MF850 Project on Feature Sets

A. Max Reppen

November 25, 2025

Group size: 3–5 students. One submission per group, with all members clearly listed.
Deliverable: A written report (PDF; not a notebook printout of your code) and any code files used to generate results (separately from the PDF).
Due date: See Blackboard.
Late submission policy: 10% penalty per day late, up to 5 days. No submissions accepted after 5 days late.

Introduction

The purpose of this assignment is to explore how different types of features — raw price/volume data, engineered technical indicators, and firm fundamentals — contribute to predicting stock returns at different horizons. You will train identical models (aside from inputs) on different feature sets and compare their performance. The goal is to understand how feature choice interacts with prediction horizon and to connect empirical results to financial theory (e.g., momentum, volatility clustering, fundamentals).

Goal

You will explore how different feature sets impact the predictive performance of a model. In particular, the goal is to give hands-on experience seeing how feature engineering and selection affect model outcomes, here in a financial context. In other words, your choice of features will impact the model's ability to predict stock returns at various horizons.

This is also an opportunity to practice avoiding pitfalls and working with other challenges in financial ML, such as point-in-time data merging, label construction, and time-series cross-validation.

Because this is a project-format, there is great flexibility in how you approach each step. The instructions below are meant as a guiding framework rather than detailed prescriptions.

1 Instructions

Follow these steps to complete the assignment:

1. **Data:** Load daily prices and quarterly fundamentals from the supplied csv files. Merge them point-in-time (as-of backward) by ticker so each daily observation uses only information available at that date.
2. **Feature sets:**
   - **Raw (optional):** Daily return (percentage change in closing price from yesterday to today) and trading volume.
   - **Engineered:** Rolling means/volatilities, momentum ratios, EMA crossovers, skewness, kurtosis, etc.
   - **Fundamentals:** EPS, profit margin, revenue growth, and other ratios derived from reported items.
   - **Fund+Eng:** Combination of engineered technical features and fundamentals.
3. **Targets (binary labels):** For each horizon $h \in \{1, 20, 60\}$ days, define

$$y_t = \begin{cases} 1 & \text{if the cumulative return over the next } h \text{ days is positive} \\ 0 & \text{otherwise} \end{cases}$$

   This ensures the model predicts whether the forward return is up or down.
4. **Modeling:** Train the same classifier across all feature sets. Use chronological splits (no shuffle) into train/validation/test and standardize features. Explore different classifiers and features.
5. **Evaluation:** Record validation and test accuracies for each feature set and horizon.
6. **Discussion:** Explain horizon-specific relevance of features (technical at short horizons; fundamentals at longer horizons; raw benefits from noise averaging). Connect to financial concepts like momentum and quarterly reporting cycles.
7. **Reflection:** Discuss consequences of pooling across firms (assumed homogeneity, loss of firm-specific dynamics, potential cross-sectional leakage).
8. **Pitfalls:** Briefly state how you avoided common mistakes (shuffling time, incorrect label shifts, rolling windows across firms, non-point-in-time merges, missing-value handling, scaling). The dataset includes COVID-era anomalies. How does that affect your modeling and results?
9. **Extensions (optional):** Try longer horizons (e.g., 120-day), sector-specific models, alternative architectures (logistic regression, tree-based), or different metrics (precision, recall, Sharpe).

2 Deliverables

By the end, present:

1. **Results table:** Validation and test accuracies for each feature set and horizon.
2. **Model and feature construction:** What was your final model and how did you reach that choice? What were the final features used and how did you choose them?
3. **Discussion and reflection (analysis write-up):** Discuss your findings on feature relevance by horizon with financial intuition.

This should include thoughts on your performance by looking different evaluation criteria, such as:

- Comparing accuracy to a baseline (e.g., random guessing or always predicting the majority class) to show whether the model adds value beyond trivial strategies.
- Running a simple trading simulation using the model's predictions and report cumulative returns, Sharpe ratio, and maximum drawdown to connect classification results to financial performance.

Further discussion points to consider:

- **Keys to success:** What were keys for you to train a successful model? Data preprocessing, feature engineering, hyperparameter tuning, etc.?
- **Pitfalls & reflection:** Reflect on potential failure modes and common mistakes (for instance, but not only, the pitfalls described earlier). Explain how and if you detected, avoided, and mitigated each.

4. **Extensions (optional):** Summary of any extra experiments.

3 Evaluation criteria and grading

Each student is expected to be able to account for any part of their process and results.

Your submission will be evaluated based on, in order of importance:

- **Communication and insight:** Clarity and organization of the write-up and presentation.
  I am looking for a well-structured report with an informative summary, clear explanations of methods, and thoughtful discussion of results, without unnecessary verbosity and jargon. Well thought-out figures and tables that effectively convey key findings are important.
  Do keep the main text concise and focused, you may include supplementary materials in an appendix.
- **Completeness:** All deliverables are provided and follow instructions.
- **Technical correctness:** Appropriate handling of financial ML challenges (point-in-time merging, label construction, time-series splits, etc.) and sound modeling choices.
- **Originality:** Creativity in feature engineering, modeling approaches, or extensions can compensate for shortcomings in other areas. Use of techniques outside of class material is encouraged, but must be well-justified and clearly explained.[1]

To summarize, the focus is not on raw performance metrics, but on demonstrating understanding of financial ML concepts, sound methodology, and clear communication of insights.

[1] For instance, in brief experiments, I found success with embedded ticker information, a topic not covered in class.

Momentum and financial theory

Momentum may appear inconsistent with the Efficient Market Hypothesis, but several frameworks reconcile it:

- **Risk-based:** Momentum portfolios are crash-prone in bear markets, consistent with earning a risk premium.
- **Behavioral:** Underreaction to news and extrapolation of trends create price continuation.
- **Information diffusion:** News is incorporated with lags across investors, producing predictable drift.
- **Market frictions:** Costs and constraints limit arbitrage, allowing momentum to persist.

Optional reading list

- Jegadeesh & Titman (1993). *Returns to Buying Winners and Selling Losers.* Journal of Finance.
- Carhart (1997). *On Persistence in Mutual Fund Performance.* Journal of Finance.
- Asness, Moskowitz, & Pedersen (2013). *Value and Momentum Everywhere.* Journal of Finance.
- Jegadeesh & Titman (2011). *Momentum.* Review of Financial Studies.