DATABRICKS-MACHINE-LEARNING-ASSOCIATE Q&As

Databricks Certified Machine Learning Associate Exam

Pass Databricks DATABRICKS-MACHINE-LEARNING-ASSOCIATE Exam with 100% Guarantee

Free Download Real Questions & Answers **PDF** and **VCE** file from:

https://www.leads4pass.com/databricks-machine-learning-associate.html

100% Passing Guarantee
100% Money Back Assurance

Following Questions and Answers are all new published by Databricks Official Exam Center

QUESTION 1

A data scientist has developed a linear regression model using Spark ML and computed the predictions in a Spark DataFrame preds_df with the following schema:

prediction DOUBLE actual DOUBLE Which of the following code blocks can be used to compute the root mean-squared-error of the model according to the data in preds_df and assign it to the rmse variable?

A.
```
rmse = BinaryClassificationEvaluator(
    predictionCol="prediction",
    labelCol="actual",
    metricName="rmse"
)
```

B.
```
rmse = RegressionEvaluator(
    predictionCol="prediction",
    labelCol="actual",
    metricName="rmse"
)
```

C.
```
rmse = RegressionEvaluator(
    predictionCol="prediction",
    labelCol="actual",
    metricName="rmse"
)
```

D.
```
classification_evaluator = BinaryClassificationEvaluator(
    predictionCol="prediction",
    labelCol="actual",
    metricName="rmse"
)
```

E.
```
rmse = classification_evaluator.evaluate(preds_df)
regression_evaluator = RegressionEvaluator(
    predictionCol="prediction",
    labelCol="actual",
    metricName="rmse"
)
```

A. Option A

B. Option B

C. Option C

D. Option D

E. Option E

Correct Answer: C

The code block to compute the root mean-squared error (RMSE) for a linear regression model in Spark ML should use theRegressionEvaluatorclass withmetricNameset to "rmse". Given the schema ofpreds_dfwith

columnspredictionandactual, the correct evaluator setup will specifypredictionCol="prediction"andlabelCol="actual". Thus, the appropriate code block (Option C in your list) that usesRegressionEvaluatorto compute the RMSE is the correct

choice. This setup correctly measures the performance of the regression model using the predictions and actual outcomes from the DataFrame.

References:

Spark ML documentation (Using RegressionEvaluator to Compute RMSE).

QUESTION 2

Which of the following hyperparameter optimization methods automatically makes informed selections of hyperparameter values based on previous trials for each iterative model evaluation?

A. Random Search

B. Halving Random Search

C. Tree of Parzen Estimators

D. Grid Search

Correct Answer: C

Tree of Parzen Estimators (TPE) is a sequential model-based optimization algorithm that selects hyperparameter values based on the outcomes of previous trials. It models the probability density of good and bad hyperparameter values and

makes informed decisions about which hyperparameters to try next. This approach contrasts with methods like random search and grid search, which do not use information from previous trials to guide the search process.

References:

Hyperopt and TPE

QUESTION 3

A data scientist has produced three new models for a single machine learning problem. In the past, the solution used just one model. All four models have nearly the same prediction latency, but a machine learning engineer suggests that the new solution will be less time efficient during inference.

In which situation will the machine learning engineer be correct?

A. When the new solution requires if-else logic determining which model to use to compute each prediction

B. When the new solution\\'s models have an average latency that is larger than the size of the original model

C. When the new solution requires the use of fewer feature variables than the original model

D. When the new solution requires that each model computes a prediction for every record

E. When the new solution\\'s models have an average size that is larger than the size of the original model

Correct Answer: D

If the new solution requires that each of the three models computes a prediction for every record, the time efficiency during inference will be reduced. This is because the inference process now involves running multiple models instead of a single model, thereby increasing the overall computation time for each record. In scenarios where inference must be done by multiple models for each record, the latency accumulates, making the process less time efficient compared to using a single model. References: Model Ensemble Techniques

---

QUESTION 4

Which of the following approaches can be used to view the notebook that was run to create an MLflow run?

A. Open the MLmodel artifact in the MLflow run paqe

B. Click the "Models" link in the row corresponding to the run in the MLflow experiment paqe

C. Click the "Source" link in the row corresponding to the run in the MLflow experiment page

D. Click the "Start Time" link in the row corresponding to the run in the MLflow experiment page

Correct Answer: C

To view the notebook that was run to create an MLflow run, you can click the "Source" link in the row corresponding to the run in the MLflow experiment page. The "Source" link provides a direct reference to the source notebook or script that

initiated the run, allowing you to review the code and methodology used in the experiment. This feature is particularly useful for reproducibility and for understanding the context of the experiment.

References:

MLflow Documentation (Viewing Run Sources and Notebooks).

QUESTION 5

A data scientist is performing hyperparameter tuning using an iterative optimization algorithm. Each evaluation of unique hyperparameter values is being trained on a single compute node. They are performing eight total evaluations across eight total compute nodes. While the accuracy of the model does vary over the eight evaluations, they notice there is no trend of improvement in the accuracy. The data scientist believes this is due to the parallelization of the tuning process.

Which change could the data scientist make to improve their model accuracy over the course of their tuning process?

A. Change the number of compute nodes to be half or less than half of the number of evaluations.

B. Change the number of compute nodes and the number of evaluations to be much larger but equal.

C. Change the iterative optimization algorithm used to facilitate the tuning process.

D. Change the number of compute nodes to be double or more than double the number of evaluations.

Correct Answer: C

The lack of improvement in model accuracy across evaluations suggests that the optimization algorithm might not be effectively exploring the hyperparameter space. Iterative optimization algorithms like Tree-structured Parzen Estimators

(TPE) or Bayesian Optimization can adapt based on previous evaluations, guiding the search towards more promising regions of the hyperparameter space.

Changing the optimization algorithm can lead to better utilization of the information gathered during each evaluation, potentially improving the overall accuracy.

References:

Hyperparameter Optimization with Hyperopt

QUESTION 6

Which of the following is a benefit of using vectorized pandas UDFs instead of standard PySpark UDFs?

A. The vectorized pandas UDFs allow for the use of type hints

B. The vectorized pandas UDFs process data in batches rather than one row at a time

C. The vectorized pandas UDFs allow for pandas API use inside of the function

D. The vectorized pandas UDFs work on distributed DataFrames

E. The vectorized pandas UDFs process data in memory rather than spilling to disk

Correct Answer: B

Vectorized pandas UDFs, also known as Pandas UDFs, are a powerful feature in PySpark that allows for more efficient operations than standard UDFs. They operate by processing data in batches, utilizing vectorized operations that leverage pandas to perform operations on whole batches of data at once. This approach is much moreefficient than processing data row by row as is typical with standard PySpark UDFs, which can significantly speed up the computation. References: PySpark Documentation on UDFs:https://spark.apache.org/docs/latest/api/python/user_guide/sql/arrow_panda s.html#pandas-udfs-a-k-a-vectorized-udfs

QUESTION 7

A data scientist is attempting to tune a logistic regression model logistic using scikit-learn. They want to specify a search space for two hyperparameters and let the tuning process randomly select values for each evaluation.

They attempt to run the following code block, but it does not accomplish the desired task:

```
distributions = dict(C=uniform(loc=0, scale=4), penalty=['l2', 'l1'])
clf = GridSearchCV(logistic, distributions, random_state=0)
search = clf.fit(feature_data, target_data)
```

Which of the following changes can the data scientist make to accomplish the task?

A. Replace the GridSearchCV operation with RandomizedSearchCV

B. Replace the GridSearchCV operation with cross_validate

C. Replace the GridSearchCV operation with ParameterGrid

D. Replace the random_state=0 argument with random_state=1

E. Replace the penalty= [\\'12\\', \\'11\\'] argument with penalty=uniform (\\'12\\', \\'11\\')

Correct Answer: A

The user wants to specify a search space for hyperparameters and let the tuning process randomly select values.GridSearchCVsystematically tries every combination of the provided hyperparameter values, which can be computationally expensive and time-consuming.RandomizedSearchCV, on the other hand, samples hyperparameters from a distribution for a fixed number of iterations. This approach is usually faster and still can find very good parameters, especially when the search space is large or includes distributions. References: Scikit-Learn documentation on hyperparameter tuning: https://scikit-learn.org/stable/modules/grid_search.html#randomized-parameter-optimization

QUESTION 8

A machine learning engineer has been notified that a new Staging version of a model registered to the MLflow Model Registry has passed all tests. As a result, the machine learning engineer wants to put this model into production by transitioning it to the Production stage in the Model Registry.

From which of the following pages in Databricks Machine Learning can the machine learning engineer accomplish this task?

A. The home page of the MLflow Model Registry

B. The experiment page in the Experiments observatory

C. The model version page in the MLflow ModelRegistry

D. The model page in the MLflow Model Registry

Correct Answer: C

The machine learning engineer can transition a model version to the Production stage in the Model Registry from the model version page. This page provides detailed information about a specific version of a model, including its metrics,

parameters, and current stage. From here, the engineer can perform stage transitions, moving the model from Staging to Production after it has passed all necessary tests.

References:

Databricks documentation on MLflow Model Registry:

https://docs.databricks.com/applications/mlflow/model-registry.html#model-version

QUESTION 9

Which of the following evaluation metrics is not suitable to evaluate runs in AutoML experiments for regression problems?

A. F1

B. R-squared

C. MAE

D. MSE

Correct Answer: A

The code block provided by the machine learning engineer will perform the desired inference when the Feature Store feature set was logged with the model at model_uri. This ensures that all necessary feature transformations and metadata

are available for the model to make predictions. The Feature Store in Databricks allows for seamless integration of features and models, ensuring that the required features are correctly used during inference.

References:

Databricks documentation on Feature Store: Feature Store in Databricks

QUESTION 10

A data scientist has created a linear regression model that useslog(price)as a label variable. Using this model, they have performed inference and the predictions and actual label values are in Spark DataFramepreds_df.

They are using the following code block to evaluate the model:

regression_evaluator.setMetricName("rmse").evaluate(preds_df)

Which of the following changes should the data scientist make to evaluate the RMSE in a way that is comparable withprice?

A. They should exponentiate the computed RMSE value

B. They should take the log of the predictions before computing the RMSE

C. They should evaluate the MSE of the log predictions to compute the RMSE

D. They should exponentiate the predictions before computing the RMSE

Correct Answer: D

When evaluating the RMSE for a model that predicts log-transformed prices, the predictions need to be transformed back to the original scale to obtain an RMSE that is comparable with the actual price values. This is done by exponentiating

the predictions before computing the RMSE. The RMSE should be computed on the same scale as the original data to provide a meaningful measure of error.

References:

Databricks documentation on regression evaluation: Regression Evaluation

---

QUESTION 11

A data scientist has developed a machine learning pipeline with a static input data set using Spark ML, but the pipeline is taking too long to process. They increase the number of workers in the cluster to get the pipeline to run more efficiently. They notice that the number of rows in the training set after reconfiguring the cluster is different from the number of rows in the training set prior to reconfiguring the cluster.

Which of the following approaches will guarantee a reproducible training and test set for each model?

A. Manually configure the cluster

B. Write out the split data sets to persistent storage

C. Set a speed in the data splitting operation

D. Manually partition the input data

Correct Answer: B

To ensure reproducible training and test sets, writing the split data sets to persistent storage is a reliable approach. This allows you to consistently load the same training and test data for each model run, regardless of cluster reconfiguration

or other changes in the environment.

Correct approach:

Split the data.

Write the split data to persistent storage (e.g., HDFS, S3). Load the data from storage for each model training session. train_df, test_df = spark_df.randomSplit([0.8,0.2], seed=42) train_df.write.parquet("path/to/train_df.parquet")

test_df.write.parquet("path/to/test_df.parquet")# Later, load the datatrain_df = spark.read.parquet("path/to/train_df.parquet") test_df = spark.read.parquet("path/to/test_df.parquet")

References:

Spark DataFrameWriter Documentation

QUESTION 12

A machine learning engineer wants to parallelize the inference of group-specific models using the Pandas Function API. They have developed theapply_modelfunction that will look up and load the correct model for each group, and they want to apply it to each group of DataFramedf.

They have written the following incomplete code block:

```
prediction_df = (df
    .groupby("device_id")
    .______(apply_model, schema=apply_return_schema)
)
```

Which piece of code can be used to fill in the above blank to complete the task?

A. applyInPandas

B. groupedApplyInPandas

C. mapInPandas

D. predict

Correct Answer: A

To parallelize the inference of group-specific models using the Pandas Function API in PySpark, you can use theapplyInPandasfunction. This function allows you to apply a Python function on each group of a DataFrame and return a

DataFrame, leveraging the power of pandas UDFs (user-defined functions) for better performance.

prediction_df = (df.groupby("device_id") .applyInPandas(apply_model, schema=apply_return_schema))

In this code:

groupby("device_id"): Groups the DataFrame by the "device_id" column. applyInPandas(apply_model, schema=apply_return_schema): Applies theapply_modelfunction to each group and specifies the schema of the return DataFrame.

References:

PySpark Pandas UDFs Documentation

QUESTION 13

A data scientist is using MLflow to track their machine learning experiment. As a part of each of their MLflow runs, they are performing hyperparameter tuning. The data scientist would like to have one parent run for the tuning process with a child run for each unique combination of hyperparameter values. All parent and child runs are being manually started with mlflow.start_run.

Which of the following approaches can the data scientist use to accomplish this MLflow run organization?

A. Theycan turn on Databricks Autologging

B. Theycan specify nested=True when startingthe child run for each unique combination of hyperparameter values

C. Theycan start each child run inside the parentrun\\'s indented code block usingmlflow.start runO

D. They can start each child run with the same experiment ID as the parent run

E. They can specify nested=True when starting the parent run for the tuningprocess

Correct Answer: B

To organize MLflow runs with one parent run for the tuning process and a child run for each unique combination of hyperparameter values, the data scientist can specifynested=Truewhen starting the child run. This approach ensures that

each child run is properly nested under the parent run, maintaining a clear hierarchical structure for the experiment. This nesting helps in tracking and comparing different hyperparameter combinations within the same tuning

process.References:

MLflow Documentation (Managing Nested Runs).

QUESTION 14

A data scientist is using the following code block to tune hyperparameters for a machine learning model:

```
num_evals = 4
trials = SparkTrials()
best_hyperparam = fmin(
    fn=objective_function,
    space=search_space,
    algo=tpe.suggest,
    max_evals=num_evals,
    trials=trials
)
```

Which change can they make the above code block to improve the likelihood of a more accurate model?

A. Increase num_evals to 100

B. Change fmin() to fmax()

C. Change sparkTrials() to Trials()

D. Change tpe.suggest to random.suggest

Correct Answer: A

To improve the likelihood of a more accurate model, the data scientist can increasenum_evalsto 100. Increasing the number of evaluations allows the hyperparameter tuning process to explore a larger search space and evaluate more

combinations of hyperparameters, which increases the chance of finding a more optimal set of hyperparameters for the model.

References:

Databricks documentation on hyperparameter tuning: Hyperparameter Tuning

QUESTION 15

A new data scientist has started working on an existing machine learning project. The project is a scheduled Job that retrains every day. The project currently exists in a Repo in Databricks. The data scientist has been tasked with improving the feature engineering of the pipeline\\'s preprocessing stage. The data scientist wants to make necessary updates to the code that can be easily adopted into the project without changing what is being run each day.

Which approach should the data scientist take to complete this task?

A. They can create a new branch in Databricks, commit their changes, and push those changes to the Git provider.

B. They can clone the notebooks in the repository into a Databricks Workspace folder and make the necessary changes.

C. They can create a new Git repository, import it into Databricks, and copy and paste the existing code from the original repository before making changes.

D. They can clone the notebooks in the repository into a new Databricks Repo and make the necessary changes.

Correct Answer: A

The best approach for the data scientist to take in this scenario is to create a new branch in Databricks, commit their changes, and push those changes to the Git provider. This approach allows the data scientist to make updates and

improvements to the feature engineering part of the preprocessing pipeline without affecting the main codebase that runs daily. By creating a new branch, they can work on their changes in isolation. Once the changes are ready and tested,

they can be merged back into the main branch through a pull request, ensuring a smooth integration process and allowing for code review and collaboration with other team members.

References:

Databricks documentation on Git integration: Databricks Repos