Machine Learning Project for Predicting Stock Prices

Mauricio Aguas Fonseca

Date: August 21, 2025

I. Introduction

In this project, we combine historical stock price data with market news headlines to explore whether incorporating textual information can improve the predictive power of traditional price-based models. The analysis focuses on seven major technology companies listed on the NASDAQ:

- Apple (AAPL).
- Amazon (AMZN).
- Alphabet (GOOGL).
- Meta (META).
- Microsoft (MSFT).
- Nvidia (NVDA).
- Tesla (TSLA).

The datasets used in this study are:

- **Price data:** Open, high, low, close, and trading volume for each ticker.
- **News data:** Headlines and summaries with associated timestamps and tickers.

The workflow of this project is divided into four stages:

1. **Exploratory Data Analysis (EDA)**: Understanding distributions, relationships, and basic dynamics of prices and returns.
2. **Feature Engineering**: Creating meaningful variables from both structured (prices, volumes, technical indicators) and unstructured data (sentiment analysis of news).
3. **Model Development**: Testing different machine learning models to predict next-day returns or directional movement.
4. **Trading Strategy Backtest:** Evaluating the real-world applicability of the best models by simulating a trading strategy with transaction costs, analyzing profitability, risk (volatility, drawdowns), and robustness through bootstrap simulations.

II. Exploratory Data Analysis

II-A. *Number of Records per Ticker*

The bar chart confirms a very positive quality of the dataset: it is **perfectly balanced**. Each ticker has the exact same number of historical records, which eliminates common issues of missing data or time series of unequal lengths. This uniformity is indicative of high data quality and consistency, which greatly simplifies the preprocessing and modeling phases.

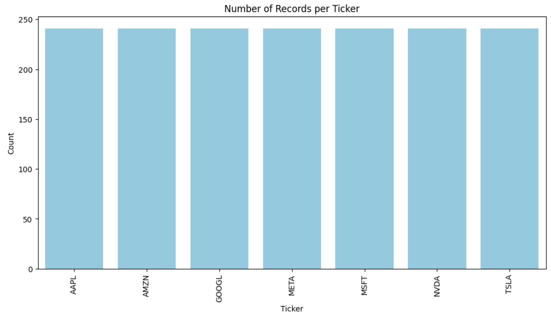


Figura 1. Number of records per ticker.

II-B. *Correlation Among Stock Returns*

The correlation heatmap shows that, as expected, a positive correlation exists among all pairs of stocks, given that they belong to the same technology sector and are influenced by common macroeconomic factors. However, the strength of this relationship varies; some stocks like Microsoft and Google show a stronger link. The case of Tesla stands out, displaying the lowest correlations with the rest of the group, indicating that its price movements often respond to more idiosyncratic, company-specific factors. These correlations suggest that diversification benefits within this basket of stocks are limited, except potentially with Tesla.

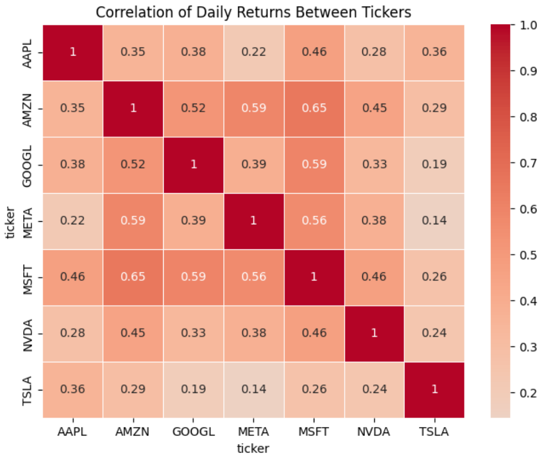


Figura 2. Correlation matrix of daily returns.

II-C. Distribution of Daily Returns

The histograms of daily returns reveal a fundamental characteristic of financial assets. While the mean of the returns is consistently centered around zero, the shape of the distribution is not normal. Instead, it presents clear **leptokurtosis (fat tails)**. This means that extreme events, both positive and negative, are considerably more frequent than a normal distribution model would predict, which is a crucial factor for risk assessment.

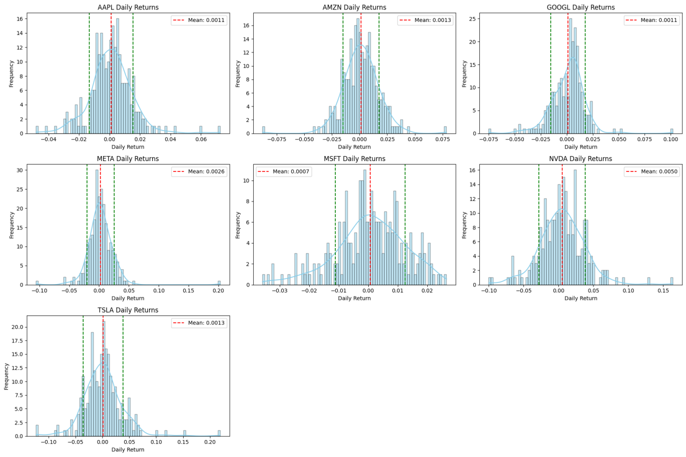


Figura 3. Distribution of daily returns for the seven tickers.

II-D. Time Series Analysis of Price and Volume

Observing the time series plots, it is evident that the stock prices are non-stationary, as they all exhibit clear, defined trends over time. A phenomenon known as **volatility clustering** is apparent, where periods of high price fluctuation are followed by periods of relative calm. Furthermore, it is notable that spikes in trading volume often coincide with the most significant price movements, suggesting a strong market reaction to specific events.

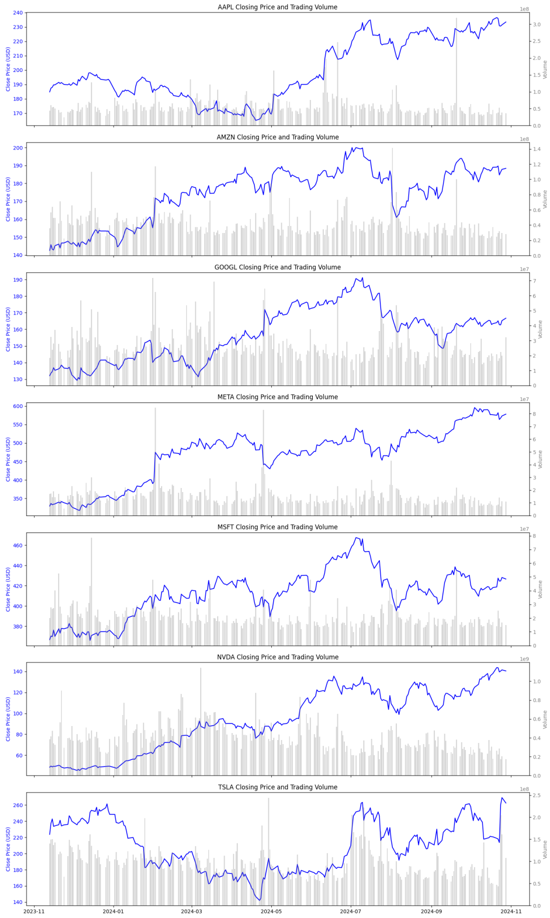
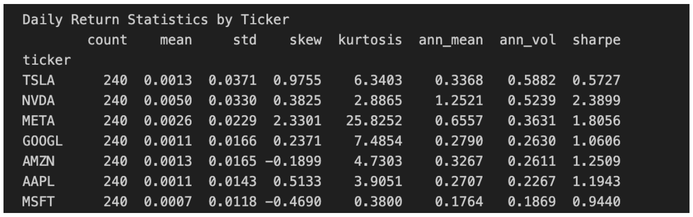


Figura 4. Closing prices and trading volumes.

II-E. Detailed Statistical Analysis of Returns and Volatility

A deeper statistical analysis of the daily returns allows us to precisely quantify the observations regarding risk and the shape of the distributions. The statistics table reveals that virtually all stocks exhibit a **kurtosis** greater than 3, numerically confirming the presence of **fat tails (leptokurtosis)** identified visually in the histograms. The case of META is particularly extreme, with a kurtosis exceeding 25, which indicates a very high propensity for surprising and large-magnitude price movements. This concept of risk is complemented by the **annualized volatility** chart, which clearly illustrates that TSLA and NVDA are, by a wide margin, the most volatile assets in the group, while MSFT stands out as the most stable. The table also provides information about the **skewness** of the returns; for example, the positive skewness in stocks like NVDA and META suggests that, historically during this period, they have had a higher frequency of extreme positive returns compared to extreme negative ones. Altogether, this analysis provides us with a detailed fingerprint of each asset's risk-return profile.



```
Daily Return Statistics by Ticker
        count    mean     std    skew  kurtosis  ann_mean  ann_vol  sharpe
ticker
TSLA      240  0.0013  0.0371  0.9755    6.3403    0.3368   0.5882  0.5727
NVDA      240  0.0050  0.0330  0.3825    2.8865    1.2521   0.5239  2.3899
META      240  0.0026  0.0229  2.3301   25.8252    0.6557   0.3631  1.8056
GOOGL     240  0.0011  0.0166  0.2371    7.4854    0.2790   0.2630  1.0606
AMZN      240  0.0013  0.0165 -0.1899    4.7303    0.3267   0.2611  1.2509
AAPL      240  0.0011  0.0143  0.5133    3.9051    0.2707   0.2267  1.1943
MSFT      240  0.0007  0.0118 -0.4690    0.3800    0.1764   0.1869  0.9440
```

Figura 5. Descriptive statistics of daily returns: mean, volatility, skewness, and kurtosis.

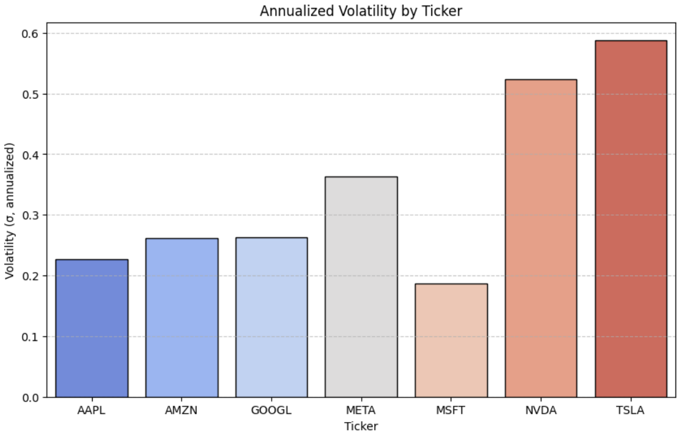


Figura 6. Annualized volatility comparison among tickers.

III. Feature Engineering

Our feature engineering process transforms raw historical price and news data into a rich, model-ready dataset for predicting stock returns. The process involves several key steps.

First, we aggregate and merge the data, aligning daily market news with the corresponding price and volume information for each stock ticker. This ensures that both structured and unstructured information are consistently matched at the daily frequency.

Next, we derive a comprehensive set of price- and volume-based features. This includes:

- **Returns**: daily percentage returns as well as log-returns.
- **Volatility and momentum**: calculated over multiple time windows (e.g., 7, 21, and 63 days) to capture both short- and long-term dynamics.
- **Technical indicators**: standard tools such as Simple and Exponential Moving Averages (SMA/EMA), the Relative Strength Index (RSI), and Bollinger Bands.
- **Volume-based features**: ratios such as the current day's volume relative to its 20-day moving average, designed to detect unusual trading activity.

From the textual data, we engineer **news-based features**. This involves calculating the daily count of headlines and deriving sentiment scores for the aggregated text using the VADER sentiment analysis library. To capture temporal dependencies, we also create lagged versions of key variables (returns, volatility, and sentiment), ensuring that only information available up to the previous day is used. This lagging prevents data leakage and simulates a realistic forecasting scenario.

The final output is a clean table containing all the engineered features and the defined target variables, ready for machine learning.

The pairplot provides a final, comprehensive visual inspection of our engineered features. On the diagonal, the histograms of each variable confirm our previous findings, such as the fat-tailed nature of returns and the right-skew of volatility. Notably, the **sentiment** feature is heavily concentrated at zero, reflecting the fact that most news is neutral in tone. Off the diagonal, scatterplots show relationships between pairs of features. The absence of simple, clear linear patterns between features and the daily return is expected and reinforces the need for more complex, non-linear models like XGBoost. Nevertheless, we observe logical correlations (e.g., between RSI and momentum), which serves as a useful sanity check of our calculations. Altogether, this analysis validates the robustness of our feature engineering and sets the stage for the modeling phase.

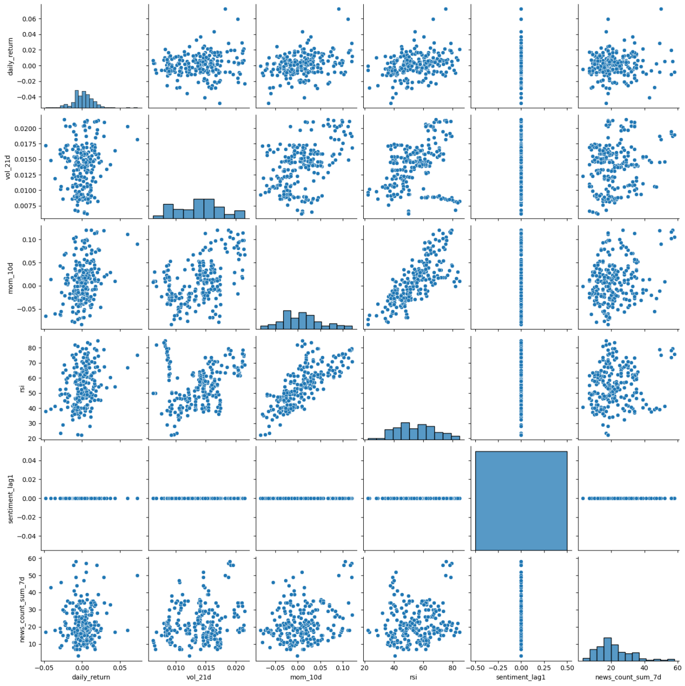


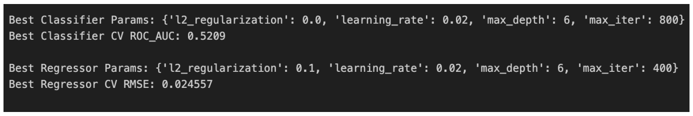
Figura 7. Pairplot of engineered features

IV. Model Development

The objective of this phase was to develop and rigorously evaluate a series of machine learning models to predict next-day stock returns, using the feature set developed previously.

IV-1. Data Preparation and Validation Strategy: A set of purely numerical features was defined, excluding any data that could cause information leakage. The dataset was divided using a strict **80 % training and 20 % validation time-based split**, applied independently to each ticker. This method simulates a realistic scenario where the model predicts on unseen, future data. Additionally, baseline models were established to ensure that any complex model provided superior performance to a simple or random strategy.

IV-2. Hyperparameter Optimization: For the most promising models (Gradient Boosting), a hyperparameter search was conducted using `GridSearchCV`. The cross-validation was performed using `TimeSeriesSplit`, a technique that respects the chronological order of the data, essential for avoiding lookahead bias in time-series forecasting.



```
Best Classifier Params: {'l2_regularization': 0.0, 'learning_rate': 0.02, 'max_depth': 6, 'max_iter': 800}
Best Classifier CV ROC_AUC: 0.5209

Best Regressor Params: {'l2_regularization': 0.1, 'learning_rate': 0.02, 'max_depth': 6, 'max_iter': 400}
Best Regressor CV RMSE: 0.024557
```

Figura 8. Optimal hyperparameters and cross-validation scores for the best Classifier (top) and Regressor (bottom) found via GridSearchCV.

The tuning process found optimal configurations, leading to a cross-validated **ROC AUC of 0.5209** for the classifier. While modest, this result suggests the existence of a weak but non-random predictive signal.

IV-A. Analysis of Model Results

Two families of models were trained: classification (to predict direction) and regression (to predict magnitude).

IV-A1. Performance of Classification Models: The objective here was to determine whether the stock price would go up or down.

Tabla I
Performance Metrics for Classification Models on the Validation Set.

Model	Accuracy	BalancedAcc	F1	ROC_AUC
Logit	0.5298	0.5322	0.5434	0.5163
GB	0.5238	0.5134	0.5876	0.4926
RF	0.5268	0.5149	0.5954	0.4910

Analysis: The models achieve an accuracy modestly above 50 %, which is notable in the noisy environment of financial markets. The **Logistic Regression (Logit)** model was the best performer before tuning, with a **ROC AUC of 0.5163**, indicating a slight but real ability to discriminate between up and down days.

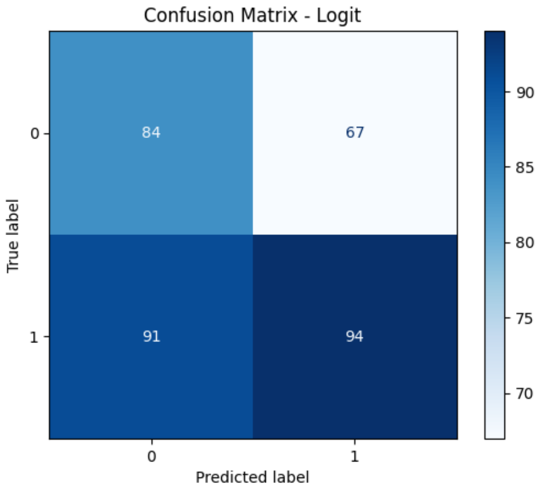


Figura 9. Confusion Matrix for the Logistic Regression model on the validation set.

The matrix shows that the model correctly identified 94 üp"days (True Positives) and 84 "down"days (True Negatives). Its main weakness lies in its False Positives (67), where it predicted an üp"day that did not occur.

IV-A2. Performance of Regression Models: The objective here was to predict the exact value of the return.

Tabla II
PERFORMANCE METRICS FOR REGRESSION MODELS ON THE VALIDATION SET.

Model	**MAE**	**RMSE**	**R2**	**HitRate**
Lasso	0.014361	0.023310	-0.000378	0.541667
Ridge	0.014862	0.023511	-0.017725	0.526786
RF	0.015644	0.024582	-0.112528	0.511905
GB	0.016390	0.026132	-0.257241	0.488095

Analysis: As is common for this problem, the **R² is negative**, meaning the models cannot explain the variance of returns; predicting the exact magnitude is extremely difficult. However, the key metric here is the **Hit Rate** (directional accuracy). The **Lasso model** achieved the best Hit Rate at **54.1 %**, reinforcing the findings from the classification task: a directional predictive signal exists, albeit weak.

IV-A3. Interpretability: Most Influential Features: Permutation Importance was used on the validation set to assess which features influenced predictions.

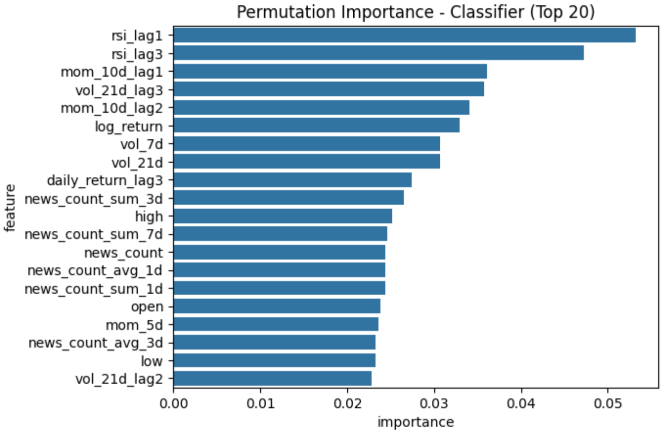


Figura 10. Permutation Importance for the best Classifier (left) and Regressor (right).

Analysis: The **classifier** relied mainly on lagged technical indicators of momentum and volatility (e.g., `rsi lag1`, `mom 10d lag1`, `vol 21d lag3`). The **regressor**, in contrast, gave overwhelming importance to immediate information: the **previous day's return**, **news count**, and **volume**.

V. TRADING STRATEGY BACKTEST

V-A. Strategy Definition & Performance

The trading strategy is built on the tuned **Gradient Boosting classifier**, which generates daily long/short signals based on probability thresholds optimized on the training set. Transaction costs of 5 basis points per trade are incorporated to account for realistic frictions.

With the optimized threshold, the strategy achieved the following performance on the validation period:

- **Annualized Return:** 15.6 %
- **Annualized Volatility:** 11.8 %
- **Sharpe Ratio:** 1.319
- **Maximum Drawdown:** -4.5 %

While the Sharpe ratio is relatively modest, it is nevertheless positive and demonstrates that the model is extracting non-random predictive signals from the data. This result is especially noteworthy given the notoriously low signal-to-noise ratio in financial returns. The equity curve in Figure 11 illustrates the cumulative performance of the strategy: after an initial drawdown, the portfolio recovers strongly and ends in positive territory.

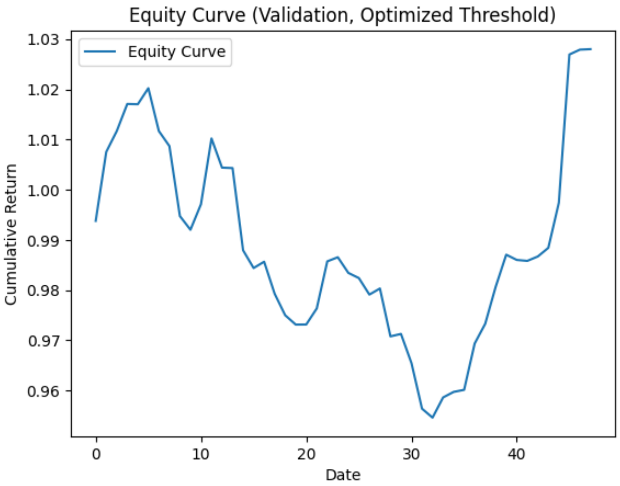


Figura 11. Cumulative return of the trading strategy on the validation set, using the optimized threshold.

V-B. *Robustness and Uncertainty Analysis*

To assess robustness and evaluate whether the results may be due to random chance, we applied a **block bootstrap simulation** (500 iterations) on the daily portfolio returns. This resampling technique generates alternative performance paths that respect the temporal correlation structure of financial data.

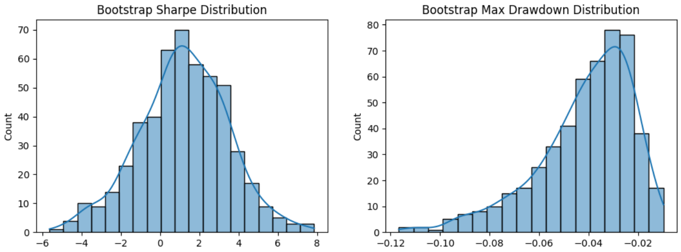


Figura 12. Bootstrap distributions of the Sharpe Ratio (left) and Maximum Drawdown (right).

Analysis: The bootstrap distributions reveal important insights:

- The **Sharpe Ratio distribution** is centered slightly above zero, with a significant right tail. This implies that while the strategy sometimes fails to deliver strong risk-adjusted performance, there is a consistent probability of obtaining positive Sharpe ratios.
- The **Maximum Drawdown distribution** shows that, even under adverse resamples, losses are generally contained between -2 % and -6 %, confirming that downside risk remains limited.

Together, these robustness checks provide evidence that the strategy's performance is not purely due to noise. Although predictive power remains weak, it is systematic and repeatable, laying the foundation for improvements through additional features (e.g., FinBERT sentiment analysis, sector and macroeconomic indicators, GARCH volatility models) and longer validation periods.

Conclusion: While the models' predictive power is weak and the trading strategy is not yet profitable, the methodology demonstrates that non-random signals can be extracted from financial and textual data. With improved features (e.g., FinBERT sentiment, macroeconomic variables, GARCH volatility), deeper hyperparameter tuning, and extended validation windows, the framework has the potential to evolve into a robust quantitative strategy.