PUBLICATION

EXAM-TM40M-POST

POST-TEST — SL 4M: ML ENGINEER

Maven Smart System (MSS) — USAREUR-AF

HEADQUARTERS
UNITED STATES ARMY EUROPE AND AFRICA
(USAREUR-AF)
Wiesbaden, Germany

26 MARCH 2026

DRAFT — UNOFFICIAL — NOT FOR OPERATIONAL USE

POST-TEST — SL 4M: ML ENGINEER

MAVEN SMART SYSTEM (MSS) — USAREUR-AF

Field	Detail
Course	SL 4M: ML Engineer
Form	Post-Test
Level	SL 4M (Specialist)
Audience	ML engineers / data scientists; prerequisite: SL 1+20+30 + Python + statistics
Time Allowed	45 minutes
Passing Score	70% (46/66)

INSTRUCTIONS

This assessment evaluates mastery of course learning objectives. A passing score of 70% is required to receive credit. Complete independently without reference to training materials.

SECTION 1 — MULTIPLE CHOICE

Circle the letter of the best answer. (2 points each)

1. When configuring a GPU-enabled Code Workspace for ML development in Foundry, environment reproducibility requires:

A. Defining a requirements file (or equivalent) that pins package versions, enabling identical environment reconstruction B. Installing all packages interactively via the terminal before each session C. Using only pre-installed Foundry-approved packages without adding new dependencies D. Reusing the same workspace session across multiple projects to maintain environment consistency

2. A feature engineering step that calculates "days since last maintenance" from a timestamp field is BEST described as:

A. A categorical encoding step B. A null imputation step C. A time-based derived feature D. A feature scaling step

3. Your pipeline computes "inspection_pass_rate" using data from the same week as the target label "failed_next_inspection." This is a feature leakage issue because:

A. The pass rate uses a rolling window that is too long B. The feature uses a different aggregation window than other features in the pipeline C. The pass rate is highly correlated with the target, which violates independence assumptions D. The inspection data from the same week as the label contains information that would not be available at prediction time

4. The USAREUR-AF minimum acceptance threshold for recall on a vehicle failure prediction model is:

A. 60% — sufficient for most operational use cases B. 75% — established threshold for safety-relevant failure prediction C. 70% — consistent with general passing standards D. 85% — required for all binary classification models

5. "Model calibration" refers to:

A. Adjusting the model's decision threshold for a specific use case B. Retraining the model on recent data to account for concept drift C. Ensuring the model's predicted probabilities reflect the true frequency of outcomes at each probability level D. Balancing precision and recall by selecting the optimal classification threshold

6. Platt scaling and isotonic regression are both techniques used to:

A. Calibrate a model's output probabilities to better reflect true likelihoods B. Reduce model complexity and prevent overfitting C. Increase the speed of model inference in production D. Handle class imbalance during model training

7. A model drift monitoring pipeline detects that the PSI (Population Stability Index) for the `mileage_since_service` feature has exceeded 0.25. This indicates:

A. A significant distribution shift — the model should be flagged for retraining review B. A moderate shift that warrants investigation — likely a data collection or operational change C. A minor shift in the feature distribution — continue monitoring D. A pipeline error — PSI above 0.2 always indicates a data ingestion failure

8. Per SL 4M, the Foundry model registry is used to:

A. Store the model's Python source code in version control B. Monitor model output quality in real-time after deployment C. Deploy models directly to production serving endpoints D. Track trained model artifacts, their versions, performance metrics, and deployment status

9. To connect a deployed model serving endpoint to an Ontology Object so predictions are available as Object properties, the correct MSS architecture is:

A. Build a Workshop widget that calls the endpoint on demand and displays results without writing to the Ontology B. Schedule a pipeline that calls the endpoint and writes results back to the Ontology via a write transaction C. Use an AIP Logic workflow to manually copy predictions into the Ontology D. Deploy the model as a Foundry function that runs at query time rather than as a stored prediction

10. Your G4 asks for a binary classifier that predicts whether a vehicle will fail its next inspection. The dataset has 5% failure rate. Which metric is MOST important for your performance report?

A. Overall accuracy — high accuracy demonstrates model quality B. Recall for the failure class — missing actual failures (false negatives) is more operationally costly than false alarms C. Precision for the failure class — false alarms create unnecessary maintenance burden D. AUC-ROC — the only metric suitable for imbalanced datasets

11. A model card required under USAREUR-AF documentation standards must include which of the following fields?

A. Model source code and hyperparameter grid search results B. The full feature importance table and SHAP value distributions C. A deployment approval signature from the MSS program office D. Intended use, out-of-scope uses, training data description, performance metrics, known limitations, and responsible AI declaration

12. You have trained and registered a new version of a predictive maintenance model. Before deploying to production, you must verify the serving endpoint with:

A. A single test prediction to confirm the endpoint is reachable B. Live inference on a held-out set of test records, validating prediction format, latency, and accuracy against acceptance thresholds C. A G4 NCOIC sign-off indicating operational readiness D. A manual review of the model's weights to confirm they updated correctly

13. When implementing drift detection in a model monitoring pipeline, the Kolmogorov-Smirnov (KS) test is used to:

A. Detect concept drift in the model's output predictions B. Test whether feature importance rankings have shifted between model versions C. Compare the empirical distribution of a continuous feature between a reference period and a monitoring period D. Validate that the model's calibration curve has not degraded

14. Per SL 4M, a write transaction from Code Workspace to a Foundry dataset is the preferred method for committing model outputs because:

A. Write transactions bypass the Foundry access control layer, enabling faster writes B. Write transactions automatically trigger the downstream pipeline to refresh C. Transactions are atomic — either all records commit or none, preventing partial writes that could corrupt downstream pipelines D. Transactions encrypt the output data before writing to the dataset

15. A compressed (quantized) version of a model shows 2% lower accuracy than the full-precision version. Per SL 4M standards, this means:

A. The compressed model must be evaluated against all acceptance thresholds independently — compression changes the model and prior approval does not transfer B. The compressed model can be deployed — 2% is within acceptable degradation tolerance C. The model must be recompressed at a lower quantization level D. The deployment must wait for a full retraining on the quantized architecture

16. Per DDOF Playbook v2.2, Phase 4 (Development) and Phase 5 (Test & Evaluation) map to the ML lifecycle. An MLE who cannot complete model training and validation within the 30-day MVP mandate should:

A. Deploy the model without Phase 5 T&E to meet the deadline and complete validation post-deployment B. Escalate to C2DAO for scope reduction before cutting T&E — a model without holdout validation and sponsor sign-off is an unauthorized deployment C. Request a 30-day extension from the mission owner, which is automatically approved for ML projects D. Switch to a pre-trained model from the model registry and skip Phases 4 and 5 entirely

17. Per UDRA v1.1 and DDOF Playbook v2.2, a trained ML model is classified as a data product and must meet VAULTIS-A compliance. Which of the following correctly describes the "Linked" dimension as applied to an ML model?

A. The model is linked to at least one downstream Workshop application that consumes its output B. Full lineage is documented from training data through the feature pipeline through the model to inference outputs C. The model's predictions are linked to ground truth labels for continuous retraining D. The model artifact is linked to the model registry via a persistent URL

18. Per SL 4M Section 9-2a, DDOF Phase 4 (Development) maps to specific ML activities. The gate output required to exit Phase 4 is:

A. A test report with accuracy, precision, and recall metrics plus sponsor sign-off B. A functional model with documented architecture C. A model card with all required sections populated and peer-reviewed D. A deployed serving endpoint with latency benchmarks

19. Per SL 4M Section 9-2b, VAULTIS-A compliance for ML models requires that the "Auditable" dimension be satisfied. For a deployed ML model, "Auditable" specifically means:

A. The model's decision logic is fully explainable to non-technical stakeholders using SHAP values B. Full training and inference logs are retained and version history is maintained in the model registry C. An external auditor has certified the model's fairness metrics before deployment D. The model's hyperparameters are documented in the model card

SECTION 2 — SHORT ANSWER

Answer in 2–5 sentences. (6 points each)

SA-1. Describe the complete feature engineering pipeline you would build for a vehicle failure prediction model. Include at least four specific feature types, the leakage check you would apply, and the scaling/encoding steps required.

SA-2. Your failure prediction model achieves 92% accuracy but only 44% recall for the failure class. The G4 asks why it is not ready for deployment. Explain the problem and describe what changes to the model training procedure would address it.

SA-3. Describe the model monitoring pipeline you would build for a deployed vehicle failure prediction model. Include the specific drift metrics you would track, the alert thresholds, and the escalation procedure when drift is detected.

SA-4. Walk through the complete model governance document (model card) sections required by USAREUR-AF for a deployed predictive maintenance model. For each section, state what information it must contain.

SA-5. After deploying your failure prediction model, you connect it to the Vehicle Ontology Object Type so that each vehicle has a predicted failure risk score as a property. Describe the architecture: how does the model output get written to the Ontology, how does it update, and how would an analyst access the prediction in Workshop?

SA-6. Describe how ML engineering capability supports two WFF functions. For each, identify the WFF track (SL 4A through SL 4F) and give a concrete example of an ML model or pipeline that supports decision-making in that function.

SCORING SUMMARY

Section	Questions	Points Each	Total Points
Multiple Choice	19	2	38
Short Answer	6	6	36
Total	—	—	**74**

Passing: 52/74 (70%) — Post-test only. Pre-test is diagnostic.

ANSWER KEY — INSTRUCTOR USE ONLY

Do not distribute to students.

Multiple Choice: 1. A — Pinned requirements file enables reproducible environment reconstruction. 2. C — Days-since-timestamp is a time-based derived feature. 3. D — Data from the same period as the label is not available at prediction time — leakage. 4. B — USAREUR-AF minimum recall threshold for failure prediction = 75%. 5. C — Calibration ensures predicted probabilities match observed frequencies. 6. A — Platt scaling and isotonic regression both calibrate probability outputs. 7. A — PSI > 0.20 = significant distribution shift; PSI 0.10-0.20 = moderate; <0.10 = stable. 8. D — Model registry tracks artifacts, versions, metrics, and deployment status. 9. B — Scheduled pipeline calls endpoint and writes predictions to Ontology via write transaction. 10. B — Recall for the failure class is most important — missing failures is the critical error type. 11. D — Model card required fields: intended use, out-of-scope use, training data, metrics, limitations, responsible AI declaration. 12. B — Live inference on held-out test

records validates format, latency, and accuracy. 13. C — KS test compares empirical distributions of continuous features between periods. 14. C — Transactions are atomic — all-or-nothing prevents partial write corruption. 15. A — Compressed model must be independently evaluated — compression changes the model. 16. B — Escalate to C2DAO for scope reduction before cutting T&E — a model without holdout validation and sponsor sign-off is an unauthorized deployment. 17. B — Full lineage documented from training data through feature pipeline through model to inference outputs satisfies the "Linked" dimension. 18. B — A functional model with documented architecture is the Phase 4 (Development) gate output per SL 4M Section 9-2a. 19. B — Full training and inference logs retained and version history maintained in the model registry satisfies the "Auditable" dimension per SL 4M Section 9-2b.

Short Answer Guidance:

SA-1. Full credit: feature types — time-based (days since last service, age of vehicle), aggregation (failure count in last 12 months, avg days between failures), categorical (vehicle type, unit assignment encoded), numeric (mileage, engine hours); leakage check: ensure all features use data available before the prediction date (no same-period labels); scaling: standardize numeric features; encoding: one-hot or ordinal for categorical; null handling: impute or flag nulls explicitly. Must include 4 feature types + leakage check + at least one preprocessing step.

SA-2. Full credit: 92% accuracy is misleading with 5% failure rate — a model predicting all "pass" achieves ~95% accuracy; recall = 44% means 56% of actual failures are missed — this is operationally unacceptable; solutions: set class_weight='balanced' in sklearn, use SMOTE oversampling, lower the decision threshold to increase recall, or use a cost-sensitive loss function; re-evaluate against the ≥75% recall acceptance threshold. Full credit requires explaining the accuracy paradox AND providing corrective steps with the acceptance threshold.

SA-3. Full credit: drift metrics — PSI for continuous features (threshold: flag >0.10, alert >0.20); KS test p-values for continuous distributions; prediction score distribution shift over time; actual vs. predicted failure rate when ground truth is available; alerts: automated email to ML engineer when PSI exceeds threshold; escalation: investigate data source changes → if data source is stable, flag model for retraining review → retrain and validate against acceptance thresholds before redeployment. Must include specific thresholds and escalation path.

SA-4. Full credit sections and required content: (1) Model Overview — name, version, date, purpose; (2) Intended Use — specific use case, authorized users; (3) Out-of-Scope Use — use cases the model must not be applied to; (4) Training Data — source, date range, known biases or gaps; (5) Performance Metrics — recall, precision, AUC on held-out test set; (6) Known Limitations — what the model cannot predict, failure modes; (7) Assumptions — stationarity of operational data, applicability conditions; (8) Responsible AI Declaration — bias evaluation, fairness considerations; (9) Maintenance — retraining schedule, drift monitoring setup. Partial credit (3 pts) for five or more sections correctly described.

SA-5. Full credit: architecture — scheduled pipeline runs model (via API call to serving endpoint or batch inference); pipeline writes predicted failure risk scores to Foundry output dataset via write transaction; Ontology write step maps `vehicle_id` to the `predicted_failure_risk` property on the Vehicle Object

Type (upsert on primary key); update frequency = pipeline schedule (daily, weekly); analyst access in Workshop — add a column for `predicted_failure_risk` to the vehicle table widget; configure conditional formatting to highlight high-risk vehicles; or build a sorted list filtered to vehicles above risk threshold. Full credit requires write transaction, Ontology property connection, and Workshop access method.

SA-6. Full credit: any two WFF tracks correctly identified with an ML example — SL 4A (Intelligence): ML classifier categorizes incoming reports by topic and threat level to assist intel analysts; SL 4B (Fires): ML model predicts target re-acquisition probability to support fires prioritization; SL 4C (Movement & Maneuver): ML model predicts route traversability from sensor data to support maneuver planning; SL 4D (Sustainment): ML failure prediction model forecasts which vehicles require maintenance before the next mission to support G4 readiness management; SL 4E (Protection): ML anomaly detection model flags unusual patterns in force protection sensor data; SL 4F (Mission Command): ML model forecasts operational tempo changes to support commander decision cycle. Each response must identify the correct SL 4 letter (A–F) and provide a concrete ML model or pipeline example for full credit.

USAREUR-AF Operational Data Team TM-40M Post-Test | Version 1.0 | March 2026