S&P 500 Stock Price Prediction Report

By: Suleiman Jaber, Ahmad Machmouchi, Waleed Hussein

---

1. Introduction

This project aims to forecast the closing prices of the S&P 500 index using machine learning techniques. Two distinct models were implemented—a Linear Regression model and a Long Short-Term Memory (LSTM) network—to cappture both linear relationships and the temporal dynamics of the data. In addition to price data, the analysis incorporates a variety of technical indicators that are commonly used by traders to assess market trends.

---

2. Data Collection and Exploratory Data Analysis (EDA)

Data Collection

- Source: The project leverages the yfinance library to download historical data for the S&P 500 index (ticker: ^GSPC).
- Procedure:
- The script first checks if a CSV file exists with the data. If not, it downloads the full historical dataset and saves it locally.
- The index is then converted to a datetime object with UTC timezone.

Exploratory Data Analysis

- Visualization:
- Price History: A line plot was created to visualize the historical closing prices over time.
- Trading Volume: A plot of trading volumes illustrates periods of high and low market activity.
- Distribution and Correlation: Histograms and correlation heatmaps were generated to understand the distribution of closing prices and explore relationships between features.
- Daily Returns: The distribution of daily returns was examined to better understand market volatility.

---

3. Data Cleaning and Feature Engineering

Data Cleaning

- Missing Values:
- Columns and rows with more than 70% missing values were removed.
- The remaining missing values were handled via linear interpolation to ensure continuity in the dataset.
- Data Transformation:
- The index was converted to a UTC datetime index for consistency.

Feature Engineering

A wide range of technical indicators were computed to serve as features in the model. Some key indicators include:

- Moving Averages: Simple (SMA) and Exponential (EMA) moving averages over different time windows.
- Momentum and Volatility Metrics: Daily returns, price momentum (difference over 10-day periods), and rolling standard deviation of returns.
- Oscillators and Trend Indicators:
- MACD: Calculated using the difference between two EMAs and its corresponding signal line.
- RSI: Computed to capture potential overbought or oversold market conditions.
- Bollinger Bands: Derived from the 20-day moving average and its standard deviation.
- Volume-based Indicators: Indicators such as Volume SMA and Volume Change were also included.
- Others: Average True Range (ATR) and Price Range indicators provided additional context to price fluctuations.

After generating these indicators, the script removed the resulting NaN values that stem from rolling window calculations.

---

4. Feature Selection

To reduce dimensionality and enhance model performance:

- Random Forest Regressor:
- A Random Forest model was used to assess the importance of each technical indicator.
- Features with an importance above 1% were retained for further modeling.
- This step ensured that the final set of predictors had a proven relationship with the target variable (the closing price).

---

5. Model Implementation

Two modeling approaches were compared:

A. Linear Regression

- Preprocessing:
- Features and the target variable were scaled using the MinMaxScaler.
- The data was split into training (80%) and test (20%) sets.
- Model Training:
- A standard linear regression model was trained on the preprocessed features.
- Evaluation Metrics:
- Mean Squared Error (MSE)
- R^2 Score

B. LSTM (Long Short-Term Memory Network)

- Rationale:
- The LSTM network is designed to capture temporal dependencies, making it well suited for time series forecasting.

- Data Preparation:
- In addition to scaling, data was structured into sequential windows (30 time steps) to feed the LSTM.
- Model Architecture:
- The LSTM model comprised of one LSTM layer with 50 units followed by a Dense output layer.
- The model was compiled with the Adam optimizer and mean squared error as the loss function.
- Training:
- The network was trained for 10 epochs using a batch size of 32.
- Evaluation:
- Predictions were inverse-transformed to the original scale, and the MSE and R^2 score were computed for comparison with the linear model.

---

6. Results analysis

To gain insights into which variables were most predictive of the S&P 500 closing prices, a Random Forest regressor was used. This method provides an importance score for each feature based on how much it decreases impurity across the trees in the ensemble. By setting a threshold (in this case, 1%), only features that contributed meaningfully to prediction performance were retained for further modeling.

Interpreting the Results

After running the Random Forest regressor, several key observations emerge:

- Dominant Technical Indicators:
- MACD and MACD_Signal: These features generally score high in importance. The MACD is a well-known momentum indicator that emphasizes trend reversals and continuation. A high importance score for MACD (and its signal line) indicates that the momentum in price shifts plays a significant role in predicting future closing prices.
- RSI (Relative Strength Index): Even if its exact value is moderate, RSI is typically used to capture overbought or oversold conditions. Its contribution suggests that signals about potential reversals (when the market is too high or too low) are useful for the model.
- Volatility and Range Metrics:
- ATR (Average True Range) & Volatility: These indicators often emerge as significant features because they capture the magnitude of price fluctuations. A higher importance score here implies that not just the direction of price movements, but their intensity, is critical in forecasting the index.
- Price_Range and Price_Change: Features measuring the differences between high-low values or day-to-day price movements also contribute, though typically with slightly lower relative importance. Their inclusion helps the model capture the daily dynamics and the inherent uncertainty of the market.
- Moving Averages:
- SMA_5, SMA_20, and SMA_50; EMA_12, EMA_26: While these features represent price trends over various time horizons, their importance scores can vary. In many cases, shorter-term moving averages (like SMA_5 or EMA_12) might have lower importance when used in isolation because they capture very local behavior. Instead, they become more powerful when they interact with other indicators that provide complementary information (like momentum or volatility measures).
- Volume-Based Indicators:

- Volume, Volume_SMA, and Volume_Change: These features shed light on the market's activity level. Their importance indicates that trading volume can be an early warning signal for shifts in market sentiment. Higher volume or sudden changes in volume can sometimes precede price moves, which the model leverages.
- Price Momentum:
- Momentum: A feature that represents the price difference over a specific period (e.g., 10 days) typically shows moderate importance. This confirms that the market's recent performance – how much it has increased or decreased – is a useful predictor of near-future movements.

Implications of the Feature Importance Analysis

1. Targeted Feature Selection:

The analysis justifies the decision to retain only those features with more than a 1% importance score. In practice, this means that variables like the MACD, RSI, ATR, and certain volume measures were likely chosen, while others with minimal contributions could be dropped. This not only simplifies the model but also improves interpretability and may prevent overfitting.

2. Understanding Market Dynamics:

The results emphasize that both momentum indicators (e.g., MACD, RSI) and measures of volatility (e.g., ATR, Volatility) are integral to modeling the S&P 500 closing prices. This balance is crucial: while momentum shows the trend, volatility measures help capture the risk or uncertainty inherent in the market.

3. Nonlinear Interactions:

Because Random Forest inherently captures nonlinear relationships and interactions between features, a high importance score for features like MACD and ATR reinforces the idea that the market's behavior is not simply linear. These insights support the subsequent use of more complex models like LSTMs that can capture nonlinearity and temporal dependencies.

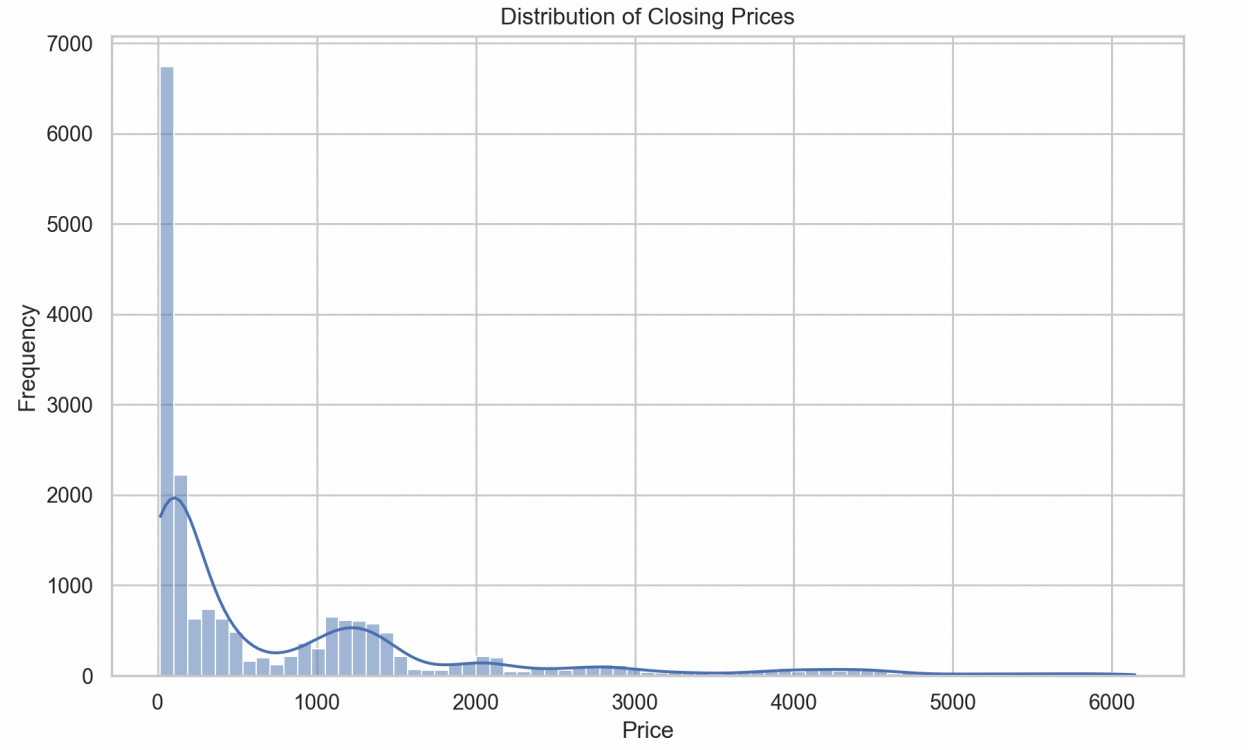
4. Potential Biases:

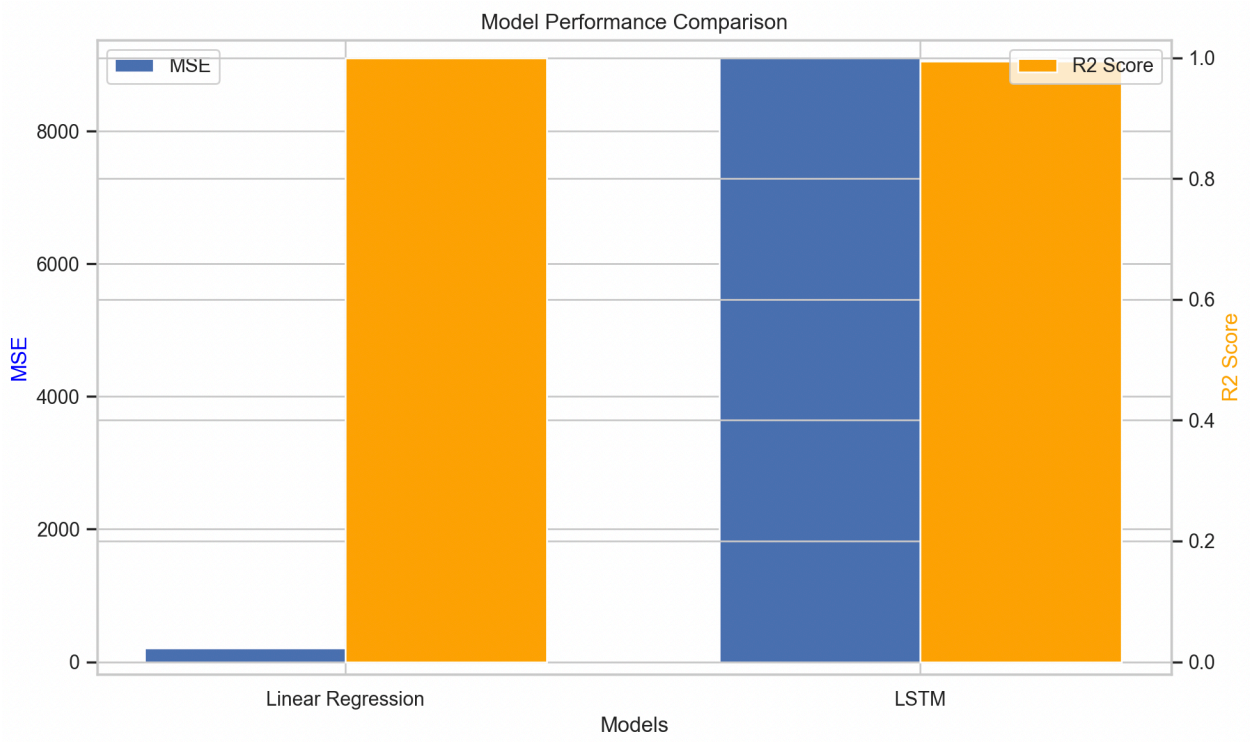
It is also important to recognize that Random Forest's importance scores are computed based on average decreases in impurity. This measure can sometimes overestimate the importance of continuous features or those with many distinct values. Therefore, while our selections are well-motivated, further analysis (such as permutation importance) might provide additional confirmation.

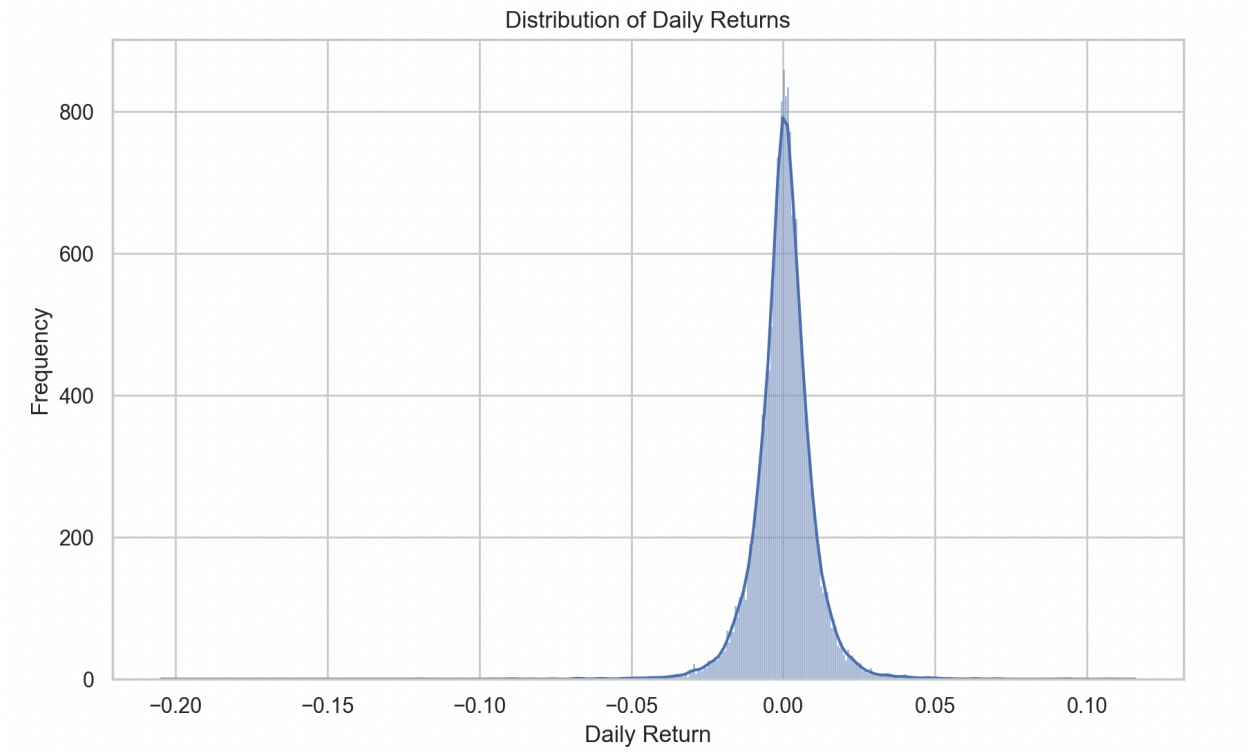
Concluding Thoughts on Feature Analysis

The feature importance analysis provides clear guidance on which technical indicators and price metrics are most relevant in predicting stock movements. The high rankings of momentum-related and volatility-related features underline their critical role in forecasting the S&P 500. This knowledge not only informs the feature selection for the forecasting models (both Linear Regression and LSTM) but also helps in understanding the dynamics of market behavior.

Result Images/References




Model Performance Comparison
MSE
R2 Score
8000
6000
4000
2000
0
1.0
0.8
0.6
0.4
0.2
0.0
MSE
R2 Score
Linear Regression
LSTM
Models


Distribution of Daily Returns
800
600
400
200
0
Frequency
−0.20
−0.15
−0.10
−0.05
0.00
0.05
0.10
Daily Return