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.geekcert.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 machine learning engineer is using the following code block to scale the inference of a single-node model on a Spark DataFrame with one million records:

```
@pandas_udf("double")
def predict(iterator: Iterator[pd.DataFrame]) -> Iterator[pd.Series]:
    model_path = f"runs:/{run.info.run_id}/model"
    model = mlflow.sklearn.load_model(model_path)
    for features in iterator:
        pdf = pd.concat(features, axis=1)
        yield pd.Series(model.predict(pdf))
```

Assuming the default Spark configuration is in place, which of the following is a benefit of using anIterator?

A. The data will be limited to a single executor preventing the model from being loaded multiple times

B. The model will be limited to a single executor preventing the data from being distributed

C. The model only needs to be loaded once per executor rather than once per batch during the inference process

D. The data will be distributed across multiple executors during the inference process

Correct Answer: C

Using an iterator in thepandas_udfensures that the model only needs to be loaded once per executor rather than once per batch. This approach reduces the overhead associated with repeatedly loading the model during the inference

process, leading to more efficient and faster predictions. The data will be distributed across multiple executors, but each executor will load the model only once, optimizing the inference process.

References:

Databricks documentation on pandas UDFs: Pandas UDFs

QUESTION 2

A data scientist uses 3-fold cross-validation when optimizing model hyperparameters for a regression problem. The following root-mean-squared-error values are calculated on each of the validation folds:

1.

10.0

2.

12.0

3.

17.0

Which of the following values represents the overall cross-validation root-mean-squared error?

A. 13.0

B. 17.0

C. 12.0

D. 39.0

E. 10.0

Correct Answer: A

To calculate the overall cross-validation root-mean-squared error (RMSE), you average the RMSE values obtained from each validation fold. Given the RMSE values of 10.0, 12.0, and 17.0 for the three folds, the overall cross-validation

RMSE is calculated as the average of these three values:

Overall CV RMSE=10.0+12.0+17.03=39.03=13.0Overall CV RMSE=310.0+12.0+17.0 =339.0=13.0

Thus, the correct answer is 13.0, which accurately represents the average RMSE across all folds.References:

Cross-validation in Regression (Understanding Cross-Validation Metrics).

QUESTION 3

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 4

Which of the following machine learning algorithms typically uses bagging?

A. IGradient boosted trees

B. K-means

C. Random forest

D. Decision tree

Correct Answer: C

Random Forest is a machine learning algorithm that typically uses bagging (Bootstrap Aggregating). Bagging is a technique that involves training multiple base models (such as decision trees) on different subsets of the data and then combining their predictions to improve overall model performance. Each subset is created by randomly sampling with replacement from the original dataset. The Random Forest algorithm builds multiple decision trees and merges them to get a more accurate and stable prediction. References: Databricks documentation on Random Forest: Random Forest in Spark ML

QUESTION 5

A data scientist has written a data cleaning notebook that utilizes the pandas library, but their colleague has suggested that they refactor their notebook to scale with big data.

Which of the following approaches can the data scientist take to spend the least amount of time refactoring their notebook to scale with big data?

A. They can refactor their notebook to process the data in parallel.

B. They can refactor their notebook to use the PySpark DataFrame API.

C. They can refactor their notebook to use the Scala Dataset API.

D. They can refactor their notebook to use Spark SQL.

E. They can refactor their notebook to utilize the pandas API on Spark.

Correct Answer: E

The data scientist can refactor their notebook to utilize the pandas API on Spark (now known aspandas on Spark, formerlyKoalas). This allows for the least amount of changes to the existing pandas-based code while scaling to handle big

data using Spark\\'s distributed computing capabilities.pandas on Sparkprovides a similar API to pandas, making the transition smoother and faster compared to completely rewriting the code to use PySpark DataFrame API, Scala Dataset API,

or Spark SQL.References:

Databricks documentation on pandas API on Spark (formerly Koalas).