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?

Answer Key (1/3)

Exercise 1: Identify the Bias

1. Four biases: (a) Survivorship bias – using current S&P 500 excludes delisted companies. (b) Look-ahead bias – "next quarter's revenue growth" uses future data. (c) Data snooping – tested 500 combinations, reported the best (multiple testing). (d) Missing transaction costs – inflates returns. Also: no slippage modeling.
2. Survivorship: overstates returns by $\sim$1–2%/yr. Look-ahead: model "sees the future", massively inflating accuracy. Data snooping: 500 tests at $p = 0.05$ yields $\sim$25 false positives. Missing costs: high-turnover strategies lose 1–3%/yr to friction.
3. Look-ahead bias is most damaging – it fundamentally breaks the information set available at decision time. The model is literally cheating.
4. Fixes: (a) Use point-in-time membership lists. (b) Use only data available as of each date. (c) Apply Bonferroni correction or use walk-forward validation. (d) Include realistic transaction costs and slippage.

Exercise 2: EMH Debate

1. Alice: rejects semi-strong form. Bob: supports semi-strong form. Carol: supports adaptive/conditional efficiency (semi-strong for large caps, weaker for small).
2. $P(\text{beat 7 years}) = 0.5^7 = 0.78\%$. Out of 1,000 managers: $1000 \times 0.0078 \approx 8$ would beat by pure luck. Not convincing.
3. AMH says efficiency varies by market, time, and competition. All three can be right simultaneously: Alice found a temporary niche, Bob's stats hold in aggregate, Carol correctly identifies where efficiency breaks down.
4. Barriers: illiquidity (can't get meaningful size), high transaction costs, limited data for modeling, capacity constraints ($<$\$50M).

Exercise 3: Regime Change Impact Analysis

1. The model learned relationships (debt ratios, income levels) calibrated to low-rate, low-default conditions. At 5% rates, different borrowers default and different features matter.
2. Sudden drift – rates moved 500 bps in 18 months (fastest tightening in 40 years). Not gradual at all.
3. Dashboard: (a) Rolling AUC on recent data (trailing 90 days). (b) Feature distribution shift (PSI – Population Stability Index). (c) Macro indicator alerts (Fed funds rate, yield curve inversion).
4. Ensemble of regime-conditional models: train separate models for low-rate and high-rate regimes, use a meta-classifier to select which model applies. Retrain quarterly with expanding window.

Answer Key (2/3)

Exercise 4: Sentiment Analysis Pipeline Design

1. Pipeline: (Step 1) Data ingestion – collect text from sources, normalize format. (Step 2) Preprocessing – tokenize, remove stop words, lemmatize. (Step 3) NER – extract companies, people, amounts, dates. (Step 4) Sentiment scoring – apply FinBERT or Loughran-McDonald dictionary per entity. (Step 5) Aggregation – compute entity-level score with time decay weighting.
2. VADER misclassifies financial terms: (a) "liability" – negative in general English, neutral in finance (it's an accounting term). (b) "volatile" – negative in general, descriptive/neutral in finance. Loughran-McDonald dictionary handles these correctly.
3. Sarcasm: requires contextual models (FinBERT handles some; rule-based negation detection helps). Hedging: phrases like "not entirely disappointing" require double-negation parsing – FinBERT handles better than dictionaries.
4. Need: (a) analyst consensus estimate (is $+0.6$ above or below expectations?), (b) volume and price action (has market already priced in?), (c) sector/market context, (d) confidence score and sample size of the sentiment measure.

Exercise 5: Walk-Forward Validation

1. Random splitting violates temporal ordering, allowing the model to train on 2020 data and test on 2015 data. This creates look-ahead bias, as the model learns from future events to predict past ones. Time series data must be split respecting chronological order.
2. Scheme: Initial training window: 5 years (2010–2014). Test window: 6 months. Rolling window (fixed 5-year training). Folds: $(2024 - 2015) \times 2 = 20$ test folds (each 6 months from 2015 through 2024).
3. Purging: removing training samples close to the test boundary to prevent label leakage. If features use 5-day look-back and labels use 1-day forward return, the last 6 training days' labels overlap with the test period.
4. (a) Regime change – first live month hits different market conditions. (b) Execution gap – live fills worse than assumed in backtest (slippage, latency). (c) Capacity/impact – live order sizes move prices.

Exercise 6: RL Trading Agent Evaluation

1. Problems: (a) No out-of-sample test (trained and tested on same data). (b) Zero transaction costs (unrealistic). (c) Perfect fills (no market impact). (d) Daily return reward encourages excessive risk/leverage.
2. Daily return reward: agent may learn to take maximum leverage since reward is linear in return but ignores risk. Could learn to "go all-in" on volatile stocks, resulting in catastrophic drawdowns.
3. Realistic setup: (a) Environment with realistic order book, 5 bps slippage, partial fills. (b) Reward: risk-adjusted (Sharpe - drawdown penalty - cost penalty). (c) Train on 2012–2019, validate 2020–2021, test 2022–2024. (d) Compare vs. buy-and-hold, equal-weight, and momentum baselines.
4. (a) Non-stationarity – live market regime differs from training. (b) Adversarial nature – other participants react to the agent's trades, changing the dynamics.

Answer Key (3/3)

Exercise 7: Alpha Decay Quantification

1. Inverse relationship: as competitors increase ($0 \rightarrow 35$), Sharpe declines ($2.4 \rightarrow 0.5$). Approximately, each 10 additional competitors reduce Sharpe by $\sim$0.5–0.6. AUM growth compounds the problem through market impact.
2. Peak Sharpe = 2.4 (Year 1). Half = 1.2. By interpolation between Year 2 (2.0) and Year 3 (1.4), the half-life is approximately 2.5 years.
3. Capacity constraint appears between Years 2–3: AUM grew from \$50M to \$200M while Sharpe dropped sharply from 2.0 to 1.4. The Sharpe/AUM ratio collapsed, indicating market impact began consuming alpha.
4. Breakeven Sharpe = 0.7 means the strategy is marginally unprofitable in Year 5 (Sharpe $0.5 < 0.7$). Should have been retired or modified in Year 4 when Sharpe (0.9) was approaching breakeven. Recommendation: retire by end of Year 4, redeploy capital to new signals.

Exercise 8: Comprehensive Model Audit

1. Audit checklist: (a) Train/test split method – is it walk-forward? (b) Survivorship bias – is the stock universe point-in-time? (c) Look-ahead bias – are all features available at prediction time? (d) Transaction costs – are realistic costs included? (e) Feature count vs. sample size – is 200 features justified? (f) Out-of-sample period – is there a true holdout?
2. Definitely present: random split bias (confirmed), missing transaction costs (confirmed), no survivorship correction (confirmed). Possibly present: look-ahead bias (200 features need individual audit), data snooping (unknown how many feature sets were tried).
3. Corrections: (a) Random split $\rightarrow$ walk-forward: typically reduces Sharpe by 20–40%, so $2.1 \times 0.7 \approx 1.47$. (b) Transaction costs: 250×15 bps $\times\ 2 = 75$ bps round-trip $\times\ 250 = 3.75\%$ annual drag, reducing return from 28% to $\sim$24%. (c) Survivorship: subtract $\sim$1.5%/yr $\rightarrow$ $\sim$22.5% return. Adjusted Sharpe: roughly 1.0–1.2. Still potentially viable but much less impressive.
4. Recommendation: "Revise before deployment. The model shows promise but the evaluation is methodologically flawed. Re-run with walk-forward validation, point-in-time stock universe, and 15 bps round-trip costs. If walk-forward Sharpe exceeds 1.0, proceed to paper trading for 3 months before live deployment."