Group A

Team Lead: Auriana Anderson

Recorder: Ross Schanck

Spokesperson: Chase Golden

November 10, 2025

Project Title: Predicting Stock Market Trends Using Historical Price Data

Background and Research Question

Motivation

The Efficient Market Hypothesis (EMH) argues that stock prices reflect all available information, so you cannot outperform the market through selective stock picking or timing. It suggests that prices only change when there is new, unpredictable information that comes out. Historical events like the 2008 financial crisis or dot-com bubble of the early 2000's clearly deviate from the assumptions outlined in the EMH. During these time periods, there is a clear overreaction, which caused prices to either sink or skyrocket, showing the market isn't always following rationale and there are in fact some short-term patterns that pop up during these times.

Research Question

Can the use of simple machine-learning models identify short-term patterns of S&P 500 stock pricing and volume to predict next-day market direction beyond random chance?

Niche

Many stock prediction projects use deep learning or complex architectures that require heavy computational resources and long training times. Our approach is to intentionally focus on simple, interpretable models like Logistic Regression and Random Forests. By limiting the complexity, we can evaluate if even basic models can identify useful patterns from short-term market data. This will provide a clear and testable foundation for exploring market predictions without the noise of overly complex models.

Why is this worthwhile?

If the model only performs slightly above random chance, it can help the financial world understand if there is any pattern to short-term market data. Data scientists can use this to test machine learning on real world time-series data and understand the limitations. Also, since the EMH was developed decades ago, this could show how it needs to be re-developed to keep up with today's market and analytics.

Hypothesis

If recent stock trends and trading volumes show consistent short-term patterns, then simple classification models like logistic regression or random forest can predict next-day stock movement more accurately than random chance.

Prediction

We expect that our Logistic Regression and Random Forest models will achieve predictive accuracy in the range 55-60%. This is slightly better than chance but will prove that short-term market data contains usable patterns. Our hope is that more historical data will improve the model's overall accuracy, however, we are aware this may not be possible due to the market not being perfectly efficient.

Data & Analysis

Data Sources

- S&P 500 Historical Data (1927 – 2022)
    - https://www.kaggle.com/datasets/henryhan117/sp-500-historical-data
- S&P 500 Stocks (daily updated)
    - https://www.kaggle.com/datasets/andrewmvd/sp-500-stocks
- The yfinance library allows you to download daily stock prices directly in python
    - https://pypi.org/project/yfinance/

Why does this data work?

These datasets offer reliable timeseries and historical data on major U.S. stocks which allows for the use of common indicators. They give a wide range of time periods, which is great for training and testing different market conditions.

Key Variables:

Target

- Next-day market directions (Binary Variable: Up if the next day's closing price is higher than today's and Down if it is lower)

Features

Daily historical indicators derived from S&P 500 data, including:

- Opening, closing, high, and low prices

- Daily trading volume
- Moving averages (3-day, 5-day, 10-day)
- Percent change from previous day
- Volatility measures (rolling standard deviation)
- Ratio-based indicators such as price-to-volume change

Analysis Plan

1. **Data Cleaning and preparation:**
    a. Download data using yfinance and Kaggle
    b. Clean the data using Python

2. **Feature Engineering:**
    a. Create new features such as movement averages and percent changes

3. **Modeling:**
    a. Split the data into training and testing sets – 80/20 split
    b. Train simple models like logistic regression and random forest to predict next-day direction
4. **Evaluation Metrics:**
    a. Evaluate performance using accuracy, precision, recall and F1-score
    b. Compare against baseline to measure improvements

c. Compare against random guessing

5. **Avoiding P-Hacking:**

   a. To avoid data snooping and p-hacking, all data processing and feature engineering steps will be defined before model training begins. We will not repeatedly tweak features or parameters based solely on test set performance. Instead, hyperparameters will be tuned using a separate validation set or cross-validation.

6. **Criteria for Success:**

   The project will be considered successful if:

   a. The model's predictive accuracy exceeds 50% (random chance) by a statistically significant margin.

   b. The models demonstrate consistent performance across different market periods

   c. The analysis provides interpretable insights into which indicators most influence short-term price movements.

Technical Details

- **Language:** Python
- **Libraries:** pandas, scikit-learn, matplotlib, yfinance, numpy, seaborn
- **Resources needed:** Kaggle datasets
- **GitHub Repository:** https://github.com/aurianaanderson/StockMarketTrends

Summary

The goal of this project is to test whether simple, interpretable machine learning models can detect short-term market trends using historical S&P 500 data. By comparing model accuracy against random guessing and a baseline strategy, we will evaluate the practical limits of using traditional algorithms for stock prediction. The results will help clarify whether short-term patterns exist in modern financial data.

References

Downey, Lucas. “Efficient Market Hypothesis (EMH): Definition and Critique.” *Investopedia*, Investopedia, www.investopedia.com/terms/e/efficientmarkethypothesis.asp. Accessed 9 Nov. 2025.

“Efficient-Market Hypothesis.” *Wikipedia*, Wikimedia Foundation, 18 Sept. 2025, en.wikipedia.org/wiki/Efficient-market_hypothesis.

Hynes, Laurent. “The Efficient Market Hypothesis Is Fatally Flawed.” *Mises Institute*, 7 Nov. 2025, mises.org/mises-wire/efficient-market-hypothesis-fatally-flawed.

Valencia, Guillermo. “The Market Isn’t Efficient, It’s a Collective Intelligence Ecosystem.” *The Market Isn’t Efficient, It’s a Collective Intelligence Ecosystem*, Macrowise Newsletter, 29 June 2025, macrowise.substack.com/p/the-market-isnt-efficient-its-a-collective.