A Case Study in Time-Series Forecasting: Sector Rotation

By Young Han

Project Overview

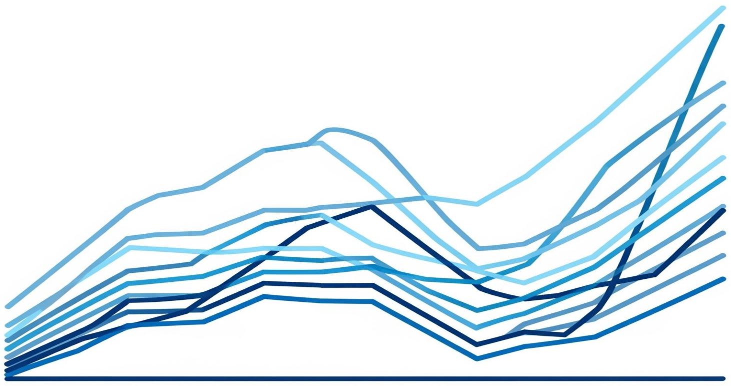
Objective:

To develop and validate a quantitative model that ranks the 11 US sector ETFs based on predicted next-month performance

The ultimate aim is to test if this active strategy can outperform a passive SPY benchmark

Why it matters:

- Provides a data-driven investigation into a popular active investment strategy for investors
- Highlights the challenges of beating the market
- Reveals how models can be biased by the market environment



Dataset

Source:

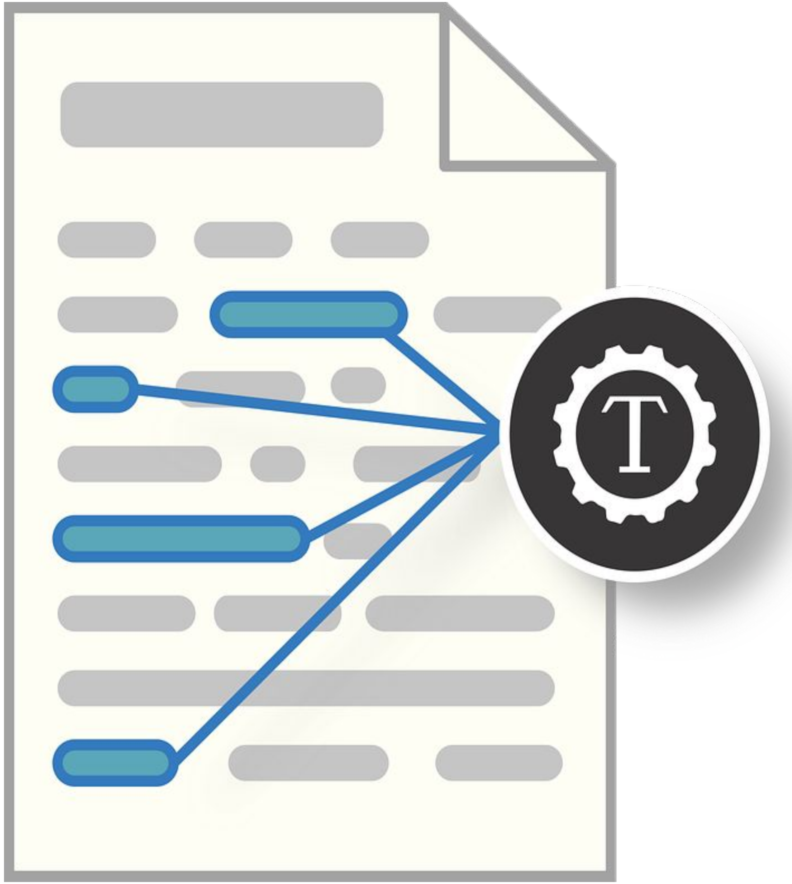
Daily price data from Yahoo Finance
Macroeconomic data from FRED

Scope:

15 years of data covering 11 sector ETFs, benchmarks, and economic indicators

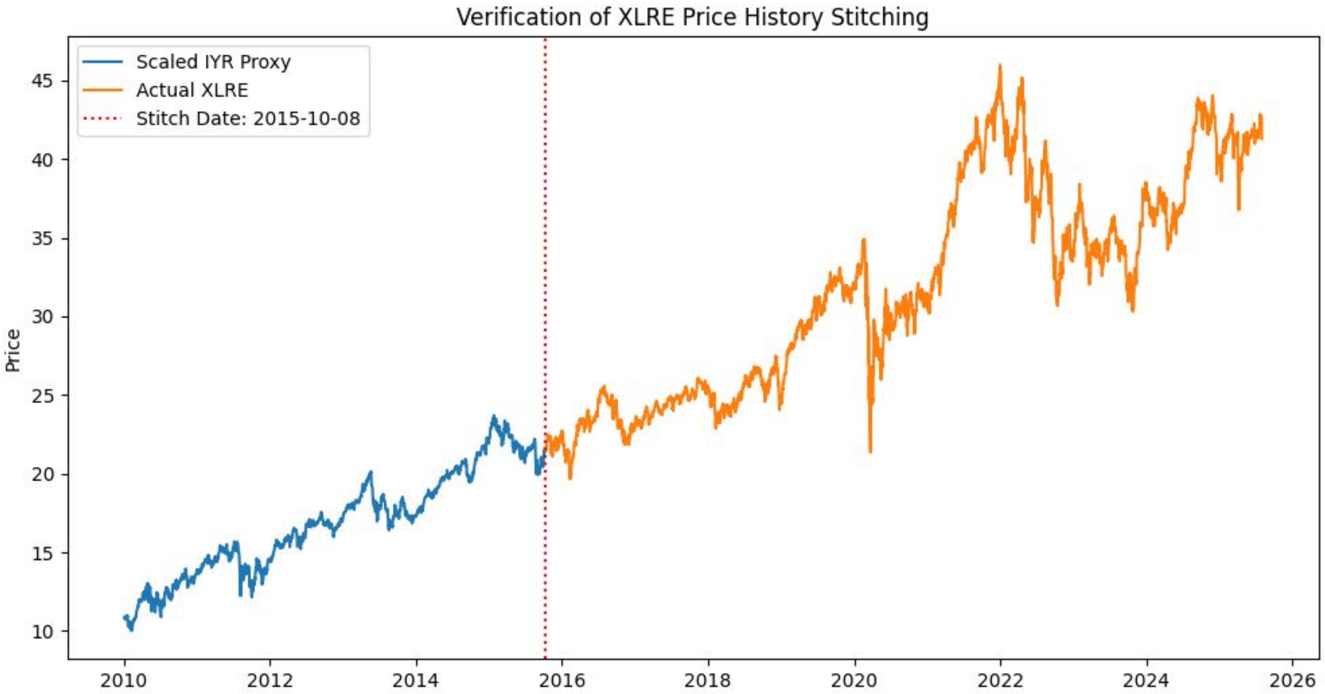
Preprocessing step:

Addressing missing data for newer ETFs by creating a continuous price history



The Data & The First Hurdle

- Like many real-world dataset, part of the data was incomplete (XLRE, XLC)
- Solved this problem by finding highly correlated 'proxy' datasets and using them to backfill the missing history



Feature & Target Engineering

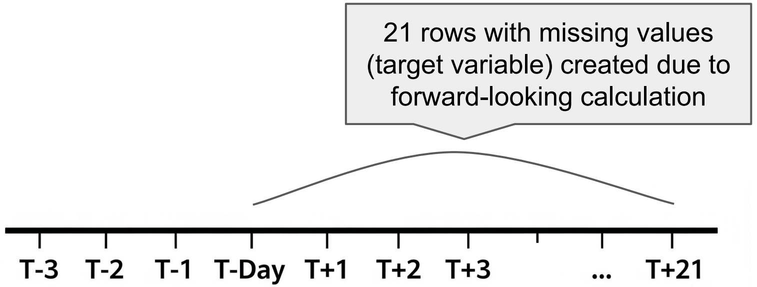
How the Model sees the Market

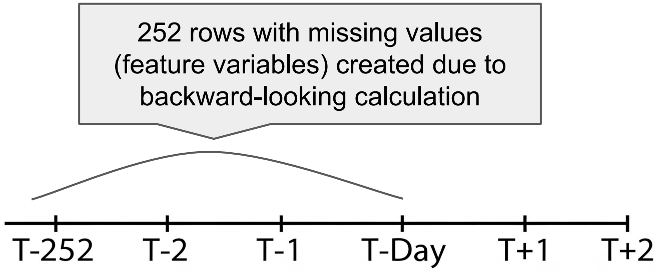
Input features: What just happened?	**Target variable:** What happened next?
Feature categories: • **Momentum**: "How has the price changed over the last n month?" • **Volatility**: "How bumpy was the ride recently?" • **Macro context**: "What was the overall market environment like?"	• Step 1: Predict Absolute Performance ◦ The actual performance over the next 21 trading days (absolute return) • Step 2: Rank for Relative Strength ◦ Individual predictions are ranked against each other to identify the top 3 expected performers ◦ Final rankings is the signal used for the portfolio strategy

Final Data Preparation

Data Cleaning

- The forward and backward looking features naturally create NA values at the start and end of the dataset
- Incomplete rows were removed for the model training





- The final analytical window for the project: From January 2011 to June 2025

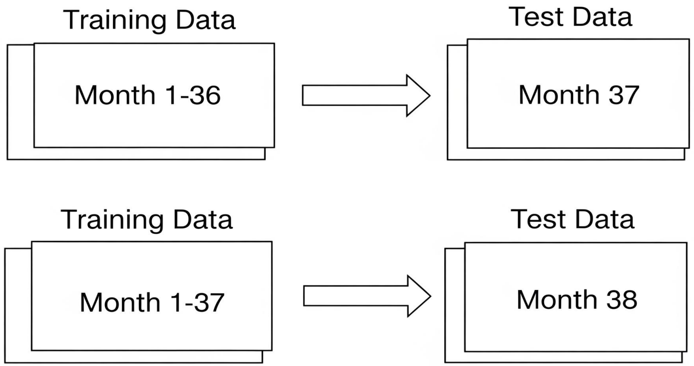
Method

T-2 Timeline

To prevent look-ahead bias, predictions were made on the second to the last business day of the month (T-2).

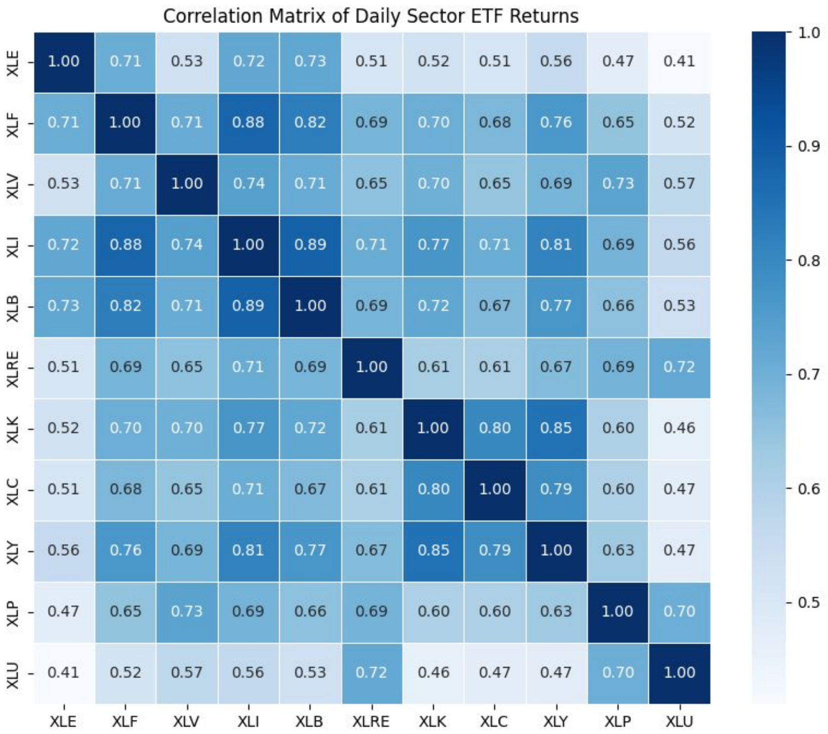
Example: for predicting September 2025
Prediction on August 27, 2025 (T-2)
Execution on August 29, 2025 (T-Day)

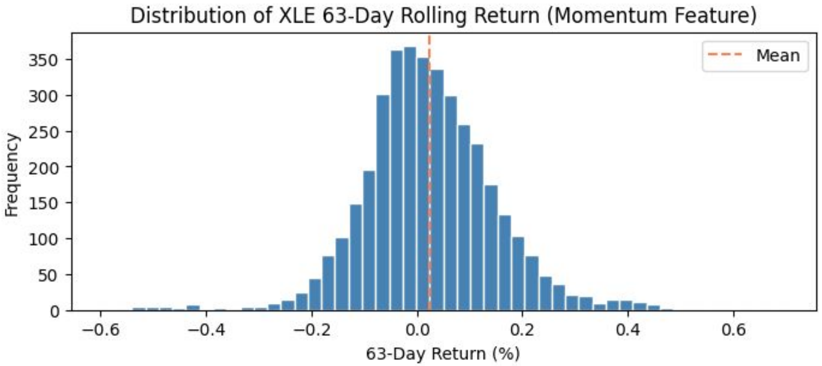
Walk-Forward Validation

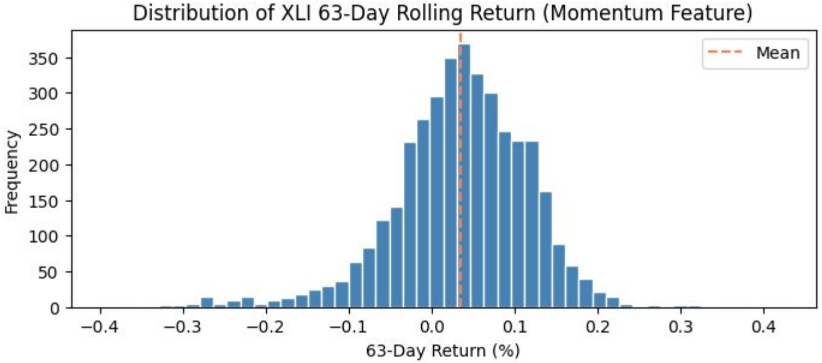


- For time-series, a standard train-test split lets the model see the future
- The method used: **Walk-Forward Validation**
- Training begins with the first 36 months and goes all the way until current month

Core Challenge

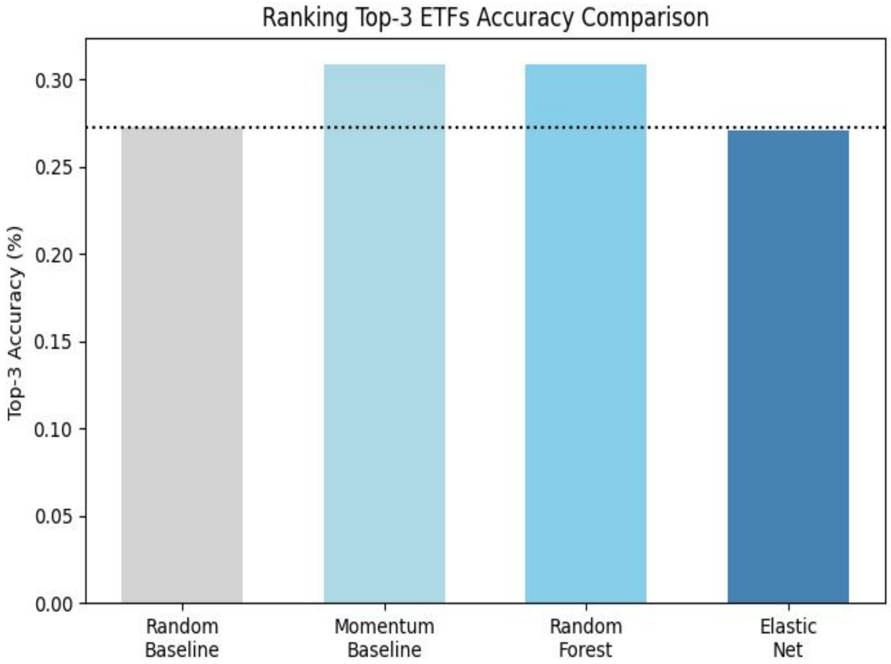




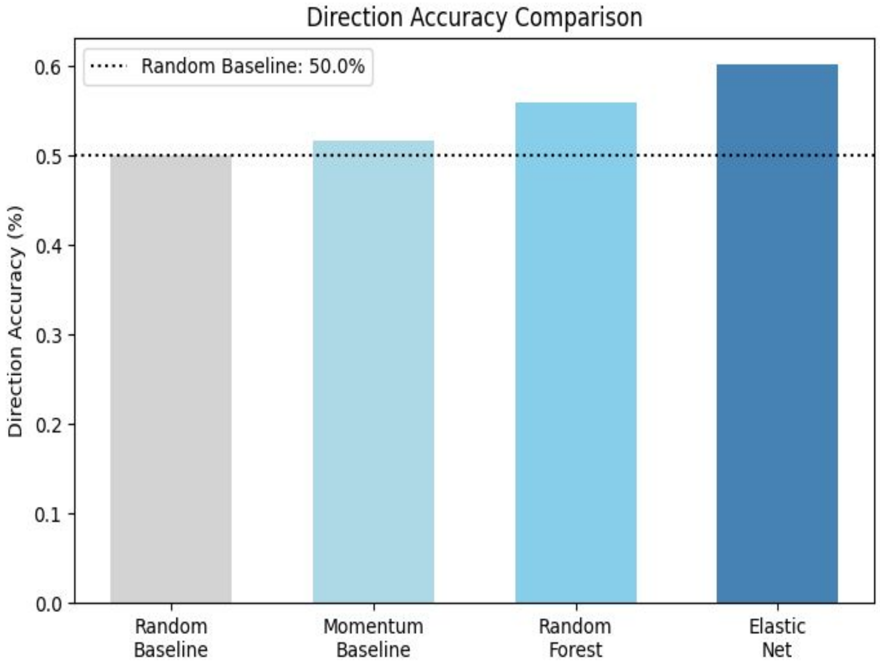


All sectors tend to move together, exhibiting high multicollinearity

Results: The Models Failed

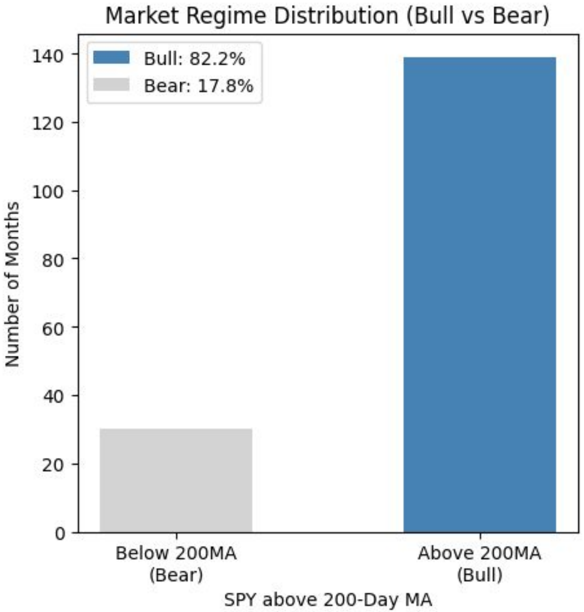


The models failed at their primary goal



A suspicious result with direction accuracy

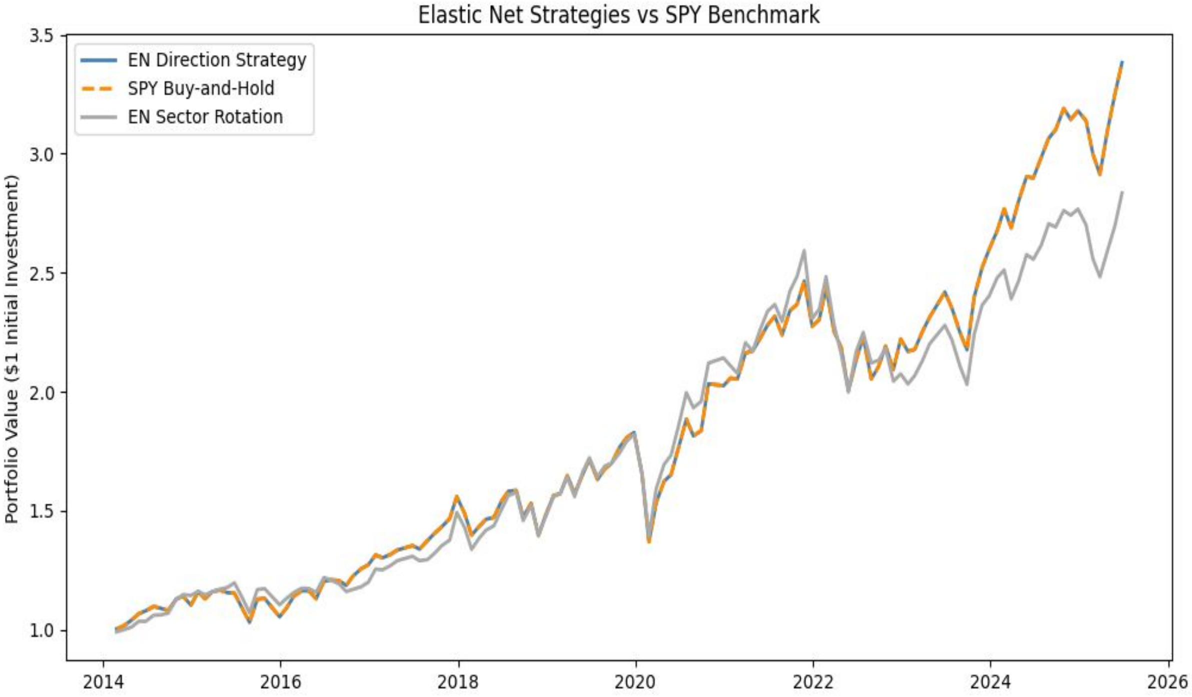
The “Lazy” Model



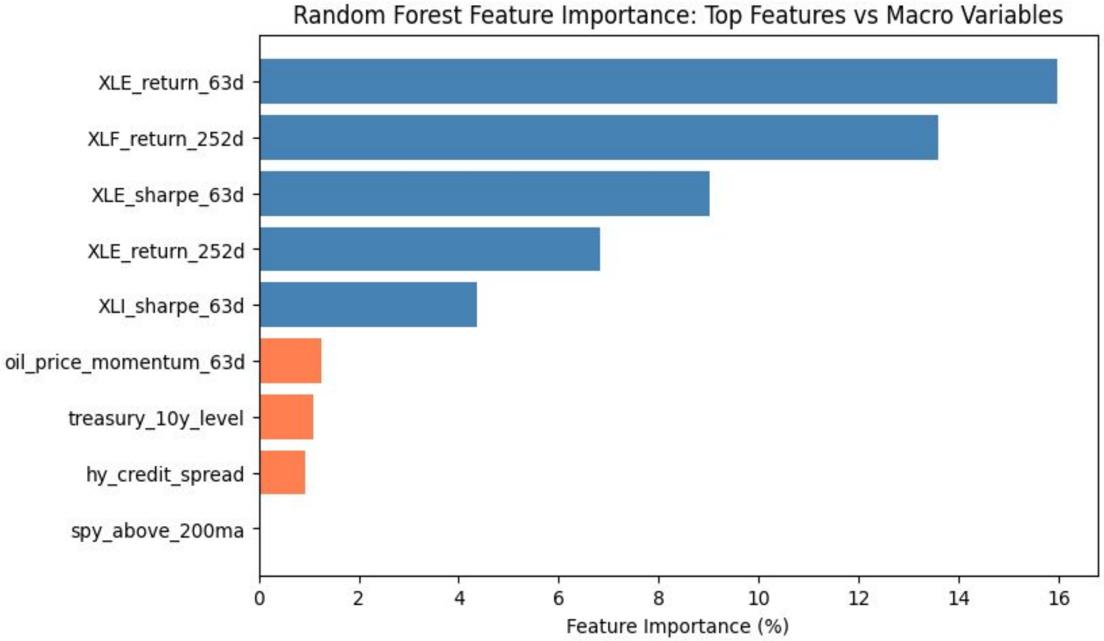
The model was trained on a dataset where the market was in a “**Bull**” regime **82.2%** of the time

Elastic Net Strategies vs

SPY Benchmark (Buy and Hold Strategy)

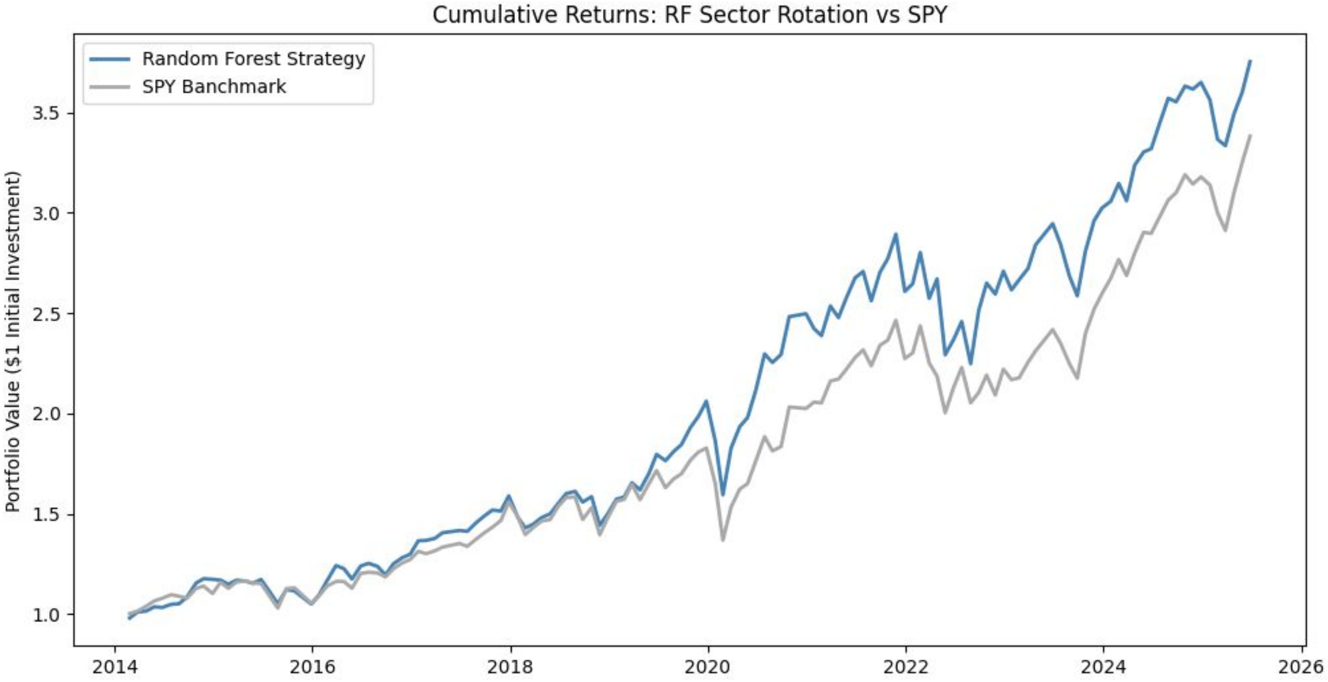


The “Obsessed” Model



- Random Forest model found strong signals - almost exclusively in the energy sector
- 3 of the top 5 features are related to XLE (energy sector)
- Macroeconomic features show little to no importance

The Verdict: Simulating the Outcome



The Random Forest strategy appeared to slightly outperform the SPY benchmark

The result came from a flawed XLE-obsessed model

The model failed to produce a viable, real-world investment strategy

Conclusion

- The linear model became “lazy,” simply following the market’s 82% bull market regime while ignoring features
- The random forest model became “obsessed,” overly relying on highly volatile patterns of the energy sector (XLE) while ignoring other sectors
- This project demonstrated how easily models can be misled by environmental biases and feature characteristics

Future work:

- Engineer more sophisticated features
- Develop a strategy focused on a smaller, less correlated portfolio of sectors
- Build a dedicated model for the energy sector (XLE) to leverage its strong patterns

Reference list
Dataset: Yahoo Finance, FRED
Libraries: scikit_learn, pandas, numpy, seaborn, matplotlib

Thank you!