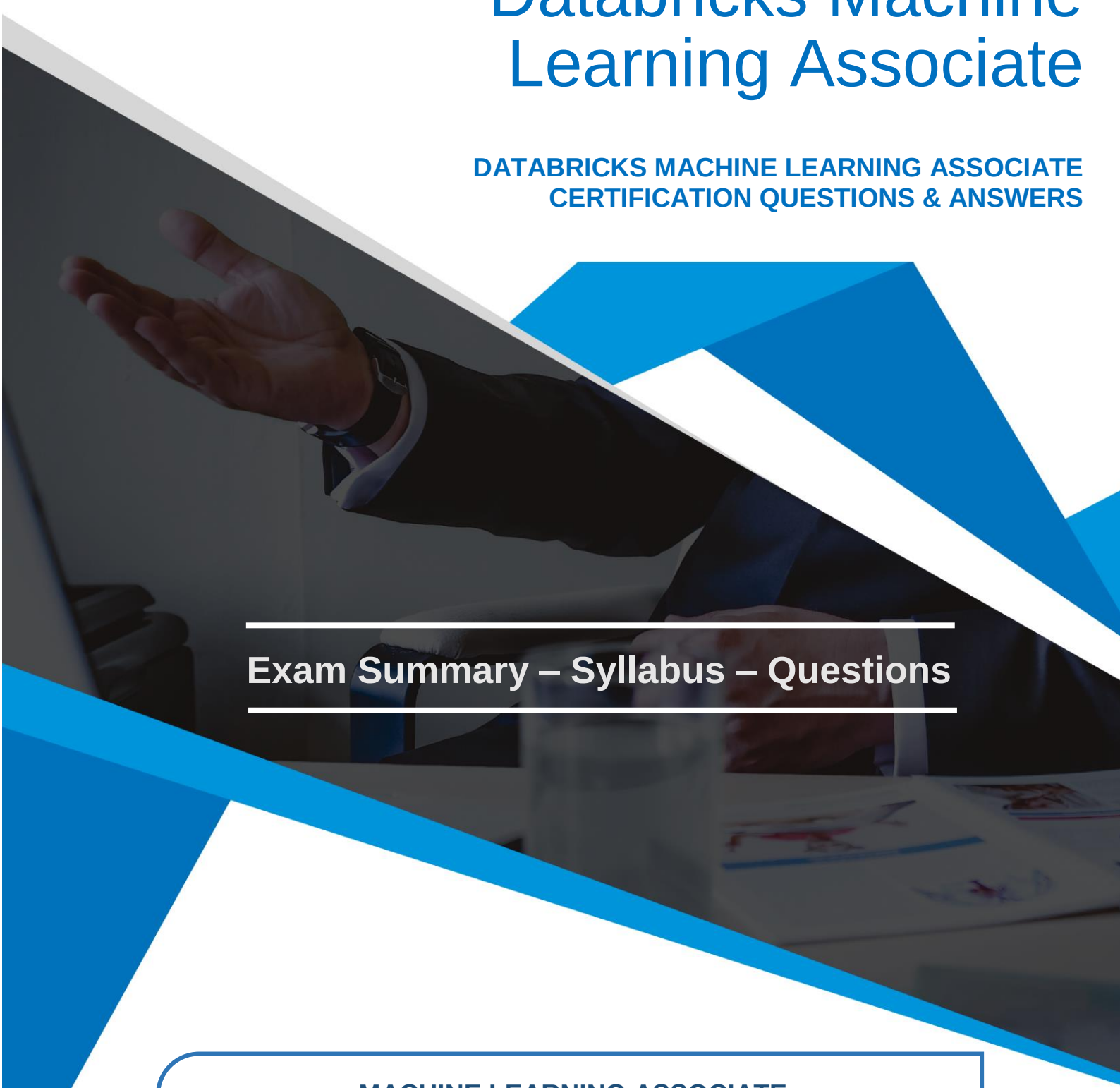
Databricks Machine Learning Associate

DATABRICKS MACHINE LEARNING ASSOCIATE CERTIFICATION QUESTIONS & ANSWERS



Exam Summary – Syllabus – Questions

MACHINE LEARNING ASSOCIATE

Databricks Certified Machine Learning Associate

45 Questions Exam – 70% Cut Score – Duration of 90 minutes

www.CertFun.com

Table of Contents

Know Your Machine Learning Associate Certification Well:

The Machine Learning Associate is best suitable for candidates who want to gain knowledge in the Databricks ML Engineer. Before you start your Machine Learning Associate preparation you may struggle to get all the crucial Machine Learning Associate materials like Machine Learning Associate syllabus, sample questions, study guide.

But don't worry the Machine Learning Associate PDF is here to help you prepare in a stress free manner.

The PDF is a combination of all your queries like-

- What is in the Machine Learning Associate syllabus?
- How many questions are there in the Machine Learning Associate exam?
- Which Practice test would help me to pass the Machine Learning Associate exam at the first attempt?

Passing the Machine Learning Associate exam makes you Databricks Certified Machine Learning Associate. Having the Machine Learning Associate certification opens multiple opportunities for you. You can grab a new job, get a higher salary or simply get recognition within your current organization.

Databricks Machine Learning Associate Certification Details:

Exam Name	Databricks Certified Machine Learning Associate
Exam Code	Machine Learning Associate
Exam Price	$200 (USD)
Duration	90 mins
Number of Questions	45
Passing Score	70%
Books / Training	**Scalable Machine Learning with Apache Spark**
Schedule Exam	**Kryterion Webassesor**
Sample Questions	**Databricks Machine Learning Associate Sample Questions**
Practice Exam	**Databricks Machine Learning Associate Certification Practice Exam**

Machine Learning Associate Syllabus:

Topic	Details	Weights
Databricks Machine Learning	- Databricks ML • Identify when a standard cluster is preferred over a single-node cluster and vice versa • Connect a repo from an external Git provider to Databricks repos. • Commit changes from a Databricks Repo to an external Git provider. • Create a new branch and commit changes to an external Git provider. • Pull changes from an external Git provider back to a Databricks workspace. • Orchestrate multi-task ML workflows using Databricks jobs. - Databricks Runtime for Machine Learning • Create a cluster with the Databricks Runtime for Machine Learning. • Install a Python library to be available to all notebooks that run on a cluster. - AutoML • Identify the steps of the machine learning workflow completed by AutoML. • Identify how to locate the source code for the best model produced by AutoML. • Identify which evaluation metrics AutoML can use for regression problems. • Identify the key attributes of the data set using the AutoML data exploration notebook. - Feature Store • Describe the benefits of using Feature Store to store and access features for machine learning pipelines. • Create a feature store table. • Write data to a feature store table. • Train a model with features from a feature store table.	**29%**

Topic	Details	Weights
	• Score a model using features from a feature store table. - Managed MLflow • Identify the best run using the MLflow Client API. • Manually log metrics, artifacts, and models in an MLflow Run. • Create a nested Run for deeper Tracking organization. • Locate the time a run was executed in the MLflow UI. • Locate the code that was executed with a run in the MLflow UI. • Register a model using the MLflow Client API. • Transition a model's stage using the Model Registry UI page. • Transition a model's stage using the MLflow Client API. • Request to transition a model's stage using the ML Registry UI page.	
ML Workflows	- Exploratory Data Analysis • Compute summary statistics on a Spark DataFrame using .summary() • Compute summary statistics on a Spark DataFrame using dbutils data summaries. • Remove outliers from a Spark DataFrame that are beyond or less than a designated threshold. - Feature Engineering • Identify why it is important to add indicator variables for missing values that have been imputed or replaced. • Describe when replacing missing values with the mode value is an appropriate way to handle missing values. • Compare and contrast imputing missing values with the mean value or median value. • Impute missing values with the mean or median value.	**29%**

Topic	Details	Weights
	• Describe the process of one-hot encoding categorical features. • Describe why one-hot encoding categorical features can be inefficient for tree-based models. - Training • Perform random search as a method for tuning hyperparameters. • Describe the basics of Bayesian methods for tuning hyperparameters. • Describe why parallelizing sequential/iterative models can be difficult. • Understand the balance between compute resources and parallelization. • Parallelize the tuning of hyperparameters using Hyperopt and SparkTrials. • Identify the usage of SparkTrials as the tool that enables parallelization for tuning single-node models. - Evaluation and Selection • Describe cross-validation and the benefits of downsides of using cross-validation over a train-validation split. • Perform cross-validation as a part of model fitting. • Identify the number of models being trained in conjunction with a grid-search and cross-validation process. • Describe Recall and F1 as evaluation metrics. • Identify the need to exponentiate the RMSE when the log of the label variable is used • Identify that the RMSE has not been exponentiated when the log of the label variable is used.	
Spark ML	- Distributed ML Concepts • Describe some of the difficulties associated with distributing machine learning models. • Identify Spark ML as a key library for distributing traditional machine learning work.	**33%**

Topic	Details	Weights
	• Identify scikit-learn as a single-node solution relative to Spark ML. - Spark ML Modeling APIs • Split data using Spark ML. • Identify key gotchas when splitting distributed data using Spark ML. • Train / evaluate a machine learning model using Spark ML. • Describe Spark ML estimator and Spark ML transformer. • Develop a Pipeline using Spark ML. • Identify key gotchas when developing a Spark ML Pipeline. - Hyperopt • Identify Hyperopt as a solution for parallelizing the tuning of single-node models. • Identify Hyperopt as a solution for Bayesian hyperparameter inference for distributed models. • Parallelize the tuning of hyperparameters for Spark ML models using Hyperopt and Trials. • Identify the relationship between the number of trials and model accuracy. - Pandas API on Spark • Describe key differences between Spark DataFrames and Pandas on Spark DataFrames. • Identify the usage of an InternalFrame making Pandas API on Spark not quite as fast as native Spark. • Identify Pandas API on Spark as a solution for scaling data pipelines without much refactoring. • Convert data between a PySpark DataFrame and a Pandas on Spark DataFrame. • Identify how to import and use the Pandas on Spark APIs. - Pandas UDFs/Function APIs • Identify Apache Arrow as the key to Pandas <-	

Topic	Details	Weights
	> Spark conversions. • Describe why iterator UDFs are preferred for large data. • Apply a model in parallel using a Pandas UDF. • Identify that pandas code can be used inside of a UDF function. • Train / apply group-specific models using the Pandas Function API.	
Scaling ML Models	- Model Distribution • Describe how Spark scales linear regression. • Describe how Spark scales decision trees. Ensembling Distribution • Describe the basic concepts of ensemble learning • Compare and contrast bagging, boosting, and stacking.	**9%**

Databricks Machine Learning Associate Sample Questions:

Question: 1

A data scientist is developing a machine learning model. They made changes to their code in a text editor on their local machine, committed them to the project's Git repository, and pushed the changes to an online Git provider. Now, they want to load those changes into Databricks. The Databricks workspace contains an out-of-date version of the Git repository.

How can the data scientist complete this task?

a) Open the Repo Git dialog and enable automatic syncing.
b) Open the Repo Git dialog and click the "Sync" button.
c) Open the Repo Git dialog and click the "Merge" button.
d) Open the Repo Git dialog and click the "Pull" button.
e) Open the Repo Git dialog and enable automatic pulling.

Answer: d

Question: 2

Which of the following steps are necessary to commit changes from a Databricks Repo to an external Git provider?

(Select two)

a) Merge changes to the master branch in the external Git provider
b) Use Databricks notebooks to push changes
c) Stage and commit changes in the Databricks workspace
d) Pull requests from the Databricks workspace to the Git provider

Answer: b, c

Question: 3

Which of the following are key components of ML workflows in Databricks?

a) Data ingestion
b) Model serving
c) Feature extraction
d) Manual model tuning

Answer: a, b, c

Question: 4

A senior machine learning engineer is developing a machine learning pipeline. They set up the pipeline to automatically transition a new version of a registered model to the Production stage in the Model Registry once it passes all tests using the MLflow Client API client.

Which operation was used to transition the model to the Production stage?

a) Client.update_model_stage
b) client.transition_model_version_stage
c) client.transition_model_version
d) client.update_model_version

Answer: b

Question: 5

A machine learning team wants to use the Python library newpackage on all of their projects. They share a cluster for all of their projects. Which approach makes the Python library newpackage available to all notebooks run on a cluster?

a) Edit the cluster to use the Databricks Runtime for Machine Learning
b) Set the runtime-version variable in their Spark session to "ml"
c) Running %pip install newpackage once on any notebook attached to the cluster
d) Adding /databricks/python/bin/pip install newpackage to the cluster's bash init script
e) There is no way to make the newpackage library available on a cluster

Answer: d

Question: 6

A data scientist has developed a two-class decision tree classifier 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 accuracy of the model according to the data in preds_df and assign it to the accuracy variable?

a) accuracy = RegressionEvaluator
predictionCol="prediction",
labelCol="actual",
metricName="accuracy"
)
b) accuracy = MulticlassClassificationEvaluator(
predictionCol="prediction",
labelCol="actual",
metricName="accuracy"
)
accuracy = classification_evaluator.evaluate(preds_df)
c) classification_evaluator = BinaryClassificationEvaluator(
predictionCol="prediction",
labelCol="actual",
metricName="accuracy"
)
d) accuracy = Summarizer(
predictionCol="prediction",
labelCol="actual",
metricName="accuracy"
)
e) classification_evaluator = BinaryClassificationEvaluator(
predictionCol="prediction",
labelCol="actual",
metricName="accuracy"
)
accuracy = classification_evaluator.evaluate(preds_df)

Answer: e

Question: 7

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 using the Feature Store Client fs can be used to accomplish this task?

a) fs.write_table(
name="features",
df=features_df,
mode="merge"
)
b) fs.write_table(
name="features",
df=features_df,
mode="overwrite"
)
c) fs.write_table(
name="features",
df=features_df,
)
d) fs.create_table(
name="features",
df=features_df,
mode="append"
)

Answer: a

Question: 8

When AutoML explores the key attributes of a dataset, which of the following elements does it typically not assess?

a) The dataset's memory footprint.
b) The potential impact of outliers on model performance.
c) The balance or imbalance of classes in classification tasks.
d) The encryption level of the dataset.

Answer: d

Question: 9

Where can you find the code that was executed with a run in the MLflow UI?

a) In the run's metadata section.
b) Inside the associated Git repository.
c) Under the "Code" tab in the run's details page.
d) It is not possible to view the executed code in the MLflow UI.

Answer: c

Question: 10

How can you identify the best run using the MLflow Client API?

a) By manually reviewing each run's metrics.
b) Utilizing the search_runs function with a specific metric sort order.
c) Comparing run IDs manually for performance metrics.
d) Using a custom Python script outside of MLflow.

Answer: b

Study Guide to Crack Databricks Machine Learning Associate Exam:

- Getting details of the Machine Learning Associate syllabus, is the first step of a study plan. This pdf is going to be of ultimate help. Completion of the syllabus is must to pass the Machine Learning Associate exam.
- Making a schedule is vital. A structured method of preparation leads to success. A candidate must plan his schedule and follow it rigorously to attain success.
- Joining the Databricks provided training for Machine Learning Associate exam could be of much help. If there is specific training for the exam, you can discover it from the link above.
- Read from the Machine Learning Associate sample questions to gain your idea about the actual exam questions. In this PDF useful sample questions are provided to make your exam preparation easy.
- Practicing on Machine Learning Associate practice tests is must. Continuous practice will make you an expert in all syllabus areas.

Reliable Online Practice Test for Machine Learning Associate Certification

Make CertFun.com your best friend during your Databricks Certified Machine Learning Associate exam preparation. We provide authentic practice tests for the Machine Learning Associate exam. Experts design these online practice tests, so we can offer you an exclusive experience of taking the actual Machine Learning Associate exam. We guarantee you 100% success in your first exam attempt if you continue practicing regularly. Don't bother if you don't get 100% marks in initial practice exam attempts. Just utilize the result section to know your strengths and weaknesses and prepare according to that until you get 100% with our practice tests. Our evaluation makes you confident, and you can score high in the Machine Learning Associate exam.

Start Online Practice of Machine Learning Associate Exam by Visiting URL

https://www.certfun.com/databricks/databricks-certified-machine-learning-associate