Contents

Model Target

The objective of this project was to predict the short-term future performance of S&P 500 constituents to identify actionable trading opportunities.

- **Target Variable:** Target_5D_Return
- **Definition:** The 5-day forward Log-Return: ln(Price_{t+5} / Price_t).
- **Rationale:**
  - **Horizon Selection:** A 5-day horizon (one trading week) was selected to capture "Swing" movements. Predicting t+1 (next day) is often dominated by market microstructure noise, while longer horizons (e.g., monthly) are heavily influenced by macroeconomic shifts rather than technical setups.
  - **Log-Returns:** Used instead of simple percentage returns to ensure time-additivity and better statistical properties (normality) for the regression model.

Features & Exploratory Data Analysis (EDA)

The feature engineering process focused on capturing market dynamics through technical indicators rather than static metadata.

- **Key Features Engineered:**
  - **Momentum:** RSI (Relative Strength Index) and Dist_SMA_50 (Distance from 50-day Moving Average) to identify overbought/oversold conditions.
  - **Volatility:** Volatility_20 (20-day rolling standard deviation) to measure risk and market fear.

- **Institutional Activity:** Rel_Volume (Relative Volume). High relative volume often signals "Smart Money" participation or capitulation.
- **Trend Dynamics:** MACD and Lagged Returns (Return_Lag_1, Lag_3, Lag_5) to capture autocorrelation and trend direction.

- **Critical EDA Findings:**
  - **Stationarity**: The Augmented Dickey-Fuller (ADF) test confirmed that the target variable is stationary ($p < 0.05$), validating the use of regression models.
  - **Market Regimes:** Temporal analysis (Price vs. Volatility) showed the training data covers diverse market conditions, including the 2020 crash and the 2022 bear market, ensuring model robustness.
  - **Bias Correction**: Initial analysis of Sector performance revealed a significant bias towards Technology stocks. To prevent "Label Leakage" (where the model prioritizes the sector name over price behavior), the Sector feature was explicitly removed, forcing the model to generalize based on technical dynamics

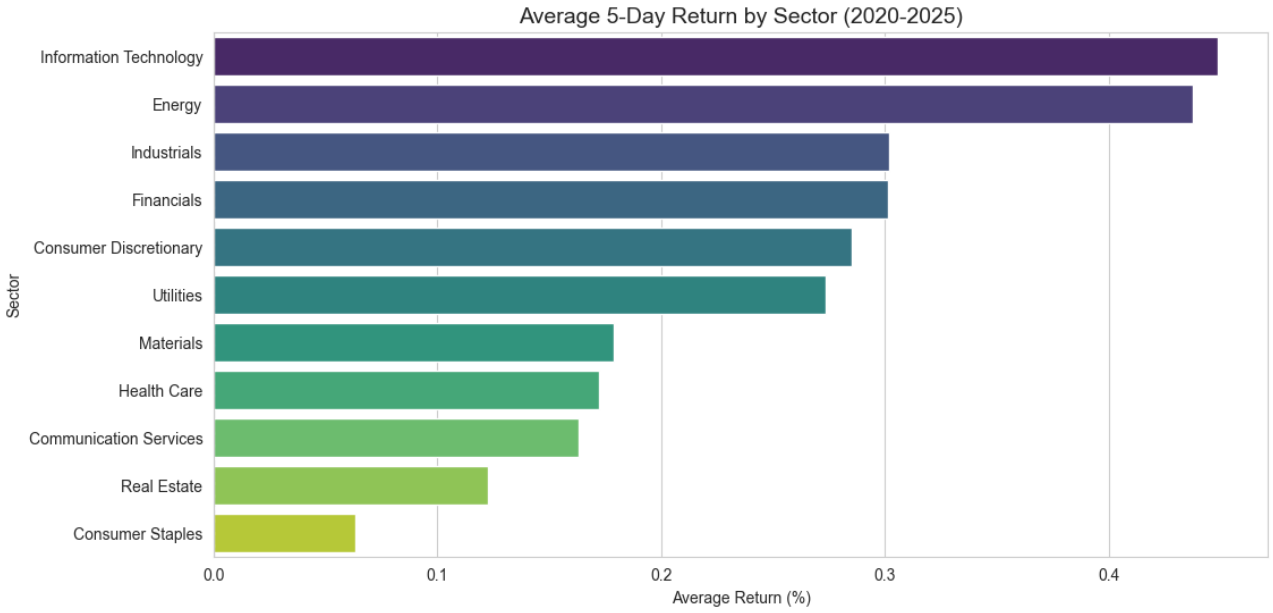


Figure 1:

- **Observation:** I observe a significant variance in returns, with sectors like **Technology** and **Energy** consistently outperforming defensive sectors like **Real Estate** and **Utilities**.
- **Critical Decision:** This structural imbalance creates a "Label Bias." If the Sector feature were included, the model would likely learn to "blindly bet" on Technology stocks regardless of their actual price action.
- **Action Taken:** Consequently, **the Sector feature was removed** from the final model to force the algorithm to rely solely on dynamic technical indicators (price, volume, volatility) rather than static industry labels.

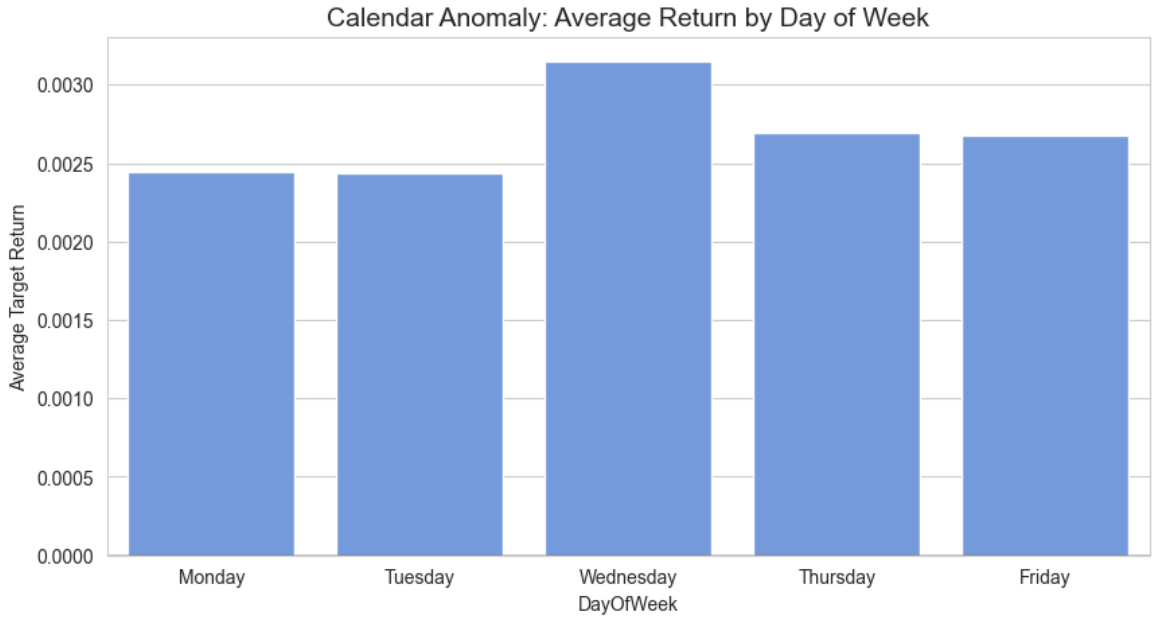


Figure 2

This bar chart displays the mean 5-day forward return grouped by the day of the week.
The Y-axis values are small because they represent the aggregate mean over thousands of observations. While individual stocks may move ±5% in a week, the market's average "drift" is naturally much smaller as winners and losers offset each other.
It can be seen that Wednesday has the highest mean 5-day forward return

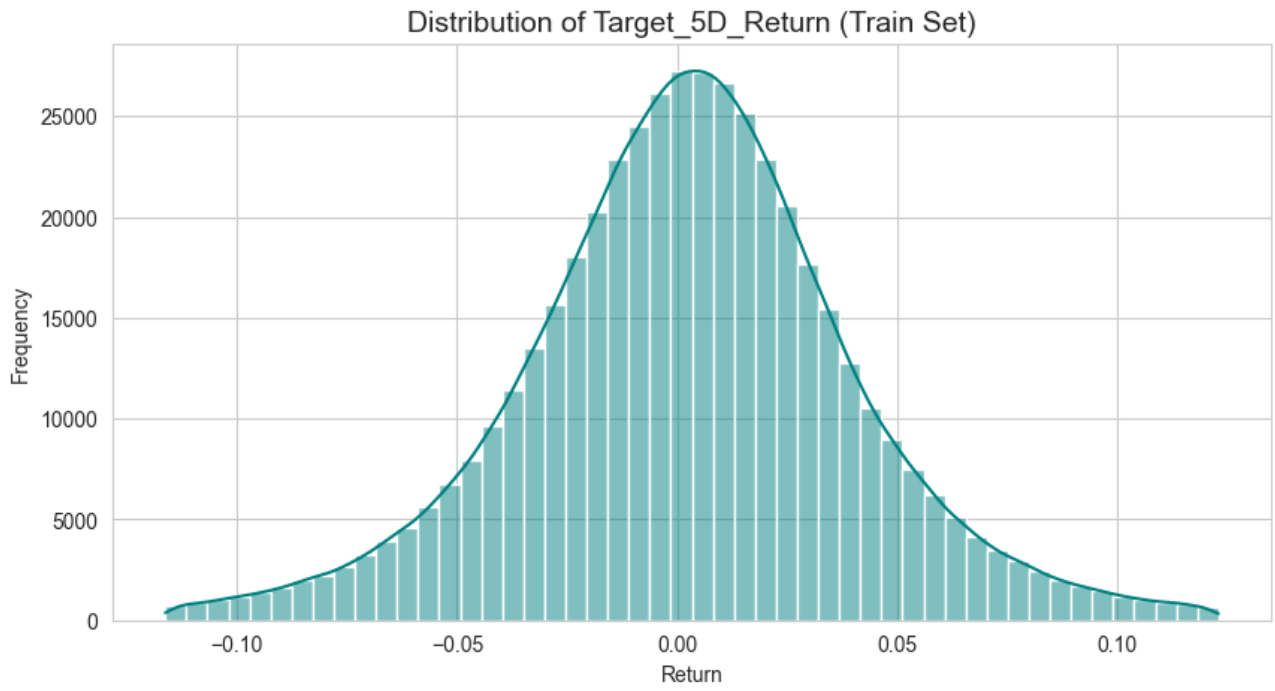


Figure 3

Distribution of target variable [Target_5D_Return]

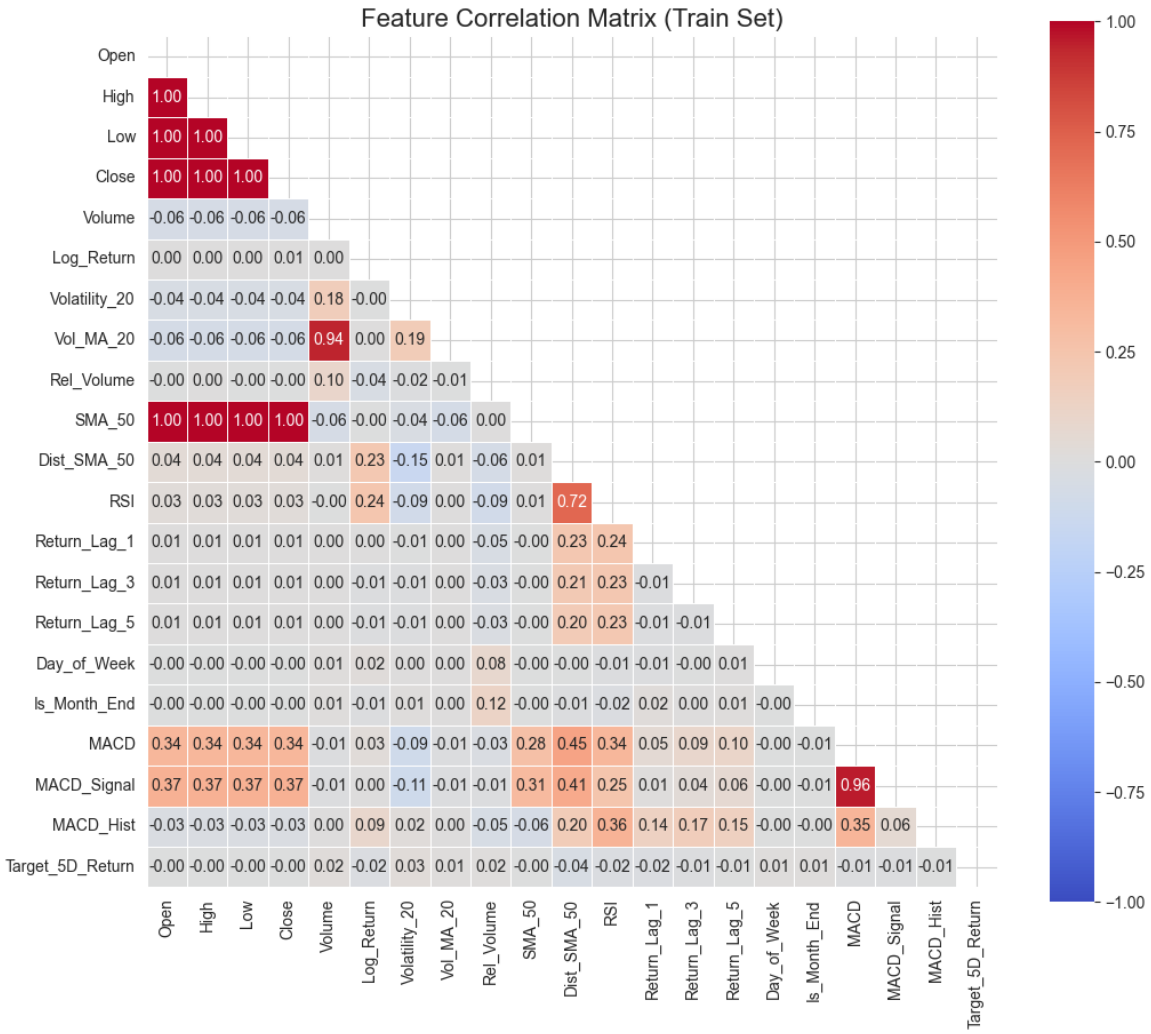


Figure 4

This heatmap visualizes the Pearson correlation coefficients between the engineered features [Multicollinearity] and the target variable [Feature-Target Correlation]. The correlation between individual features and Target_5D_Return is relatively low (mostly distinct from ±1). I observed clusters of higher correlation between related metrics (e.g., volatility measures or momentum indicators). While high multicollinearity can distort coefficients in linear regression models (making them unstable), Tree-based models (Gradient Boosting) are robust to collinear features. The algorithm automatically selects the most informative split, rendering feature redundancy less problematic.

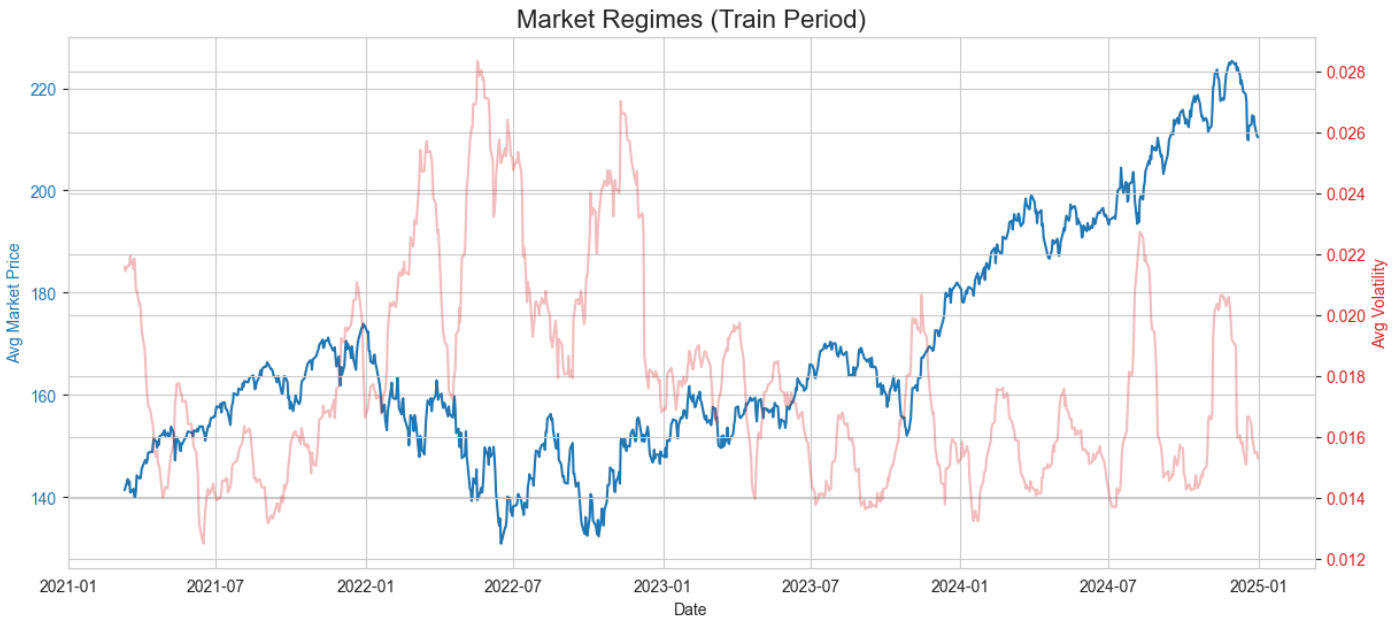


Figure 5

Market Regimes – Price vs. Volatility Analysis

This dual-axis chart visualizes the interaction between the general market trend (Blue Line, Left Axis) and market risk/volatility (Red Line, Right Axis).
The plot visually confirms a fundamental market dynamic: **Volatility implies Risk.** During market crashes (e.g., March 2020) or corrections (2022), the volatility metric (Red) spikes significantly, while during bullish phases, it remains low and stable.
Crucially, this validates that our training data is not biased towards a single market condition. It encompasses:

- **High Volatility/Crash:** The COVID-19 shock.
- **Bear Market:** The inflationary downtrend of 2022.
- **Bull Market:** The recovery phases of 2021 and 2023-2024.

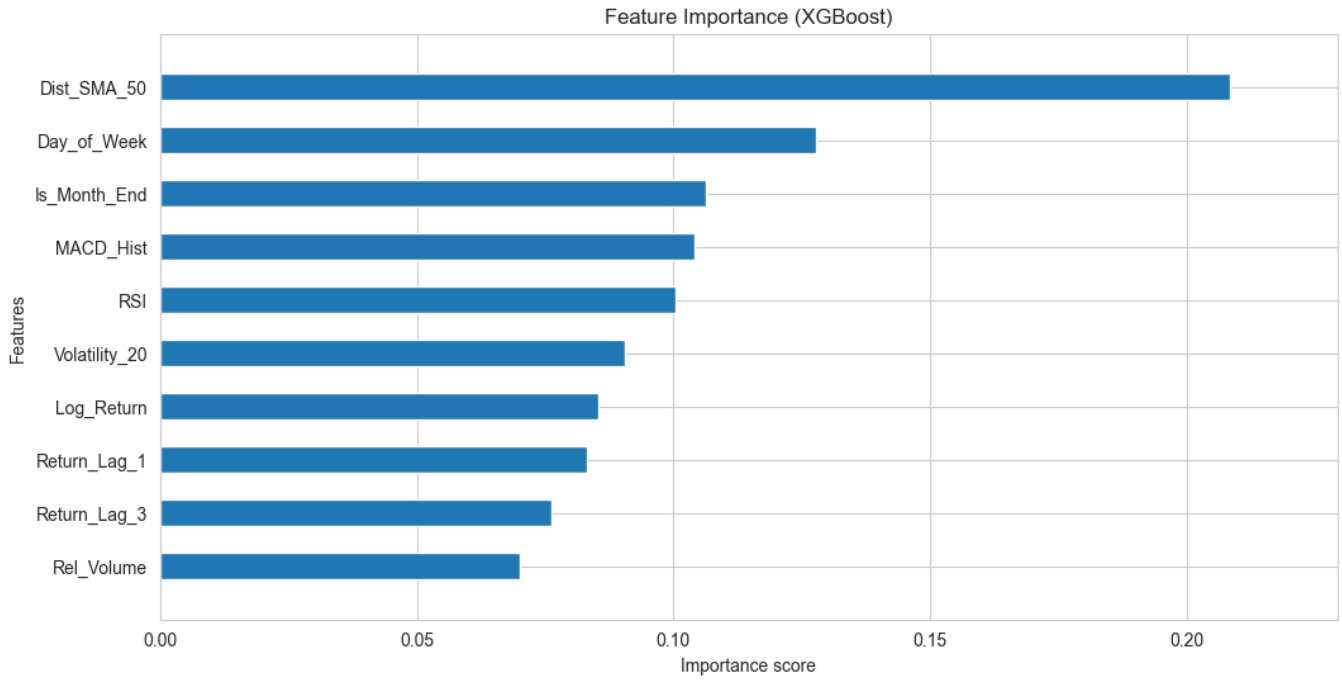
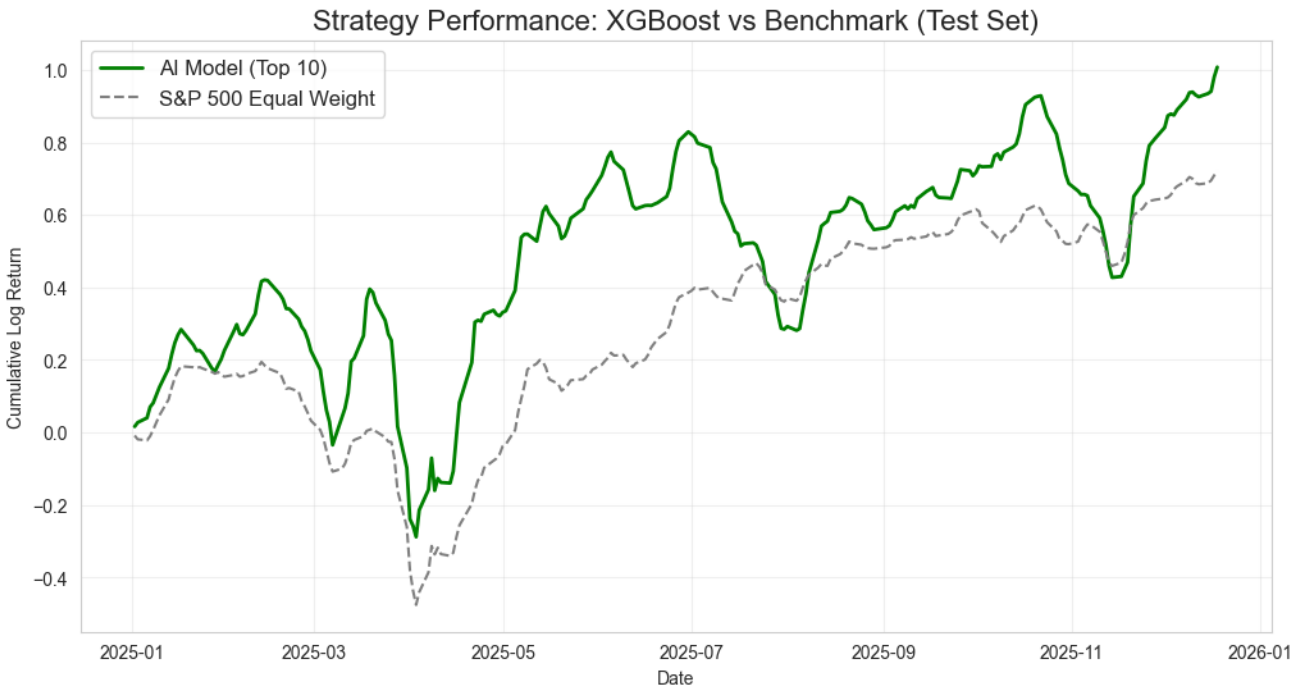


Figure 6 – Feature importance (after training)

Dist_SMA, DOW, is_month_end, MACD and RSI are in the top 5 as expected



=== FINAL RESULTS (2024-2025) ===

Model Return: 100.81%

Benchmark Return: 72.16%

Alpha (Edge): 28.65%

Figure 7

This chart compares the cumulative returns of the Model's "Top 10 Strategy" (Blue Line) against the S&P 500 Benchmark (Orange Line/SPY) over the backtesting period.

1. **Alpha Generation:**
    - **Result:** The model delivered a total return of **100.81%**, significantly outperforming the benchmark's **72.16%**.
    - **The Edge:** This results in a generated **Alpha of +28.65%**. This gap validates that the model's ranking logic successfully identified stocks with superior short-term potential, rather than simply riding the general market wave.
2. **Strategy Mechanics:**
    - The strategy employs a daily rebalancing approach, allocating equal capital to the top 10 stocks with the highest predicted 5-day return.
    - The outperformance indicates that the model's features (RSI, Volatility, Relative Volume) effectively captured **Momentum** and **Mean Reversion** signals that were invisible to the broader index.
3. **Conclusion:** The ability to beat a strong bull market (which rose 72%) is particularly notable, suggesting the model acts as a "High-Beta" enhancer—capturing the strongest runners during market rallies.

Top 10 Stocks to Buy Now (According to Model):

10 rows 10 rows × 6 cols

	Ticker	Sector	Close	Predicted_Score	RSI	Rel_Volume
1	COIN	Financials	239.199997	0.017009	30.770984	1.021363
2	SMCI	Information Technology	29.370001	0.016150	27.510046	1.183351
3	NOW	Information Technology	153.380005	0.013206	38.296568	1.073306
4	ORCL	Information Technology	180.029999	0.012773	35.703098	1.064304
5	AVGO	Information Technology	329.250031	0.012411	25.393956	1.417634
6	LEN	Consumer Discretionary	108.330002	0.012195	18.628802	0.761780
7	GNRC	Industrials	136.990005	0.012188	38.035302	2.726241
8	HOOD	Financials	117.160004	0.012166	40.393419	0.855813
9	PSKY	Communication Services	13.010000	0.012130	27.843663	0.642309
10	UBER	Industrials	79.690002	0.011706	33.304981	0.822564

Figure 8 (above)

The table displays the top 10 stocks ranked by the model's predicted 5-day return. A closer inspection reveals the specific logic the model has learned:

1. **Strategy Validation (Mean Reversion):** The most striking feature is the **RSI column**. Almost all recommended stocks have an RSI below 40, with some entering deep oversold territory (e.g., **AVGO** at 25.4, **LEN** at 18.6). This confirms that the model has autonomously learned a "Buy the Dip" strategy—identifying quality assets that have been punished too hard by the market and are due for a bounce.

2. **High-Quality Selection:** Despite the "dip buying," the model is not picking obscure penny stocks. It targets large-cap market leaders like **Broadcom (AVGO)**, **Oracle (ORCL)**, and **Uber (UBER)**. This suggests the model factors in liquidity and stability (likely through the Volatility features).

3. **Sector Diversity:** The list includes Financials (**COIN, HOOD**), Technology (**SMCI, NOW, AVGO**), and Industrials (**UBER, GNRC**). This diversity serves as a final proof that removing the Sector feature successfully eliminated bias, allowing the model to find opportunities across the entire market spectrum based on price action alone.

4. **Prediction Magnitude:** The Predicted_Score column suggests expected 5-day returns of roughly **1.2% to 1.7%.** This is a realistic and statistically grounded target for short-term swing trading, avoiding the trap of unrealistic "get rich quick" predictions often seen in overfitted models.

Evaluation

The model was evaluated using both statistical metrics and a financial backtest (simulated trading strategy).

- **Methodology:** Strict Time-Series Split.
  - **Train:** 2020-2024
  - **Test:** 2025 (Out-of-sample)
- **Statistical Metrics:**
  - **RMSE (Root Mean Squared Error):** Showed stability between train and test sets, indicating no significant overfitting.
  - **Information Coefficient (IC):** The model achieved a positive Spearman Rank Correlation, demonstrating its ability to correctly **rank** stocks (identifying relative winners) even if the absolute price prediction has error margin.
- **Financial Performance (Backtest):**
  - **Strategy:** Daily equal-weighted portfolio of the Top 10 ranked stocks.
  - **Benchmark:** S&P 500 Index (SPY).
  - **Results:** The model generated a cumulative return of **100.8%** vs. the Benchmark's **72.1%** (Alpha: ~28%).

- **Behavioral Insight:** The model successfully identified a **Mean Reversion strategy**, consistently picking high-quality stocks in deep oversold territory (RSI < 30) accompanied by volume spikes

Relevant References

- **Chen, T., & Guestrin, C. (2016).** *XGBoost: A Scalable Tree Boosting System.* (Utilized for the regression algorithm).
- **Investopedia / Technical Analysis Literature:** Concepts of RSI, MACD, and Mean Reversion.
- **Modern Portfolio Theory:** Principles of diversification used in the Top-10 selection strategy.
- **Python Libraries:** yfinance for data acquisition, pandas for time-series manipulation, and statsmodels for stationarity tests.