PUBLICATION

EXAM-TM40M-PRE

PRE-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

PRE-TEST — SL 4M: ML ENGINEER

MAVEN SMART SYSTEM (MSS) — USAREUR-AF

Field	Detail
Course	SL 4M: ML Engineer
Form	Pre-Test
Level	SL 4M (Specialist)
Audience	ML engineers / data scientists; prerequisite: SL 1+20+30 + Python + statistics
Time Allowed	30 minutes
Passing Score	N/A — diagnostic only

INSTRUCTIONS

This diagnostic assessment establishes your baseline knowledge before training. Your score does not affect course eligibility. Answer honestly — results help the instructor tailor instruction to gaps.

SECTION 1 — MULTIPLE CHOICE

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

1. In a supervised machine learning problem, the "training set" is used to:

A. Evaluate the final model's performance before deployment B. Simulate model behavior under production conditions C. Tune hyperparameters without overfitting to the test set D. Learn the model parameters by fitting to labeled examples

2. "Feature leakage" in a machine learning pipeline occurs when:

A. A feature column contains too many null values to be useful B. The test set is inadvertently included in the training data during cross-validation C. Model weights from the training phase are exposed to end users D. Information from the future (relative to the prediction target) or from the target itself is inadvertently included in the training features

3. In a classification problem, "precision" measures:

A. The fraction of actual positive cases that the model correctly identifies B. The model's performance relative to a random baseline C. The overall fraction of all cases the model classifies correctly D. The fraction of predicted positive cases that are truly positive

4. "Recall" (also called sensitivity) measures:

A. The fraction of predicted positive cases that are truly positive B. The total accuracy of the model across all classes C. The fraction of actual positive cases correctly identified by the model D. The model's false positive rate

5. A ROC-AUC score of 0.50 indicates:

A. A model performing at chance (no better than random guessing) B. A perfect classifier C. A model with 50% accuracy D. A severely overfit model

6. "Cross-validation" in machine learning is primarily used to:

A. Validate that the training and test sets have the same class distribution B. Estimate the model's generalization performance using multiple train/test splits of the data C. Cross-check the model's feature importances against domain knowledge D. Validate that the model's predictions are within expected bounds

7. "Hyperparameter tuning" refers to:

A. Selecting the best values for configuration parameters (learning rate, tree depth, etc.) that are set before training B. Adjusting the model's learned weights during training C. Retraining the model on new data after deployment D. Reducing the number of features to improve model interpretability

8. A model trained on 2022 data is deployed in 2025 to predict equipment failures. Performance has degraded. The most likely cause is:

A. Data drift — the statistical distribution of the input features has changed since training B. The model's hyperparameters are no longer optimal C. The model was overfit to the 2022 test set D. The model's feature importances have been altered by the deployment environment

9. "Feature scaling" (e.g., standardization or normalization) is important for which type of model?

A. Decision trees and random forests — they are sensitive to feature scale B. K-nearest neighbors and support vector machines — they use distance metrics that are scale-dependent C. Linear regression only — other models are scale-invariant D. All models require feature scaling regardless of type

10. A model card is a document that:

A. Summarizes a model's intended use, performance metrics, limitations, training data, and ethical considerations B. Describes the hardware configuration required to train the model C. Contains the model's serialized weights for portability D. Lists all API endpoints exposed by the model serving infrastructure

11. "Data poisoning" in ML security refers to:

A. Using low-quality or biased training data that degrades model performance B. Exposing the training data to unauthorized users C. Adversarially injecting malicious training examples to manipulate model behavior D. Retraining a model on data that has already been used for evaluation

12. In an imbalanced classification dataset (e.g., 95% negative, 5% positive), a model that always predicts "negative" will achieve:

A. 5% accuracy B. 50% accuracy by chance C. 95% accuracy but 0% recall for the positive class D. A ROC-AUC of 0.95

13. "Platt scaling" is a technique used to:

A. Reduce model training time by scaling the feature matrix B. Scale the model's gradient updates during training C. Calibrate a classifier's output probability scores so they reflect true probabilities D. Normalize the model's output predictions to a [0,1] range

14. "Model versioning" in an ML deployment pipeline ensures:

A. That the same model version is always used regardless of retraining B. That each trained model artifact is uniquely identified and retrievable, enabling rollback to prior versions C. That model versions are synchronized across all user workspaces D. That model weights are compressed before storage

15. The "Population Stability Index" (PSI) is a metric used to:

A. Measure class imbalance in a training dataset B. Detect distribution shift in input features between a reference period and a monitoring period C. Evaluate whether a model's calibration is stable over time D. Measure the correlation between feature importance rankings across model versions

SECTION 2 — SHORT ANSWER

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

SA-1. Explain the difference between a model's training accuracy and its test accuracy. When you observe a large gap between the two, what does this indicate, and what corrective steps would you take?

SA-2. Describe what "feature engineering" means in the context of a predictive maintenance model. Give two examples of features you might engineer from raw maintenance log data that would be useful for predicting equipment failure.

SA-3. You are building a model to predict which vehicles will fail their next maintenance inspection. The dataset has 4,000 pass records and 200 fail records. Describe how class imbalance will affect model training and what techniques you would use to address it.

SA-4. Explain what a "model serving endpoint" is and describe the high-level steps required to deploy a trained scikit-learn model to a serving endpoint in a production environment.

SA-5. Describe two ML governance requirements that should be documented in a model card before an ML model is deployed for operational use in an Army data system.

SCORING SUMMARY

Section	Questions	Points Each	Total Points
Multiple Choice	15	2	30
Short Answer	5	6	30
Total	—	—	**60**

Passing: N/A — Pre-test is diagnostic only.

ANSWER KEY — INSTRUCTOR USE ONLY

Do not distribute to students.

Multiple Choice: 1. D — Training set is used to learn model parameters. 2. D — Feature leakage = future or target-derived information leaks into training features. 3. D — Precision = TP / (TP + FP) — fraction of predicted positives that are truly positive. 4. C — Recall = TP / (TP + FN) — fraction of actual positives correctly identified. 5. A — AUC = 0.50 = random chance performance. 6. B — Cross-validation estimates generalization performance via multiple train/test splits. 7. A — Hyperparameter tuning selects configuration parameters set before training. 8. A — Data drift is the most common cause of post-deployment performance degradation. 9. B — KNN and SVM use distance metrics that are scale-dependent; decision trees are not. 10. A — Model card summarizes intended use, performance, limitations, training data, and ethics. 11. C — Data poisoning = adversarial injection of malicious training examples. 12. C — Always-negative classifier achieves 95% accuracy but 0% recall for the positive class. 13. C — Platt scaling calibrates probability outputs to reflect true probabilities. 14. B — Model versioning uniquely identifies artifacts and enables rollback. 15. B — PSI detects distribution shift between reference and monitoring periods.

Short Answer Guidance:

SA-1. Full credit: training accuracy = performance on data the model was trained on (can be artificially high due to memorization); test accuracy = performance on held-out data the model has not seen; large gap = overfitting; corrective steps: regularization (L1/L2/dropout), reduce model complexity, increase training data, or use ensemble methods. Partial credit (3 pts) for correct explanation without corrective steps.

SA-2. Full credit: feature engineering creates new predictive variables from raw data; maintenance log examples: days since last service (time-based feature), cumulative failure count (aggregation over history), failure rate over last 90 days (rolling window), mileage since last repair (computed from mileage

readings). Any two valid engineered features with explanation for full credit.

SA-3. Full credit: class imbalance causes the model to optimize for the majority class (pass), effectively ignoring the minority class (fail) — produces high accuracy but near-zero recall for failures; techniques: oversample minority class (SMOTE), undersample majority class, use class-weight parameter in sklearn (class_weight='balanced'), or use precision-recall AUC instead of accuracy as the primary metric. Partial credit (3 pts) for identifying the problem without techniques.

SA-4. Full credit: a serving endpoint is an API that accepts feature inputs and returns model predictions in real time; deployment steps: train and serialize the model (pickle/joblib); register the model in a model registry; configure the serving endpoint with model artifact and input/output schema; deploy and validate with test requests; monitor for performance degradation. Partial credit (3 pts) for correct concept without deployment steps.

SA-5. Full credit: any two from — intended use and out-of-scope use cases; training data description (source, date range, known limitations); performance metrics on validation data; known failure modes or biases; responsible AI declaration; assumptions and limitations; maintenance and retraining plan. Each item must include a brief explanation of why it is required for full credit.

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