Final Written Deliverable

Predicting Next-Day S&P 500 Direction Using Price and Volume Data

Group A

Auriana Anderson, Chase Golden, Ross Schanck

Each author contributed equally to the design, coding & development, analysis, and writing of this project

December 17, 2025

Elevator Pitch

The assumption that financial markets are efficient suggests that it is very difficult to predict price changes in the very near term. Therefore, the purpose of this research is to test whether applying machine learning algorithms can uncover short-term historical patterns in S&P 500 price and volume data that will help predict which direction the market will trend the next trading day. We utilized daily price and volume data for all five years starting from 2010, created a variety of technical indicators, and tested several classification models: Logistic Regression, Random Forest, XGBoost and LightGBM. Despite multiple methods of preprocessing, feature engineering and hyperparameter tuning, none of the tested models produced consistent outperformance versus a naive baseline of always predicting upward movement. The results of this study indicate not only the significant difficulty of predicting price movement over one day but also that there is substantial evidence supporting market efficiency for daily trading and thus, that modern machine learning techniques cannot improve predictions or results.

Background & Question

According to the Efficient Market Hypothesis, all the available information about an asset is used to price it, meaning that investors cannot consistently achieve excess returns through timing or predicting the market. The debate over the validity of the EMH continues, but many people believe predicting short-term price movements is extremely difficult because of noise, volatility, and the rapid incorporation of information into prices. However, advances in machine learning have led to renewed interest in exploring more complex patterns that may exist in financial time series data.

This project aims to confirm whether both simple and advanced machine-learning techniques can produce meaningful forecast signals from historical S&P 500 price and volume data. Specifically, we will investigate whether machine-learning models can produce a more accurate forecast of the S&P 500's next-day direction than a random chance or a baseline strategy that assumes a market environment. We believe that, if these short-term patterns exist in momentum, volatility, and trading activity, we should be able to build models that will level the playing field and achieve predictability above the 50 per cent mark. We will ultimately conclude that the features we engineer will yield to models with an accuracy range of 55 to 60 per cent but recognize that the market is efficient and this is a limitation on possible returns.

Data

Data Acquisition

Using the yfinance API, daily S&P 500 indexes were retrieved through Yahoo Finance, for periods starting January 2010 to current day. The information was comprised of open, high, low, close, and volume throughout the trading period. This dataset was determined to be an excellent source for the research question because it can be acquired easily and is known to be very accurate. The response variable indicates on what direction the market will go tomorrow; if the price of the S&P 500 Index on the following day's market, it is labeled “Up.” Otherwise, if the next day's market is below the current day’s price, it is labelled “Down.” However, even though this dataset looks at all relevant market data, it does not provide economic, investor, or even intraday indicators which can potentially lead to significant and unknown amounts of variability on the data collected.

Data Cleaning

When we perform feature engineering steps; including lagged feature creation and rolling-window feature calculations, we also introduce missing values at the start of our time series. All missing observations at the start of our time series were removed to maintain temporal integrity. All other missing values were imputed using median imputation; this is a method that is less sensitive to outliers, which are typically present in financial datasets. We performed all our cleaning procedures to prevent look ahead bias and to allow for repeatability. Assumptions made when making these decisions include that the feature distributions are stationary and that the use of median imputation for skewed financial variables was appropriate.

Data Exploration

The exploratory analysis indicated that there is a small degree of class imbalance with a positive skew toward upward moving markets. This skew corresponds to the long-term growth trend of the equity market. The volatility and return distributions also exhibited heavy tails; therefore, we still need to perform robust preprocessing methods for our dataset. The correlation analysis indicated that the correlations between individual features and the target variable were not linear. This led us to the conclusion that we would need to use nonlinear modelling methods rather than linear ones. We have provided some plots and tables that provide a summary of the above-mentioned findings, located in the appendix.

Models

Preprocessing & Feature Engineering

We researched many different types of indicators for the development of features which would help us to predict short-term market trends, including lagged returns, rolling volatilities, simple moving averages, momentum-based indicators, OBV, Bollinger Bands, RSI, and MACD. These

indicators have been created through research and are widely known throughout the technical analysis community. We utilized standardization techniques with some indicators so that we could use models that are sensitive to the scale of the input data.

Unsupervised Methods

We did assess PCA as an unsupervised way of reducing feature dimensionality because we thought it might help solve the problem of multicollinearity across engineered features. Although PCA allowed us to reduce feature dimensionality, it did not help with downstream model performance or interpretability, thus PCA was ultimately excluded from the final modelling pipeline but the assessment of PCA guided some of our feature selection decisions.

Algorithm Selection

The four supervised learning models we used were Logistic Regression, Random Forest, XGBoost and LightGBM. Logistic Regression was selected as a model for baseline interpretability of the other models. Random Forest, XGBoost and LightGBM were selected because they were tree-based and boosting methods meaning they can capture non-linear relationships. The modelling process included use of TimeSeriesSplit Cross Validation, to maintain a time series structure. The model assumptions evaluated included stationarity and independence, and class imbalance. Therefore, weighted loss functions and alternative evaluation metrics were implemented.

Final Model & Evaluation

An investigation into Hyperparameter tuning for the Random Search approach showed that ROC-AUC was used as the main optimization metric. This reflects the imbalance between the classes of this project and the limitations of using accuracy as a performance metric. The

evaluation of model performance was completed using metrics such as accuracy, ROC-AUC, confusion matrices and plot charts for evaluation on the hold-out test set. Test-set accuracy across all three models ranged from approximately 43% to approximately 47% with ROC-AUC in the vicinity of 0.48. All three of the models tested performed worse than the baseline strategy of predicting all Up moves which achieved approximately 55%. Analysis of variable importance showed that among all variables used to develop the models, features involving lagged returns and the measure of volatility were the most relevant. However, the predictive power of each feature was limited. The results suggest that the direction taken in a short time frame is generally driven by random noise rather than by any predicting power that can be exploited.

Conclusions

The study conducted investigated if machine learning models could accurately forecast the daily direction of the S&P 500 index market direction based on historical data related to price and volume. The results from all stages of pre-processing, feature engineering, tuning, and selection phases of the machine learning models produced no model that could successfully predict the daily price changes within the S&P 500 consistently better than a naive model. This supports the efficient market hypothesis at daily intervals of time and demonstrates the difficulty in identifying useful short-term signals using historical aggregate market data. Machine Learning models are unable to give a trader enough confidence to use their predictions when developing a strategy based on momentum or volatility even if those machine learning models successfully capture some components of momentum and volatility.

Discussion & Next Steps

According to results, daily market direction cannot yet be accurately predicted based on technical analysis alone. Additional variables such as macroeconomics, sentiment or market regime switching will need to be considered to enhance prediction models for market structure changes over time. There is a need to further research the potential benefit of increasing the length of time predictions are being made. Potentially exploring methods and alternative loss functions to predict direction of stock prices through technical analysis will also provide useful insight. Furthermore, this project supports the need for developing and maintaining a rigorous evaluation process of research findings as well as the need for an open reporting method even if research does not support the initial hypothesis or claims.

GitHub Link: https://github.com/aurianaanderson/StockMarketTrends.git