Lesson 5.3: The Limits of Prediction – Practice Exercises

Digital Finance v4

Exercise 1: Identify the Bias

A data scientist presents the following backtest results for an ML stock-picking model:

Setup:

- Universe: current S&P 500 members (as of January 2025)
- Training period: 2010–2022, Test period: 2023–2024
- Features include "next quarter's revenue growth" (from restated financials)
- Tested 500 feature combinations, reported the best one
- No transaction costs included

Reported result: Annual return 42%, Sharpe ratio 3.8, max drawdown -4%

Questions:

1. Identify *at least four* distinct biases or errors in this backtest.
2. For each bias, explain *how* it inflates the reported performance.
3. Explain which single bias is likely the most damaging and why.
4. Describe the correct methodology to fix each bias.

Exercise 2: EMH Debate

Three investment professionals make the following claims:

Alice (Hedge Fund PM): "My momentum strategy has beaten the market for 7 consecutive years. EMH is wrong."
Bob (Index Fund Manager): "85% of active managers underperform over 15 years. EMH is right – just buy the index."
Carol (Quant Researcher): "EMH is mostly right for large caps, but micro-cap markets and emerging markets have exploitable inefficiencies."

Questions:

1. Which form of EMH (weak, semi-strong, strong) does each person's view most closely align with?
2. Is 7 years of outperformance statistically convincing? Calculate how many managers out of 1,000 would beat the market 7 years in a row by pure luck (assume 50% chance each year).
3. How does the Adaptive Market Hypothesis (Lo, 2004) reconcile these three views?
4. If Carol is right, what are the practical barriers to profiting from micro-cap inefficiencies?

Exercise 3: Regime Change Impact Analysis

You manage an ML credit risk model trained on 2015–2019 data (low rates, low defaults).

Scenario: In 2022, central banks raise interest rates from 0.25% to 5.25% in 18 months.

Metric	Training Period	2022–2023
Default rate	1.2%	3.8%
Avg. interest rate	1.5%	4.8%
Model AUC	0.89	0.64

Questions:

1. Why did the model's AUC drop from 0.89 to 0.64?
2. Is this a "gradual drift" or "sudden drift" scenario? Explain.
3. Design a monitoring dashboard with 3 specific metrics that would have detected this regime change *before* the model failed.
4. Propose an adaptive strategy (ensemble, retraining, or online learning) to handle future regime changes.

Exercise 4: Sentiment Analysis Pipeline Design

You are building a sentiment analysis pipeline for trading earnings announcements.

Data sources:

- Earnings call transcripts (CEO/CFO remarks + analyst Q&A)
- SEC 8-K filings (earnings announcements)
- Financial news articles (Reuters, Bloomberg)
- Social media posts (Twitter/X, StockTwits)

Questions:

1. Design a 5-step pipeline: specify each NLP step and its input/output.
2. Why would a general-purpose sentiment dictionary (e.g., VADER) give misleading results on financial text? Give 2 specific examples of words with different sentiment in finance.
3. How would you handle *sarcasm* and *hedging* in analyst language? (e.g., "The results were not entirely disappointing.")
4. Your pipeline produces a sentiment score of $+0.6$ for a company. What additional information do you need before making a trading decision?

Exercise 5: Walk-Forward Validation

You have daily return data from January 2010 to December 2024 (15 years).

A colleague proposes: "Let's use 80/20 random split for train/test."

Questions:

1. Explain in 3 sentences why random splitting is wrong for this problem.
2. Design a walk-forward validation scheme:
   - Specify initial training window size
   - Specify test window size
   - Specify whether the window is expanding or rolling
   - How many test folds will you produce?
3. What is "purging" and why is it necessary at the train/test boundary?
4. Your model achieves Sharpe 1.8 in walk-forward validation but 0.4 in the first month of live trading. List 3 possible explanations that are *not* overfitting.

Exercise 6: RL Trading Agent Evaluation

A startup claims their RL trading agent "outperforms the S&P 500 by 15% annually."

Their methodology:

- Trained on 10 years of historical 1-minute bar data
- Simulated environment with zero transaction costs and perfect fills
- Reward function: daily portfolio return
- Tested on the same 10 years of data (no out-of-sample period)

Questions:

1. Identify 4 specific methodological problems with this evaluation.
2. Explain why "daily portfolio return" is a problematic reward function. What could the agent learn to do?
3. Design a more realistic evaluation: specify (a) environment, (b) reward function, (c) train/test split, and (d) baseline comparison.
4. Even with perfect methodology, name 2 fundamental reasons why RL may still fail in live markets.

Exercise 7: Alpha Decay Quantification

A quantitative fund discovers a pairs trading signal with the following track record:

Year	**1**	**2**	**3**	**4**	**5**
Sharpe Ratio	2.4	2.0	1.4	0.9	0.5
AUM ($M)	10	50	200	500	800
Known Competitors	0	2	8	20	35

Questions:

1. Plot or describe the relationship between Sharpe decay and number of competitors.
2. Estimate the alpha half-life (time for Sharpe to halve from its peak).
3. At what AUM level did the strategy become capacity-constrained? How can you tell?
4. If the fund's breakeven Sharpe (after costs) is 0.7, when should the strategy be retired or significantly modified?

Exercise 8: Comprehensive Model Audit

You are asked to audit an ML model before it goes into production for a bank's trading desk.

Model details:

- XGBoost with 200 features predicting next-day stock returns
- Trained on 2012–2023 with random 80/20 split
- Backtest Sharpe: 2.1, Annual return: 28%, Max drawdown: -8%
- No survivorship correction, no transaction costs

Questions:

1. Write a structured audit checklist with at least 6 items to verify.
2. Which biases are definitely present? Which are possibly present?
3. Estimate the "true" Sharpe ratio after correcting for: (a) random split bias, (b) transaction costs (assume 15 bps round-trip, 250 trades/year), (c) survivorship bias.
4. Write a 3-sentence recommendation to the trading desk: deploy, revise, or reject?