

PRACTICE QUESTIONS

Snowflake

# SnowPro Advanced: Data Scientist

SHARED BY [u/cosmicseal2002](#) on aiexamcerts

---

EDITION Free Set 2

---

QUESTIONS 10

---

SHARED October 15, 2026

---

EXAM Snowflake SnowPro Advanced: Data Scientist

---

Shared by [u/cosmicseal2002](#) on aiexamcerts

10 practice questions for the Snowflake SnowPro Advanced: Data Scientist exam, each followed by its answer and a short explanation. More free sets and the full practice exam:

[www.aiexamcerts.com/snowflake/snowpro-advanced-data-scientist](https://www.aiexamcerts.com/snowflake/snowpro-advanced-data-scientist). Spotted a mistake? Email [support@aiexamcerts.com](mailto:support@aiexamcerts.com) with the exam name and the question number.

1. Which Snowflake capability provides a hosted notebook experience inside Snowsight for Python and SQL data science work?
  - A External functions with API integrations
  - B Secure views
  - C Snowflake Notebooks
  - D Resource monitors

**Answer: C**

Snowflake Notebooks run inside Snowsight and let data scientists mix Python and SQL cells against Snowflake data. External functions call remote services from SQL, which is an integration pattern rather than an authoring environment. Secure views and resource monitors govern data access and credit usage.

2. Cross-validation with shuffled folds always gives an unbiased estimate of performance on future data when rows are strongly ordered by time.
  - A True
  - B False

**Answer: B (False)**

The statement is false because shuffled folds let future rows train models that score past rows. Time-ordered validation such as expanding-window splits is needed when the data has a time structure that deployment will face.

3. A pipeline must join a 2 billion row events table to a 5 thousand row lookup table for one date range. Which approach is most efficient in Snowflake?
  - A Cross join both tables, then filter on the event timestamp
  - B Copy the events table into a Python list on the client and merge the lookup values in memory
  - C Filter the events to the date range and join on the key so pruning and join plans stay efficient
  - D Export both tables to local CSV files and join them offline

**Answer: C**

Filtering early and joining on keys lets Snowflake prune micro-partitions and choose efficient join plans. A cross join creates an enormous intermediate result before the filter applies. Moving two billion rows into Python or local CSV files abandons the warehouse engine and is slow.

4. A \_\_\_\_\_ table is defined by a query and a target lag, and Snowflake refreshes it automatically to keep results within that lag.

**Answer: dynamic**

Dynamic tables let a pipeline be declared as a query, with Snowflake managing refresh scheduling against a TARGET\_LAG. They replace many hand-built stream and task chains for simple incremental transformations.

5. Which two practices help a team compare two candidate models fairly? (Choose two.)

- A Report only the best single fold from cross-validation
- B Document the tuning budget and settings used for each model
- C Use different random seeds for each model without recording them
- D Choose the model after looking at the final test set repeatedly
- E Evaluate both on the same held-out split with the same metric definitions

**Answer: B, E**

Shared evaluation data and metrics make the comparison meaningful, and documenting the tuning budget for each model keeps the contest fair. Reporting only the best fold, unrecorded seeds, and repeated peeking at the test set all inflate apparent performance.

6. Which Snowpark ML preprocessing transformer scales numeric columns to a fixed range such as 0 to 1?

- A OneHotEncoder
- B MinMaxScaler
- C SimpleImputer
- D KFold

**Answer: B**

MinMaxScaler rescales each numeric column into a chosen range, which is what the question describes. OneHotEncoder converts categories into indicator columns, SimpleImputer fills missing values, and KFold generates validation splits rather than transforming features.

7. How should target encoding be computed so that a row's own label does not leak into its encoded value?

- A Using the mean over the full column including the row itself
- B Copying the row's own label into the feature
- C Using out-of-fold statistics computed from other folds
- D Drawing a random value per row that is independent of the label

**Answer: C**

Out-of-fold or smoothed encodings compute each row's value from other rows only, so the feature cannot reveal its own label. Including the row itself or copying its label leaks the target directly, and a random value removes the useful signal.

8. Adding more trees to a random forest always improves its test accuracy without limit.
- A True
  - B False

**Answer: B (False)**

The statement is false. Adding trees usually reduces variance with diminishing returns, and the gains flatten out while cost grows. Accuracy does not improve indefinitely, and other hyperparameters still matter.

9. What does a calibration check compare for a classifier's predicted probabilities?
- A Training accuracy against validation accuracy only
  - B Feature names against the entries in the data dictionary and schema documentation
  - C Warehouse size against query runtime
  - D Predicted probabilities against observed event rates within probability bins

**Answer: D**

A calibrated model predicting 30 percent should see roughly 30 percent of those cases turn out positive, which binned comparison reveals. Training versus validation accuracy measures overfitting, and the other two comparisons do not evaluate probabilities at all.

10. A regression model reports a root mean squared error of 40 on prices averaging 400. Which interpretation is most useful for stakeholders?
- A Typical prediction error is about 10 percent of the average price
  - B The model is perfect because the error is below the average price
  - C Every row is off by exactly forty dollars
  - D The error only matters if the target is categorical

**Answer: A**

Relating the RMSE to the scale of the target makes the error interpretable, so 40 on 400 is roughly ten percent. Calling the model perfect misreads the metric, and RMSE does not mean every row is off by the same amount. Regression errors matter regardless of target type.