

## Lesson 5.4: MLOps and Production ML in Finance – Quiz

Module 5: Automation and Infrastructure

Digital Finance

## Question 1

**A credit scoring model trained on 2022 data shows  $PSI = 0.25$  in production (2025). Customer incomes have risen 15% due to inflation, but default behavior for a given income level has not changed. What type of drift is this?**

- A. Data drift (covariate shift) –  $P(X)$  changed while  $P(Y|X)$  is constant
- B. Concept drift – the relationship between income and default changed
- C. Prediction drift – the model's output distribution shifted
- D. Label drift – the target variable definition changed

## Question 1

**A credit scoring model trained on 2022 data shows  $PSI = 0.25$  in production (2025). Customer incomes have risen 15% due to inflation, but default behavior for a given income level has not changed. What type of drift is this?**

- A. Data drift (covariate shift) –  $P(X)$  changed while  $P(Y|X)$  is constant
- B. Concept drift – the relationship between income and default changed
- C. Prediction drift – the model's output distribution shifted
- D. Label drift – the target variable definition changed

*[Answer hidden – compile with `\solutionstrue` to reveal]*

Income distributions shifted ( $P(X)$  changed), but the relationship between income and default ( $P(Y|X)$ ) remains the same. This is classic covariate shift.

## Question 2

**During COVID-19, a bank's fraud detection model continued receiving normal transaction volumes, but fraud patterns shifted from card-present to card-not-present. What type of drift occurred?**

- A. Data drift only – transaction volumes didn't change
- B. No drift – the model should still work fine
- C. Concept drift – same features, different fraud/non-fraud relationship
- D. Prediction drift – the model's confidence scores shifted

## Question 2

**During COVID-19, a bank's fraud detection model continued receiving normal transaction volumes, but fraud patterns shifted from card-present to card-not-present. What type of drift occurred?**

- A. Data drift only – transaction volumes didn't change
- B. No drift – the model should still work fine
- C. Concept drift – same features, different fraud/non-fraud relationship
- D. Prediction drift – the model's confidence scores shifted

*[Answer hidden – compile with \solutionstrue to reveal]*

The relationship  $P(Y|X)$  changed: transactions with similar features now have different fraud probabilities because criminals adapted their tactics. This is concept drift.

## Question 3

A feature has the following bin distributions:

Bin	Expected	Actual
Low	0.30	0.25
Medium	0.50	0.45
High	0.20	0.30

Which PSI contribution is largest?

- A. All three bins contribute equally
- B. High bin:  $(0.30 - 0.20) \times \ln(0.30/0.20)$
- C. Medium bin:  $(0.45 - 0.50) \times \ln(0.45/0.50)$
- D. Low bin:  $(0.25 - 0.30) \times \ln(0.25/0.30)$

## Question 3

A feature has the following bin distributions:

Bin	Expected	Actual
Low	0.30	0.25
Medium	0.50	0.45
High	0.20	0.30

Which PSI contribution is largest?

- A. All three bins contribute equally
- B. High bin:  $(0.30 - 0.20) \times \ln(0.30/0.20)$
- C. Medium bin:  $(0.45 - 0.50) \times \ln(0.45/0.50)$
- D. Low bin:  $(0.25 - 0.30) \times \ln(0.25/0.30)$

*[Answer hidden – compile with \solutionstrue to reveal]*

High bin:  $(0.10) \times \ln(1.5) = 0.10 \times 0.405 = 0.0405$ . Low bin: 0.0093. Medium bin: 0.0056. The High bin has the largest absolute shift and contributes most to PSI.

## Question 4

**A bank's model monitoring dashboard shows  $PSI = 0.08$  across all features, but AUC-ROC dropped from 0.91 to 0.82. What is the most likely explanation?**

- A. Data drift – the PSI confirms significant distribution shifts
- B. The dashboard has a bug – PSI and AUC cannot diverge
- C. Concept drift – relationships changed without input distribution changes
- D. Prediction drift – the model outputs are unreliable

## Question 4

**A bank's model monitoring dashboard shows  $PSI = 0.08$  across all features, but AUC-ROC dropped from 0.91 to 0.82. What is the most likely explanation?**

- A. Data drift – the PSI confirms significant distribution shifts
- B. The dashboard has a bug – PSI and AUC cannot diverge
- C. Concept drift – relationships changed without input distribution changes
- D. Prediction drift – the model outputs are unreliable

*[Answer hidden – compile with \solutionstrue to reveal]*

PSI < 0.10 means input distributions are stable, yet performance dropped significantly. This means  $P(Y|X)$  changed while  $P(X)$  stayed the same – the hallmark of concept drift.

## Question 5

**Under SR 11-7, who should validate a model before production deployment?**

- A. The engineering team that will deploy it
- B. Qualified personnel independent of the development team
- C. The data scientist who built the model
- D. External regulators from the Federal Reserve

## Question 5

**Under SR 11-7, who should validate a model before production deployment?**

- A. The engineering team that will deploy it
- B. Qualified personnel independent of the development team
- C. The data scientist who built the model
- D. External regulators from the Federal Reserve

*[Answer hidden – compile with \solutionstrue to reveal]*

SR 11-7 mandates “effective challenge” by qualified staff who are independent of the model development process. This is the second line of defense.

## Question 6

**A feature store provides both an “online store” and an “offline store.” Why are both needed?**

- A. Online for internal users, offline for external customers
- B. Online for new features, offline for deprecated features
- C. Online for production predictions (low latency), offline for model training (high throughput)
- D. Online during business hours, offline at night

## Question 6

**A feature store provides both an “online store” and an “offline store.” Why are both needed?**

- A. Online for internal users, offline for external customers
- B. Online for new features, offline for deprecated features
- C. Online for production predictions (low latency), offline for model training (high throughput)
- D. Online during business hours, offline at night

*[Answer hidden – compile with \solutionstrue to reveal]*

The online store (e.g., Redis) serves features in <10ms for real-time inference. The offline store (e.g., S3) provides historical features for batch training with point-in-time correctness.

## Question 7

**In a champion-challenger test, the challenger model shows +3% AUC improvement but +40ms latency increase (from 50ms to 90ms). The SLA requires <100ms. What should you do?**

- A. Promote the challenger – it meets the SLA and has better accuracy
- B. Reject because champion-challenger cannot measure latency
- C. Reject the challenger – any latency increase is unacceptable
- D. Deploy as shadow model indefinitely

## Question 7

**In a champion-challenger test, the challenger model shows +3% AUC improvement but +40ms latency increase (from 50ms to 90ms). The SLA requires <100ms. What should you do?**

- A. Promote the challenger – it meets the SLA and has better accuracy
- B. Reject because champion-challenger cannot measure latency
- C. Reject the challenger – any latency increase is unacceptable
- D. Deploy as shadow model indefinitely

*[Answer hidden – compile with \solutionstrue to reveal]*

The challenger meets the <100ms SLA (90ms) and delivers meaningful accuracy improvement (+3% AUC). Latency increased but remains within acceptable bounds. Promote with monitoring.

## Question 8

**A team detects  $PSI = 0.35$  on the “transaction amount” feature. Before retraining, what should they check first?**

- A. Check data quality – is the shift real, or caused by a pipeline bug?
- B. Switch to a different algorithm
- C. Immediately retrain the model with new data
- D. Remove the drifted feature from the model

## Question 8

**A team detects  $PSI = 0.35$  on the “transaction amount” feature. Before retraining, what should they check first?**

- A. Check data quality – is the shift real, or caused by a pipeline bug?
- B. Switch to a different algorithm
- C. Immediately retrain the model with new data
- D. Remove the drifted feature from the model

*[Answer hidden – compile with \solutionstrue to reveal]*

The retraining decision tree starts with data quality validation. A PSI spike could be caused by an upstream ETL bug, a schema change, or a data source outage – not real drift.

## Question 9

**What is “training-serving skew” and why is it dangerous?**

- A. When feature computation logic differs between training and production, causing silent accuracy loss
- B. When the model trains faster than it can serve predictions
- C. When training data is biased toward certain demographics
- D. When training accuracy is higher than test accuracy (overfitting)

## Question 9

**What is “training-serving skew” and why is it dangerous?**

- A. When feature computation logic differs between training and production, causing silent accuracy loss
- B. When the model trains faster than it can serve predictions
- C. When training data is biased toward certain demographics
- D. When training accuracy is higher than test accuracy (overfitting)

*[Answer hidden – compile with \solutionstrue to reveal]*

Training-serving skew occurs when features are computed differently (e.g., Python in training, SQL in production). The model sees different inputs than expected, degrading accuracy silently.

## Question 10

**Which deployment strategy runs the new model in parallel but does NOT serve its predictions to users?**

- A. A/B testing
- B. Canary deployment
- C. Blue-green deployment
- D. Shadow deployment

## Question 10

**Which deployment strategy runs the new model in parallel but does NOT serve its predictions to users?**

- A. A/B testing
- B. Canary deployment
- C. Blue-green deployment
- D. Shadow deployment

*[Answer hidden – compile with \solutionstrue to reveal]*

Shadow deployment runs the new model alongside production, logging predictions for offline comparison without affecting any user-facing decisions.

## Question 11

**A bank has 500+ models in production. The CRO asks: “How many models have not been validated in the past 12 months?” What system answers this question?**

- A. Feature store
- B. Data warehouse
- C. Experiment tracker
- D. Model inventory / registry

## Question 11

**A bank has 500+ models in production. The CRO asks: “How many models have not been validated in the past 12 months?” What system answers this question?**

- A. Feature store
- B. Data warehouse
- C. Experiment tracker
- D. Model inventory / registry

*[Answer hidden – compile with \solutionstrue to reveal]*

The model inventory tracks every model's status, risk tier, last validation date, and next review date. This is a core SR 11-7 requirement for governance.

## Question 12

**Which of the following is a valid trigger for automated model retraining?**

- A. A competitor launches a new product
- B. The model has been in production for exactly 6 months
- C. The CEO requests it during a board meeting
- D. PSI exceeds 0.20 on three or more features simultaneously

## Question 12

**Which of the following is a valid trigger for automated model retraining?**

- A. A competitor launches a new product
- B. The model has been in production for exactly 6 months
- C. The CEO requests it during a board meeting
- D. PSI exceeds 0.20 on three or more features simultaneously

*[Answer hidden – compile with \solutionstrue to reveal]*

Automated retraining triggers should be data-driven: drift metrics (PSI > 0.20), performance drops (AUC decline), or business KPI breaches. Calendar-based triggers are scheduled, not event-driven.

## Question 13

**In the “three lines of defense” governance framework, what is the role of the second line?**

- A. Audit the entire process annually
- B. Independently validate the model and challenge its assumptions
- C. Deploy the model to production
- D. Build and train the model

## Question 13

**In the “three lines of defense” governance framework, what is the role of the second line?**

- A. Audit the entire process annually
- B. Independently validate the model and challenge its assumptions
- C. Deploy the model to production
- D. Build and train the model

*[Answer hidden – compile with \solutionstrue to reveal]*

The second line (Model Validation) conducts independent review, challenges assumptions, validates performance, and approves or rejects models. The first line develops, the third line audits.

## Question 14

**A fraud detection model's predicted fraud rate suddenly jumps from 2% to 8%. What should the team investigate first?**

- A. Increase the prediction threshold to reduce flagged transactions
- B. Assume fraud has actually increased and alert law enforcement
- C. Immediately retrain the model
- D. Check the data pipeline for upstream failures or schema changes

## Question 14

**A fraud detection model's predicted fraud rate suddenly jumps from 2% to 8%. What should the team investigate first?**

- A. Increase the prediction threshold to reduce flagged transactions
- B. Assume fraud has actually increased and alert law enforcement
- C. Immediately retrain the model
- D. Check the data pipeline for upstream failures or schema changes

*[Answer hidden – compile with \solutionstrue to reveal]*

Prediction drift is a symptom, not a diagnosis. The first step is to rule out data quality issues (pipeline failure, missing features, schema change) before assuming real-world changes.

**What does CI/CD/CT stand for in the MLOps context?**

- A. Continuous Integration, Continuous Delivery, Continuous Training
- B. Code Integration, Code Deployment, Code Testing
- C. Continuous Inference, Continuous Data, Continuous Tuning
- D. Continuous Improvement, Continuous Development, Continuous Testing

**What does CI/CD/CT stand for in the MLOps context?**

- A. Continuous Integration, Continuous Delivery, Continuous Training
- B. Code Integration, Code Deployment, Code Testing
- C. Continuous Inference, Continuous Data, Continuous Tuning
- D. Continuous Improvement, Continuous Development, Continuous Testing

*[Answer hidden – compile with \solutionstrue to reveal]*

CI validates code and data automatically, CD automates model packaging and deployment, and CT adds automated model retraining triggered by drift or performance degradation.

## Question 16

**A loan default model has a “label delay” of 12 months. How can you monitor for drift before ground truth labels arrive?**

- A. Use the model's own confidence scores as ground truth
- B. Manually review every prediction
- C. Wait 12 months – there is no way to detect problems earlier
- D. Use proxy metrics: PSI on features, prediction distribution shift, data quality checks

## Question 16

**A loan default model has a “label delay” of 12 months. How can you monitor for drift before ground truth labels arrive?**

- A. Use the model's own confidence scores as ground truth
- B. Manually review every prediction
- C. Wait 12 months – there is no way to detect problems earlier
- D. Use proxy metrics: PSI on features, prediction distribution shift, data quality checks

*[Answer hidden – compile with \solutionstrue to reveal]*

When labels are delayed, proxy metrics (feature PSI, prediction drift, data quality) serve as early warning signals until ground truth becomes available.

## Question 17

**What is the main risk of a “feedback loop” in ML systems?**

- A. The model becomes too accurate over time
- B. Model predictions influence future training data, creating self-reinforcing bias
- C. The model runs out of training data
- D. Feedback loops only affect supervised learning

## Question 17

**What is the main risk of a “feedback loop” in ML systems?**

- A. The model becomes too accurate over time
- B. Model predictions influence future training data, creating self-reinforcing bias
- C. The model runs out of training data
- D. Feedback loops only affect supervised learning

*[Answer hidden – compile with \solutionstrue to reveal]*

Example: A fraud model flags transactions for review. Only flagged transactions get investigated (and labeled). The model never learns from false negatives it missed, creating biased training data.

## Question 18

**A monitoring alert fires at 2:00 AM: “Model accuracy dropped 1.5% over the past hour.” This is classified as a WARNING, not CRITICAL. Why?**

- A. Because it happened at night when no one is working
- B. Because monitoring alerts are never critical
- C. Because the drop is within the WARNING threshold (2–5%) and not yet CRITICAL (>5%)
- D. Because accuracy is not an important metric

## Question 18

**A monitoring alert fires at 2:00 AM: “Model accuracy dropped 1.5% over the past hour.” This is classified as a WARNING, not CRITICAL. Why?**

- A. Because it happened at night when no one is working
- B. Because monitoring alerts are never critical
- C. Because the drop is within the WARNING threshold (2–5%) and not yet CRITICAL (>5%)
- D. Because accuracy is not an important metric

*[Answer hidden – compile with \solutionstrue to reveal]*

Alert tiers are defined by threshold severity. A 1.5% drop is concerning (WARNING → Slack/email) but not yet an emergency (CRITICAL → PagerDuty/phone). This prevents alert fatigue.

## Question 19

**A model has experienced gradual concept drift over 6 months. Which retraining strategy is most appropriate?**

- A. Roll back to the original model
- B. Rebuild the model from scratch with a new algorithm
- C. Retrain with recent data and consider adding new features
- D. Ignore it – gradual drift is not actionable

## Question 19

**A model has experienced gradual concept drift over 6 months. Which retraining strategy is most appropriate?**

- A. Roll back to the original model
- B. Rebuild the model from scratch with a new algorithm
- C. Retrain with recent data and consider adding new features
- D. Ignore it – gradual drift is not actionable

*[Answer hidden – compile with \solutionstrue to reveal]*

Gradual concept drift means relationships are slowly evolving. Retraining on recent data with potential new features captures the evolved patterns. Rebuilding from scratch is for sudden, severe drift.

**Which statement best captures the “90/10 rule” of production ML?**

- A. 90% of models pass validation on the first attempt
- B. Building a model is 10% of the work; deploying, monitoring, and governing it is 90%
- C. 90% of drift is data drift and 10% is concept drift
- D. 90% of models use 10% of available data

**Which statement best captures the “90/10 rule” of production ML?**

- A. 90% of models pass validation on the first attempt
- B. Building a model is 10% of the work; deploying, monitoring, and governing it is 90%
- C. 90% of drift is data drift and 10% is concept drift
- D. 90% of models use 10% of available data

*[Answer hidden – compile with \solutionstrue to reveal]*

The core lesson of MLOps: the model is the easy part. Production infrastructure, monitoring, governance, retraining, and compliance constitute the vast majority of effort and cost.