kod szkolenia: DBX-MLPREP / ENG DL 1d

Databricks – Data Preparation for Machine Learning

This intensive one-day training is aimed at participants already familiar with the Databricks environment who want to organize and deepen their skills in preparing data for machine learning. During the course, participants will learn to perform exploratory data analysis, handle missing values and categorical variables, standardize features, build complex transformation pipelines, and log and version results with MLflow and Feature Store.

Training recipients

The training is aimed at:

– Data scientists and ML engineers who want to learn proper data preparation for models.

– Teams designing ML pipelines and MLOps specialists managing data quality and logic.

– Data engineers and DataOps professionals preparing datasets for training and deploying models.

– Participants familiar with Databricks (SQL/PySpark) and ML basics, wishing to extend their toolbox with advanced data preparation.

Benefits

– Solid foundation in data preparation: participants will cover EDA, imputation, encoding, standardization, and learn to create modular transformation pipelines using Spark MLlib.

– Integration with MLflow and Feature Store: exercises showing how to log metrics, hyperparameters, and models in MLflow, as well as create and register feature sets in Feature Store to ensure training-serving consistency.

– Modular approach to data: the training shows how to combine DataFrames, Delta Lake, MLflow, and Feature Store in one process, facilitating ML implementation.

– Comprehensive feature engineering: participants will learn extraction, selection, and creation of new features, including window functions and VectorAssembler.

Training program

-1.Data exploration (EDA):
– Using display, summary, and profile commands to explore distributions and dependencies.
– Creating visualizations in Databricks notebooks; identifying outliers.
-2.Data splitting:
– Splitting into training, validation, and test sets; random and stratified sampling.
– Cross-validation and time-based split concept.
-3.Imputation of missing values:
– Methods for dropping (dropna) and filling (fillna) missing values.
– Using MLlib Imputer to replace with mean/median and create missing flags.
-4.Feature encoding and transformation:
– Encoding categories with StringIndexer and OneHotEncoder; introduction to target encoding.
– Scaling numeric variables with StandardScaler, MinMaxScaler, RobustScaler.
– Window functions: lag, lead, row_number, rolling average for sequential features.
5. Feature engineering and selection:
– Difference between extraction and selection; creating new variables via aggregations, log transforms, interactions.
– Combining variables into a vector with VectorAssembler.
-6.Building ML pipelines:
– Defining Pipeline stages (imputation, encoding, scaling, model).
– Fitting models and evaluation; logging metrics, hyperparameters, and models in MLflow.
-7.Feature Store & MLflow:
– Creating and registering feature tables in Databricks Feature Store; versioning and sharing features for multiple models.
– Using mlflow.log_model and mlflow.register_model to track experiments and ensure consistency between training and serving.
-8.Best practices and testing:
– Modular notebook design and unit testing of transformation functions.
– Documenting pipelines, monitoring data quality, and managing lineage.

Expected preparation of the participant

Participants should know basic ML concepts (e.g., regression and classification), use PySpark or SQL, and

have experience with data. Completing the Data Engineering module or equivalent skills is recommended.

Training Includes

- access to Altkom Akademia student

Product Includes
– Duration: 1 day / 8 h
– Lecture (30%) – theory and examples of data preparation including EDA, transformations, MLflow logging, and Feature Store feature management.
– Workshops (30%) – practical exercises in imputation, encoding, standardization, and ML pipeline building with CrossValidator/TrainValidationSplit.
– Exercises (40%) – complete data preparation process, pipeline logging in MLflow, and feature registration in Feature Store.
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.

Czas trwania

1 dni / 8 godzin

Language

Training: English
- Materials: English