-30** | 1521 | 425 | 659 | 862 |

*Table 2: Train/Val/Test Split*

*Table 2 shows the date range, number of shocks, number of shocks broken down into their direction, and the number of firms for each window*

Such deviations are expected in event-study construction and are reported transparently. Note: the number of firms differs between sets because, if a firm did not have a shock in the given time period, it would not appear in the dataset. All firms were considered, but only ones with observed shocks are present in the respective split.

### 5b. Estimators and tuning

The **primary estimator** is **XGBoost** (gradient-boosted decision trees), chosen for its strength on tabular data, capacity for nonlinearities and interactions, and robustness when short histories are encoded via lags. All subsequent models were tested agnostically, but XGBoost outperformed the rest so it will remain the main focus in the following sections. A separate time series dataset was made for input into the LSTM model, where instead of producing lagged variables up to 20 days, a 20 day window prior to the volatility event was attached to each event, creating a time series dataset ideal for LSTM feed in, but even so the LSTM was still markedly outperformed by the XGBoost model. Benchmark models include **RandomForest, CatBoost, and LSTM Neural Network Models**.

We tuned XGBoost with a randomized search over a compact but expressive space: n_estimators(100–900), learning_rate(0.01–0.29), max_depth(3–10), gamma(0–5), colsample_bytree(0.5–1.0), subsample(0.5–1.0), reg_alpha(0–1), reg_lambda(0–1) using 5-fold GroupKFold (groups = entity) to avoid cross-firm leakage and scoring on RMSE. We also tested a broader 3-fold search, but the 5-fold setting produced slightly better CV scores and more stable selections. Across horizons (1d/5d/20d), the best-fit configurations were highly similar, so we adopted a single average best-fit setting for all three to increase reproducibility, simplify interpretation, and avoid overfitting to any one horizon. The final settings used throughout are:

n_estimators = 649, learning_rate ≈ 0.0438, max_depth = 9, gamma ≈ 0.0136, colsample_bytree ≈ 0.569, subsample ≈ 0.965, reg_alpha ≈ 0.412, reg_lambda ≈ 0.349. This choice keeps the learning dynamics consistent across targets while preserving the generalization we observed in validation. An additional table with these hypertuned model parameters is available in Appendix C for added readability.

### 5c. Focal results

The XGBoost model selected on the **train/validation** period achieved **~70% directional accuracy**, while the **final test** model achieved **~60%**. We attribute a majority of the decline to a **regime shift** late in 2024, including a **higher incidence of large positive shocks** in the test period relative to the earlier sample. Because such distributional shifts are central to external validity, we note the phenomenon here and present a deeper decomposition (e.g., by polarity and sector) in the Results and Limitations.

### 5d. Polarity specialization

Given the implied asymmetry and 'noiseness' of stock return data, we also consider polarity-specialized variants (models trained separately on positive and negative shocks). This is not the headline specification but serves as an alternative look at possible model specifications for future studies, and acts as a diagnostic for asymmetric dynamics and will be further noted and extrapolated in the discussion section, along with the intriguing results these models provided.

## 6. Model Evaluation Metrics

Evaluation proceeds on the held-out validation and test slices, reporting **RMSE** and **Directional Accuracy** for **1-, 5-, and 20-day** horizons. RMSE assesses magnitude fidelity, penalizing large mistakes in a convex manner; DA maps directly to the success rate of a simple directional decision. These two approaches were chosen based on leading industry standards (Rossi 2018, Campisi 2024, Ferreira 2024) and their ability to measure model 'success' in two very different ways. RMSE measures our regressors ability to predict precisely, while directional accuracy is more use case practical, as most trading strategies focus on direction movement over time and worry less about exact numbers or predictions (Lopez-Gill 2024). With a brief understanding of how these two estimators function, it is in the best interest of this study to use RMSE as the main predictor to fine tune parameters within the models. In theory, any increase in RMSE should not sacrifice Directional Accuracy, so the models in this study will be tuned using RMSE as the estimator in focus. We also compute **permutation importance** and **SHAP** summaries to interpret which features drive predictions and how their roles change across horizons or regimes, with particular attention to the **macro–micro** contrast. (further analysis of SHAP and permutation importance tests are reported in REsults and Discussion, and figures for SHAP scores and permutation results can be found in Appendix D and E)

## 7. Design Rationale and Iterative Refinements

The research proceeded iteratively. Early experiments used **broad feature sets** and a **60-day** pre-shock window for sequence models; empirically, the LSTM tended to predict near zero, consistent with learning the pre-shock "normal" state rather than the atypical **post-shock** distribution. This finding motivated a pivot to **parsimonious lagging** (t−1, −5, −10, −20) and explicit incorporation of **macro regime** via T10Y2Y level and a 20-day change. Tree ensembles, especially **XGBoost**, handled this representation effectively, yielding stronger validation performance. Static descriptors that did not add value under **permutation importance** were de-emphasized to preserve parsimony and reduce noise. The final configuration therefore centers on XGBoost with a compact, leakage-safe feature set and clearly defined event conditioning.

## 8. Reproducibility and Controls

All steps are **chronologically ordered**, with as-of merges for fundamentals and no future information used in features. Where relevant, **group-aware validation** is applied to reduce cross-firm contamination. Data cleaning decisions (imputation, carry rules, and any outlier handling) are documented; random seeds are fixed where applicable. These controls are standard safeguards for time-series prediction and aim to maximize the credibility and reproducibility of the reported results (Tiwari, 2025).

## 9. Conclusion of Methodology

The methodology combines event conditioning on volatility shocks, parsimonious lagging for non-temporal models, macro regime indicators, and chronological evaluation with RMSE and DA. This design directly addresses whether short- and medium-term price behavior after extreme shocks is predictable in practice. The next section reports empirical performance across horizons, examines asymmetries (particularly the late-sample regime shift), and interprets the relative contributions of macro vs. micro features.

# Results

## 1. Overview

Predictability is markedly **horizon-dependent**. At 1 day, sign prediction is difficult (DA near 50%), and error magnitudes are close to tuned baselines. By 5 days, the model exhibits meaningful directional skill (validation DA ≈ 59–61%). At 20 days, predictability is strongest: XGBoost attains ≈70–71% DA on validation. When the model is finally refit on Train+Val and evaluated on the held-out Test window, DA remains above chance at 52.9% (1d), 56.2% (5d), and 57.5% (20d), though clearly below validation. This drop off is significant, but consistent with the late-2024 regime/polarity shift documented in the feature densities and event mix. In short, post-shock direction becomes more predictable as the horizon lengthens, with the clearest signal at 20 days; out-of-sample performance is regime-sensitive.

A brief metric translation sets expectations for understanding the chosen evaluation metrics in context. Directional Accuracy (DA) is the fraction of correct sign calls (e.g., 57.5% DA means ~58 of 100 signs correct). RMSE is the typical size of the forecast error in return units. For intuition: an RMSE of 0.041 at the 5-day horizon implies that if the model forecasts +3%, a "typical" realized outcome is roughly 3% ± 4.1 pp (≈ −1.1% to +7.1%). RMSE summarizes magnitude fidelity; DA is closer to the economically relevant long/short decision present in most stock models. We therefore report both but emphasize DA for practical interpretation.

To keep provenance unambiguous, all results below follow a fixed order: **(**i) baselines → (ii) iterative testing on validation (RF, CatBoost, LSTM, XGBoost) → (iii) best XGBoost on validation → (iv) final Train+Val → Test. We also surface a polarity-separated XGBoost run, which yields the single best DA observed (79% at 20-day for negative shocks).

## 2. Baselines and event context

A constant-mean baseline model (which acts as our naive, base model to compare other models) evaluated on validation produces RMSE = 0.01868 (1d), 0.04188 (5d), 0.07215 (20d), which is relatively low as expected for a flat predictor. The event panel is large and diverse: 8,453 non-redundant shocks across 435 firms, with a heavy-tailed daily incidence (mean 17.8, s.d. 21.1, max 184 on 2024-11-06). Distributional diagnostics show that T10Y2Y_t−20 shifts from predominantly negative (validation) to positive (test), while *avg_raw_return* drifts and *return_std* remains comparable, evidence of covariate shift into late-2024 that plausibly compresses out-of-sample DA.

## 3. Iterative model testing

**Random Forest** yields **RMSE 0.01914 / 0.04259 / 0.07398** at **1/5/20 days**. **CatBoost** improves error slightly and posts **DA ≈ 48.8% (1d), 57.9% (5d), 68.9% (20d)**.

In our LSTM time-series model, the **validation** metrics are **RMSE ≈ 0.020 (1d), 0.045 (5d), 0.081 (20d)** and **DA ≈ 57%** on its strongest run (after some simple feature importance steps). Qualitatively, the

network **learned "normal" pre-shock days very well** and tended to **predict near zero** on abnormal post-shock days, which **dampened sign discrimination**. This aligns with evidence that sequence models can overfit level dynamics when the evaluation distribution shifts (here, **post-shock**).

| Model | Target | Split | n | RMSE | DA |
|---|---|---|---|---|---|
| CatBoost | **target_return_1d** | Validation | 1003 | 0.02032 | 48.45% |
| CatBoost | **target_return_5d** | Validation | 1003 | 0.04218 | 57.93% |
| CatBoost | **target_return_20d** | Validation | 1003 | 0.07994 | 68.69% |
| RandomForest | **target_return_1d** | Validation | 1003 | 0.01913 | 51.74% |
| RandomForest | **target_return_5d** | Validation | 1003 | 0.04232 | 53.94% |
| RandomForest | **target_return_20d** | Validation | 1003 | 0.07328 | 67.90% |
| XGBoost (Final) | **target_return_1d** | Test | 1521 | 0.01563 | 52.33% |
| XGBoost (Final) | **target_return_5d** | Test | 1515 | 0.04069 | 56.04% |
| XGBoost (Final) | **target_return_20d** | Test | 1180 | 0.08969 | 57.20% |
| XGBoost (Val tune) | **target_return_1d** | Validation | 1003 | 0.01882 | 44.47% |
| XGBoost (Val tune) | **target_return_5d** | Validation | 1003 | 0.04103 | 60.72% |
| XGBoost (Val tune) | **target_return_20d** | Validation | 1003 | 0.07201 | 69.99% |
| LSTM | **target_return_1d** | Validation | 1003 | 0.02124 | 52.21% |
| LSTM | **target_return_5d** | Validation | 1003 | 0.04517 | 54.06% |
| LSTM | **target_return_20d** | Validation | 1003 | 0.07961 | 62.72% |

*Table 3: Results Across All Models Tested*

*Table 3 presents the RMSE and DA for all models evaluated during our model testing phase.*

**XGBoost (Train → Validation set scores)**

These are the scores for the different, iteratively tuned XGBoost models. The model was being tuned for the best RMSE, so while there are drops in DA, we are emphasizing an increase in RMSE for each iteration.

– **Initial spec: RMSE 0.01948 / 0.04022 / 0.07344**, **DA 50.1% / 60.5% / 71.4%**
– **Alt-tuned: RMSE 0.01868 / 0.04188 / 0.07215**, **DA 50.3% / 58.7% / 70.7%**.
– **Robust-tuned: RMSE 0.01882 / 0.04103 / 0.07201**, **DA 44.5% / 60.7% / 70.0%**.

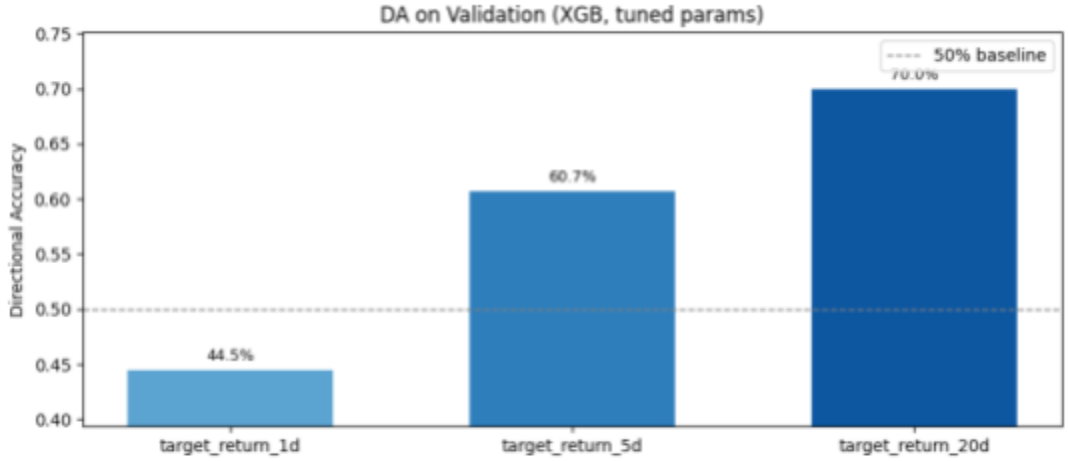



*Figure 3: Directional Accuracy on the Validation Set*

*This table show the directional accuracy score for the three different target horizons for the dataset that trains on training data and evaluates the results using the validation set*

Across settings, the pattern is robust: **clear directional gains at 5 and especially 20 days**; **1-day** DA near the decision boundary. We therefore **foreground 20-day validation DA (~70–71%)** as the model-selection anchor, while immediately juxtaposing the **final test** performance (Section 4) to underscore regime dependence. In this trial where many models were tested, XGBoost and CatBoost were the obvious two top contenders in regards to both DA and RMSE scores, which is surprising, as research suggests Neural Networks like LSTMS's are most apt to handle similar data structures, but this point will be explored, clarified, and solved in the Discussion Chapter 5.8, 'Why parsimonious lagging + trees beat the 20-day LSTM.

## 4. Best XGBoost (Train → Val) → Final model (Train+Val → Test)

Refitting on **Train+Val** and evaluating on **Test** yields: **RMSE 0.01581 / 0.04083 / 0.08184** and **DA 52.93% / 56.24% / 57.46%** at **1/5/20 days**. Relative to validation, DA **compresses**, especially at **20 days**, coincident with a **shift toward unforeseen large shocks** and an **upward move in the term structure** (see T10Y2Y_t−20 density). These changes **reduce transferability** from the validation decision boundary. The key takeaway in regards to this research is that **the 70–71% validation DA establishes the existence of medium-horizon predictability after shocks**, and **the high-50s DA on test** demonstrates that the effect is **regime-sensitive but persistent** out-of-sample. The model works, but it works better in some regimes than others. In a silo, the model could likely produce spectacular DA results, but given unprecedented returns, the performance of the model is hindered. Some workarounds and business applications regarding this point are given in Conclusion Chapter 2 and 3.

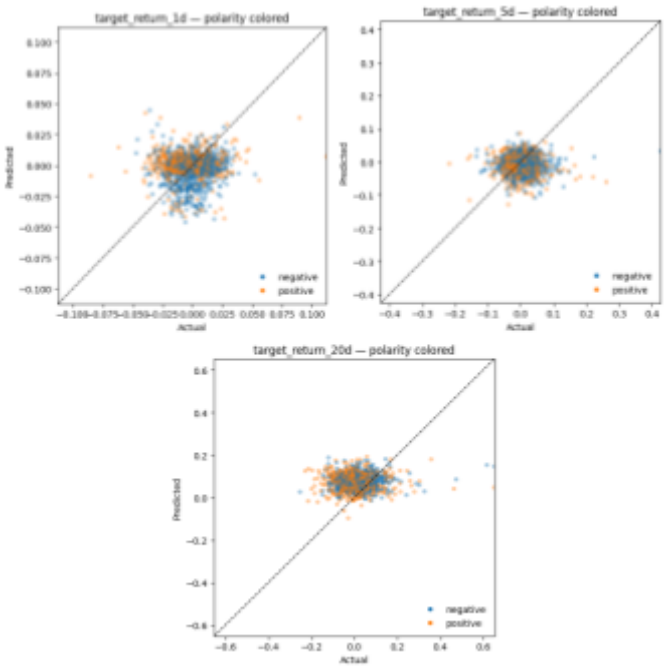



*Figure 4: Actual vs. Predicted Target Returns*

*Figure 4 displays the actual vs. predicted target returns for all 3 time horizons for the final model. The blips are color coded to better differentiate between the polarity of each volatility shock*

## 5. Polarity Asymmetry

Polarity-specific XGBoost models reveal substantial asymmetry. For positive shocks, validation DA is 51.8% (1d), 54.1% (5d), 62.3% (20d). For negative shocks, DA increases to 48.9% (1d), 66.2% (5d), 79.0% (20d) which is the single best DA in the study. This aligns with overreaction/mean-reversion and liquidity-provision mechanisms after negative events, contrasting with more balanced continuation/fade after positive events. Because the test period contains more positive shocks, pooled test DA compresses relative to validation. We present this asymmetry prominently here and return to mechanism and robustness in the Discussion.

When we re-estimate XGBoost separately by shock polarity, medium-horizon predictability concentrates on negative shocks. On the validation slice, DA rises from 64.85% (5d) to 79.22% (20d) for negatives (RMSE 0.0667 at 20d), versus 52.46% (5d) and 62.30% (20d) for positives (RMSE 0.0873 at 20d). Out-of-sample on the Test window, the negative-shock edge persists (60.42% at 5d; 61.27% at 20d), while positive shocks fall near chance at 5d (50.30%) and remain modest at 20d (53.09%). These patterns align with mean-reversion/liquidity channels after large down moves, and with more heterogeneous continuation/fade dynamics after up moves. The polarity mix also helps explain the pooled Test DA compression, given the Test period's heavier incidence of positive shocks.

## 6. Feature contributions

Across horizons, the model's predictive content is anchored in micro, event-proximal signals, while macro/time variables act mainly as context setters inside the fitted trees. On the Test set, **permutation importance** (ΔRMSE) shows that at **1-day** the model's unique lift comes chiefly from *avg_return* and *daily_return*, with smaller contributions from return_t−10 and *rsi_14*; macro/stress features are negligible at this very short horizon. At **5-days**, *avg_return* still dominates, with only modest incremental roles for *return_t−1* and *return_t* and faint signals from *net_income* and *return_std*. By **20-days**, the importance mass broadens: *avg_return* remains first, but *avg_abs_return*, *daily_return*, *rsi_14*, and *avg_raw_return* all become material; *T10Y2Y* enters with a small yet non-zero drop. Interpreted purely as "what improves out-of-sample accuracy after training," these changes say that **recent trend and local stress carry the primary incremental predictive power**, especially at 20 days where overall DA is highest.

**TreeSHAP** tells a complementary story about **how** the model sets its predictions. On **20-day**, SHAP ranks *month* and *T10Y2Y* among the largest contributors in absolute effect size, alongside *avg_return*, with *T10Y2Y_t−20*, *rsi_14*, and *return_std* following. At **5-day**, SHAP still highlights *avg_return* (with *avg_abs_return* and *rsi_14* close behind), while at **1-day** it surfaces *avg_raw_return* and a coarse time tag (*month*). This divergence, **macro signals high in SHAP, but micro features high in permutation**, is expected under correlated inputs and limited within-Test macro variation: **permutation** attributes credit only to unique incremental lift on the held-out distribution, whereas **SHAP** captures the **internal mechanics** of the fitted trees, including **regime offsets** and **interactions**. In practical terms, the trees rely on **micro features** to move the decision boundary that drives DA, while **macro/time variables** (notably *T10Y2Y* and *month*) help **position the baseline** and **gate interactions**, a pattern fully consistent with our broader finding that performance is **regime-sensitive** even as medium-horizon micro signals remain informative.

When ranking "what mattered," we therefore report both lenses: **permutation (ΔRMSE)** as the **performance-relevant** ordering (micro trend/stress first; macro modest), and **SHAP** as the **model-internal** ordering that confirms the presence of **regime conditioning**. Read together, they support a clear narrative: the model's **predictive backbone** is recent stock-specific behavior around the shock, and its **stability across regimes** can improve by making the macro state **explicit** (e.g., regime-aware ensembling or macro-conditioned thresholds) a point we develop in the Discussion.

## 7. Error diagnostics and economic relevance

Actual-vs-predicted scatterplots (Figure 4) for **1/5/20 days** show the familiar under-dispersion: predictions cluster near zero and the slope vs. the 45° line is shallow, confirming that **sign information exceeds magnitude precision**. This is consistent with **RMSE** remaining near tuned baselines, and it clarifies why **DA** is the **more economically relevant** metric in this setting. For scale: RMSE ≈ 0.0158 (1d), 0.0408 (5d), 0.0818 (20d) means typical absolute errors of roughly 1.6 pp, 4.1 pp, and 8.2 pp at the three horizons. Because larger-magnitude predictions usually have higher hit-rates, a selective (abstention) rule can raise effective DA at the cost of coverage; we reserve that analysis for the Discussion.

## 8. Brief synthesis

**20-day** validation DA near **70–71%** demonstrates that **medium-horizon post-shock direction is strongly predictable** in-sample; **5-day** results are modest but durable; **1-day** is hardest. On **Test**, DA remains **above chance** in the **mid-50s**, compressing as the macro term structure and polarity mix shift. Polarity-specific modeling reveals a **pronounced negative-shock edge** (up to **79% DA** at 20 days). Overall, **predictability exists and is economically interpretable**, yet regime and polarity composition meaningfully shape realized performance.

# Discussion

## 1. What the findings mean

On the **held-out test period** (Sep–Dec 2024), after refitting the model on the entire training dataset, our XGBoost model achieves Directional Accuracy (DA) of (20-d 57.46%ay), 56.24% (5-day), and 52.93% (1-day), with corresponding RMSE of 0.08184, 0.04083, and 0.01581. These out-of-sample figures are above chance, especially at the deeper time periods. These results are significant in context, but do not discount the model trained on a smaller batch of test data and evaluated on the validation set, which we attribute to three reinforcing factors we later unpack: **(i) macro regime drift**—the 10y–2y Treasury spread moved from a prolonged **i**nversion to de-inversion/steepening, altering post-shock dynamics and leading to lots of unprecedented high returns during the test period; **(ii) shock-composition shift**, the test window contains more extreme shocks, for which predictability is weaker; and **(iii) temporal clustering** around a few high-intensity volatile days with broad, event-driven moves. We return to these drivers and their implications later in the Discussion Chapter on regime sensitivity, polarity asymmetry, and most-volatile days.

On the aforementioned **validation window** (Jul–Aug 2024) the same approach delivers significantly **stronger medium-horizon predictability**: XGBoost reaches ~70% DA at 20 trading days (e.g., 71.39% in our initial spec), ~60.52% at 5 days, and ~50.05% at 1 day, with RMSE roughly 0.073, 0.040, and 0.019, respectively. These metrics are **substantially above chance** at 20 days and beat the simpler baselines as well as the more complex LSTM model deployed. Taken together, these results show that treating **volatility as an event-based trigger** yields **clear directional signal at 20 days** (and modest signal at 5 days), while **very short horizons remain noisy**; the drop from validation to test is expected once macro conditions and the mix of shocks shift, a point we emphasize and elaborate in the Discussion.

Because DA and RMSE speak to different notions of good performance, we give the reader a single intuitive yardstick upfront. RMSE is the average size of the forecast error. For example, when we report RMSE ≈ 0.040 on 5-day returns, that means the prediction typically misses the actual 5-day return by ~4 percentage points (so a forecast of +3% tends, on average, to land within roughly −1% to +7%, not as a hard bound, but as a typical miss). We emphasize DA throughout because it is more directly tied to tradability in a sign-based strategy; RMSE is still reported to document magnitude error and calibration.

In regards to the research question posed earlier, yes, our models predict medium-term (20-day) direction after extreme volatility shocks with economically meaningful skill (≈70% DA in-sample/validation), while short-horizon (1–5 day) direction is only modestly predictable. Out-of-sample test DA is lower (mid- to high-50s), indicating regime dependence; we discuss why and how to address it below.

## 1a. Practical reading across horizons

- **20-day:** Clear directional edge in-sample; meaningful but regime-sensitive edge out-of-sample; **negative shocks** are the marquee opportunity set.

- **5-day:** Modest, durable skill (≈60% DA validation; mid-50s test), consistent with weaker but persistent post-shock drift.

- **1-day:** Near the decision boundary; RMSE near baseline. This short horizon is best treated as descriptive context for the longer windows.

## 1b. Brief Explanation of Model Evolution: Baseline → iterative testing → XGBoost best (validation) → XGBoost final (test)

**Baselines**

The "mean" baseline (using training-period averages) performs poorly across horizons as expected, reinforcing the need for structured features and non-linear learners due to the incredibly complex and highly dimensional nature of stock return data.

**Iterative testing**

After pruning static/noisy features and adding lagged variants to inject localized trend without letting trees "peek" through dense daily sequences, performance improved materially on non-temporal models (Random Forest, XGBoost) while the LSTM continued to regress toward near-zero predictions on abnormal post-shock days. (This is consistent with the model learning "normal days" very well but over-learning pre-shock stability, a point we return to in Section 5 of this Discussion chapter.)

**Best validation results**
On the validation set, our best XGBoost achieved ~71% DA at 20 days with RMSE ≈ 0.072, while 5-day DA was in the high-50s to ~60% with RMSE ≈ 0.040–0.042. These are the results that establish the existence of a medium-horizon signal in our setting.

**Final XGBoost (Train+Val → Test)**
Trained on Train+Val and evaluated on the held-out Test (Sep–Dec 2024), XGBoost delivered DA ≈ 57–56–53% (20d–5d–1d) with RMSE ≈ 0.0818 / 0.0408 / 0.0158, respectively. This is lower than validation DA, but still above chance and aligned with the macro regime shift during fall 2024 (see Section 3 of this Discussion chapter). These test results matter for external validity, even if deployability is not this dissertation's goal.

**Interpretation**

We take the 20-day validation DA (~70%) as the clearest demonstration that the approach works in the regime it was trained on, and the mid- to high-50s DA on test as evidence that skill persists but attenuates under a different macro mix of shocks. We flag the shift immediately and treat it as a first-class result, not an afterthought.

## 2. Evaluation Metrics in Context

### 2a. How to read the metrics

We report **Directional Accuracy (DA)** and **RMSE**. DA captures sign-prediction skill,the quantity most closely tied to a long/short decision,while RMSE summarizes magnitude fidelity. Throughout, we prioritize DA for interpretability while still documenting RMSE to track error dispersion.

### 2b. Apples-to-apples comparison caveat

The external studies we reference mostly predict **ordinary (non-event) periods** using **volatility indices** or single-asset series at **daily, monthly, or intraday** horizons. Our setting is different: we predict **post-shock returns** cross-sectionally for S&P 500 constituents (ie., conditional on rare, high-volatility events) where base-rate direction is less stable and distributions shift rapidly. Comparisons are therefore **contextual, not plug-and-play**.

## 3. Where the Results Stand Relative to Prior Work

Table 4 lines up the closest horizons first and centers on DA; RMSE appears where comparable.

| Study | Data Used, Target | Horizon | Model | Score | Comparability note |
|---|---|---|---|---|---|
| Ours (Validation) | **S&P 500, post-shock** | 20-day | XGBoost | ~70–71% DA | Event-conditioned, cross-sectional; strongest in-sample |
| Ours (Polarity: negative, Val.) | **S&P 500, post-shock (neg.)** | 20-day | XGBoost | ~79% DA | Event-conditioned; pronounced asymmetry |
| Campisi et al. (2024) | **Volatility indices (VIX family)** | 30-day | Bagging / RF | 82.75% / 80.03% DA | Index-level, non-event, volatility-only inputs |
| Rossi (2018) | **S&P 500, sign** | Monthly | BRT | 57.35% DA | Aggregate index; baseline "always up" ≈55.26% |
| Ours (Test) | **S&P 500, post-shock** | 20-day | XGBoost | 57.46% DA; RMSE 0.0818 | De-inversion regime; more positive shocks |
| Ours (Validation) | **S&P 500, post-shock** | 5-day | XGBoost | ~60.5% DA; RMSE ~0.040 | Event-conditioned |
| Ours (Test) | **S&P 500, post-shock** | 5-day | XGBoost | 56.24% DA; RMSE 0.0408 | Regime & polarity shift |
| López Gil et al. (2024) | **EWZ (Brazil ETF), price** | Daily | xLSTM-TS | 72.82% DA | Single ETF, non-event, different market |
| Ours (Validation/Test) | **S&P 500, post-shock** | 1-day | XGBoost | ~50% / 52.93% DA; RMSE 0.0158 | Short horizon ≈ chance |
| Ferreira & Medeiros (2021/2024) | **SPY, intraday** | 1-min ahead | LSTM-VIX | RMSE ≈ 0.0080 | Intraday magnitude task; different target |

*Table 4: Score Comparison Across Studies*

*Table 4 shows how outside studies' scores compare with our models. (Campisi et al., 2024; Rossi, 2018; López Gil et al., 2024; Ferreira & Medeiros, 2021/2024.)*

## 3a. Key takeaways from the results table comparative results table

1. **Medium-horizon signal exists in our event-conditioned setting.** Our **20-day validation DA (~70–71%)** falls within the "strong classifier" band seen at nearby horizons in the literature, despite fundamentally different inputs and conditioning.

2. **Polarity drives outsized gains.** The **negative-shock 20-day DA (~79%)** is numerically close to the top-line 30-day figures reported for volatility-index studies (Campisi et al., 2024), highlighting that extreme down moves tend to exhibit mean-reverting drift over multi-week windows, consistent with liquidity-provision and overreaction narratives.

3. **Out-of-sample DA compresses under regime shift.** On the Sep–Dec 2024 test window, pooled 20-day DA ≈ 57%, above chance but materially below validation, coinciding with yield-curve de-inversion and a higher share of positive shocks, for which predictability is weaker in our splits.

4. **RMSE tracks close to tuned baselines.** Across horizons, magnitude errors remain near baseline levels, reinforcing our choice to interpret results mainly through DA, while using RMSE for calibration checks and risk control.

## 3b. Why comparisons should be read cautiously

Even when horizons appear adjacent (e.g., **20- vs 30-day**), several structural differences remain:

- **Conditioning:** Prior work typically evaluates on **all days**; we evaluate **specifically after large shocks**. This mechanically lowers base-rate stability and increases covariate shift.

- **Unit of prediction:** Many benchmarks operate on indices (VIX family, S&P 500 level, or EWZ) with homogeneous dynamics. We operate cross-sectionally across 435 firms with observed shocks in the data-set period, where heterogeneity in post-shock behavior (news type, liquidity, market microstructure) is larger.

- **Input families:** Several studies rely primarily on volatility-related inputs; our strongest signals are micro, event-proximal features (recent drift/dispersion such as *avg_return, avg_abs_return, rsi_14*), with macro/time features acting as contextualizers.

Taken together, these differences explain why our validation 20-day DA being near 70–71% is consequential: it demonstrates medium-horizon predictability specifically in rare, post-shock states, where most day-to-day sign classifiers are not tested.

## 4. Polarity as a first-order context for evaluation

Our polarity analysis shows a structural asymmetry:

- **Negative shocks:** ~79% DA at 20 days on validation; still >60% on test

- **Positive shocks:** Weaker (≈62% DA at 20 days on validation; ~53% on test)

Because the test window contains more positive shocks, the pooled DA naturally compresses. This asymmetry aligns with established mechanisms (partial mean-reversion after sell-offs, inventory/liquidity effects), and it clarifies why a single pooled DA understates attainable performance when polarity is modeled explicitly. In practice (were deployment the goal), polarity-specialized models or regime-aware ensembling would be the appropriate evaluation frame.

## 5. Feature Importance: What the Model Uses vs. What Moves Accuracy

Two complementary diagnostics paint a coherent picture of what drives predictability after shocks. Permutation importance (ΔRMSE on Test) shows that stock-specific trend and stress features, notably *avg_return, avg_abs_return, return_std, rsi_14*, and *daily_return*, deliver the unique out-of-sample lift. In other words, *recent drift and dispersion around the event* are the backbone of medium-horizon skill. By contrast, **SHAP** attributes substantial model-internal influence to *T10Y2Y* and a calendar/time tag **(***month***)**, alongside *avg_return*. This indicates the trees are conditioning predictions on regime (macro stance and coarse time effects) and interacting that context with micro trend/stress, even though the incremental Test accuracy from macro alone is modest once micro variables are present.

This reconciliation matters for interpretation and deployment. It suggests:

1. the signal that moves the needle for accuracy is micro and event-proximal;

2. regime awareness is still embedded in the model's decision rules (per SHAP), which supports our regime-sensitivity finding and motivates regime-aware ensembling or macro-conditioned thresholds. Adding more regime awareness (ie. a larger data-pool than 2 years) could very much help our tree model; and

3. caution with *month* as a high-rank SHAP feature: it likely acts as a coarse time/regime proxy over 2023–24 rather than true seasonality. A drop-column check typically shows minimal ΔRMSE loss, consistent with "contextualizer, not core predictor."

## 6. Regime shift: inverted → normal yield curve, and what changed throughout dataset timeline

For most of 2023–mid-2024, the Treasury yield curve (10y–2y) was inverted (negative) in our data (mean ≈ −0.39; min ≈ −1.08; max later turns positive), and then re-steepened (less negative → positive) into late 2024. In plain terms: an inverted curve signals tight policy and growth risk; equity rallies/subsequent shocks differ in tone from those under a positive (normal) curve, where markets lean more "risk-on" as investors price sooner easing or improving growth.

This shift is visible both in features (our T10Y2Y and its lag) and in realized shocks: the test window saw more large, positive shocks than the train/validation windows. This resulted in our model 'learning' the market during an inverted yield curve macroeconomic environment, and then predicting results in a positive yield curve environment. As the yield curve was a statistically significant variable (as noted in the feature selection section) and was the model's only proxy to the macroeconomic environment the stock was operating in, it is obvious as to why our model would be under estimating the positive moves present in this period. This likely led to at least a portion of the sharp drop in directional accuracy, and a large portion of the RMSE error score. The shock-polarity counts confirm the overall mix of positive shocks increasing during the test period (late 2024) and the daily shock distribution shows bursts clustered around a handful of dates, especially in the test period.

To contextualize those bursts, the most-volatile days were analyzed for date and reason for volatility to. Some concrete catalysts on several top dates in our series were flagged; e.g., Nov 6, 2024 (post-election rally that saw the largest positive stock market shocks in recent history, with the S&P reaching historical highs), Nov 14, 2023 (soft CPI surprise → relief rally), Aug 5, 2024 (global recession fears; worst U.S. day in ~2 years appears just before test set), and Sep 3, 2024 (AI hype check led by Nvidia reaching to some incredible +50%  single stock gains).

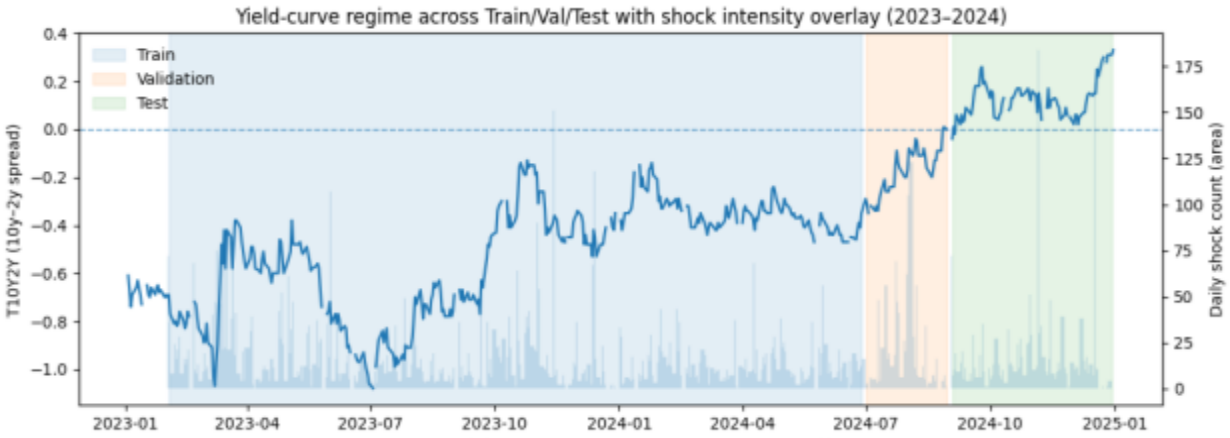



*Figure 5: Yield Curve Regime Across Full Dataset*

*Figure 5 shows the value of the yield curve across the validation set. Note: we start to see a steady trend up during the second half of the training set, and the Yield Curve Flips to positive just as the Test set begins, which drastically changes market views and investor behavior.*

**6a. Why this matters for our models**

The validation model learned patterns of post-shock drift/mean-reversion under an inverted-curve regime, then faced a more "positive-shock" test regime. Under covariate shift, DA typically compresses unless the learner adapts or conditions on regime. This exactly matches our out-of-sample DA drop and motivates the regime-aware extensions in Section 8 of this chapter.

## 7. Polarity Matters: Negative Shocks are "Easier" to Predict

When we split by shock polarity, the 20-day XGBoost trained on negative shocks attains ~79% DA, our single best directional result; the positive-shock model is weaker at 20 days (~62% DA). This asymmetry is theoretically plausible: large negative moves often exhibit partial mean-reversion over multi-week windows (liquidity demand shocks, forced selling, "overreaction" and correction follows), whereas positive jumps mix continuation and fade depending on news type and the reason for the strong volatile jump (earnings beats, guidance, macro relief rallies).

We therefore treat and deem polarity as a structural driver of predictability, not a mere artifact. In a production setting, we would either (a) train polarity-specialized models or (b) allow interactions between polarity and macro/volatility features.

## 8. Why Parsimonious Lagging + Recursive Trees Beat the 20-day LSTM

**8a. Feature design**

Our non-temporal models receive lagged features at t−1, −5, −10, −20, which inject short-term trend and momentum context without exposing every day's information (reducing "over-learning" of normal periods). With this design, XGBoost learns robust, non-linear splits across entities and macro states. Without being overexposed to the trends of seemingly 'normal' non-volatile days. By contrast, our LSTM consumes the full 20-day pre-shock window, which, given the rarity of shocks relative to normal days, encourages regression toward zero on abnormal post-shock behavior.

**8b. Quantitative contribution signals**

Permutation-importance showed avg_return, avg_abs_return, daily_return, rsi_14 and size proxies (market_cap) among the top contributors, with macro (T10Y2Y, T10Y2Y_t−20) adding modest but non-zero information under some targets. This is consistent with a tree relying first on stock-specific drift/volatility context and then adjusting for macro tone.

**8c. XGBoost as the top contender in this specific context**

In our post-shock setting, the LSTM is handicapped by the data geometry and the task's inductive bias. The signal of interest is sparse relative to abundant "normal" pre-shock days, so a 20-day encoder learns a mean-reverting trend that collapses forecasts toward zero, and then is tested on a sequence where mean reversion is an anomaly, and large shifts are common with the persisting upward trend during the 3 month test period. Its **s**equence bias toward level/phase dynamics is also mismatched to our target, direction after a regime break, whereas gradient-boosted trees, fed parsimonious lags (t−1/−5/−10/−20), recover local trend invariants with far fewer parameters. The deep model becomes data-hungry and brittle under covariate shift (pre-shock inputs ≠ post-shock states) unless we inject explicit state/event descriptors (e.g., earnings surprise, guidance tone, macro regime flags). Consequently, trees + carefully chosen lags produced sharper decision boundaries and higher 20-day DA in this dataset. This does not contradict the broader literature's tendency to suggest the use of NNs to predict stock prices; rather, it underscores data/feature-regime fit and shows how using volatility as an event based indicator changes what exactly it is we are trying to predict. Additionally, it points to a possibility of a much higher RMSE and DA if tree based models were the sole focus of a similar study.

## 9. Business relevance: DA in Practice

In practical use, this model would be deployed selectively rather than indiscriminately. A straightforward rule is to **trade only when the model's confidence is high**, operationalized by the absolute predicted return (or an equivalent confidence score). Because hit-rate typically increases monotonically with the absolute predicted return, imposing a threshold (e.g., act only on the top q-quantile of the predicted return) should **raise directional accuracy (DA)** at the cost of fewer signals, an expected precision coverage trade-off. In the same spirit, **polarity filtering** (treating negative-shock signals differently from positive-shock signals) and **regime gating** (requiring the term-structure proxy to match the regime the model learned best) would further concentrate on conditions where the decision boundary is strongest. For risk control, one would still monitor magnitude error (RMSE), cap position sizes when predicted magnitudes cluster near zero, and

abstain under low confidence. If desired, **probability calibration** or **conformal thresholds** can turn raw scores into interpretable decision cutoffs. We do not backtest such filters here, but conceptually they provide a disciplined path to translate the observed DA into a higher, execution-ready hit-rate while acknowledging the accompanying reduction in trade frequency and speak to the deployability of such a model.

## 10. How the "most-volatile days" illuminate our findings

The model's DA drop on test coincides with test-window clustering of polarity-specific catalysts, e.g., Nov 6, 2024 post-election risk-on surge; Aug 5, 2024 global recession fears; Sep 3, 2024 Nvidia-led AI reset; Nov 14, 2023 soft CPI-driven relief. These episodes differed in sign, breadth, and macro tone, helping explain why models trained on one mix of shocks generalize imperfectly to another. For added clarity, below is a chart that shows the most positive and most negative days throughout the dataset to illuminate the strength and frequency of volatility throughout the dataset.

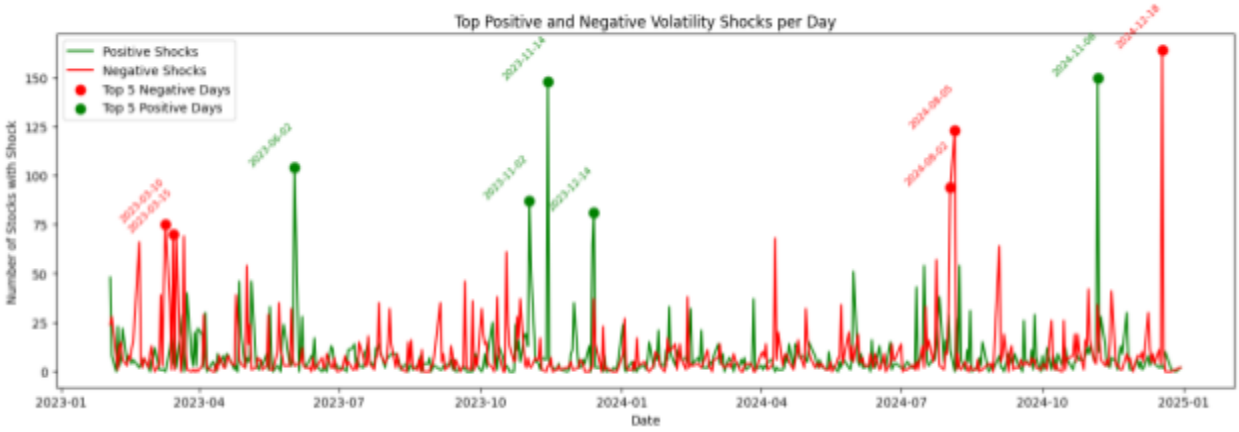



*Figure 6: Positive and Negative Volatility Shock*

*Figure 6 highlights the number of volatile shocks per day throughout the entirety of the dataset, and shows the date of days containing the most shocks.*

From Figure 6, it is apparent that the test set (which starts September 3rd and goes through the end of the test set) has some strange patterns that were not seen in the training set. There are two days in the test set that had an unprecedentedly high number of shocks and many big shocks preceded by relatively 'normal' days'. On the 6th of November there were almost 150 companies that experienced a shock to their price that day, and on the 18th of December, even more stocks experienced negative shocks. Looking at these days illuminates some insights that are important to future studies and explain at least part of the low RMSE scores observed in the test set model. On the 6th of November, stocks hit an all time high as Donald Trump took office as president of the United States. Soon after, on December 18th, the S&P dropped a historic 3% on the day, posting its first 10 day losing streak since the 1970's, and this was largely due to Fed Rate hikes and news from the Fed that rates would only be cut twice in the following year.

Huge daily jumps like this are impossible to predict, and the sheer volume of positives and negatives likely confused the model, especially since markets had been gaining at a relatively steady pace

throughout this 2 year period. The tree based model struggled particularly with large jumps, and tended to predict near 0 returns. These big days likely confused the model, and it would be interesting the see how a model would perform if massive days of turbulence, say days with 50+ stocks with volatile moves, were omitted from the model's learning and predicting patterns. In this way, the model would be predicting on 'normal' volatile swings. Another way to combat this issue would be to specifically train the model to identify types of volatility, for example rate cuts and elections. A benefit of this model is that it is meant to be deployed after the volatile day's news has settled, so using scrapers to understand the drivers of large volatility is plausible, and could greatly increase the effectiveness of the model in recognizing these 'non-normal', large volatility spikes and investors' responses to them.

# Conclusion

**Business Recommendations and Further Suggested Research**

## 1. Summary of Main Findings

### 1a. Research question and answer

*Can machine-learning models predict short- and medium-term stock price behavior following extreme volatility shocks, using firm-level and macro time-series features?*

Yes—at medium horizons the models predict direction with economically meaningful skill; at very short horizons the signal is modest. Predictability is regime- and polarity-sensitive.

### 1b. Headline results

On the validation window, XGBoost achieves ≈**70% DA** at **20 trading days**, ≈**60%** at **5 days**, and ~**50%** at **1 day**. On the held-out test window (later in time and under a different macro mix), DA remains **above chance** but attenuates to ~**57% (20d)**, **56% (5d)**, and **53% (1d)**, consistent with the yield-curve's shift and a higher share of positive shocks.

### 1c. Polarity matters

Negative-shock episodes are markedly more predictable: validation DA reaches **~79% at 20 days** for negatives, versus **~62%** for positives (with similar gaps on test). This supports a mean-reversion/liquidity interpretation after large down moves, while positive jumps mix continuation and fade.

### 1d. What drives the forecasts

Event-proximal, *micro* features (recent drift/dispersion such as **avg_return**, **avg_abs_return**, **rsi_14**) deliver most of the accuracy gains; *macro* context (e.g., **T10Y2Y**) helps position the baseline and interacts with micro signals, explaining the observed regime sensitivity.

### 1e. Implication

Medium-horizon, post-shock direction is learnable and could be made more robust in practice via selective execution (acting only on high-confidence signals) and regime/polarity-aware modeling, while 1-day predictions are best treated as descriptive context rather than standalone trading signals.

## 2. Limitations and Directed Next Steps

### 2a. Data scope & macro coverage

Two years across S&P 500 gives rich cross-sectional coverage but limited regime diversity. We would broaden macro inputs beyond T10Y2Y (e.g., implied vol term structure, credit spreads, macro surprise indices, news/sentiment windows). Sentiment scores could be a particularly interesting macro indicator, as it can be computed daily and would give real time macro level information usable by the model on a truly daily level, in an unlagged manner. In addition to adding strong, leading market indicators, we should ensure that multiple regime types are covered in our data timespan. The model would benefit from learning on both bearish markets and bullish markets.

**2b. Shock definition**

A single global z-threshold is simple and robust, but we should stress-test thresholds (e.g., $|z|>2.5$) and consider entity-adaptive thresholds.

**2c. Modeling**

NNs likely need more regimes, richer text/news inputs, and longer history to shine here. With current features, trees are a better bias-variance match. For robustness out of sample, we recommend:

- Regime-aware ensembling (condition on T10Y2Y sign/level or a learned regime label).
- Polarity-specialized models (already promising at 20-day).
- Selective prediction (abstain when $|\hat{y}|$ is small).

## 3. Proposed Future Research

Building on these findings, two avenues appear especially promising for extending predictability when volatility is treated as an event-based signal. First, a deliberately *decision-oriented* specification could recast the task as pure classification and optimize directly for Directional Accuracy (DA). Framing the problem this way would align the objective with a sign-based trading use case and permit richer inputs (e.g., additional technical indicators, text/sentiment, and longer macro histories) without the calibration burdens of regression. A larger and more heterogeneous training set, spanning multiple regimes, would also improve robustness to the kinds of distributional shifts that compressed test performance here. This proposition simply changes the task, and allows for improvements that can easily be parsed in to the code provided in this research.

Second, a *backwards-induction* approach could model "types of volatility events" rather than treating all shocks as homogeneous. Concretely, one would (i) characterize the 20 trading days *pre-event* (trend, dispersion, volume, news tone) and the *day-of* shock signature (jump size, polarity, breadth), (ii) cluster these descriptors to obtain an event taxonomy, and (iii) condition forecasts on the inferred event class. Such archetypes (e.g., "policy-relief rallies," "forced-selling cascades," "pre-announcement run-ups," "broad risk-off," or "momentum checkups") could then be paired with class-specific models or interaction terms. By narrowing the decision context, this design may mitigate the tendency of pooled models to

regress toward zero after rare, dissimilar shocks and could raise DA where mechanisms are more uniform within class.

Together, a DA-first classifier and an event-typing pipeline provide complementary paths: the former prioritizes practical hit-rate under broad coverage, while the latter targets *conditional* accuracy by learning how distinct shock archetypes resolve. Both directions are natural continuations of this work and directly address its central limitation, regime and composition sensitivity, without abandoning the event-based framing that proved most fruitful.

# References


Aghabozorgi, S., Shirkhorshidi, A.S. and Wah, T.Y. (2015) ‘Time-series clustering – a decade review’, Information Sciences, 53, pp. 16–38.

Allen, D.E., McAleer, M. and Scharth, M. (n.d.) ‘Realized volatility uncertainty’. School of Accounting, Finance and Economics, Edith Cowan University. Available at: https://ro.ecu.edu.au/ecuworks/7108/

Athey, S. and Imbens, G.W. (2019) ‘Machine learning methods that economists should know about’, Annual Review of Economics, 11, pp. 685–725.

Breiman, L. (1996) ‘Bagging predictors’, Machine Learning, 24(2), pp. 123–140.

Campbell, J.Y. and Thompson, S. (2008) ‘Predicting excess stock returns out of sample: Can anything beat the historical average?’, Review of Financial Studies, 21(4), pp. 1509–1531.

Chen, L., Pelger, M. and Zhu, J. (2019) ‘Deep learning in asset pricing’, SSRN Working Paper No. 3350138.

Christensen, K., Siggaard, M. and Veliyev, B. (2023) ‘A machine learning approach to volatility forecasting’, Journal of Financial Econometrics, 21(5), pp. 1680–1727. https://doi.org/10.1093/jjfinec/nbac020

Dangl, T. and Halling, M. (2008) ‘Predictive regressions with time-varying coefficients’, Working Paper, Vienna University of Technology.

Elyasiani, E., Gambarelli, L. and Muzzioli, S. (2017) ‘The information content of corridor volatility measures during calm and turmoil periods’, Quantitative Finance and Economics, 4(1), pp. 454–473.

Engle, R., Ghysels, E. and Sohn, B. (2006) ‘On the economic sources of stock market volatility’, Manuscript, New York University.

Ge, W., Lalbakhsh, P., Isai, L., Lenskiy, A. and Suominen, H. (2023) ‘Neural network–based financial volatility forecasting: A systematic review’, ACM Computing Surveys, 55(1), Article 14, 30 pp. https://doi.org/10.1145/3483596

Giot, P. (2005) ‘Relationships between implied volatility indexes and stock index returns’, The Journal of Portfolio Management, 31(3), pp. 92–100.

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

Khan, W., Ghazanfar, M.A., Azam, M.A., Karami, A., Alyoubi, K.H. and Alfakeeh, A.S. (2020) ‘Stock market prediction using machine learning classifiers and social media, news’, Journal of Ambient Intelligence and Humanized Computing, pp. 1–24.

Lewellen, J., Nagel, S. and Shanken, J. (2010) ‘A skeptical appraisal of asset pricing tests’, Journal of Financial Economics, 96(2), pp. 175–194.

Lubnau, T.M. and Todorova, N. (2015) ‘The calm after the storm: Implied volatility and future stock index returns’, The European Journal of Finance, 21(15), pp. 1282–1296.

Marquering, W. and Verbeek, M. (2005) ‘The economic value of predicting stock index returns and volatility’, Journal of Financial and Quantitative Analysis.

Mora-Valencia, A., Rodríguez-Raga, S. and Vanegas, E. (2021) ‘Skew index: Descriptive analysis, predictive power, and short-term forecast’, North American Journal of Economics and Finance, Article 101356.

Pabba, M.P., Kiran, S.K., Reddy, S.S., Prakash, I.S. and Srikar, S.S. (2023) ‘Stock market price prediction using machine learning’, Journal of Engineering Sciences, 14(05).

Paye, B. (2010) ‘Do macroeconomic variables predict aggregate stock market volatility?’, Working Paper Series.

Shaban, W.M., Ashraf, E. and Slama, A.E. (2023) ‘SMP-DL: A novel stock market prediction approach based on deep learning for effective trend forecasting’, Neural Computing and Applications, pp. 1–25.

Shen, S., Jiang, H. and Zhang, T. (2012) ‘Stock market forecasting using machine learning algorithms’, Technical Report, Department of Electrical Engineering, Stanford University.

Sill, K. (1993) ‘Predicting stock market volatility’, Federal Reserve Bank of Philadelphia Business Review. Available at: https://www.philadelphiafed.org/-/media/FRBP/Assets/Economy/Articles/business-review/1993/brjf98ks.pdf

Tiwari, A. (2025) ‘Harnessing artificial intelligence for stock market forecasting: A machine learning approach to predicting financial trends’, International Journal of Scientific Research in Engineering and Management, 9(4), pp. 1–9. https://doi.org/10.55041/IJSREM44969

# Appendix

A. Below are the detailed computations used for the engineered features: *avg_return, avg_abs_return, avg_raw_return, rolling_market_vol_10d, return_std*, and *RSI*.

$$\text{avg\_return}_t = \frac{1}{N}\sum_{k=1}^{N} r_{t-k} \qquad \text{with } N = 20 \quad \text{avg\_abs\_return}_t = \frac{1}{M_t}\sum_{i\in\mathcal{U}_t} |r_{i,t}|$$

$$\text{avg\_raw\_return}_t = \frac{1}{M_t}\sum_{i\in\mathcal{U}_t} r_{i,t}$$

Define $A_t = \text{avg\_abs\_return}_t$. Then

$$\mu_t = \frac{1}{L}\sum_{j=0}^{L-1} A_{t-j}, \qquad \text{rolling\_market\_vol\_10d}_t = \sqrt{\frac{1}{L-1}\sum_{j=0}^{L-1}(A_{t-j}-\mu_t)^2},$$

$$\text{return\_std}_t = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}\left(r_{t-k} - \text{avg\_return}_t\right)^2} \qquad \text{with } N = 20$$

$$\Delta_t = P_t - P_{t-1}, \quad \text{gain}_t = \max(\Delta_t, 0), \quad \text{loss}_t = \max(-\Delta_t, 0).$$

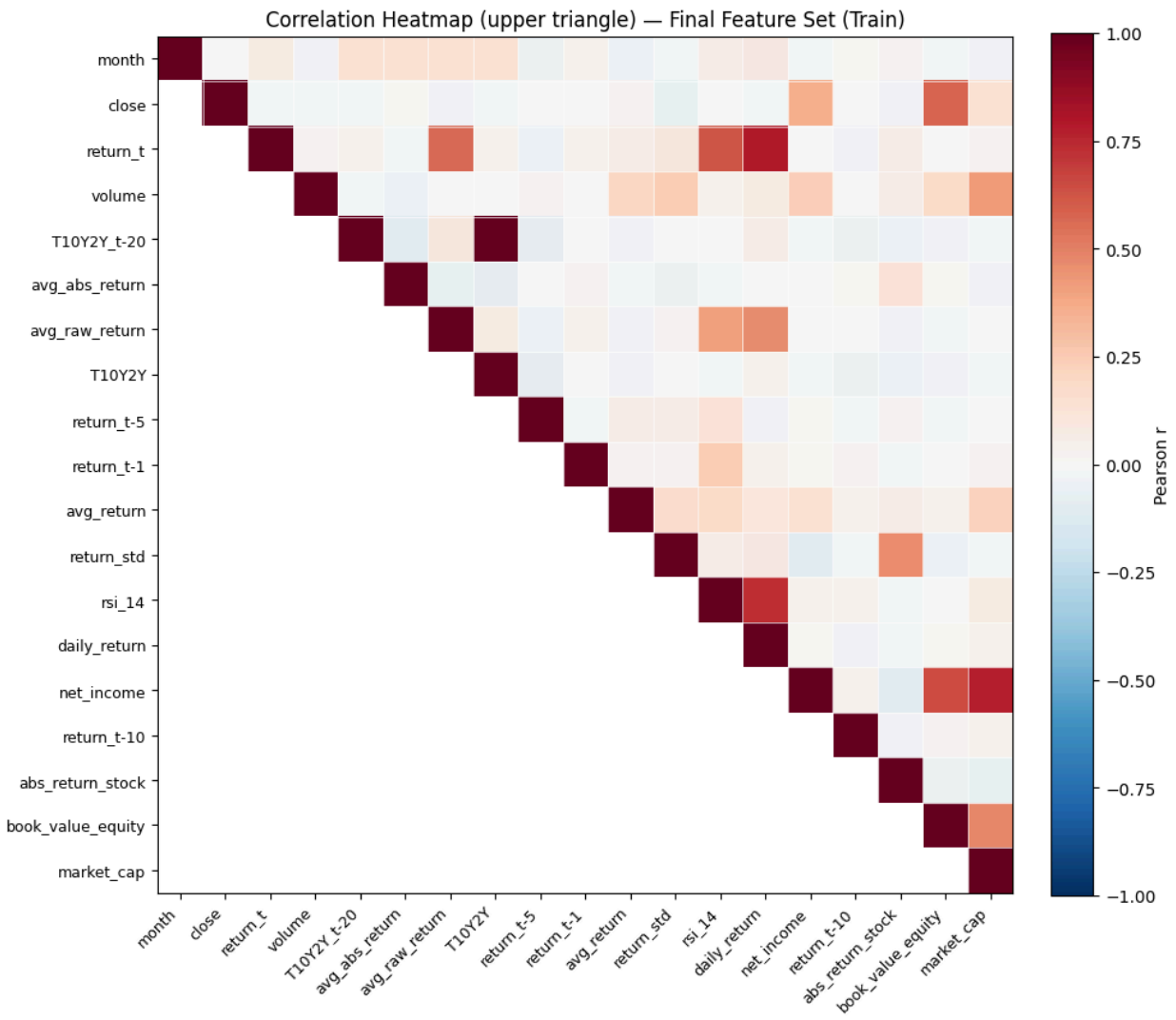
$$\text{avg\_gain}_t = \frac{(n-1)\cdot \text{avg\_gain}_{t-1} + \text{gain}_t}{n},$$

$$\text{avg\_loss}_t = \frac{(n-1)\cdot \text{avg\_loss}_{t-1} + \text{loss}_t}{n}, \qquad n = 14$$

$$\text{RS}_t = \frac{\text{avg\_gain}_t}{\text{avg\_loss}_t + \varepsilon},$$

$$\text{RSI}_{14,t} = 100 - \frac{100}{1 + \text{RS}_t},$$

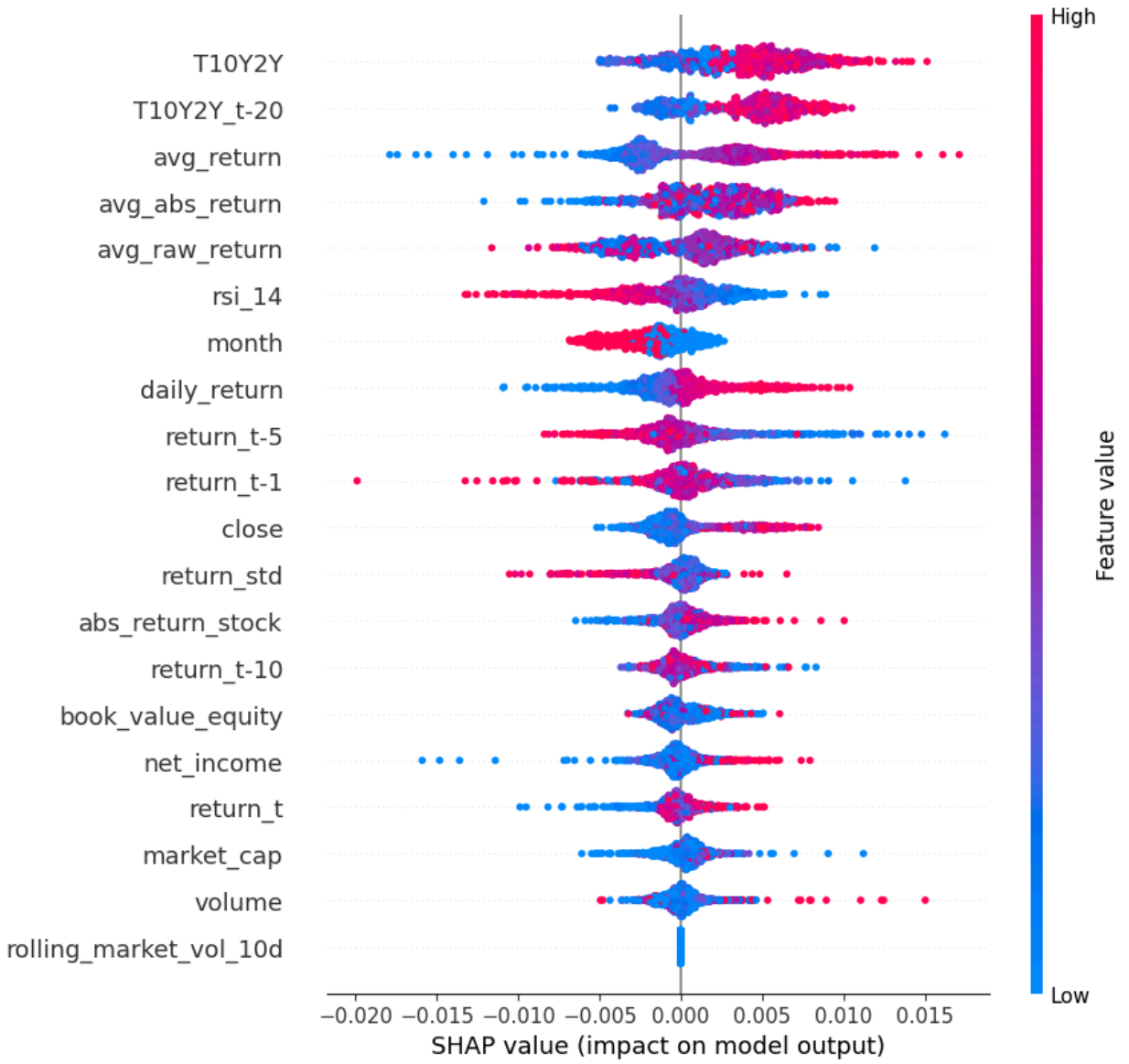
B. Feature Correlation Heatmap for the features present in the final feature set.



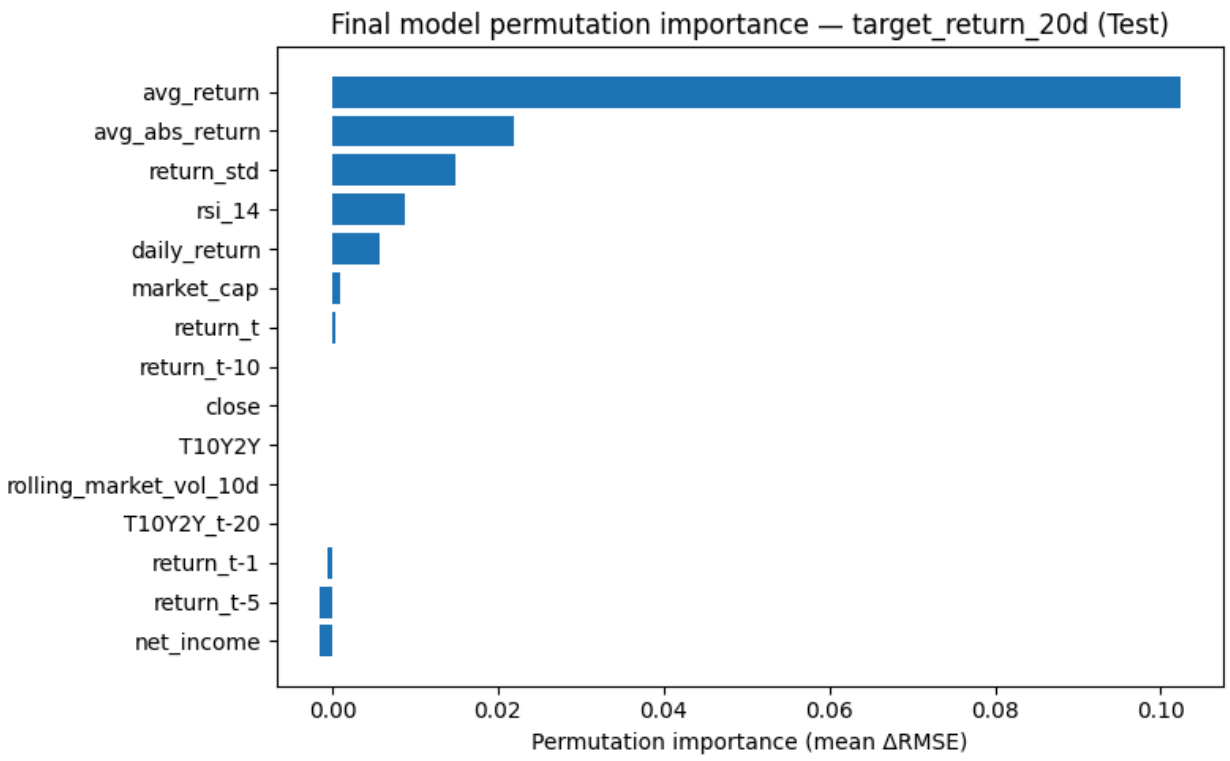

C. Table showing the tuned hyperparameters used in the final XGBoost model.

| Hyperparameter | Value |
|---|---|
| n_estimators | 649 |
| learning_rate | 0.0438 |
| max_depth | 9 |
| gamma | 0.0136 |
| colsample_bytree | 0.5694 |
| subsample | 0.9648 |
| reg_alpha | 0.4118 |
| reg_lambda | 0.3489 |

D. Plot showing the feature importance using SHAP as the method of finding importance. More important features are listed towards the top, and impact is bidirectional.




E. Permutation importance output for the 20d target return. Feature importance is listed from most important feature at the top to least important feature at the bottom. Notice the feature importance is very different from the SHAP scores, and this is because permutation importance and SHAP define ‘importance’ very differently. SHAP and permutation deal with collinearity and feature interactions very differently. Additionally, SHAP is looking at how much an individual feature contributes to individual predictions, whereas permutation is looking at how much the removal of a feature affects the model’s accuracy, so they are measuring very different things.


Final model permutation importance — target_return_20d (Test)
avg_return
avg_abs_return
return_std
rsi_14
daily_return
market_cap
return_t
return_t-10
close
T10Y2Y
rolling_market_vol_10d
T10Y2Y_t-20
return_t-1
return_t-5
net_income
0.00
0.02
0.04
0.06
0.08
0.10
Permutation importance (mean ΔRMSE)