training code: DBX-AI&ML / ENG DL 1d / EN

Machine Learning with Azure Databricks: From Principles and Model Development to Optimization

This original instructor-led course guides participants through the complete machine learning journey on Azure Databricks - from understanding what learning from data means, through data preparation and building the first model, to a managed, deployed, and optimized production model. The program combines solid theoretical foundations with the current capabilities of the platform: Databricks Runtime ML, Spark MLlib, MLflow 3, Models in Unity Catalog, Model Serving, Optuna, and AutoML.

The course focuses on the decisions that determine model quality in practice: formulating the problem, avoiding data leakage, selecting a metric that reflects the business cost of errors, diagnosing underfitting and overfitting through the bias-variance perspective, and consciously choosing among manual tuning, Optuna, and AutoML. The materials reflect recent platform changes, including the transition from Hyperopt to Optuna, model aliases replacing registry stages, and AutoML requirements related to Databricks Runtime ML versions.

The training consists of 3 modules, 4 demonstrations using real telecommunications data, and 5 step-by-step labs completed in a single Azure Databricks environment created in Module 1.

Training recipients

This training is intended for:

- Data analysts and aspiring data scientists, who know Python and SQL and want to move from data analysis to building, evaluating, and deploying machine learning models at scale.
- Data engineers working with Azure Databricks, who prepare data and features for ML teams and want to understand the complete model lifecycle, from Medallion Architecture to serving predictions.
- Python developers, who want to enter machine learning in a structured way, with theory where it is needed and code that runs on a distributed Spark platform.
- Data architects and team leaders, who are responsible for model-related standards, experiment tracking, governance in Unity Catalog, model versioning, and deployment.
- Professionals migrating from other ML tools, who want to learn current Databricks-recommended practices, including MLflow 3, Models in Unity Catalog, Optuna, and AutoML.

Benefits

- After completing the training, participants will gain the following benefits:
- Strong machine learning foundations – understand why learning from a sample works, how the training loop operates, and how to recognize underfitting, overfitting, and extrapolation risks.
- Model-ready data – design a Medallion Architecture in Unity Catalog, automate data quality validation with Great Expectations and Lakeflow pipeline expectations, and build features without data leakage.
- Features as managed data products – create feature tables in Unity Catalog, use FeatureLookup, and generate text embeddings with models hosted in Databricks.
- Spark MLlib pipelines – build repeatable data preparation and model training pipelines for regression and classification that behave consistently during training and inference.
- Informed model evaluation – select metrics that reflect the business cost of errors, including RMSE, MAE, R-squared, precision, recall, F1, and AUC, and compare models fairly on the same data.
- Complete model lifecycle – track experiments in MLflow, register models in Unity Catalog with champion and challenger aliases, and expose them through Model Serving.
- Systematic optimization – diagnose a model before tuning, use Optuna with TPE for hyperparameter search and trial scaling, and use AutoML as a fast, transparent benchmark.
- In practice, participants leave the course able to independently deliver an ML project on Azure Databricks, from raw data and managed features to tracked experiments and a registered, deployed, and optimized model.

Training program

Module 1 – Principles of Machine Learning

- What machine learning is: Mitchell's definition (task, experience, performance measure), feasibility of

learning from a sample, the training loop, and a model as a function.

- Underfitting and overfitting, matching model complexity to data, model boundaries, and extrapolation risks.
- The ML model lifecycle and reference architecture on Azure Databricks; Medallion Architecture (bronze, silver, gold) in Unity Catalog.
- Data quality: common issues, Great Expectations (GX Core), and Lakeflow pipeline expectations.
- Feature engineering: missing-value imputation, normalization, discretization, one-hot and target encoding, feature hashing, TF-IDF, embeddings, dataset balancing with SMOTE, and train/validation/test splitting without data leakage.
- Feature tables in Unity Catalog using FeatureEngineeringClient and FeatureLookup, and embedding models using Foundation Model APIs and ai_query.
- Types of ML tasks; linear regression, the cost function, and MAE, MSE, RMSE, and R-squared metrics; training with MLflow tracking.
- Azure Databricks for ML: architecture, Databricks Runtime ML, serverless and AI Runtime, notebooks and Genie Code, and libraries including scikit-learn, Spark MLlib, XGBoost, and PyTorch.
- Demonstrations: Data architecture (raw to bronze to silver), Feature store and embeddings, and Regression model using telecommunications data.
- Lab: Machine Learning in Azure Databricks – Regression.

Module 2 – Model Development Workflow in Azure Databricks

- Spark MLlib (pyspark.ml): transformers, estimators, Pipeline, and PipelineModel, ensuring consistency between training and inference.
- Feature preparation in MLlib: StringIndexer, OneHotEncoder, VectorAssembler, MinMaxScaler, StandardScaler, and Imputer.
- Classification algorithms: logistic regression and decision thresholds, decision trees, bootstrapping, random forests, boosting with GBT, XGBoost, and LightGBM, and neural networks.
- Classifier evaluation: confusion matrix, accuracy, precision, recall, F1, ROC curve, and AUC; MLlib evaluators, CrossValidator, and ParamGridBuilder.
- MLflow 3: experiments and runs, autologging and explicit logging, and Logged Models.
- Models in Unity Catalog: versions, champion and challenger aliases, permissions; Model Serving with real-time endpoints, batch inference using ai_query, and AI Gateway.
- Labs: Machine Learning in Azure Databricks – Multiclass Classification and Binary Classification with MLlib Pipelines.

Module 3 – Optimizing Model Performance

- Diagnosis before optimization: model-quality levers, the bias-variance trade-off, and interpretation of training and validation error.
- Reliable evaluation: k-fold cross-validation and its cost, selecting the optimization metric, and fair use of the test set.
- Hyperparameters versus model parameters; the CASH problem; search strategies including grid search, random search, and Bayesian optimization with TPE.
- Optuna on Azure Databricks: objective function, search space, Study and Trial, MLflow integration, and trial scaling with MlflowSparkStudy and Ray Tune; Hyperopt as a solution being phased out.

- AutoML: the scope of automation, requirements including Databricks Runtime ML and Dedicated access mode, configuration through the UI and Python API with classify, regress, and forecast, and generated notebooks.
- Choosing an approach: manual tuning, Optuna, or AutoML; fair champion/challenger comparison.
- Demonstration: Optimize the telecommunications model with the AutoML wizard.
- Labs: Optimize Hyperparameters in Azure Databricks and Train a Model with AutoML.

Expected preparation of the participant

- Basic knowledge of Python, including variables, functions, and working with libraries; pandas experience is beneficial.
- Basic knowledge of SQL, including SELECT queries, filtering, and aggregations.
- Basic understanding of descriptive statistics, including mean, standard deviation, and distribution.
- General familiarity with Microsoft Azure, including the portal, subscriptions, and resource groups.
- Previous exposure to Apache Spark or Jupyter/Databricks notebooks is beneficial.
- Previous machine learning experience is not required; ML principles are introduced from the ground up in Module 1.
- Completion of AI-901 or DP-900, or equivalent knowledge, is beneficial.

To increase the comfort of work and training's effectiveness we suggest using an additional monitor. The lack of additional monitor does not exclude participation in the training, however, it significantly influences the comfort of work during classes.
You can find information and requirements of participation in Distance Learning trainings at:
https://www.altkomakademia.pl/en/how-it-works/about-courses/

Training Includes

access to Altkom Akademia student portal

- Lecture (30%)
- Workshops (20%)
- Exercises (50%)

Main teaching tools: PowerPoint presentations, a practical Azure Databricks lab environment, Microsoft Learn resources, and Azure Databricks documentation.
Note: During selected training sessions, the Academy provides Participants with access to paid tools and services necessary to complete the training program (in particular Azure, Fabric, and other cloud resources). The Participant is required to use the provided resources solely within the scope of the

training agenda and in accordance with the trainer's instructions. Use of these resources beyond the training scope—particularly for private projects, testing additional services, modifying the environment configuration, or undertaking activities that generate costs unrelated to the exercises—will result in the Participant being charged for costs exceeding the allocated limit, based on expenses incurred by the Academy.

Language

- **Training**: English
- **Materials**: English

Examination method

The training does not conclude with an exam. However, it prepares participants for the Databricks Certified Machine Learning Professional exam.
Exam description: https://www.databricks.com/learn/certification/machine-learning-professional

Duration

1 days / 7 hours

Examination description

The training does not conclude with an exam. However, it prepares participants for the Databricks Certified Machine Learning Professional exam.
Exam description: https://www.databricks.com/learn/certification/machine-learning-professional