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 engineering team has a Job with three successive tasks. Each task runs a single notebook. The team has been alerted that the Job has failed in its latest run.

Which of the following approaches can the team use to identify which task is the cause of the failure?

A. Run each notebook interactively

B. Review the matrix view in the Job\\'s runs

C. Migrate the Job to a Delta Live Tables pipeline

D. Change each Task\\'s setting to use a dedicated cluster

Correct Answer: B

To identify which task is causing the failure in the job, the team should review the matrix view in the Job\\'s runs. The matrix view provides a clear and detailed overview of each task\\'s status, allowing the team to quickly identify which task

failed. This approach ismore efficient than running each notebook interactively, as it provides immediate insights into the job\\'s execution flow and any issues that occurred during the run.

References:

Databricks documentation on Jobs: Jobs in Databricks

QUESTION 2

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 3

Which of the following tools can be used to parallelize the hyperparameter tuning process for single-node machine learning models using a Spark cluster?

A. MLflow Experiment Tracking

B. Spark ML

C. Autoscaling clusters

D. Autoscaling clusters

E. Delta Lake

Correct Answer: B

Spark ML (part of Apache Spark\\'s MLlib) is designed to handle machine learning tasks across multiple nodes in a cluster, effectively parallelizing tasks like hyperparameter tuning. It supports various machine learning algorithms that can be

optimized over a Spark cluster, making it suitable for parallelizing hyperparameter tuning for single-node machine learning models when they are adapted to run on Spark.

References:

Apache Spark MLlib Guide:https://spark.apache.org/docs/latest/ml-guide.html

Spark ML is a library within Apache Spark designed for scalable machine learning. It provides tools to handle large-scale machine learning tasks, including parallelizing the hyperparameter tuning process for single-node machine learning

models using a Spark cluster. Here\\'s a detailed explanation of how Spark ML can be used:

Hyperparameter Tuning with CrossValidator: Spark ML includes theCrossValidatorandTrainValidationSplitclasses, which are used for hyperparameter tuning. These classes can evaluate multiple sets of hyperparameters in parallel using a

Spark cluster. from pyspark.ml.tuning import CrossValidator, ParamGridBuilder from pyspark.ml.evaluation import BinaryClassificationEvaluator

Define the model

model = ...

Create a parameter grid

paramGrid = ParamGridBuilder() \

addGrid(model.hyperparam1, [value1, value2]) \

addGrid(model.hyperparam2, [value3, value4]) \

build()

Define the evaluator

evaluator = BinaryClassificationEvaluator()

Define the CrossValidator

crossval = CrossValidator(estimator=model,

estimatorParamMaps=paramGrid,

evaluator=evaluator,

numFolds=3)

Parallel Execution: Spark distributes the tasks of training models with different hyperparameters across the cluster\\'s nodes. Each node processes a subset of the parameter grid, which allows multiple models to be trained simultaneously.

Scalability: Spark ML leverages the distributed computing capabilities of Spark. This allows for efficient processing of large datasets and training of models across many nodes, which speeds up the hyperparameter tuning process significantly

compared to single-node computations.

References:

Apache Spark MLlib Documentation

Hyperparameter Tuning in Spark ML

QUESTION 4

A data scientist is working with a feature set with the following schema:

```
customer_id STRING,
spend DOUBLE,
units INTEGER,
loyalty_tier STRING
```

Thecustomer_idcolumn is the primary key in the feature set. Each of the columns in the feature set has missing values. They want to replace the missing values by imputing a common value for each feature.

Which of the following lists all of the columns in the feature set that need to be imputed using the most common value of the column?

A. customer_id, loyalty_tier

B. loyalty_tier

C. units

D. spend

E. customer_id

Correct Answer: B

For the feature set schema provided, the columns that need to be imputed using the most common value (mode) are typically the categorical columns. In this case, loyalty_tieris the only categorical column that should be imputed using the

most common value.customer_idis a unique identifier and should not be imputed, whilespendandunits are numerical columns that should typically be imputed using the mean or median values, not the mode.

References:

Databricks documentation on missing value imputation: Handling Missing Data If you need any further clarification or additional questions answered, please let me know!

QUESTION 5

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