CERTIFICATION BOARD

# Certified Machine Learning Professional (CMLP)

## Ten Sample Questions

These ten questions are written at the difficulty of the live exam. Six carry a figure, a data table, a configuration or a code listing; four are plain scenario questions, as on the live paper. Every one is a scenario a working machine learning engineer would meet, and none of them appears in the live question bank.

The live exam is multistage adaptive: how you score on one stage selects the difficulty of the next, so no two candidates sit the same paper. It is a performance-based exam: each form can carry up to 5 to 12 performance-based questions alongside the multiple-choice questions, which means you build, sequence or complete something rather than only choosing from a list. Answers and full explanations follow each question here; the live exam does not disclose item-level answers.

### Question 1 — Domain 3: Algorithms and Model Selection — DATA

Emma Wilson matches problems to a fitting model.

| Problem | Fitting model |
|---|---|
| Interpretable classification baseline | Logistic regression |
| Strong accuracy on tabular data | Gradient boosting |
| Images with plenty of data | Neural network |
| Group unlabelled customers | k-means clustering |

Table 1.1 Problems and models.

Which statement about these choices is correct?

A. A neural network is always the best choice, so it should replace logistic regression, gradient boosting and k-means in every one of these cases.

B. Model choice should follow the data and needs: linear for interpretable baselines, boosting for tabular, neural nets for rich data, clustering for unlabelled grouping.

C. k-means should be used for the interpretable classification baseline, because clustering is the most explainable of all the machine learning methods.

D. Gradient boosting is only for images, so it should be swapped with the neural network, which is better suited to structured tabular data instead.

**Correct answer: B**
There is no single best algorithm; the right choice follows the data type and needs. Logistic regression is a good interpretable baseline, gradient boosting excels on tabular data, neural networks suit rich data like images with enough data, and k-means groups unlabelled data. The other options misapply these.

## Question 2

Domain 2: Data Preparation and Feature Engineering

CONFIG

James Carter reviews a churn model that looks excellent offline but fails live.

```
feature: "account_closed_date" is used as an input
note   : this field is only set after a customer churns
result : near-perfect offline accuracy, poor in production
```

Listing 2.1 A suspicious feature.

What is the problem, and what should be done?

A. The model is simply too small, so training a larger model on the same features will make the offline accuracy hold up in production too.

B. Data leakage: the feature is only known after churn happens, so it must be removed; features must use only information available at prediction time.

C. The offline metric was miscalculated, so recomputing the offline accuracy to a lower number will make the model perform correctly when it is deployed.

D. The production data is wrong, so the fix is to add the account_closed_date field to the live data feed so the model can use it at prediction time.

**Correct answer: B**
The account-closed date is only set after a customer churns, so using it as an input is data leakage that inflates offline accuracy and fails live where the field is absent. It must be removed; features may use only information available at prediction time. A bigger model, recomputed metrics, or trying to feed post-event data live do not fix leakage.

## Question 3

Domain 4: Training, Tuning and Evaluation

DATA

Olivia Davis evaluates a fraud classifier where fraud is 1% of cases.

| Metric | Value |
|---|---|
| Overall accuracy | 99% |
| Recall (fraud caught) | 8% |
| Behaviour | Predicts "not fraud" almost always |
| Cost of a missed fraud | Very high |

Table 3.1 Classifier results.

Why is 99% accuracy misleading here, and what should guide evaluation?

A. It is not misleading at all; the 99% accuracy figure shows the fraud classifier is performing excellently and no further metric is needed.

B. Class imbalance makes accuracy deceptive; predicting "not fraud" scores 99% while missing most fraud, so use recall and precision-recall here.

C. The accuracy is too low, so training the classifier for more epochs until accuracy reaches 100% is the correct way to make it useful for fraud.

D. The metric is fine but the data is too small, so simply collecting more non-fraud examples will make the 99% accuracy figure meaningful for fraud.

**Correct answer: B**
With fraud at 1%, a model that almost always predicts "not fraud" scores 99% accuracy while catching only 8% of fraud, so accuracy is deceptive under imbalance. Given the high cost of a missed fraud, recall and precision-recall should guide evaluation and threshold choice. More epochs or more non-fraud data does not fix the wrong metric.

## Question 4

Domain 4: Training, Tuning and Evaluation

CONFIG

Liam Anderson describes how a colleague evaluated a model.

```
process:
  1. train the model
  2. check the test set
  3. tweak settings, check the test set again
  4. repeat many times until the test score looks great
```

Listing 4.1 The evaluation process used.

Why is the reported test score untrustworthy, and what is the fix?

A. It is trustworthy, because repeatedly checking the test set and improving the score is the standard way to tune a machine learning model.

B. The test set was effectively used for tuning, so it no longer estimates unseen performance; tune on a validation set and use the test set once at the end.

C. The only problem is too few iterations, so tuning against the test set even more times would eventually produce a trustworthy performance estimate.

D. The score is untrustworthy only because the model is too simple, so switching to a larger model while keeping the same process fixes the estimate.

**Correct answer: B**
Repeatedly tuning against the test set leaks it into the selection process, so the score becomes optimistic and no longer reflects unseen data. The fix is to tune on a validation set (or cross-validation) and reserve the test set for a single final check. More iterations against the test set, or a bigger model, does not restore validity.

## Question 5

Domain 6: Responsible ML and Practice

CONFIG

Ava Thompson reviews a hiring model built to be fair.

```
mitigation: removed the "gender" field
result    : selection rate still much lower for one group
features  : include a variable strongly correlated with gender
```

Listing 5.1 The model and its features.

Why does the disparity persist, and what is the right response?

A. Removing the gender field made the model fair, so the remaining disparity is a coincidence that requires no further investigation or action at all.

B. A proxy feature still encodes gender, so bias persists; test outcomes across groups, mitigate the disparity, and reconsider the model for this high-impact use.

C. The model needs more historical hiring data, because training on more of the company's past decisions will remove the disparity between the groups.

D. The disparity is acceptable because the gender field was removed, so the outcome is fair by definition and the model can be deployed as it stands.

**Correct answer: B**
Dropping the gender field does not remove bias when a correlated proxy still encodes it, so the disparity persists. The right response is to test outcomes across groups, mitigate the disparity, and for a high-impact hiring use reconsider the model. Assuming fairness because a field was removed, or training on more biased history, does not address it.

## Question 6

Domain 5: Deployment and Monitoring

DATA

A team plans how to release a new model version to production.

| Option | Approach |
|---|---|
| A | Deploy to all users at once, no monitoring |
| B | Canary rollout with monitoring and rollback |
| C | Shadow then A/B test before full switch |
| D | Never deploy; keep it offline |

Table 6.1 Release options.

Which approach best balances safety and real-world validation?

A. Option A, because releasing the new model to everyone immediately with no monitoring is the fastest way to realise its benefits in production.

B. Option B (optionally with C), because a canary rollout with monitoring and rollback, and A/B testing, catches problems and measures real impact safely.

C. Option D, because keeping the model offline permanently is the only way to guarantee that a new model version never causes a production incident.

D. Option A combined with D, deploying to everyone at once but keeping the model offline as a backup so that no monitoring or rollback is ever required.

**Correct answer: B**
A canary rollout with monitoring and rollback limits the blast radius of a bad model, and shadow or A/B testing measures real-world impact before a full switch, balancing safety and validation. Deploying to everyone with no monitoring is reckless, and never deploying forgoes all value.

## Question 7 Domain 1: ML Foundations and Problem Framing

A team applies a complex machine learning model to a problem that a simple threshold rule already solves perfectly. What principle does this violate?

A. None; using the most sophisticated machine learning model available is always preferable to relying on a simple rule for any problem.

B. Prefer the simplest solution that meets the need; ML adds cost, complexity and maintenance when a rule already solves the problem well.

C. The team should have used an even more complex model, because a problem worth solving always justifies the most advanced approach available.

D. The only issue is that the rule was not written in the same programming language as the model, which is easily fixed by rewriting the rule.

**Correct answer: B**
When a simple rule solves the problem, machine learning adds cost, complexity and maintenance for no benefit. A good practitioner matches the tool to the problem and prefers the simplest solution that meets the need. More complexity is not inherently better, and the programming language is irrelevant.

## Question 8 Domain 2: Data Preparation and Feature Engineering

An engineer fits a feature scaler on the entire dataset, including the test set, before splitting. Why is this a problem?

A. It is not a problem, because fitting the scaler on all the data available gives the most accurate scaling and therefore the best model.

B. It leaks test-set information into training, giving an optimistic and invalid performance estimate; fit preprocessing on the training set only.

C. The only issue is that scaling is slower on the full dataset, so fitting the scaler on all the data merely wastes a little compute time.

D. It means the model cannot use the test set at all, so the fix is to also train the model on the test set to make full use of the data.

**Correct answer: B**
Fitting the scaler on all data, including the test set, lets test information influence training, producing an optimistic and invalid estimate. Preprocessing must be fit on the training data only and then applied to validation and test. It is a validity problem, not a speed one, and the test set must never be used for training.

## Question 9 Domain 3: Algorithms and Model Selection

Two models achieve almost identical validation performance, but one is far simpler, faster and easier to explain. Which is usually the better production choice, and why?

A. The more complex model, because a more sophisticated model is always the better production choice regardless of how it compares on performance.

B. The simpler, faster model, because for similar performance it is cheaper to run and easier to maintain, explain and debug in production.

C. It makes no difference which is chosen, because the two models perform similarly and every other consideration is irrelevant to production use.

D. The slower model, because a model that takes longer to produce each prediction is generally more thorough and therefore more reliable in production.

**Correct answer: B**
When performance is similar, the simpler, faster model is usually the better production choice: it is cheaper to run and easier to maintain, explain and debug. Complexity is not inherently better, the choice does matter, and being slower is a drawback, not a mark of reliability.

## Question 10 Domain 5: Deployment and Monitoring

A model improved on offline metrics, but an A/B test in production shows no business benefit. What does this teach?

A. Offline metrics are all that matter, so the model should be rolled out to everyone regardless of what the production A/B test showed.

B. Offline gains do not always translate to real impact, so changes must be validated with live experiments tied to the actual business objective.

C. A/B tests are unreliable, so the correct response is to ignore the test result and trust the offline improvement when deciding whether to deploy.

D. The offline metric must have been calculated incorrectly, so recomputing it is the right way to reconcile it with the production A/B test result.

**Correct answer: B**
An offline improvement that shows no benefit in an A/B test demonstrates that offline gains do not always translate to real-world impact, so changes must be validated with live experiments tied to the business objective. Offline metrics alone are not sufficient, the A/B result should not be dismissed, and the offline metric is not necessarily wrong.

Published by the CertLabz Certification Board, a division of CertLabz LLC. Companion to objectives document CMLP-001. https://certlabz.com/certifications/certified-machine-learning-professional.html