CAS Applied Data Science, University of Bern

**Łukasz Maciaś**

lukasz.macias.01@gmail.com

# Final Project Report — CAS Applied Data Science

*"Forecasting S&P 500 Index Daily Movements with Machine Learning. A Comparative Study of Random Forest and XGBoost"*

*Zürich, May 2025*

## Abstract

This project focuses on forecasting the next-day price movement of the S&P 500 Index using two machine learning algorithms — Random Forest and XGBoost — applied to a diverse set of low-correlated features. The main objectives are: (1) to predict the magnitude of the S&P 500's daily price changes, and (2) to predict the direction of the movement (up or down). Both models were trained on a dataset containing daily closing prices along with a variety of engineered financial and macroeconomic features. Although XGBoost had a slightly higher Mean Absolute Error compared to Random Forest, it demonstrated better performance in capturing the variance of daily price change magnitudes. In terms of predicting the direction of price movements, XGBoost correctly predicted the direction on 54% of days, while Random Forest achieved 53%. These findings are significant because they show that both models consistently outperform random guessing — a notable achievement given that short-term stock market predictions are widely regarded as extremely difficult.

## Table of Contents

# 1. Introduction

## 1.1 Goal of the analysis

The objective of this project is to compare the outcomes of two machine learning algorithms - Random Forest and XGBoost - in forecasting the S&P 500 index's next-day price based on a comprehensive set of diverse and low correlated features. In particular, the study focuses on predicting:

- The magnitude of the daily price change
- The direction of the daily price movement (whether the price will increase or decrease)

## 1.2 What is S&P 500 index and its significance in the global economy

The S&P 500 Index, or Standard & Poor's 500, is a market-capitalization-weighted benchmark comprising 500 leading U.S. companies that collectively represent the large-cap segment of the U.S. equity market. Constituents are selected based on size, liquidity, domicile, and sector balance, ensuring broad coverage of the U.S. economy rather than simply the largest firms by market cap.

As one of the most widely followed equity indices, the S&P 500 is one of the primary gauges of U.S. stock-market performance and investor sentiment. Its float-weighted structure and depth make it important for portfolio benchmarking, index-fund and ETF construction, derivatives markets, and macroeconomic analysis, driving investment decisions and economic insights around the world.[1]

## 1.3 Factors influencing S&P 500 valuation

The stock market offers a reliable and regulated environment for participants to efficiently trade financial instruments and allocate capital. However, markets are complex systems, heavily influenced by political, economic, and psychological factors. To provide the context for this analysis the list of potential factors influencing S&P 500 index valuation has been listed below.

[1] Kenton, W. (2024, June 12). *S&P 500 Index: What it is and why it's important in investing*. Investopedia. https://www.investopedia.com/terms/s/sp500.asp

Key factors influence the valuation of the S&P 500[2]:

- **Earnings Growth:** The profits generated by companies that are part of the index are a fundamental driver of the S&P 500 valuation. Strong earnings growth supports higher valuations, while weak earnings may indicate that the index is overvalued.
- **Interest Rates and Monetary Policy:** Low interest rates benefit stocks by making them more attractive relative to bonds. In contrast, high interest rates tend to pressure stock prices downward, as fixed-income investments become more profitable.
- **Inflation:** Moderate, stable inflation typically supports higher stock valuations. When inflation rises sharply, it usually erodes corporate profits and dampens consumer spending, negatively impacting stock prices.
- **Market Sentiment:** Investor perception of the stock market plays a significant role. Optimism and confidence can drive stock prices higher, often beyond reasonable levels.
- **Oil prices** Changes in oil prices have a significant and non-linear effect on the S&P 500. Oil supply shocks are usually harming stock prices by increasing production costs. When oil price increases it is reflected in stronger economic growth and can sometimes positively impact the S&P 500, depending on broader market conditions.[3]
- **"Black Swan" events in stock markets** are unpredictable events outside of normal market expectations and can have potentially severe consequences, such events are for example: COVID-19 epidemic, current Trump's tariff war, the "dot.com" bubble burst, the crisis of 2009 etc.[4]

The aim of this analysis is to incorporate the broadest possible set of factors as features in the machine learning algorithm, in order to capture the complex relationships between these factors and the S&P 500 valuation, ultimately improving the prediction of future price movements.

# 2 Data

## 2.1 Data Sources

The features used in this project can be categorized into three categories: (1) Stock-Specific Features, (2) Technical Analysis Indicators, and (3) Macroeconomic Indicators. Data for categories (1) and (2) was sourced from Yahoo Finance using the *yfinance* python library. All technical analysis indicators were computed based on the stock attributes retrieved via *yfinance*. For category (3), Macroeconomic Indicators, the data was partially sourced from Yahoo Finance,

[2] Caparrini, A., Arroyo, J., & Escayola Mansilla, J. (2024). S&P 500 stock selection using machine learning classifiers: A look into the changing role of factors. *Research in International Business and Finance* https://doi.org/10.1016/j.ribaf.2024.102336

[3] Aloui, M., Hammoudeh, R., & Nguyen, D. K. (2019). The dynamic impact of crude oil price changes on the S&P 500: Evidence from the U.S. economy. *International Review of Economics & Finance* https://doi.org/10.1016/j.iref.2019.05.005

[4] Julius Baer. (n.d.). *Is the S&P 500 overvalued?* Julius Baer. Retrieved 05.05.2025, from https://www.juliusbaer.com/en/insights/market-insights/markets-explained/is-the-sp-500-overvalued/

while features such as the Consumer Price Index, Consumer Sentiment Index, Employment Level, and Interest Rates were obtained through the US Federal Reserve's FRED API (library *fredapi*).

## 2.2 Data Collection

The dataset utilized in this project covers a period of over nine years, from 1 January 2016, to 27 February 2025. The end date was selected due to a major "black swan" event - the introduction of trade tariffs by the Trump administration - which caused significant disruption in the financial markets and represented a type of event that no prediction algorithm could reasonably be expected to anticipate.


Figure 1: Data Flow (own work)

The process begins with retrieving data from Yahoo Finance for the GSPC ticker (S&P 500 Index) over the relevant period. Next, macroeconomic data from the U.S. Federal Reserve API is integrated into the dataset. Finally, additional features, such as technical analysis indicators, are calculated, followed by various preprocessing steps to prepare the data for analysis.

## 2.3 Data Quality

Overall, the quality of the data retrieved from both Yahoo Finance and the FRED API was satisfactory. However, there were two instances where the original data required augmentation:

1. Missing Values: Occasionally, data for a single day was missing. In such cases, the Pandas *ffill()* function was used to forward-fill the missing value with the most recent known value.

2. Monthly Data from FRED API: Certain attributes, such as the Consumer Price Index and Consumer Sentiment Index, were reported only once a month (unusually on the first day). For the remaining days of the month, these values were forward-filled using the Pandas *ffill()* function as well.

## 2.4 Feature Engineering

**Stock-Specific Features**

The stock-specific features used in this analysis provide key insights into stock price movements. The Daily OHLC Prices (Open, High, Low, Close) are fundamental for technical analysis, as they reflect the market's range and direction each day. The Daily Price Movement

captures the change in the closing price from one day to the next, helping to identify trends, momentum, and return behavior. Daily Volatility is calculated by measuring the difference between daily highest and lowest price. Trading Volume tracks the number of shares traded each day, serving as an indicator of investor participation, with significant volume spikes often signaling potential market turning points or breakouts. Finally, Dividend Payments represent the cash paid to shareholders, which can impact the total return on an investment.

| Category | Feature | Typical Parameters | Intuition / Why It Helps |
|---|---|---|---|
| **Price Action** | **(1) Daily OHLC Prices** | Open, High, Low, Close (Yahoo Finance) | Basic building blocks for all technical analysis; show market range and direction. |
| | **(2) Daily Price Movement** | Close_t – Close_(t-1), % change | Captures daily trend, momentum, and return behavior. |
| **Volatility** | **(3) Daily Volatility** | Difference between daily high and low price | Tracks daily extremes of price changes |
| **Volume / Flow** | **(4) Trading Volume** | Daily volume (Yahoo Finance) | Tracks investor participation; spikes often signal turning points or breakouts. |
| **Returns to Investors** | **(5) Dividend Payments** | Cash dividend per share (Yahoo Finance) | Impacts total return; relevant for valuation |

Figure 2: Stock-specific features (own work)

**Technical Analysis Indicators**

Technical analysis examines patterns and trends in trading activity - primarily price and volume - to find out potential investment opportunities. Unlike fundamental analysis, which assesses a security's intrinsic value through financial metrics such as revenues and earnings, technical analysis relies only on market behavior to forecast future price movements.[5]

The technical analysis indicators used in this analysis focus on market trends, momentum, volume, and broader market structures to evaluate stock price movements. Moving Averages (5-, 10-, 20-, 50-, 200-day) capture the prevailing trend and provide a lagged view of momentum, helping to identify the direction of the market. The Relative Strength Index[6] (RSI 14-days) is an overbought/oversold oscillator that signals whether a stock is potentially overbought or oversold, aiding in the identification of reversal points. The Rate of Change[7] (ROC 5-days) measures the percentage change in price, acting as a momentum gauge that helps to identify fast price movements. On-Balance Volume[8] (OBV) combines volume and price direction, confirming whether price moves are supported by volume, giving insights into the strength of a trend. The VWAP[9] Deviation measures the price relative to the volume-weighted average price, helping traders understand the price action within the context of volume. The High-Low & Close Location Value[10] (CLV) shows where the close price sits within the day's range, indicating control between

[5] Hayes, A. (2024, July 4). *Technical analysis: What it is and how to use it in investing*. Investopedia. Retrieved 08.05.2025, from https://www.investopedia.com/terms/t/technicalanalysis.asp
[6] Investors Underground. (n.d.). *RSI (Relative Strength Index)*. Retrieved 08.05.2025, from https://www.investorsunderground.com/rsi-relative-strength-index/
[7] TradingView. (n.d.). *Rate of Change (ROC)*. TradingView. Retrieved 22.04.2025, from https://www.tradingview.com/support/solutions/43000502343-rate-of-change-roc/
[8] Hayes, A. (2024, August 30). *On-Balance Volume (OBV): Definition, formula, and uses as indicator*. Investopedia. Retrieved 08.05.2025, from https://www.investopedia.com/terms/o/onbalancevolume.asp
[9] Corporate Finance Institute. (n.d.). *Volume Weighted Adjusted Price (VWAP)*. Retrieved 09.05.2025, from https://www.corporatefinanceinstitute.com/resources/career-map/sell-side/capital-markets/volume-weighted-adjusted-price-vwap/
[10] Chen, J. (2023, June 30). *Accumulation/Distribution Line (A/D): Definition and Formula*. Investopedia. Retrieved 10.05.2025, from https://www.investopedia.com/articles/trading/08/accumulation-distribution-line.asp

buyers and sellers. The S&P 500 vs. Nasdaq-100 price ratio captures the relative strength of a stock against these major benchmarks, helping assess its performance relative to the broader market. Finally, Implied Volatility (VIX level) is a sentiment and risk-aversion proxy, influencing tech stocks' movements based on overall market sentiment.

| Category | Feature | Typical Parameters | Intuition / Why It Helps |
|---|---|---|---|
| **Trend & Momentum** | **(1) Moving Averages** | 5-, 10-, 20-, 50-, 200-day | Captures prevailing trend and lagged momentum. |
| | **(2) Relative Strength Index (RSI)** | 14-day (or 7/21) | Overbought/oversold oscillator (0-100). |
| | **(3) Rate of Change (ROC)** | 5- or 10-day | Percentage change in price; pure momentum gauge. |
| **Volume/Flow** | **(4) On-Balance Volume (OBV)** | cumulative | Combines volume and direction; confirms price moves. |
| **Market Structure** | **(5) VWAP Deviation** | intraday → daily close | Price relative to volume-weighted average price. |
| | **(6) High-Low & Close Location Value (CLV)** | daily | Where the close sits inside day's range; buyers vs. sellers control. |
| **Cross-Asset / Macro** | **(7) S&P500 vs. Nasdaq-100** | price ratio | Normalized Relative strength to benchmark. |
| | **(8) Implied Volatility (VIX level)** | daily | Sentiment & risk-aversion proxy that influences tech stocks. |

Figure 3: Technical Analysis Indicators (own work)

## Macroeconomic Indicators

The macroeconomic indicators used in this analysis provide insights into the broader economic environment, influencing S&P 500 valuations. Interest Rates (Fed Funds Rate) play an important role in determining borrowing costs, which affects the valuation of the stocks by discounting future earnings. The Consumer Price Index (CPI) measures inflation trends, which influence central bank policy and the real returns investors can expect. The Employment Level (Nonfarm Payrolls) reflects the health of the labor market, with strong employment indicating a strong economy and increased demand in the economy, while weak employment can lead to negative sentiments. The US Dollar Index (DXY) measures the strength of the dollar, which affects global revenue of the US companies, especially those with significant exports, such as Apple. Crude Oil Prices are a proxy for global demand and inflation pressures, indirectly influencing Federal Reserve policies. The Consumer Sentiment Index captures consumer confidence, a forward-looking indicator of potential tech adoption and spending behavior. The Money Supply (M1 / M2) tracks the available liquidity in the economy, which supports asset price growth, while Treasury Bond Yields serve as a benchmark for the risk-free rate, impacting discount rates, equity valuations, and investors' risk appetite.

| Category | Feature | FRED/Yahoo ID / Typical Data | Intuition / Why It Helps |
|---|---|---|---|
| **Monetary Policy** | **(1) Interest Rates** | Fed Funds Rate (FEDFUNDS) | Drives borrowing costs, discounting future earnings — critical for tech valuations. |
| **Inflation** | **(2) Consumer Price Index (CPI)** | CPIAUCSL | Captures inflation trends, affecting central bank policy and real returns. |
| **Labor Market** | **(3) Employment Level** | PAYEMS (Nonfarm Payrolls) | Strong labor = strong economy = demand for tech; weak = risk-off sentiment. |
| **Currency Strength** | **(4) US Dollar Index** | DXY (Yahoo Finance) | Affects global tech revenue (e.g., Apple); strong dollar = headwind for exports. |
| **Commodities** | **(5) Crude Oil Prices** | WTI Crude (CL=F on Yahoo) | Proxy for global demand & inflation pressure; influences Fed policy indirectly. |
| **Consumer Behavior** | **(6) Consumer Sentiment Index** | UMCSENT (Univ. of Michigan) | Forward-looking indicator of consumer confidence, tech adoption, and spending. |
| **Liquidity** | **(7) Money Supply (M1 / M2)** | M1SL, M2SL | Measures available liquidity in the economy; supports asset price growth. |
| **Fixed Income Market** | **(8) Treasury Bond Yields** | 10-Year Treasury Yield (TNX on Yahoo) | Benchmark for risk-free rate; impacts discount rates, equity valuations, and investor risk appetite. |

Figure 4: Macroeconomic Indicators (own work)

## 2.5 Data Preprocessing

In the data preprocessing phase for the Random Forest model, specific features such as 'open', 'high', 'low', 'volume', measures of volatility, technical indicators (e.g., moving averages), and macroeconomic variables (e.g., VIX, Treasury yields, crude oil prices) were normalized using percentage change (Pandas function *pct_change()*). This transformation was necessary because Random Forest Regressor algorithms partition data based on feature thresholds. Using absolute values can lead the model to overfit to specific price levels or market conditions, which may not remain consistent over time. By converting these variables into relative changes, the model focuses on the magnitude of movements rather than static levels, thereby improving its ability to generalize across different periods and market regimes.[11]

In contrast, for the XGBoost model, this preprocessing step was not required. XGBoost, being a gradient boosting algorithm, builds trees sequentially based on minimizing prediction errors and can inherently manage varying scales and distributions of input features. Consequently, it can effectively utilize raw feature values without introducing bias towards absolute levels, maintaining robust performance even without transforming features into percentage changes.[12]

# 3 Exploratory Data Analysis

Before building predictive models, it is essential to explore and understand the underlying structure, trends, and patterns in the data. This chapter provides an overview of the dataset used, key statistical properties, and initial insights that will guide the model development.

## 3.1 S&P 500 Index price development trend

The dataset used in this project covers a period of 9 years and 2 months, spanning from January 1, 2016, to February 27, 2025, with weekend days excluded. It is a single, daily-aggregated dataset where the date serves as the index of the DataFrame. In total, the dataset contains multiple columns, of which 21 attributes are later selected as features for the predictive models. The target variable is the daily percentage change in the S&P 500 closing price, specifically the difference between next day's closing price and current day closing price.

[11] Arora, S. (2019, January 3). *How data normalization affects your random forest algorithm*. Medium. Retrieved 09.05.2025, from https://medium.com/data-science/how-data-normalization-affects-your-random-forest-algorithm-fbc6753b4ddf

[12] Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.

The chart below illustrates the development of the S&P 500's index price, daily averages, and trading volume over the analysis period. Overall, a clear upward trend is visible, with occasional downturns such as the sharp drop in March 2020, driven by the COVID-19 crisis, which was quickly reversed, and the S&P 500 reached new all-time highs by autumn 2020. During the analysis period, only one "bear market" occurred in 2022, when the index declined by approximately 23% from its previous highs. (A "bear market" is typically defined as a decline of more than 20% from a recent peak.)

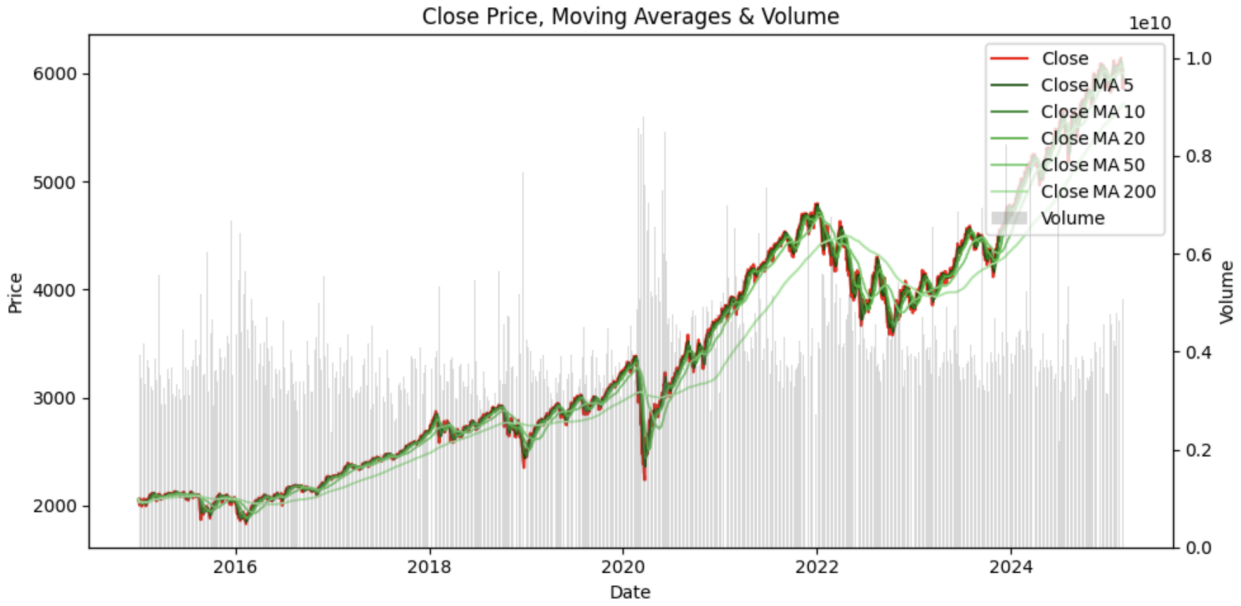



Figure 5: S&P 500 Index price (incl. moving avg.) and volume development (source: Yahoo Finance API)

## 3.2 S&P 500 Index returns

Historically, since its inception in 1926, the S&P 500 has delivered an average annual return of approximately 10%.[13] During the analyzed period, however, the index performed even better, achieving a total average annual return of 12.4%. Yearly returns varied significantly: some years saw exceptional gains, such as 2019 with a return of 29% and 2021 with 27%, while others experienced notable declines, most prominently in 2022 (a "bear market" year) when the S&P 500 fell by 19%.

[13] Investopedia. (2024, November 4). *S&P 500 average returns and historical performance*. Retrieved 12.05.2025, from https://www.investopedia.com/ask/answers/042415/what-average-annual-return-sp-500.aspInvestopedia

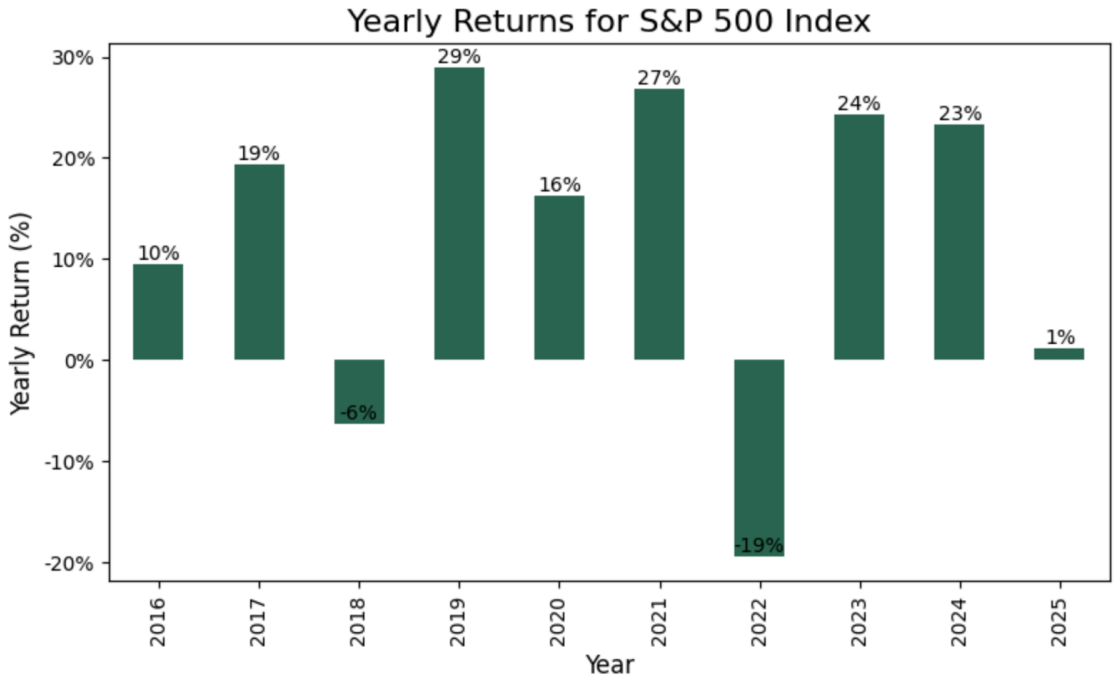



Figure 6: yearly S&P 500 Index returns (own work based on: Yahoo Finance API)

The goal of this project is to predict the daily percentage change in the S&P 500 closing price. Therefore let's examine the descriptive statistics of this attribute per day in the analyzed period. The dataset spans a total of 2,299 trading days. On average, the daily movement of the S&P 500 is very low and circles around 0.05%, with a standard deviation of 1.13%. The best-performing day saw a growth of 9.38%, while the worst experienced a decline of -11.98%. This highlights that, despite the modest average movement, the S&P 500 can exhibit significant volatility on individual days. On most days, the S&P 500 saw a positive price change the following day, occurring on 1,252 days (54%). In contrast, the price decreased on 1,046 days (46%) due to a negative change.

| | Daily Change (%) |
|---|---|
| count | 2299.0 |
| mean | 0.05% |
| std | 1.13% |
| min | -11.98% |
| 25% | -0.37% |
| 50% | 0.07% |
| 75% | 0.58% |
| max | 9.38% |

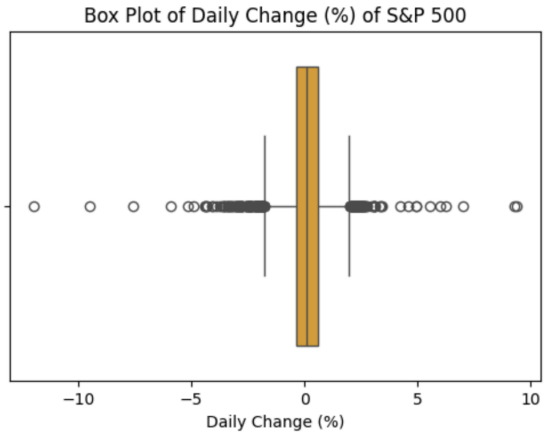
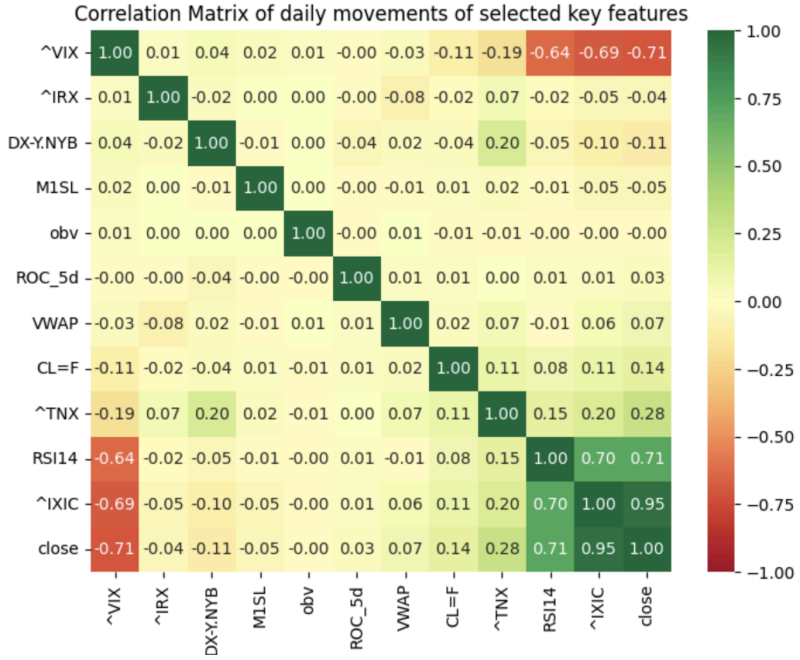



Figure 6: Daily S&P 500 Index price change (own work based on: Yahoo Finance API)

## 3.3 Correlation Analysis

The correlation matrix below presents the relationships between the daily relative changes of key features in the dataset. A strong positive correlation is observed between the S&P 500 closing price, the NASDAQ Composite Index (^IXIC), and the 14-day Relative Strength Index (RSI14). This is expected, as RSI14 is a technical indicator that tracks price momentum over a 14-day period, and the NASDAQ Index shares significant overlap with many stocks listed in the S&P 500. The strongest negative correlations are found between the Volatility Index (^VIX) and the S&P 500, NASDAQ (^IXIC), and RSI14. This occurs because the VIX measures expected market fear and uncertainty — it tends to rise when stock prices fall and decline when markets are stable and rising, leading to a strong inverse relationship. For the remaining features, correlations are relatively weak (ranging between -0.3 and +0.3). This is a positive sign, as it suggests that most features provide unique information, which can potentially help machine learning models make more accurate predictions by reducing redundancy.




| Ticker | Explanation |
|---|---|
| close | Closing price of S&P500 |
| ^VIX | Volatility Index, market fear gauge |
| ^IXIC | NASDAQ Composite Index, tech stocks |
| ^IRX | 13-week Treasury Bill yield |
| ^TNX | 10-year Treasury Yield |
| CL=F | Crude Oil futures price |
| DX-Y.NYB | U.S. Dollar Index (USD strength) |
| M1SL | U.S. Money Supply (M1) |
| ROC_5d | 5-day Rate of Change in price |
| RSI14 | 14-day Relative Strength Index (momentum) |
| VWAP | Volume Weighted Average Price |
| obv | On-Balance Volume (volume momentum) |

Figure 7: Correlation matrix of daily changes in key features (own work based on: Yahoo Finance API and FRED API)

# 4 S&P 500 daily price change prediction using Random Forest

## 4.1 Random Forest Algorithm

Random Forest is an ensemble machine learning algorithm that builds and combines multiple decision trees to create more accurate and stable predictions. In this project it is implemented to solve a regression task. The algorithm constructs a "forest" of decision trees, where each tree is trained on a different random subset of the original data. This approach introduces two main sources of randomness, which contribute to the robustness of Random Forest. The first source of randomness is bootstrap sampling, also known as bagging. In this process, each tree is trained on a random sample of the data selected with replacement. As a result, each tree is exposed to slightly different observations, which helps to reduce overfitting and increase diversity among the trees. The second source of randomness is random feature selection. When constructing each decision split within a tree, only a random subset of the available features is considered instead of evaluating all possible features. This technique further decorrelates the trees, making the combined predictions of the forest more reliable. In regression tasks, it outputs the average of the predictions from all individual trees.

Random Forest offers several important advantages. It significantly reduces overfitting compared to a single decision tree, making it better suited for complex datasets. It is capable of handling high-dimensional data. Moreover, it provides feature importance scores, which are useful for interpreting which variables contribute most to the model's predictions.

However, because Random Forest splits the data using simple threshold rules, it can be sensitive to the absolute magnitude of features. If the scale of features carries unintended meaning, it may impact the model's performance. This is particularly important in domains like financial data, where large numerical values can vary drastically. To address this, features are often transformed into relative terms, such as percentage changes, before training the model. This preprocessing step helps improve the model's stability and generalization by ensuring that the model focuses on changes and patterns rather than absolute magnitudes.[14]

## 4.2 Data Preprocessing, Hyperparameter Tuning, and Random Forest Model Training

The dataset utilized in this analysis comprises the daily closing prices of the S&P 500 Index, along with all features described in Section 2.4, *Feature Engineering*. The target variable was defined as the relative change between the current day's closing price and the following day's closing price, representing the value the model is intended to predict. Subsequently, as detailed in Section 2.5, *Data Preprocessing*, the features were transformed from absolute values

[14] Van der Plas, J. (2017). *Python Data Science Handbook*. O'Reilly Media.

into relative (percentage) changes to better align with the working characteristics of the Random Forest algorithm. Following preprocessing, the dataset was divided into training and testing subsets. The test set was allocated 249 days, approximately equivalent to one year of trading data. Finally, a Random Forest Regressor was trained on the data, utilizing 300 estimators to build the ensemble of decision trees.

## 4.3 Prediction Outcomes and Model Performance Evaluation

The Random Forest algorithm developed in this project is capable of predicting both the percentage change in the S&P 500 Index between today's and the next day's closing price, as well as the direction of the movement (i.e., whether the price will increase or decrease). Depending on the intended application, different aspects of the prediction (magnitude or direction) may be of particular interest.

The model's performance has been evaluated using standard regression metrics for the magnitude of difference between the prediction and by confusion matrix to evaluate the predicted direction of price change. The Mean Absolute Error (MAE), which measures the average magnitude of the errors between the predicted and actual values without considering their direction, is 0.64%. This metric reflects the typical size of the prediction errors in absolute terms. Additionally, the Root Mean Squared Error (RMSE), which also measures the average prediction error but places greater emphasis on larger errors by squaring the deviations before averaging, is 0.84%. The charts below presents the prediction results in absolute terms and as % of price change.

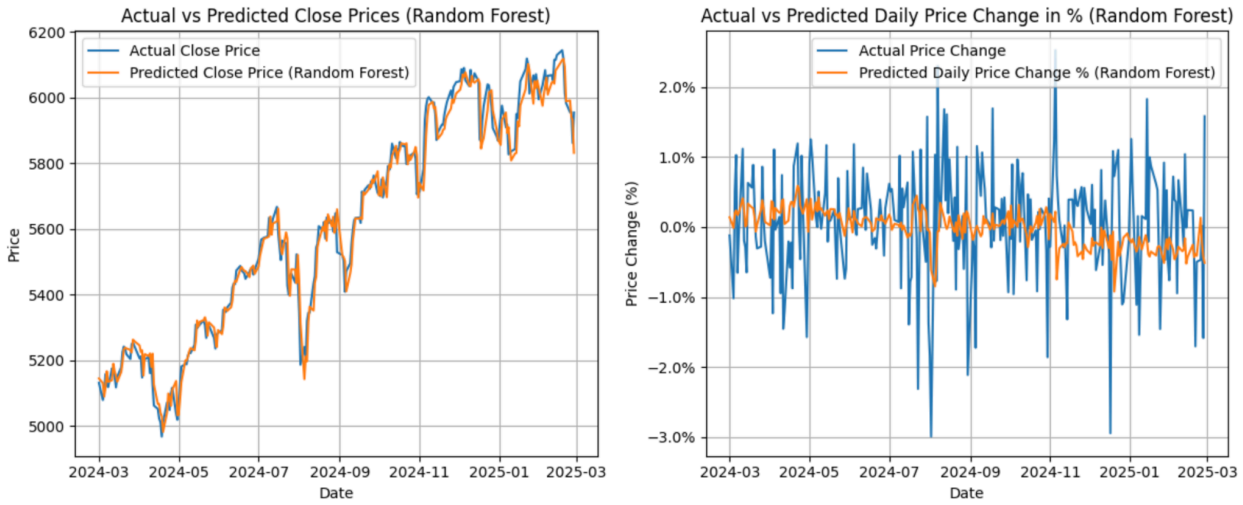



Figure 8: Prediction Outcomes vs Actual Values (own work)

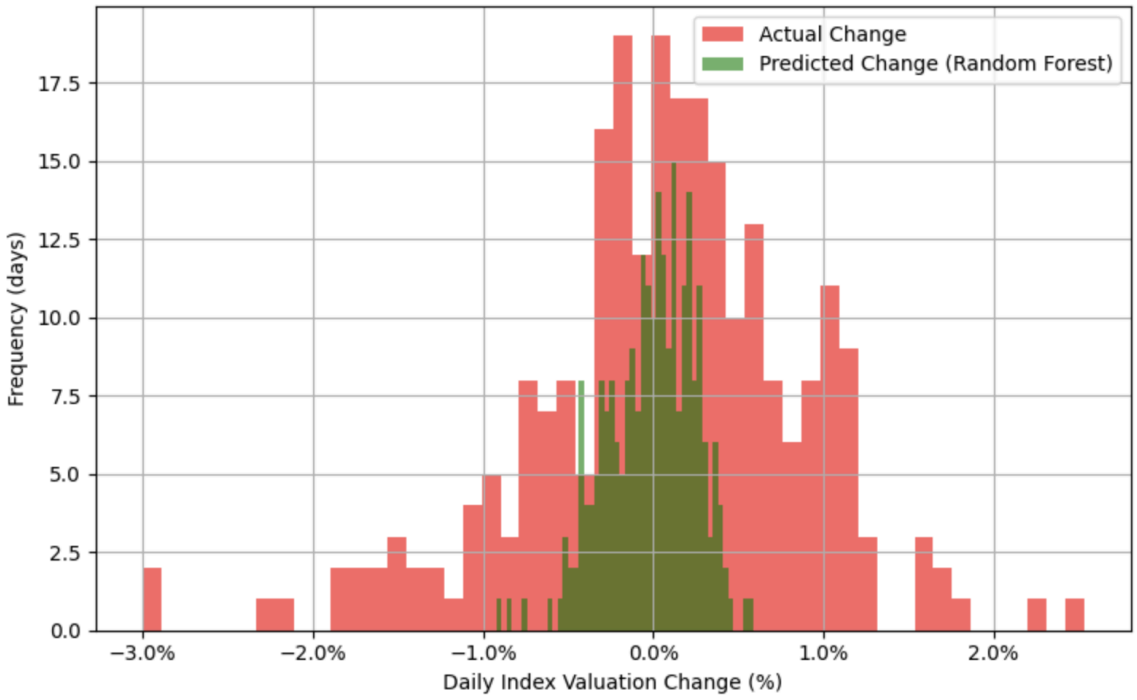



Figure 8: Distribution of Prediction Outcomes vs Actual Values (own work)

The values of both MAE and RMSE indicate that the model tends to be conservative in its predictions, often underestimating the magnitude of actual price movements. This observation is further illustrated by the histogram in Figure 8: the predicted daily price changes are tightly clustered around 0%, with a minimum prediction of -0.92% and a maximum of +0.59%, whereas the actual daily price changes ranged much more broadly from -3.00% to +2.53%.

| | Actual Daily Price Change | Predicted Daily Price Change |
|---|---|---|
| **count** | 249 | 249 |
| **mean** | 0.06% | -0.01% |
| **std** | 0.82% | 0.26% |
| **min** | -3.0% | -0.92% |
| **25%** | -0.3% | -0.18% |
| **50%** | 0.1% | 0.02% |
| **75%** | 0.55% | 0.18% |
| **max** | 2.53% | 0.59% |

Figure 9: Predicted vs Actual daily price change in % (own work)

In addition to evaluating the magnitude of the predicted price changes compared to the actual changes, it is valuable to assess the model's ability to predict the direction of price movements. The model achieved an overall accuracy score of 53%, correctly predicting the direction in 77 cases of price increases (true positives) and 54 cases of price decreases (true

negatives) out of a total of 249 observations. More specifically, the model correctly predicted an upward movement in 30.9% of the cases and a downward movement in 21.7% of the cases.

The precision score measuring the proportion of predicted upward movements that were correct was 58%, calculated as 77 true positives divided by the sum of 77 true positives and 55 false positives.

Finally, the recall score representing the proportion of actual upward movements that were correctly identified was 55%, calculated as 77 true positives divided by the sum of 77 true positives and 63 false negatives.

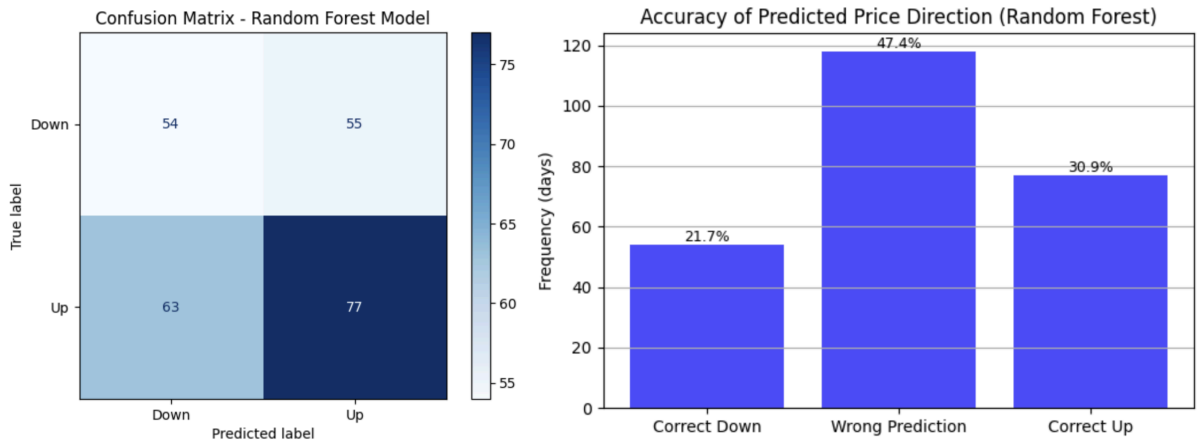



Figure 10: Confusion Matrix and Prediction Accuracy of the price change direction (own work)

## 4.4 Feature Importance

In a Random Forest model, feature importance measures how much each feature contributes to the overall predictive power of the model. The algorithm determines feature importance based on how much each feature improves the purity of the splits across all trees in the forest. Specifically, when a tree splits a node based on a feature, it results in a decrease in a chosen impurity measure which is variance (for regression). Each time a feature is used to split a node, the algorithm records the amount by which the impurity decreases. These decreases are then averaged across all trees and normalized so that the importances sums to 1. Therefore, features that are frequently used in splits and that lead to large reductions in impurity will be assigned higher importance scores. Features that are rarely used or lead to smaller improvements will have lower importance scores.[15]

[15] Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.

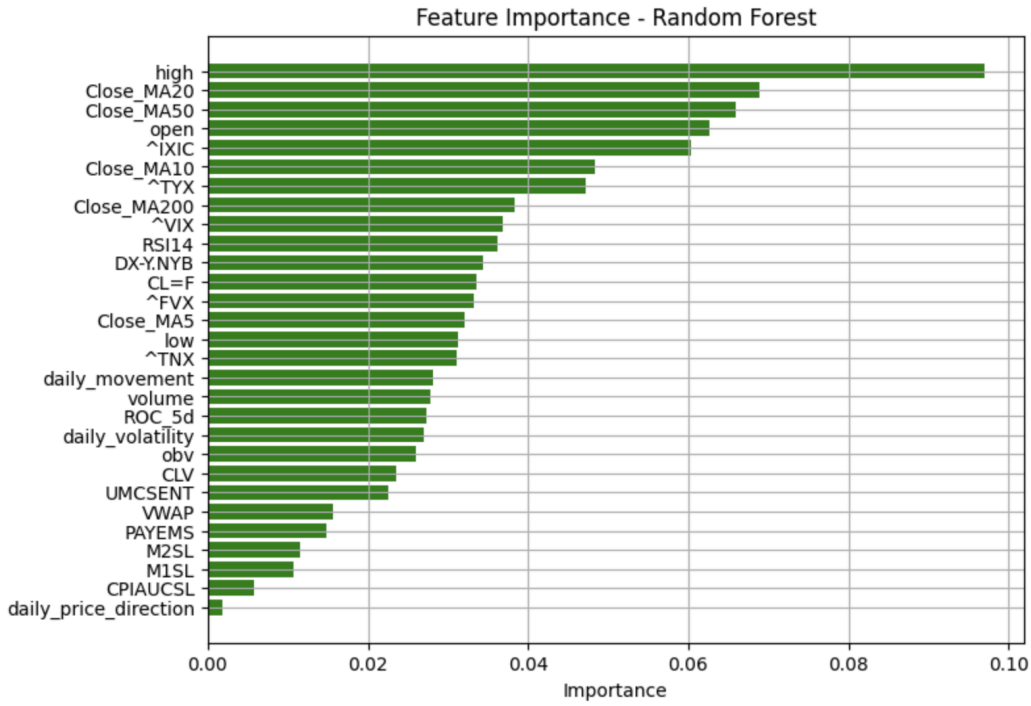



Figure 11: Random Forest Feature Importance Scores (own work)

The key takeaway from the feature importance scores for Random Forest is that stock-related features, including the daily highest price, daily open price, and various technical indicators such as moving averages, are the most influential factors. Additionally, the Nasdaq index (IXIC) and the treasury bond yield (TYX) also play a significant role, with a relatively high level of importance attributed to them as well.

# 5 S&P 500 daily price change prediction using XGBoost

## 5.1 XGBoost Algorithm

XGBoost, short for Extreme Gradient Boosting, is a powerful machine learning algorithm based on the gradient boosting framework. It builds decision trees sequentially, where each new tree is trained to correct the errors made by the previous trees. The process starts with an initial prediction, which could be as simple as the mean of the target variable for regression problems or a uniform probability for classification tasks. The first decision tree is trained to predict the residuals, meaning the differences between the actual target values and the initial predictions. Once the first tree is created, the model updates its predictions by adding the new tree's output to the previous predictions, improving its overall accuracy. Each subsequent tree is built to predict the new residuals, focusing on areas where the model still performs poorly.

XGBoost uses gradient descent to minimize a specific loss function, such as mean squared error for regression. This optimization process ensures that each tree added to the model helps reduce the overall error in the predictions. A key feature of XGBoost is its use of regularization, meaning it penalizes complex trees to avoid overfitting and improve generalization to new data. Regularization techniques make XGBoost particularly robust compared to traditional gradient boosting methods.[16]

## 5.2 Data Preprocessing, Hyperparameter Tuning, and XGBoost Model Training

The dataset used to train the XGBoost regressor was the same as that described in Section 4.2 ("Data Preprocessing, Hyperparameter Tuning, and XGBoost Model Training"), with one key exception: XGBoost does not require transforming features into relative values as outlined in Section 2.5. The data was split into training and test subs, with the test set fixed at 249 trading days, roughly one year of market data. An XGBoost Regressor was then trained on the remaining data. Thanks to the broad hyperparameters selection, XGBoost allows fine-grained control over model behavior. Through an iterative trial-and-error process, the final selected hyperparameter configuration yielded the best predictive performance.

| Parameter | Description |
|---|---|
| `n_estimators=300` | Number of boosting rounds, meaning how many trees the model will build sequentially. |
| `learning_rate=0.01` | Controls how much each tree corrects the previous errors; lower values slow learning but often improve performance. |
| `max_depth=8` | Maximum depth of each decision tree, controlling model complexity and how finely it can split data. |
| `subsample=0.8` | Fraction of the training data randomly sampled to grow each tree, helping prevent overfitting. |
| `colsample_bytree=0.8` | Fraction of features randomly selected for each tree, improving model robustness and diversity. |
| `random_state=1` | Fixes the random number generator seed for reproducibility of results. |
| `objective='reg:squarederror'` | Specifies the learning task as regression, minimizing the squared error between predicted and actual values. |

Figure 12: XGBoost hyperparameters selection

## 5.3 Prediction Outcomes and Model Performance Evaluation

The XGBoost algorithm developed in this project, similar to the previously described Random Forest model, predicts both the percentage change in the S&P 500 Index between today's and the next day's closing price, as well as the direction of the price movement (i.e., whether the price will increase or decrease).

The Mean Absolute Error (MAE), which measures the average magnitude of the errors between the predicted and actual values without considering their direction, is 0.72%. This metric reflects the typical size of the prediction errors in absolute terms. The Root Mean Squared Error

[16] NVIDIA Corporation. (n.d.). *What is XGBoost and why does it matter?* In *NVIDIA Glossary*. Retrieved 28.05.2025, from https://www.nvidia.com/en-us/glossary/xgboost/

(RMSE), which also measures the average prediction error but places greater emphasis on larger errors by squaring the deviations before averaging, is 0.95%. The charts below presents the prediction results in absolute terms and as % of price change.

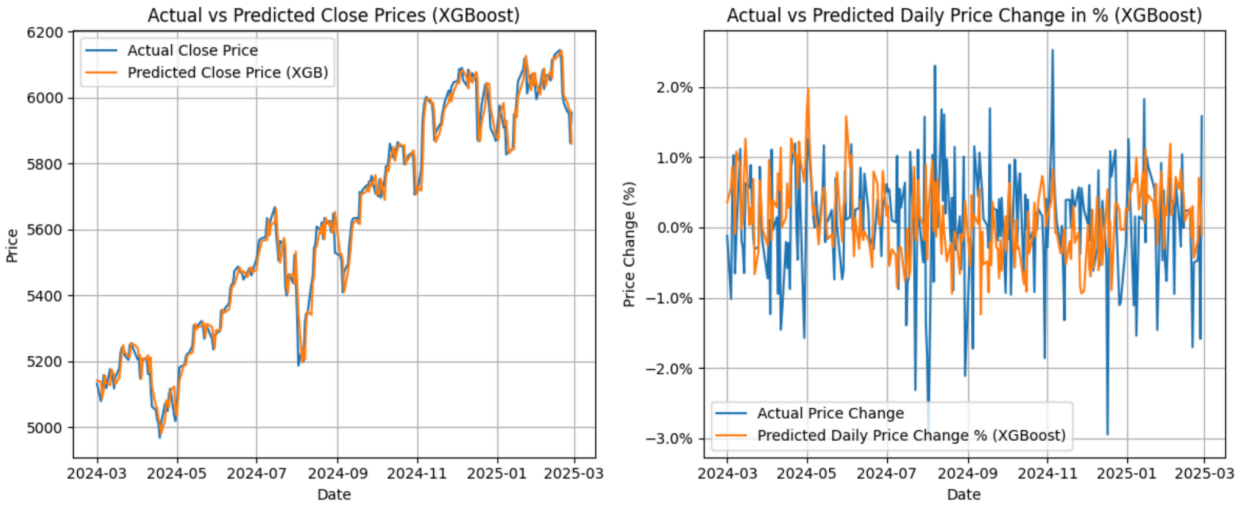



Figure 13: Prediction Outcomes vs Actual Values (own work)

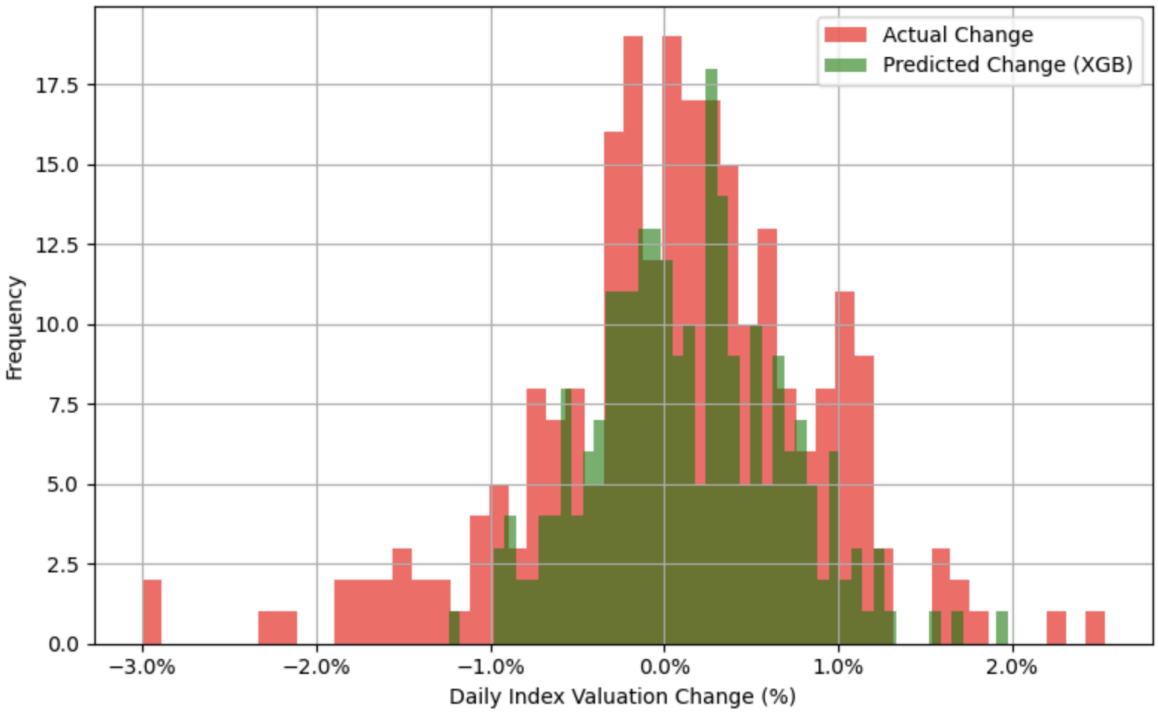



Figure 14: Distribution of Prediction Outcomes vs Actual Values (own work)

The standard deviation of the predicted daily changes is 0.54%, compared to 0.82% for the actual daily changes. While the predicted variability remains lower than the actual, it is a noticeable improvement over the Random Forest model (std = 0.26%).

| | Actual Daily Price Change | Predicted Daily Price Change |
|---|---|---|
| **count** | 249 | 249 |
| **mean** | 0.06% | 0.14% |
| **std** | 0.82% | 0.54% |
| **min** | -3.0% | -1.24% |
| **25%** | -0.31% | -0.24% |
| **50%** | 0.1% | 0.11% |
| **75%** | 0.55% | 0.5% |
| **max** | 2.53% | 1.97% |

Figure 15: Predicted vs Actual daily price change in % (own work)

The XGBoost model produced the following performance metrics for predicting the daily direction of price changes. It achieved an overall accuracy of 54%, correctly identifying 84 instances of price increases (true positives) and 51 instances of price decreases (true negatives) out of a total of 249 observations.

In more detail, the model accurately predicted upward movements in 33.7% of cases and downward movements in 20.5% of cases. The precision score indicating the proportion of predicted upward movements that were correct was 59%, calculated as 84 true positives divided by the sum of 84 true positives and 58 false positives.

The recall score measuring the proportion of actual upward movements that were correctly identified was 60%, calculated as 84 true positives divided by the sum of 84 true positives and 56 false negatives.

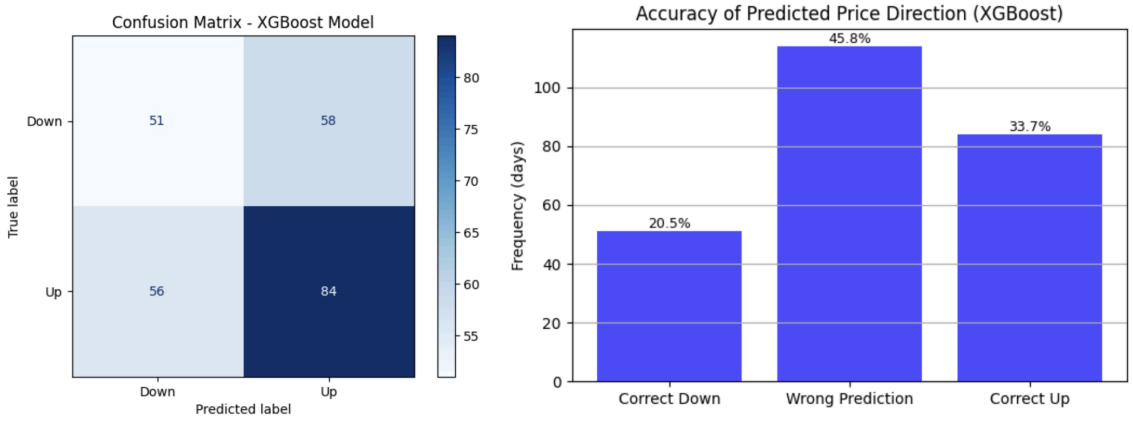



Figure 16: Confusion Matrix and Prediction Accuracy of the price change direction (own work)

## 5.4 Feature Importance

In XGBoost, gain measures how much a feature helps to reduce prediction error when it is used to split the data. Each time a feature is used for a split, the improvement in model performance is recorded. The more a feature improves the model, the more important it is considered. XGBoost adds up these improvements across all trees and normalizes them to show each feature's overall importance.[17]

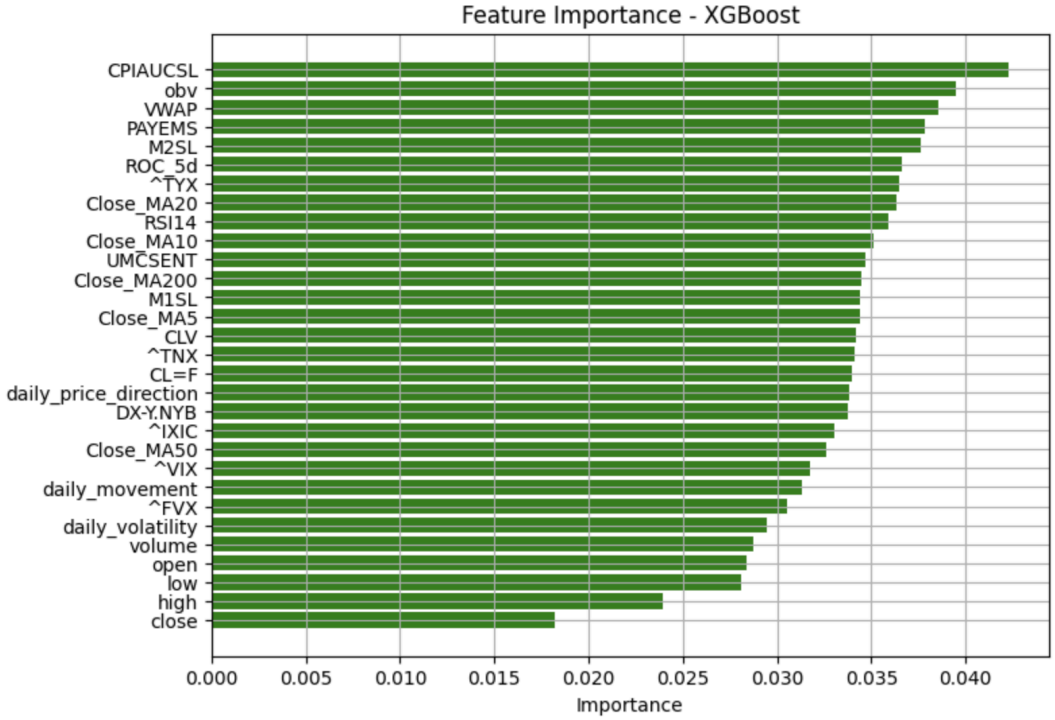



Figure 17: XGBoost Feature Importance Scores (own work)

Compared to Random Forest, the distribution of feature importance in XGBoost appears more balanced. The ranking of features is also noticeably different. In the case of XGBoost, stock-specific variables such as the daily high, low, close, and open prices show the lowest importance. Trend indicators like moving averages hold medium importance. The most influential factors are macroeconomic indicators, including the Consumer Price Index (CPIAUCSL), Employment Level (PAYEMS), and Money Supply (M2SL), along with technical indicators such as On-Balance Volume (OBV), Price Relative to Volume-Weighted Average Price (VWAP), the 5-day Rate of Change (ROC_5d), and the 14-day Relative Strength Index (RSI14). Treasury bond yields (TYX) also rank relatively high in importance.

[17] Lawrence, R. (2023, September 1). *How to use feature importance with XGBoost*. evolvingDev. Retrieved 29.05.2025, from https://www.evolvingdev.com/post/xgboost-model-feature-importance

# 6 Results Discussion - Comparative Analysis of Random Forest and XGBoost Results

Short-term stock price prediction is widely considered as an extremely difficult task. In this context, achieving results even slightly better than random chance should already be seen as a success. In this context both the Random Forest and XGBoost models delivered satisfactory prediction results. However, based on most of the performance metrics, XGBoost outperformed Random Forest in this particular use case.

| Model | Mean Absolute Error (MAE) | Root Mean Squared Error (RMSE) |
|---|---|---|
| Random Forest | 0.67% | 0.84% |
| XGBoost | 0.73% | 0.96% |

| Metric | Random Forest | XGBoost |
|---|---|---|
| **Precision Score** | 58% | 59% |
| **Recall Score** | 55% | 60% |
| **Accuracy Score** | 53% | 54% |

Figure 18: Comparison of performance metrics (own work)

While the MAE and RMSE are lower for the Random Forest model compared to XGBoost, this is largely due to Random Forest's failure to capture the true magnitude of the actual variance in daily price changes. This becomes evident when examining the standard deviation: for the actual daily price changes, the standard deviation is 0.82%, for XGBoost it is 0.54%, but for Random Forest it is only 0.26%. In essence, XGBoost provides a better representation of the true variability in stock price movements.

In practice, especially in day trading, one of the most critical aspects of a prediction model is its ability to correctly forecast the direction of price changes. The exact magnitude of the prediction matters less if the direction is incorrect — a wrong directional forecast can lead to immediate financial losses. In this regard, XGBoost performs slightly better than Random Forest. Overall, Random Forest predicted the wrong direction of price change on 47.5% of days, compared to 45.8% for XGBoost. It is worth noting that in both cases, the models performed better than random guessing (which would be expected to be wrong about 50% of the time). Although the difference between the two models may seem small, even a marginal improvement in directional accuracy can have a significant economic impact when applied in real trading environments.

The feature importance analysis shows interesting differences between the prediction bases of the Random Forest and XGBoost models. Random Forest relies heavily on stock-specific variables such as daily price levels (open, high, etc.) and technical indicators like moving averages, while also giving notable weight to broader market indicators like the Nasdaq index (IXIC) and treasury yields (TYX). In contrast, XGBoost presents a more balanced distribution of feature importance, with macroeconomic indicators including the Consumer Price Index (CPI), Employment Level (PAYEMS), and Money Supply (M2SL) acting as the most influential. Technical indicators such as OBV, VWAP, ROC_5d, and RSI14 also play a significant role. Overall, XGBoost appears to capture a broader range of economic drivers beyond stock-specific factors,

suggesting a more comprehensive understanding of market dynamics compared to Random Forest.

## 7 Conclusions

The objective of this project was to develop and compare two machine learning models, Random Forest and XGBoost, in forecasting the next-day price movement of the S&P 500 Index using a broad, diversified set of features. The analysis was based on over nine years of daily data incorporating stock-specific attributes, technical analysis indicators, and macroeconomic variables.

Key findings:

- Both Random Forest and XGBoost were able to capture patterns in the data and provide meaningful predictions about the magnitude and direction of daily S&P 500 price movements. XGBoost achieved slightly better accuracy and lower prediction error compared to Random Forest.
- Both prediction methods demonstrated predictive power by performing better than a random chance
- The analysis confirmed that the S&P 500's price movements are influenced by a complex combination of internal market dynamics and external macroeconomic factors because not a single feature dominated predictions, justifying the need for a bread and diversified features set. Furthermore, each method utilized a different hierarchy of feature importance to generate its predictions.

## 8 Outlook

Looking forward, the following directions could be pursued to enhance and expand this project:

- Adding additional non-traditional features such as news sentiment analysis, social media trends (e.g., X, Yahoo Finance comments section, Reddit), or options market data (e.g., put/call ratios) could enrich the feature set and further improve predictive power
- Future work could explore deep learning methods designed specifically for time series forecasting, such as Long Short-Term Memory networks (LSTM) which can capture time series related dependencies better than tree-based methods.
- Extending the analysis to forecast multi-day or weekly returns rather than focusing solely on the next day could reduce noise and improve forecast stability, better aligning with the investment horizons of many real-world investors.
- A logical next step would be the development of a live system to continuously retrain models and generate forecasts in real-time, combined with thorough backtesting to assess the viability of trading strategies based on model outputs.

## Statement

Ich erkläre hiermit, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls die Arbeit als nicht erfüllt bewertet wird und dass die Universitätsleitung bzw. der Senat zum Entzug des aufgrund dieser Arbeit verliehenen Abschlusses bzw. Titels berechtigt ist. Für die Zwecke der Begutachtung und der Überprüfung der Einhaltung der Selbstständigkeitserklärung bzw. der Reglemente betreffend Plagiate erteile ich der Universität Bern das Recht, die dazu erforderlichen Personendaten zu bearbeiten und Nutzungshandlungen vorzunehmen, insbesondere die schriftliche Arbeit zu vervielfältigen und dauerhaft in einer Datenbank zu speichern sowie diese zur Überprüfung von Arbeiten Dritter zu verwenden oder hierzu zur Verfügung zu stellen.

Date:

31.05.2025

Signature:

## GITHUB Repository

https://github.com/lukaszmacias01/CAS_UniBern_Applied_Data_Science/Final_Project

## References and Bibliography

Aloui, M., Hammoudeh, R., & Nguyen, D. K. (2019). The dynamic impact of crude oil price changes on the S&P 500: Evidence from the U.S. economy. International Review of Economics & Finance https://doi.org/10.1016/j.iref.2019.05.005

Arora, S. (2019, January 3). How data normalization affects your random forest algorithm. Medium. Retrieved 09.05.2025, from https://medium.com/data-science/how-data-normalization-affects-your-random-forest-algorithm-fbc6753b4ddf

Caparrini, A., Arroyo, J., & Escayola Mansilla, J. (2024). S&P 500 stock selection using machine learning classifiers: A look into the changing role of factors. Research in International Business and Finance https://doi.org/10.1016/j.ribaf.2024.102336

Chen, J. (2023, June 30). Accumulation/Distribution Line (A/D): Definition and Formula. Investopedia. Retrieved 10.05.2025, from https://www.investopedia.com/articles/trading/08/accumulation-distribution-line.asp

Corporate Finance Institute. (n.d.). Volume Weighted Adjusted Price (VWAP). Retrieved 09.05.2025, from https://www.corporatefinanceinstitute.com/resources/career-map/sell-side/capital-markets/volume-weighted-adjusted-price-vwap/

Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). O'Reilly Media.

Hayes, A. (2024, August 30). On-Balance Volume (OBV): Definition, formula, and uses as indicator. Investopedia. Retrieved 08.05.2025, from https://www.investopedia.com/terms/o/onbalancevolume.asp

Hayes, A. (2024, July 4). Technical analysis: What it is and how to use it in investing. Investopedia. Retrieved 08.05.2025, from https://www.investopedia.com/terms/t/technicalanalysis.asp

Investopedia. (2024, November 4). S&P 500 average returns and historical performance. Retrieved 12.05.2025, from https://www.investopedia.com/ask/answers/042415/what-average-annual-return-sp-500.aspInvestopedia

Investors Underground. (n.d.). RSI (Relative Strength Index). Retrieved 08.05.2025, from https://www.investorsunderground.com/rsi-relative-strength-index/

Julius Baer. (n.d.). Is the S&P 500 overvalued? Julius Baer. Retrieved 05.05.2025, from https://www.juliusbaer.com/en/insights/market-insights/markets-explained/is-the-sp-500-overvalued/

Kenton, W. (2024, June 12). S&P 500 Index: What it is and why it's important in investing. Investopedia. https://www.investopedia.com/terms/s/sp500.asp

Lawrence, R. (2023, September 1). How to use feature importance with XGBoost. evolvingDev. Retrieved 29.05.2025, from https://www.evolvingdev.com/post/xgboost-model-feature-importance

NVIDIA Corporation. (n.d.). What is XGBoost and why does it matter? In NVIDIA Glossary. Retrieved 28.05.2025, from https://www.nvidia.com/en-us/glossary/xgboost/

TradingView. (n.d.). Rate of Change (ROC). TradingView. Retrieved 22.04.2025, from https://www.tradingview.com/support/solutions/43000502343-rate-of-change-roc/

Van der Plas, J. (2017). Python Data Science Handbook. O'Reilly Media.