Predicting S&P 500 Direction with Ensemble Methods

Christian Weißmeier Farkas Tallos

Statistical and Machine Learning (2025/26)

November 14, 2025

Agenda

Motivation

"Stock returns are predictable, but not by much."

— John H. Cochrane, Asset Pricing (2005)

- Forecasting stock market direction is one of the most classical and challenging tasks in finance.
- Weak predictability can matter for portfolio allocation and risk management.

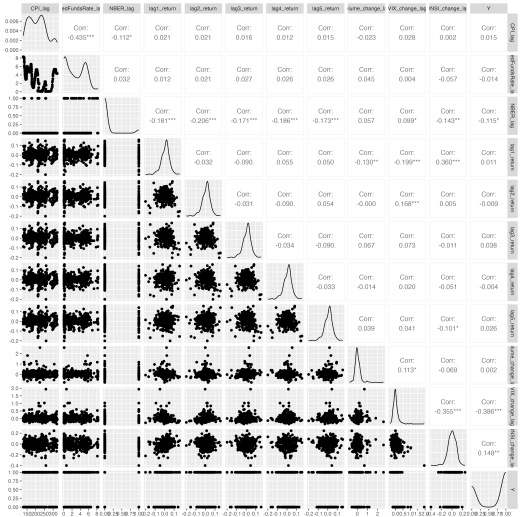
Our Set-Up:

- We focus on predicting whether the S&P 500 index goes **Up** or **Down** next month.
- We evaluate statistical learning methods in a time-series context.
- Introduce the role of regularization and nonlinear models.
- Compare linear vs. nonlinear classification models (Elastic Net vs. Random Forest).

Data Overview

- We created our own monthly dataset from S&P 500, FRED, and FRBSF data banks (1990–today).
- **Target:** Monthly S&P 500 market direction (**UP** or **DOWN**) in $t + 1$.
- **Predictors:**
  1. **Market data:** Lagged S&P 500 returns (up to five months) and trading volume changes.
  2. **Macroeconomic indicators:** CPI, Federal Funds Rate, NBER recession dummy.
  3. **Volatility:** Lagged changes in the VIX index.
  4. **Sentiment:** Daily News Sentiment Index.
- All features are lagged to avoid look-ahead bias.

Exploratory Analysis: Feature Relationships



Pairwise correlations and marginal distributions.

Why these predictors may matter:

- **Macro indicators** (CPI, Fed Funds Rate, NBER) affect discount rates and expected returns.
- **Lagged returns** reflect momentum and reversal effects.
- **Volatility** (VIX) captures uncertainty and risk sentiment.
- **Sentiment** (DNSI) reflects investor expectations and behavioral biases.

Empirical observation:

- Weak linear correlations $\Rightarrow$ low-signal, non-linear problem.
- Several predictors show significant **multicollinearity**.
- Motivates using **Elastic Net**, **Random Forests**, and **Gradient Boosting**.

Stationarity Considerations

- Most macro-financial time series are **non-stationary** in levels.
- We therefore use:
  - **Lagged returns** instead of prices.
  - **Changes** in VIX and sentiment rather than levels.
  - **Changes in macro variables** to preserve temporal causality.
- This transformation makes features approximately stationary, ensuring:
  - Stable model coefficients over time.
  - Valid cross-validation across time periods.

Hyperparameter Tuning: Time-Series Cross-Validation

Why not standard cross-validation?

- Random k-fold CV assumes i.i.d. data.
- Time-series data exhibit autocorrelation $\Rightarrow$ temporal dependence.
- Random shuffling lets the model "see the future" $\Rightarrow$ data leakage and over-optimism.

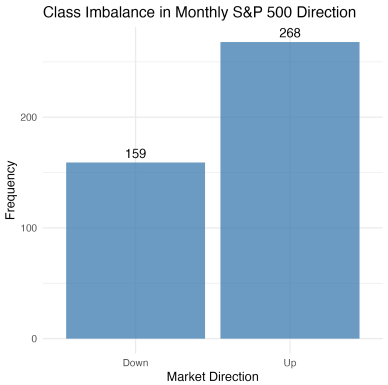
Our method: Rolling-window time-series CV

- Construct a hyperparameter grid:
  - Elastic Net: α and λ grid.
  - Random Forest: `mtry`, `min.node.size`, `max.depth`, `sample.fraction`.
- Use a chronological window structure:
  - Initial training window: 60 months (5 years).
  - Validation horizon: 12 months (1 year).
  - Fixed-length rolling window moving forward in time.
- For each fold:
  - Fit model only on past data and predict the next 12 months.
  - Record accuracy and AUC on the validation slice.

Result:

- Select the hyperparameters achieving the **highest mean AUC across all rolling folds**.
- Prevents look-ahead bias and mimics real-time forecasting.

Class Imbalance, Threshold Adjustment & AUC



Class imbalance in monthly S&P 500 direction

- Dependent variable `UP_DOWN` is imbalanced.
- A naive classifier that would always predict "Up" would achieve high accuracy.
- A simple threshold of 0.5 reduces the accuracy.
  $\Rightarrow$ **Accuracy alone is misleading.**

Threshold adjustment (Youden's J):

$$t^* = \arg\max_t \left\{ \text{Sensitivity}(t) + \text{Specificity}(t) - 1 \right\}$$

- We tune t^* on the **training set** using the ROC curve.
- Then apply this fixed threshold to the **test set** for true OoS evaluation.

AUC is robust here:

- ROC/AUC depends on the *ranking* of predicted probabilities, not on class proportions.
  $\Rightarrow$ **AUC is invariant to class imbalance** and a suitable metric for our setting.

Elastic Net Logistic Regression

Model Intuition:

- We model the probability of an **Up** move using a logistic function:

$$P(Y_t = 1 \mid X_t) = \frac{1}{1 + e^{-(\beta_0 + X_t \beta)}}$$

- Coefficients β are maximum likelihood estimates under a regularization penalty to prevent overfitting and perform variable selection.

Elastic Net Regularization:

$$\min_{\beta} \left[-\ell(\beta) + \lambda \left((1-\alpha) \frac{\|\beta\|_2^2}{2} + \alpha \|\beta\|_1 \right) \right]$$

- $\ell(\beta)$: log-likelihood of the logistic model.
- λ: overall penalty strength controlling coefficient shrinkage.
- α: mixes the two types of regularization:
  - Ridge (L2): smooth shrinkage.
  - Lasso (L1): sets some coefficients exactly to zero.

Elastic Net: Out-of-Sample Performance

Predictive performance (test set):

- Accuracy: **64.3%**
- AUC: **0.79** (strong probability ranking ability)
- Threshold tuned : $t^* = 0.6743 \rightarrow$ The model avoids overly optimistic Up predictions

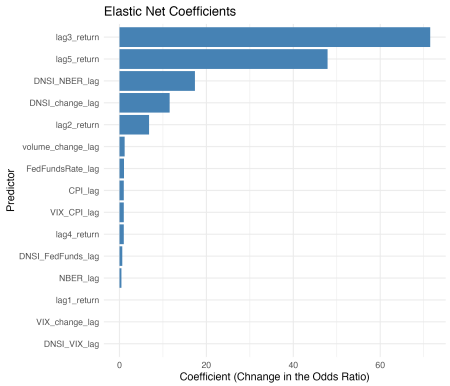
Confusion matrix:

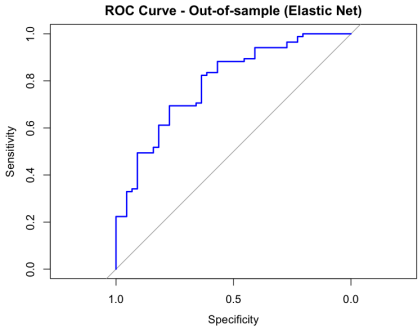
	Actual 0	Actual 1
Predicted 0	36	38
Predicted 1	8	47

Interpretation:

- **Specificity** (correct Down predictions): $36/44 \approx$ **82%** $\Rightarrow$ we correctly flag most Down months.
- **Sensitivity** (correct Up predictions): $47/85 \approx$ **55%** $\Rightarrow$ we still capture the majority of Up months.
- The model trades a small loss in raw accuracy for a much better **balance** between detecting costly Down markets and still identifying Up markets, which is desirable in an asset-management context.

Elastic Net: Selected Predictors and Probability Ranking





- Coeff. represent changes in the **odds** of an "Up" month.
- Elastic Net highlights variables with consistent signal.
- Momentum is a major driver of Up probabilities.
- Sentiment matters more during recession periods.

- **AUC = 0.79**: strong discriminative ability despite low-signal, noisy financial data.
- ROC curve well above diagonal ⇒ model ranks Up vs. Down months effectively.

Random Forest: Out-of-Sample Performance

Predictive performance (test set):

- Accuracy: **71.3%**
- AUC: **0.753** (strong non-linear discriminative ability)
- RF produces well-calibrated probability scores without threshold tuning

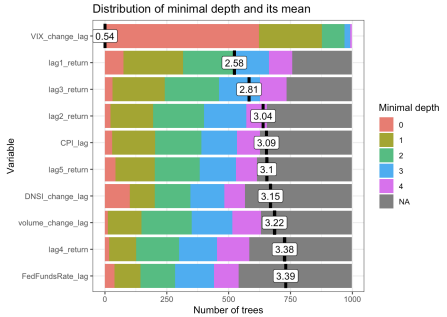
Confusion matrix:

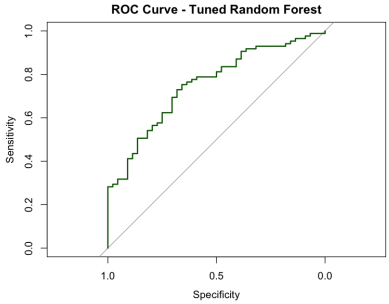
	Actual 0	Actual 1
Predicted 0	29	22
Predicted 1	15	63

Interpretation:

- **Specificity** (correct Down predictions): $29/44 \approx$ **66%**. RF detects a meaningful fraction of negative weeks.
- **Sensitivity** (correct Up predictions): $63/85 \approx$ **74%**. RF identifies most positive-return weeks.
- **Overall**: RF captures important **nonlinear return patterns**, outperforming Elastic Net in accuracy and maintaining a strong AUC.

Random Forest: Minimal Depth, Interpretation, and AUC





Explanation and Interpretation of Minimal Depth

- Min. depth measures how **early** a variable is used in the tree.
- **VIX_change_lag** is the dominant signal
- **Lagged returns** are meaningful momentum predictors.
- Overall: RF relies mainly on **volatility shocks** and **recent price dynamics**.

AUC

- **AUC = 0.753** $\Rightarrow$ strong discriminative ability and reliable ranking of Up vs. Down weeks.

Takeaway

- Random Forest effectively captures nonlinear interactions between **volatility movements** and **short-term momentum**.

The Problem: Time-Series Data

Standard K-fold CV shuffles data randomly, "peeking into the future" and violating temporal order. This leads to overly optimistic results.

3. Methodology: Model Assessment Strategy

The most critical methodological slide

The Problem: Time-Series Data

Standard K-fold CV shuffles data randomly, "peeking into the future" and violating temporal order. This leads to overly optimistic results.

Our Solution: Two-Level Chronological Split

- **Level 1: Train/Test Split (for Assessment)**
  - Chronological 70/30 split.
  - **Train:** 1990–2014 (used for tuning).
  - **Test:** 2015–2025 (used once for final evaluation).

3. Methodology: Model Assessment Strategy

The most critical methodological slide

The Problem: Time-Series Data

Standard K-fold CV shuffles data randomly, "peeking into the future" and violating temporal order. This leads to overly optimistic results.

Our Solution: Two-Level Chronological Split

- **Level 1: Train/Test Split (for Assessment)**
  - Chronological 70/30 split.
  - **Train:** 1990–2014 (used for tuning).
  - **Test:** 2015–2025 (used once for final evaluation).
- **Level 2: Rolling-Window CV (for Tuning)**
  - Within the 70% training set, perform rolling-window validation.
  - Simulates real-world use: train on past, predict the future.