Lesson 5.4: MLOps and Production ML in Finance – Exercises

Module 5: Automation and Infrastructure

Digital Finance

Exercise 1: PSI Calculation

A credit scoring model was trained on the following income distribution. After 6 months in production, the distribution has shifted.

Income Bin	Expected (Training)	Actual (Production)
\$0–30k	0.20	0.12
\$30k–60k	0.35	0.30
\$60k–90k	0.25	0.32
\$90k–120k	0.12	0.16
\$120k+	0.08	0.10

Tasks:

1. Calculate the PSI contribution for each bin using $(A_i - E_i) \times \ln(A_i/E_i)$
2. Calculate the total PSI
3. Interpret the result using standard thresholds ($<$0.10, 0.10–0.20, $>$0.20)
4. What action would you recommend?

Exercise 2: Drift Type Identification

For each scenario, identify the drift type (data drift, concept drift, or prediction drift) and explain your reasoning.

1. A mortgage model's average applicant income rises from \$65k to \$85k over 2 years. Default rates for each income bracket remain unchanged.
2. A fraud detection model maintains normal feature distributions, but the false negative rate doubles because criminals developed new attack vectors that mimic legitimate transactions.
3. A churn prediction model's average predicted churn probability rises from 12% to 22%, though no changes are visible in the input features.
4. Interest rates rise from 3% to 7%. Both loan application features and default behavior shift simultaneously.

Exercise 3: Monitoring Dashboard Design

You are designing a monitoring dashboard for a real-time fraud detection model that processes 5 million transactions per day at a payment processor.

Tasks:

1. List 3 **operational metrics** you would track and their alert thresholds
2. List 3 **model performance metrics** and their alert thresholds
3. List 2 **business impact metrics** and their alert thresholds
4. For each alert, specify the tier (CRITICAL / WARNING / INFO)
5. Describe one situation where a WARNING should automatically escalate to CRITICAL

Constraints: The model must respond in $<$50ms. Maximum acceptable fraud loss is \$500k/month. False positive rate must stay below 3%.

Exercise 4: SR 11-7 Governance Scenario

Scenario: A mid-size US bank wants to deploy a machine learning model for automated loan approval (replacing a scorecard-based system). The model uses XGBoost with 150 features.

Tasks:

1. According to SR 11-7, what risk tier should this model be classified as? Justify your answer.
2. List 5 specific items that the **model documentation** (model card) must include.
3. What should the **independent validation team** (2nd line) test before approving deployment?
4. The model uses a feature called "zip code." What governance concern does this raise, and how would you address it?
5. After deployment, what should the **annual review** process include?

Exercise 5: Champion-Challenger Analysis

A bank runs a 4-week champion-challenger test for its credit scoring model. Results:

Metric	Champion (v2.1)	Challenger (v3.0)
AUC-ROC	0.86	0.89
Accuracy	83%	85%
Default rate (approved loans)	4.1%	3.5%
Approval rate	67%	64%
Avg. latency	38ms	95ms
Latency SLA	$<$100ms	
Fairness (demographic parity ratio)	0.85	0.82
Fairness threshold	≥ 0.80	

Tasks:

1. Should you promote the challenger? Justify with specific metrics.
2. What is the business trade-off of the approval rate decrease?
3. What concerns does the latency increase raise, even though it meets SLA?
4. Propose a gradual rollout plan with specific traffic percentages and durations.

Exercise 6: Retraining Decision

Your credit scoring model shows the following monitoring data over 8 weeks:

Metric	W1	W2	W3	W4	W5	W6	W7	W8
AUC-ROC	0.89	0.89	0.88	0.87	0.86	0.85	0.84	0.83
Avg. PSI	0.04	0.05	0.07	0.09	0.12	0.15	0.19	0.23
Data quality	OK	OK	OK	OK	OK	OK	OK	OK

Tasks:

1. At which week should a WARNING alert have fired? Why?
2. At which week should a CRITICAL alert have fired? Why?
3. Is this data drift, concept drift, or both? Justify using the metrics.
4. Using the retraining decision tree, what action should the team take?
5. After retraining, how would you validate the new model before production deployment?

Exercise 7: Feature Store Design

Scenario: A bank uses a "30-day average transaction amount" feature in three models: fraud detection (real-time), credit scoring (batch), and customer segmentation (weekly batch).

Currently:

- Fraud team calculates it in a Redis Lua script
- Credit team calculates it in a SQL query against the data warehouse
- Segmentation team calculates it in a PySpark job
- Each team uses slightly different definitions (e.g., business days vs. calendar days)

Tasks:

1. Identify 3 specific problems with the current approach
2. Design a feature store solution: which store type (online/offline) does each model need?
3. How does point-in-time correctness prevent data leakage in the credit scoring model?
4. What happens if the fraud model needs the feature in $<$5ms but the credit model can tolerate 5-second latency?

Exercise 8: MLOps ROI Analysis

A fintech company runs 12 ML models in production. Current state (no MLOps):

- Manual deployment: 3 days per model update, 4 updates/year per model
- Data scientist time: $120/hour, 24 hours per manual deployment
- Undetected drift costs: $30k per incident, 8 incidents per year
- Time to detect drift: average 6 weeks (manual review)

Proposed MLOps platform:

- Setup cost: $150,000 (one-time)
- Annual platform cost: $60,000 (tooling + infrastructure)
- Reduces deployment to 4 hours (automated)
- Drift detection: $<$1 week, reduces incidents to 2/year
- Requires 1 MLOps engineer at $130k/year

Tasks:

1. Calculate the annual cost of the current manual approach
2. Calculate the Year 1 total cost of the MLOps approach (including setup)
3. Calculate the annual cost savings from Year 2 onward
4. At what month does the investment break even?