Vendor: Databricks

Exam Code: Certified-Machine-Learning-Professional

Exam Name: Databricks Certified Machine Learning Professional

Version: DEMO

QUESTION 1
A data scientist wants to examine the data in the Feature Store table table from the database dev as a Spark DataFrame. They have access to Feature Store Client fs. Which line of code can be used to gel the data from table as a Spark DataFrame?

A. fs.get_table("table)
B. fs.read_table("dev.table")
C. fs.create_table("dev.table")
D. fs.get_table("dev.table")

Answer: B
Explanation:
To read data from a Feature Store table as a Spark DataFrame, the correct method is fs.read_table("dev.table"). This retrieves the contents of the table in the dev database using the Feature Store Client (fs) and returns it as a Spark DataFrame.

QUESTION 2
A data scientist has computed updated rows that contain new feature values for primary keys already stored in the Feature Store table features. The updated feature values are stored in the DataFrame features_df. They want to update the rows in features if the associated primary key is in features_df. If a row's primary key is not in features_df, they want the row to remain unchanged in features. Which code block can they use to perform this task using the Feature Store Client fs?

A.
```
fs.write_table(
    name="features",
    df=features_df,
    mode="append"
)
```

B.
```
fs.write_table(
    name="features",
    df=features_df,
    mode="overwrite"
)
```

C.
```
fs.refresh_table(
    name="features",
    df=features_df,
    mode="overwrite"
)
```

D.
```
fs.write_table(
    name="features",
    df=features_df,
    mode="merge"
)
```

Answer: D
Explanation:
To update existing rows based on primary keys while leaving other rows unchanged, the correct

mode to use with fs.write_table() is "merge". This performs an upsert operation - updating rows where keys match and keeping others intact - making it ideal for updating feature values in a Feature Store table.

QUESTION 120
A machine learning engineer has developed a machine learning pipeline that produces a scikit-learn model `model` and computes the RMSE rmse. MAE mae, and R-squared r2 values for the test set. They now want to log these values with the MLflow run. These values are stored in the dictionary metrics.

They run the following code block:

```
with mlflow.start_run(experiment_id<exp_id, run_name=run_name) as run:
    # Log metrics
    mlflow.log_metric(metrics)
```

The code block produces an error.

Which changes to the code block will successfully complete the task?

A. Replace mlflow.log_metric with mlflow.sklearn.log_metric
B. Replace metrics with model
C. Replace log_metric with log_metrics
D. Replace metrics with rmse, mae, r2

Answer: C
Explanation:
The method mlflow.log_metric() logs a single metric, while mlflow.log_metrics() is used to log multiple metrics at once from a dictionary. Since metrics is a dictionary containing rmse, mae, and r2, the correct function is mlflow.log_metrics(metrics).

QUESTION 3
A Data Scientist is developing a model training pipeline on Databricks and needs to track custom performance metrics during training. They want to log a custom evaluation score (team_score), a single hyperparameter, and a confusion matrix plot as part of their MLflow experiment. Which code snippet correctly logs all three types of information in MLflow?

A. mlflow.log_param("max_depth", 5)
mlflow.log_metric("team_score", 0.92)
mlflow.log_artifact(open("confusion_matrix.png"))
B. mlflow.log_param("max_depth", 5)
mlflow.log_metric("team_score", 0.92)
mlflow.log_artifact("confusion_matrix.png")
C. mlflow.log_param("max_depth", "5")
mlflow.log_metric("team_score", "0.92")
mlflow.log_artifact("confusion_matrix.png")
D. mlflow.log_metric({"team_score": 0.92})
mlflow.log_param(["max_depth", 5])
mlflow.log_file("confusion_matrix.png")

Answer: B

Explanation:
This snippet correctly uses MLflow's APIs to log each item in its expected form: a single hyperparameter with log_param, a numeric custom metric with log_metric, and a file-based artifact such as a confusion matrix image by passing its file path to log_artifact. This is the standard and correct way to track parameters, metrics, and artifacts in an MLflow experiment.

QUESTION 4
A Machine Learning Engineer is using Lakehouse Monitoring to track the performance of ML models deployed in their environment. They want to monitor significant distributional drift in categorical features with a metric bounded on [0,1] for easy interpretation. Which statistical method should they use?

A. Jensen-Shannon Distance
B. Wasserstein Distance
C. Kolmogorov-Smirnov Test
D. Kullback-Leibler Distance

Answer: A
Explanation:
Jensen-Shannon Distance is well suited for measuring distributional drift in categorical features. It is symmetric, numerically stable, and bounded between 0 and 1, which makes it easy to interpret and ideal for monitoring categorical feature drift in Lakehouse Monitoring.

QUESTION 5
A Machine Learning Engineer is considering moving their functions and unit tests from notebooks into separate Python files (e.g., modules and test scripts) to take advantage of the numerous benefits of this approach like automated execution, code reusability, and version control. Which challenge should the engineer consider with this approach?

A. This change would prevent the use of any non-Python environments like Scala.
B. Separating code and tests often leads to decreased reliability and poor code quality.
C. Managing a more complex project structure can be harder to maintain and navigate.
D. It makes functions harder to import and reuse across different notebooks.

Answer: C
Explanation:
Moving code and unit tests into separate Python modules introduces a more structured project layout, which can increase complexity. Engineers must manage directories, dependencies, and imports carefully, making the project slightly harder to navigate and maintain compared to simple notebook-based workflows, especially for teams new to this structure.

QUESTION 6
A Machine Learning Engineer has automated a model retraining job in Databricks. Each scheduled run trains multiple candidate models with new sales data and logs all runs with MLflow. The goal is to select and register the best-performing model at the end of each cycle to ensure optimal forecast accuracy. Which approach will meet this goal?

A. Register the first model that has an evaluation metric better than the previously deployed model.
B. Register the best model based on the primary evaluation metric.
C. Register the model with the lowest training loss regardless of validation metrics.

D. Register all models from the run since all of them are trained on newer data and hence will be more accurate.

Answer: B
Explanation:
Selecting and registering the model that performs best on the primary evaluation metric ensures that only the highest-quality model is promoted at each retraining cycle. This approach aligns with MLOps best practices by basing promotion decisions on objective performance criteria rather than training order or assumptions about data freshness.

QUESTION 7
A Machine Learning Engineer is conducting hyperparameter tuning for multiple XGBoost models using Ray Tune on Databricks. They want to integrate MLflow tracking to monitor their experiments and need to ensure proper authentication. The engineer has Ray 2.41 installed and wants to use both Ray Tune and MLflow together in their distributed tuning workflow. They have to configure Databricks to run the hyperparameter optimization with MLflow integration. Which set of configuration steps will do this?

A. Install the MLflow Ray plugin using %pip install mlflow-ray and configure the workspace connection.
B. Configure DATABRICKS_HOST and DATABRICKS_TOKEN environment variables before calling setup_ray_cluster().
C. Enable MLflow autologging with mlflow.ray.autolog() and set the tracking server URI.
D. Set MLFLOW_TRACKING_URI and MLFLOW_EXPERIMENT_TD environment variables before initializing Ray.

Answer: B
Explanation:
When using Ray Tune with MLflow on Databricks, Ray workers must be able to authenticate back to the Databricks workspace to log runs to MLflow Tracking. Setting the DATABRICKS_HOST and DATABRICKS_TOKEN environment variables before initializing the Ray cluster ensures all Ray processes can securely communicate with Databricks and correctly log MLflow experiments during distributed hyperparameter tuning.

QUESTION 8
A Machine Learning Engineer needs to deploy a production ML workflow that includes an MLflow experiment for tracking model training runs, a registered model in Unity Catalog for version management, and a model serving endpoint for real-time inference. The team requires a unified configuration approach that ensures consistent deployment across development and production environments while adhering to infrastructure-as-code best practices. Which approach should the Machine Learning Engineer use to define all three components together?

A. Use MLflow's deployment tools to create individual deployment configurations for each component, then orchestrate them using Databricks Jobs.
B. Use Terraform to provision the infrastructure and then manually configure each ML component through the Databricks UI.
C. Define experiments, registered_models, and model_serving_endpoints resources in a Databricks Asset Bundle (DAB) configuration file.
D. Create separate REST API calls for each component and run them sequentially using a shell script.

Answer: C

Explanation:
Databricks Asset Bundles allow experiments, Unity Catalog–registered models, and model serving endpoints to be defined declaratively in a single configuration. This provides a unified, version-controlled, infrastructure-as-code approach that ensures consistent deployment across environments and aligns with MLOps best practices.

QUESTION 9
A Machine Learning Engineer uses Lakehouse Monitoring to track their credit scoring model's performance. The existing profile metrics table contains three aggregate metrics:

```
- adefault_risk_score
- payment_history_score
- credit_utilization_score
```

They need to:

1. Create a composite risk rating that combines these three scores using weights of 0.5, 0.3, and 0.2 respectively.
2. Monitor drift of this composite score against an established baseline.

Which approach should be used to implement both requirements within Lakehouse Monitoring?

A. Create two separate aggregate metrics: one for the composite score calculation and another for drift detection.
B. Use a scheduled notebook to calculate the composite score and manually insert both the score and drift values into the profile metrics table.
C. Create a derived metric for the composite_risk_rating calculation and create a drift metric on this derived metric.
D. Create an aggregate metric for composite_risk_rating and configure a separate drift metric to monitor changes.

Answer: C
Explanation:
Lakehouse Monitoring supports derived metrics that are computed from existing profile metrics using custom expressions. By defining a derived metric for the composite_risk_rating using the specified weights, the composite score becomes a first-class metric in the monitoring framework. A drift metric can then be directly configured on this derived metric to compare current values against the baseline, fulfilling both the composite calculation and drift monitoring requirements in a native, governed way.

QUESTION 10
A Machine Learning Engineer has deployed a fraud detection model that processes 10,000 transactions per hour. The model was trained on data from Q1 2024, but it's now Q4 2024. The ML team notices three concerning trends: (1) the model's precision has dropped from 92% to 78% over the past month, (2) the average transaction amount in recent data has increased from $150 to $220 and (3) the relationship between transaction frequency and fraud likelihood has weakened significantly due to new payment methods being introduced. The engineer needs to implement a monitoring solution that can detect the root cause of the performance degradation, identify why the precision dropped, and be able to do this at the scale needed. Which monitoring pipeline component will do this?

A. Model Health Monitoring Pipeline because it ensures the infrastructure can handle the 10,000

transactions per hour load.

B. Drift Detection Pipeline because it can identify both the data drift (transaction amount changes) and concept drift (weakened relationship between features and target).
C. Logging and Monitoring Pipeline because it captures all prediction requests and can identify patterns in the fraud predictions.
D. Model Performance Monitoring Pipeline because it directly tracks precision metrics and can alert when thresholds are breached.

Answer: B
Explanation:
A drift detection pipeline is designed to diagnose the root causes of model performance degradation at scale. It can detect data drift, such as changes in the distribution of transaction amounts, and concept drift, where the relationship between input features and the fraud label changes due to new payment methods. This directly explains why precision dropped and provides actionable insight beyond simply observing metric degradation.

QUESTION 111
A machine learning engineer is converting a Hyperopt-based hyperparameter tuning process from manual MLflow logging to MLflow Autologging. They notice that not all details and objects are automatically logged, and they will need to manually log some things. Which of the following will need to be manually logged when performing nested runs with Hyperopt and MLflow Autologging?

A. Trial models
B. Trial status
C. Best trial evaluation metric
D. Hyperparameter values
E. Evaluation metrics

Answer: C
Explanation:
When using MLflow Autologging with Hyperopt and nested runs, the best trial evaluation metric is not automatically logged and must be logged manually. Autologging captures trial-level details like hyperparameters and evaluation metrics, but summarizing and logging the overall best trial's result is a manual responsibility of the engineer.

QUESTION 112
A data scientist wants to track the runs of their random forest model. The data scientist is changing the number of trees and the maximum depth of the trees in the forest across each run.

They write the following code block:

```
with mlflow.start_run(experiment_id=exp_id, run_name=run_name) as run:
    rf = RandomForestRegressor(**params)
    rf.fit(X, y)
    predictions = rf.predict(X_test)
    mlflow.log_params(params)
    return run.info.run_id
```

Which Python object type does params need to be an instance of?

A. array
B. PySpark DataFrarne
C. dict
D. list

Answer: C
Explanation:
The params variable must be a dictionary (dict) because mlflow.log_params() expects a dictionary where each key-value pair represents a parameter name and its corresponding value. Additionally, the model instantiation RandomForestRegressor(**params) also requires params to be a dictionary to unpack the parameters correctly.

QUESTION 13
A machine learning engineer wants to move their model version model_version for the MLflow Model Registry model model from the Staging stage to the Production stage using MLflow Client client. Which code block can they use to accomplish the task?

A.
```
client.transition_model__stage(
    name=model,
    version=model_version,
    from="Staging",
    to="Production"
)
```

B.
```
client.transition_model_version_stage(
    name=model,
    version=model_version,
    from="Staging",
    to="Production"
)
```

C.
```
client.transition_model_stage(
    name=model,
    version=model_
    version, stage="Production"
)
```

D.
```
client.transition_model_version_stage(
    name=model,
    version=model_version,
    stage="Production"
)
```

Answer: D
Explanation:
The correct method for transitioning a model version to a new stage using the MLflow Client is client.transition_model_version_stage(name, version, stage). This updates the stage of a specific model version in the Model Registry, such as moving it from "Staging" to "Production". The stage parameter specifies the target stage only -- there is no from argument.

Thank You for Trying Our Product

Lead2pass Certification Exam Features:

★ More than **99,900** Satisfied Customers Worldwide.

★ Average **99.9%** Success Rate.

★ **Free Update** to match latest and real exam scenarios.

★ **Instant Download** Access! No Setup required.

★ Questions & Answers are downloadable in **PDF** format and **VCE** test engine format.

★ Multi-Platform capabilities - **Windows, Laptop, Mac, Android, iPhone, iPod, iPad**.

★ **100%** Guaranteed Success or **100%** Money Back Guarantee.

★ **Fast**, helpful support **24x7**.

View list of all certification exams: http://www.lead2pass.com/all-products.html

10% Discount Coupon Code: **ASTR14**