Databricks Machine Learning Associate Certification Training

OEM: Apache • Duration: 3 Days (24 hrs) • Code: DBX-ML-A

COURSE MODULES & TOPICS

Section 1: Databricks Machine Learning (38%)

- MLOps strategy best practices
- ML Runtimes (Databricks Runtime for ML) advantages
- AutoML for model and feature selection
- Feature Store: tables in Unity Catalog vs. workspace level
- Create, write to, and score with feature store tables
- Offline vs. online feature tables
- MLflow: identify best run, manually log metrics/artifacts/models
- Register models using MLflow Client API in Unity Catalog registry
- Promoting code vs. promoting models — when to use each
- Set/remove model tags; promote challenger to champion using aliases

Section 2: Data Processing (19%)

- Compute summary statistics on Spark DataFrames (.summary())
- Remove outliers based on standard deviation or IQR
- Create visualizations for categorical and continuous features
- Compare two categorical or two continuous features
- Imputing missing values: mean, median, mode — compare and apply
- One-hot encoding for categorical features (when appropriate and not)
- Log scale transformation — when to apply

Section 3: Model Development (31%)

- Select appropriate algorithm for a given scenario
- Methods to mitigate data imbalance in training data
- Estimators vs. transformers in Spark ML Pipelines
- Develop a training pipeline end-to-end
- Hyperopt fmin for hyperparameter tuning
- Random, grid, and Bayesian hyperparameter search
- Parallelize single-node models for hyperparameter tuning
- Cross-validation vs. train-validation split — benefits and trade-offs
- Classification metrics: F1, Log Loss, ROC/AUC

- Regression metrics: RMSE, MAE, R-squared
- Bias-variance trade-off and model complexity

Section 4: Model Deployment (12%)

- Batch, real-time, and streaming inference — differences and advantages
- Deploy a custom model to a model serving endpoint
- Batch inference with pandas
- Streaming inference with Delta Live Tables
- Deploy and query model for real-time inference
- Split traffic between endpoints (A/B testing / canary deployments)