ML Associate - Mock Test 3

Question 1 of 65

What is a feature store in the context of Databricks?

- A tool for data visualization
- A tool for data cleaning
- A centralized repository that enables data scientists to find and share features
- A tool for data extraction

✓ Answer >

Correct Answer: A centralized repository that enables data scientists to find and share features

Question 2 of 65

(From Exam Guide) A data scientist wants to create a feature table to use in their models. They are working in a workspace with Unity Catalog enabled. What is the correct way of creating this feature table?

- Create a Delta table with data in it, as usual, then use the register_table method from the FeatureStoreClient.
- Create an empty Delta table on Unity Catalog with the AS FEATURE STORE clause via SQL, then write data to it.
- Use the create_table method of the Feature EngineeringClient in Python to create the table, then write data to it.
- Create a Delta table with data in it in Unity Catalog then use the ALTER TABLE command in SQL to configure it as a feature table.

✓ Answer >

Correct Answer: Use the create_table method of the Feature EngineeringClient in Python to create the table, then write data to it.

Question 3 of 65

What is the role of Unity Catalog in the context of Databricks Feature Store?

- It automates data ingestion and ETL processes for the Feature Store
- It acts as a centralized logging system for tracking ML experiments
- It becomes the feature store when the workspace is enabled for Unity Catalog
- It is used primarily for managing notebook versions across teams

✓ Answer >

Correct Answer: It becomes the feature store when the workspace is enabled for Unity Catalog

Question 4 of 65

What is the structure of a feature table name in Unity Catalog?

- schema-name.table-name
- table-name
- catalog-name.table-name
- catalog-name.schema-name.table-name

✓ Answer >

Correct Answer: catalog-name.schema-name.table-name

Question 5 of 65

What actions can be performed using the Model Registry UI page?

- To quantify the proportion of the variance in the dependent variable.
- To register, version, and manage machine learning models throughout their lifecycle.
- To calculate the total sum of squares in the target variable.
- To track the time complexity of the regression algorithm.

✓ Answer >

Correct Answer: To register, version, and manage machine learning models throughout their lifecycle.

Question 6 of 65

How does AutoML contribute to reproducibility in machine learning experiments?

- By restricting the customization and deployment options for models
- By automating the entire machine learning project without user intervention
- By integrating with MLflow Tracking and creating an experiment listing all generated models
- By providing training code only for selected trial runs

✓ Answer >

Correct Answer: By integrating with MLflow Tracking and creating an experiment listing all generated models

---

Question 7 of 65

What is one key advantage of AutoML in the context of machine learning projects?

- It promotes transparency and control by providing editable notebooks that include training code and baseline models
- It automates the process of generating complex machine learning algorithms
- It only supports classification problems
- It restricts users from customizing and deploying models

✓ Answer ›

Correct Answer: It promotes transparency and control by providing editable notebooks that include training code and baseline models

---

Question 8 of 65

What does AutoML offer in terms of model explainability?

- It generates complex models without any explanation
- It limits customization of the training process for better interpretability
- It integrates with MLflow but does not provide any visualization for model interpretability
- It includes SHAP plots in the generated notebooks to visualize feature importance and enhance explainability

✓ Answer ›

Correct Answer: It includes SHAP plots in the generated notebooks to visualize feature importance and enhance explainability

---

Question 9 of 65

What types of machine learning problems does Databricks AutoML support?

- Only classification problems
- Only regression problems
- Various problems, including regression, classification, and time series forecasting
- None of the above

✓ Answer ›

Correct Answer: Various problems, including regression, classification, and time series forecasting

---

Question 10 of 65

What is the primary focus of DataOps in MLOps?

- Developing deep learning algorithms
- Managing data pipelines and ensuring data quality
- Orchestrating DevOps workflows
- Building machine learning models

✓ Answer ›

Correct Answer: Managing data pipelines and ensuring data quality

---

Question 11 of 65

Which tool within the Databricks ecosystem is primarily used for continuous integration and continuous deployment (CI/CD) of machine learning models?

- MLflow
- Databricks Delta
- Databricks SQL Analytics
- Databricks Workspace

✓ **Answer** ›

Correct Answer: MLflow

Question 12 of 65

What is the primary purpose of Databricks Workflows?

- To build and monitor machine learning models
- To analyze SQL queries
- To perform ETL processes
- To orchestrate and automate general-purpose tasks

✓ **Answer** ›

Correct Answer: To orchestrate and automate general-purpose tasks

---

Question 13 of 65

What is the purpose of Continuous Integration and Continuous Deployment (CI/CD) in MLOps?

- To perform manual model testing
- To analyze data in real-time
- To manage data governance and security
- To automate the deployment and testing of code and models

✓ **Answer** ›

Correct Answer: To automate the deployment and testing of code and models

---

Question 14 of 65

What is a Databricks Asset Bundle (DAB)?

- A collection of Databricks machine learning models
- A specific type of SQL query used in Databricks
- A set of artifacts and assets for automation and CI/CD processes
- A pre-built machine learning model available in Databricks

✓ Answer >

Correct Answer: A set of artifacts and assets for automation and CI/CD processes

---

Question 15 of 65

What is the function of orchestration services in MLOps?

- To handle user authentication and authorization
- To write and execute machine learning algorithms
- To manage the scheduling and execution of data and ML workflow
- To store large volumes of data

✓ Answer >

Correct Answer: To manage the scheduling and execution of data and ML workflow

---

Question 16 of 65

Which stage in the MLOps lifecycle is primarily responsible for ensuring the model is ready for deployment in a live environment?

- Staging
- Infrastructure
- Production
- Development

✓ Answer >

Correct Answer: Staging

Question 17 of 65

What type of model drift occurs when the statistical properties of the input data change over time, potentially reducing prediction accuracy?

- Concept Drift
- Model Quality Drift
- Bias Drift
- Data Drift

✓ Answer >

Correct Answer: Data Drift

Question 18 of 65

What is the purpose of preparing data for machine learning projects?

- To make the data look more appealing
- To reduce the size of the dataset
- To ensure the data fits into a specific database
- To ensure a solid foundation for successful machine learning model deployment

✓ Answer >

Correct Answer: To ensure a solid foundation for successful machine learning model deployment

---

Question 19 of 65

What is the purpose of one-hot encoding categorical features in machine learning?

- To make the data more visually appealing
- To reduce the size of the dataset
- To standardize the features in a training set
- To convert categorical data into a binary format

✓ Answer >

Correct Answer: To convert categorical data into a binary format

---

Question 20 of 65

What does feature selection involve in the context of machine learning projects?

- Standardizing the features in a training set
- Adding new columns/rows
- Replacing missing values in columns/rows
- Dropping unnecessary columns/rows

✓ Answer >

Correct Answer: Dropping unnecessary columns/rows

Question 21 of 65

Which method is most suitable for imputing missing values in a categorical variable?

- Random imputation
- Mode imputation
- Median imputation
- Mean imputation

✓ Answer ›

Correct Answer: Mode imputation

Question 22 of 65

(From Exam Guide) A data scientist needs to impute the missing values in a continuous feature. They want to do this with the least amount of effort but with correct results. Which strategy will do this?

- Use sklearn SimpleImputer, which automatically selects the best methodology
- Examine the distribution of the values and select the appropriate imputation upon review
- Use .mean(), which is the most appropriate imputation on continuous columns
- Use .mode(), which is the most appropriate imputation on continuous columns

✓ Answer ›

Correct Answer: Examine the distribution of the values and select the appropriate imputation upon review

---

Question 23 of 65

What is a major drawback of using one-hot encoding for categorical features in machine learning?

- Inability to handle high cardinality variables.
- Difficulty in interpreting the significance of original features
- Unsuitability for continuous variables
- Computational expense and potential for sparse matrices

✓ Answer ›

Correct Answer: Computational expense and potential for sparse matrices

---

Question 24 of 65

When and why do we need data standardization?

- When all features in the dataset are on the same scale
- When features in the dataset have varying scales and the model being used is scale-sensitive
- When the dataset is too large
- When the dataset contains only categorical data

✓ Answer ›

Correct Answer: When features in the dataset have varying scales and the model being used is scale-sensitive

---

Question 25 of 65

Which of these models is sensitive to the scale of the features?

- Naive Bayes
- Random Forest
- Decision Trees
- Support Vector Machines

✓ Answer >

Correct Answer: Support Vector Machines

Question 26 of 65

What is the primary purpose of evaluation metrics in machine learning?

- Compatibility with Python-based machine learning libraries
- Optimized model training algorithms
- To provide a numerical representation of how well a model is performing.
- To generate visual representations of model performance.

✓ Answer >

Correct Answer: To provide a numerical representation of how well a model is performing.

Question 27 of 65

How do evaluation metrics contribute to the machine learning development process? (Select all that apply)

- To guide the fine-tuning of models by adjusting hyperparameters.
- To provide a numerical representation of how well a model is performing.
- Automated data collection
- To enable comparison between different models or versions of the same model.

✓ Answer >

Correct Answers:

- To guide the fine-tuning of models by adjusting hyperparameters.
- To provide a numerical representation of how well a model is performing.
- To enable comparison between different models or versions of the same model.

Question 28 of 65

What is the role of R-squared in a regression model?

- To quantify the proportion of the variance in the independent variables

- To quantify the proportion of the variance [in the target variable explained by the model]
- To calculate the total sum of squares in the target variable
- To measure the average error of the predictions

✓ Answer >

Correct Answer: To quantify the proportion of the variance [in the target variable explained by the model]

Question 29 of 65

Which evaluation metrics are most appropriate for assessing classification models with imbalanced datasets? (Select 2)

- Accuracy
- Recall
- RMSE
- Precision

✓ Answer >

Correct Answers:

- Recall
- Precision

Question 30 of 65

What characterizes the use of accuracy as an evaluation metric in machine learning? (Select 2)

- Accuracy is not affected by class imbalances in the dataset.
- Accuracy is expressed as a percentage, and higher values indicate better performance.
- Accuracy is primarily used in regression models.
- Accuracy is commonly employed in binary and multiclass classification problems.

✓ Answer >

Correct Answers:

- Accuracy is expressed as a percentage, and higher values indicate better performance.
- Accuracy is commonly employed in binary and multiclass classification problems.

Question 31 of 65

What key characteristics define precision as an evaluation metric in classification models?

- True Positives (TP) are instances incorrectly predicted as positive.
- Precision focuses specifically on the accuracy of the model when it predicts positive instances.
- Precision is independent of recall.

- Precision is not affected by the trade-off relationship with other metrics.

✓ Answer ›

Correct Answer: Precision focuses specifically on the accuracy of the model when it predicts positive instances.

Question 32 of 65

What defines recall as an evaluation metric in classification models?

- Recall centers on the model's accuracy in predicting positive instances.
- True Positives (TP) represent instances incorrectly predicted as positive.
- Recall emphasizes capturing all true positive instances accurately.
- Recall remains uninfluenced by trade-off relationships with other metrics.

✓ Answer ›

Correct Answer: Recall emphasizes capturing all true positive instances accurately.

Question 33 of 65

Under what circumstances might prioritizing recall over accuracy be a sensible choice in machine learning?

- If the cost of a false negative is high
- When the dataset is perfectly balanced between two classes.
- If the consequences of false negatives are negligible.
- In datasets where one class is significantly more prevalent than the other.

✓ Answer ›

Correct Answer: If the cost of a false negative is high

Question 34 of 65

What is the primary purpose of hyperparameter tuning in machine learning?

- To configure the model's architecture during training
- To control the learning process and enhance model performance
- To define the input features for the machine learning model
- To determine the output classes in a clustering algorithm

✓ Answer ›

Correct Answer: To control the learning process and enhance model performance

---

Question 35 of 65

Why is grid search considered computationally expensive?

- Because it requires a powerful GPU for model training
- Due to the need for extensive cross-validation and evaluating all possible hyperparameter combinations
- Due to its limitation in handling large datasets
- Because it only works with a small set of hyperparameter values

✓ Answer ›

Correct Answer: Due to the need for extensive cross-validation and evaluating all possible hyperparameter combinations

---

Question 36 of 65

What is a key advantage of using random search over grid search in hyperparameter tuning?

- Random search is ideal for exploring narrow hyperparameter ranges
- Random search requires fewer data points for each combination
- Random search is less efficient than grid search
- Random search is more efficient... as it explores a broader range without exhaustive search

✓ Answer ›

Correct Answer: Random search is more efficient... as it explores a broader range without exhaustive search

---

Question 37 of 65

(From Exam Guide) A data scientist is working on a machine learning project to develop a model that predicts whether a customer will churn. The dataset is highly imbalanced (10% churn). Which strategy directly mitigates the model's bias towards the non-churn customers?

- Normalize the features to ensure they are on the same scale, improving model performance.
- Use cost-sensitive learning by assigning a higher misclassification cost to the minority class.
- Increase the size of the training dataset by collecting more data on non-churn customers.
- Use a simpler model to reduce overfitting, ensuring it generalizes better to the minority class.

✓ Answer ›

Correct Answer: Use cost-sensitive learning by assigning a higher misclassification cost to the minority class.

Question 38 of 65

(From Exam Guide) A data scientist is tuning an SVM model using 5-fold cross-validation and GridSearchCV. The parameter grid includes: C with values [0.1, 1, 10], kernel with choices ['linear', 'rbf'], and gamma with values [0.01, 0.1, 1]. How many different models will be trained in total?

- 90
- 18
- 1
- None of the above

✓ Answer ›

Correct Answer: 90

Question 39 of 65

What are the pros and cons of using embeddings to represent categorical features?

- Pros: They can be computationally inexpensive; Cons: They can decrease the model's accuracy
- Pros: They can increase the model's accuracy; Cons: They can lead to underfitting
- Pros: They can decrease the dimensionality of the dataset; Cons: They can lead to overfitting
- Pros: They can capture complex relationships; Cons: They can be computationally expensive

✓ Answer ›

Correct Answer: Pros: They can capture complex relationships; Cons: They can be computationally expensive

---

Question 40 of 65

Why is batch deployment considered a common model deployment strategy?

- Because it handles large volumes of pre-stored data efficiently.
- Because it is resource-intensive.
- Because it allows real-time processing.
- Because it supports delayed results.

✓ Answer ›

Correct Answer: Because it handles large volumes of pre-stored data efficiently.

Question 41 of 65

What are the advantages of deploying a model via batch processing? (Select 2)

- Efficiently handles large volumes of pre-stored data.
- Resource-efficient in real-time scenarios.
- Enables parallel processing, enhancing overall performance.
- Seamlessly integrates with tools like MLflow.

✓ Answer ›

Correct Answers:

- Efficiently handles large volumes of pre-stored data.
- Enables parallel processing, enhancing overall performance.

Question 42 of 65

(From Exam Guide) A company has a podcast platform... [with] an anomaly detection algorithm... on a 10-minute running window of user events. A machine learning engineer wants to deploy this model into a production data pipeline that needs to handle up to tens of thousands of events per second... [and] compute to be resized dynamically. Which pipeline design approach meets these requirements?

- Create a Delta Live Tables pipeline that applies the algorithm as a Spark UDF.
- Create a Structured Streaming Job that applies the algorithm as a Spark UDF.
- Create a model serving endpoint... [and a] DLT pipeline that calls a custom UDF which invokes the endpoint.
- Create a model serving endpoint... [and a] Structured Streaming job that calls a custom UDF which invokes the endpoint.

✓ **Answer** ›

Correct Answer: Create a Delta Live Tables pipeline that applies the algorithm as a Spark UDF.

Question 43 of 65

In what situations is real-time deployment of machine learning models typically required?

- When dealing with historical data analysis
- When offline predictions are sufficient
- When focusing on batch processing
- When immediate predictions and low latency are crucial

✓ **Answer** ›

Correct Answer: When immediate predictions and low latency are crucial

---

Question 44 of 65

What does Databricks Model Serving offer for deploying MLflow models?

- Batch processing capabilities
- Scalable REST API endpoints for deployment
- Exclusively offline deployment options
- Limited model exposure

✓ **Answer** ›

Correct Answer: Scalable REST API endpoints for deployment

Question 45 of 65

What are some crucial features provided by Databricks Model Serving?

- Limited availability and security
- Batch processing and reduced infrastructure costs
- High availability, scalability, and security with permissions
- Default concurrency limits and reduced scalability

✓ Answer >

Correct Answer: High availability, scalability, and security with permissions

---

Question 46 of 65

Which MLflow Client API method is used to manually log a trained model as an artifact within an active run?

- mlflow.log_artifact()
- mlflow.sklearn.log_model()
- mlflow.log_metric()
- mlflow.register_model()

✓ Answer >

Correct Answer: mlflow.sklearn.log_model()

Question 47 of 65

What is the primary difference between an online and an offline feature store table?

- Online tables are used for batch inference, and offline tables are used for real-time inference.
- Online tables are optimized for low-latency lookups, while offline tables are optimized for large-scale training.
- Online tables can only be used by models in production, while offline tables are for development.
- Online tables are stored in Unity Catalog, while offline tables are stored in the workspace registry.

✓ Answer >

Correct Answer: Online tables are optimized for low-latency lookups, while offline tables are optimized for large-scale training.

Question 48 of 65

How can you promote a newly validated model version to be the 'champion' model in the Unity Catalog Model Registry?

- By manually renaming the model version to 'champion'.
- By setting the 'champion' tag on the model version.
- By assigning the 'champion' alias to the new model version.

- By deleting all other model versions that are not the champion.

✓ Answer ›

Correct Answer: By assigning the 'champion' alias to the new model version.

Question 49 of 65

Beyond finding the best model, how does Databricks AutoML primarily facilitate the model development process?

- By automatically deploying the best model to a serving endpoint.
- By generating a 'glass-box' notebook with the source code for the best run, promoting transparency and reproducibility.
- By automatically collecting and cleaning all required data from source systems.
- By guaranteeing the model will never overfit the training data.

✓ Answer ›

Correct Answer: By generating a 'glass-box' notebook with the source code for the best run, promoting transparency and reproducibility.

Question 50 of 65

What is a key benefit of registering models in the Unity Catalog (UC) registry compared to the older workspace registry?

- Models in the UC registry can be trained much faster.
- The UC registry is the only one that supports model aliases.
- Models in the UC registry can be shared and accessed across multiple workspaces.
- Only models in the UC registry can be logged using MLflow.

✓ Answer ›

Correct Answer: Models in the UC registry can be shared and accessed across multiple workspaces.

Question 51 of 65

You are imputing a numerical feature that has a highly skewed distribution with significant outliers. Which imputation method is generally more robust?

- Mean imputation
- Median imputation
- Mode imputation
- Dropping the rows

✓ Answer ›

Correct Answer: Median imputation

Question 52 of 65

What is the most appropriate type of visualization to examine the relationship between two continuous variables, such as 'Age' and 'Income'?

- A bar chart
- A scatter plot
- A pie chart
- A histogram

✓ Answer ›

Correct Answer: A scatter plot

Question 53 of 65

In which of the following scenarios is one-hot encoding (OHE) a *poor* choice for feature engineering?

- A categorical feature 'Marital_Status' with 5 unique values (e.g., 'Single', 'Married', 'Divorced').
- A categorical feature 'Zip_Code' with 30,000 unique values.
- A numerical feature 'Age' with values from 18 to 65.
- A categorical feature 'Day_of_Week' with 7 unique values.

✓ Answer ›

Correct Answer: A categorical feature 'Zip_Code' with 30,000 unique values.

Question 54 of 65

In which scenario would applying a log scale transformation to a feature be most appropriate?

- When the feature is categorical, like 'Color'.
- When the feature has a normal (bell-shaped) distribution.
- When the feature is highly right-skewed, like 'Income' or 'File_Size'.
- When the feature contains negative values.

✓ Answer ›

Correct Answer: When the feature is highly right-skewed, like 'Income' or 'File_Size'.

Question 55 of 65

Which of the following is a correct way to compute summary statistics (like mean, stddev, min, max) for a Spark DataFrame named `sdf`?

- sdf.stats()
- dbutils.data.summarize(sdf)
- sdf.info()
- sdf.metrics()

✓ Answer >

Correct Answer: dbutils.data.summarize(sdf)

Question 56 of 65

Which open-source library is integrated into Databricks to facilitate distributed hyperparameter tuning using Bayesian search methods?

- GridSearchCV
- SparkML
- Hyperopt
- MLflow

✓ Answer >

Correct Answer: Hyperopt

Question 57 of 65

Besides cost-sensitive learning, what is another common strategy to mitigate data imbalance in a training dataset?

- Using one-hot encoding for all categorical features.
- Applying a log transform to the target variable.
- Resampling the dataset, such as oversampling the minority class or undersampling the majority class.
- Using a more complex model, like a deep neural network.

✓ Answer >

Correct Answer: Resampling the dataset, such as oversampling the minority class or undersampling the majority class.

Question 58 of 65

A model with very high training accuracy but very low testing accuracy is most likely suffering from what?

- High bias and high variance
- Low bias and low variance
- High bias and low variance (Underfitting)

- Low bias and high variance (Overfitting)

✓ Answer ›

Correct Answer: Low bias and high variance (Overfitting)

Question 59 of 65

You are building a regression model to predict house prices. Your dataset is known to have some extreme outliers (e.g., multi-million dollar mansions). Which evaluation metric would be *most* sensitive to errors in predicting these outliers?

- R-squared (R^2)
- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- F1-Score

✓ Answer ›

Correct Answer: Root Mean Squared Error (RMSE)

Question 60 of 65

In a scikit-learn or SparkML pipeline, what is the key difference between an Estimator and a Transformer?

- Estimators use a `.transform()` method, and Transformers use a `.fit()` method.
- Estimators (like a model) learn parameters from data using `.fit()`, while Transformers (like a scaler) use `.transform()` to change data.
- Estimators are only used for classification, and Transformers are only used for regression.
- Estimators can only be at the end of a pipeline, and Transformers can only be at the beginning.

✓ Answer ›

Correct Answer: Estimators (like a model) learn parameters from data using `.fit()`, while Transformers (like a scaler) use `.transform()` to change data.

Question 61 of 65

How can you use a trained MLflow model (e.g., a pandas-optimized model) to perform efficient, distributed batch inference on a large Spark DataFrame?

- By calling `mlflow.pyfunc.predict()` on the entire Spark DataFrame.
- By converting the Spark DataFrame to pandas with `.toPandas()` and then using the model.
- By using `mlflow.pyfunc.spark_udf()` to create a Spark UDF and applying it to a column in the Spark DataFrame.
- By manually exporting the model and loading it on each executor node.

✓ Answer ›

Correct Answer: By using `mlflow.pyfunc.spark_udf()` to create a Spark UDF and applying it to a column in the Spark DataFrame.

Question 62 of 65

What is a key advantage of real-time model deployment over batch deployment?

- It is more cost-effective for scoring large, historical datasets.
- It allows for immediate predictions with low latency for interactive applications.
- It simplifies data preprocessing, as no feature store is needed.
- It can handle larger volumes of data per request than batch processing.

✓ Answer ›

Correct Answer: It allows for immediate predictions with low latency for interactive applications.

Question 63 of 65

How is streaming inference typically performed using Delta Live Tables (DLT)?

- By defining a real-time serving endpoint within the DLT pipeline.
- By loading the model and applying it as a transformation (e.g., a UDF) within a DLT streaming query.
- By scheduling the DLT pipeline to run once per day in batch mode.
- By connecting the DLT pipeline directly to the MLflow Model Registry UI.

✓ Answer ›

Correct Answer: By loading the model and applying it as a transformation (e.g., a UDF) within a DLT streaming query.

Question 64 of 65

When deploying a custom model to a Databricks Model Serving endpoint, what must you do to handle custom logic or dependencies?

- You cannot deploy custom models; you can only deploy models trained with AutoML.
- Log the model using `mlflow.pyfunc.log_model()`, specifying a custom `python_model` class and a `conda_env` file.
- Manually install the custom libraries on the serving cluster before deploying.
- Embed all custom logic into a single large SQL UDF.

✓ Answer ›

Correct Answer: Log the model using `mlflow.pyfunc.log_model()`, specifying a custom `python_model` class and a `conda_env` file.

Question 65 of 65

What is the primary purpose of splitting traffic between different model versions (e.g., 90% to the champion, 10% to a challenger) on a serving endpoint?

- To reduce the computational cost of the endpoint.
- To perform A/B testing or a canary rollout in a live production environment.
- To meet regulatory requirements for model governance.
- To train the challenger model using the 10% of live traffic.

✓ **Answer** ›

Correct Answer: To perform A/B testing or a canary rollout in a live production environment.