kod szkolenia: DBX-MLPREP / PL DL 1d

Databricks – Data Preparation for Machine Learning

To intensywne, jednodniowe szkolenie jest skierowane do osób, które znają już środowisko Databricks i chcą uporządkować oraz pogłębić umiejętności przygotowywania danych pod uczenie maszynowe. W czasie kursu uczestnicy nauczą się wykonywać eksploracyjną analizę danych, radzić sobie z brakami i zmiennymi kategorycznymi, standaryzować cechy, tworzyć złożone pipeline'y transformacyjne oraz logować i wersjonować wyniki w MLflow i Feature Store.

Odbiorcy szkolenia

Szkolenie jest adresowane do:

- Data scientistów i inżynierów ML, którzy chcą nauczyć się właściwego przygotowania danych do modeli.
- Zespołów projektujących pipeline'y ML oraz specjalistów MLOps, którzy zarządzają jakością i logiką danych.
- Inżynierów danych i DataOps odpowiedzialnych za przygotowanie zestawów danych do uczenia i wdrażania modeli.
- Uczestników znających Databricks (SQL/PySpark) i podstawy modelowania ML, chcących rozszerzyć warsztat o zaawansowane przygotowanie danych.

Korzyści

- Solidny fundament data preparation: uczestnicy zgłębią EDA, imputację, kodowanie i standaryzację oraz nauczą się tworzyć modularne pipeline'y transformacyjne z wykorzystaniem Spark MLlib.

- Integracja z MLflow i Feature Store: ćwiczenia pokazujące logowanie metryk, hiperparametrów i modeli w MLflow oraz tworzenie i rejestrację zestawów cech w Feature Store, co zapewnia spójność między treningiem a serwowaniem modeli.
- Modularne podejście do danych: szkolenie pokazuje, jak łączyć DataFrames, Delta Lake, MLflow i Feature Store w jednym procesie, co ułatwia implementację modeli ML.
- Kompleksowa inżynieria cech: uczestnicy poznają techniki ekstrakcji, selekcji i tworzenia nowych cech, w tym użycie funkcji okienkowych i VectorAssemblera.

Program szkolenia

1.Eksploracja danych (EDA):
- Użycie polecenia display, summary i profile do eksploracji rozkładów i zależności.
- Tworzenie wizualizacji w notebookach Databricks; identyfikacja odstających obserwacji.

2.Podział danych:
- Podział na zbiory treningowe, walidacyjne i testowe; losowy i warstwowy sampling.
- Cross-validation i koncepcja time-based split.

3.Imputacja brakujących wartości:
- Metody usuwania braków (dropna) i uzupełniania (fillna).
- Użycie klasy Imputer z MLlib do zastępowania wartości średnią/medianą oraz tworzenia flag braków.

4.Kodowanie i transformacja cech:
- Kodowanie kategorii za pomocą StringIndexer i OneHotEncoder; wprowadzenie do kodowania docelowego.
- Skalowanie zmiennych numerycznych przy użyciu StandardScaler, MinMaxScaler i RobustScaler.
- Funkcje okienkowe (window functions): lag, lead, row_number, rolling average dla tworzenia cech sekwencyjnych.

5.Feature engineering i selekcja:
- Różnica między ekstrakcją cech a selekcją; tworzenie nowych zmiennych przez agregacje, transformacje logarytmiczne, interakcje.
- Łączenie zmiennych w wektor za pomocą VectorAssembler

6.Budowa pipeline’ów ML:
- Definiowanie etapów Pipeline (imputacja, kodowanie, skalowanie, model).
- Dopasowanie modeli i ewaluacja; logowanie metryk, hiperparametrów i modeli w MLflow.

7.Feature Store & MLflow:
- Tworzenie i rejestrowanie tabel cech w Databricks Feature Store; utrzymanie wersji i udostępnianie cech dla wielu modeli.
- Zastosowanie mlflow.log_model i mlflow.register_model do śledzenia eksperymentów oraz gwarancja spójności między treningiem a serwowaniem modeli.

8.Dobre praktyki i testowanie:

- Projektowanie modularnych notebooków oraz wykorzystanie testów jednostkowych dla funkcji transformacyjnych.
- Dokumentowanie pipeline’ów, monitorowanie jakości danych i zarządzanie lineage.

Oczekiwane przygotowanie uczestnika

Uczestnik powinien znać podstawowe koncepcje uczenia maszynowego (np. regresja i klasyfikacja), posługiwać się PySpark lub SQL oraz mieć doświadczenie w pracy z danymi. Zaleca się ukończenie modułu Data Engineering lub posiadanie równoważnych umiejętności

Szkolenie obejmuje

* dostęp do portalu słuchacza Altkom Akademii
Produkt zawiera
Czas trwania: 1 dzień / 8h
Wykład (30%) – teoria i pokaz przykładów przygotowania danych, w tym EDA, transformacje, logowanie w MLflow i zarządzanie cechami w Feature Store.
Warsztaty (30%) – praktyczne ćwiczenia z imputacji, kodowania, standaryzacji oraz budowy pipeline’u ML z wykorzystaniem CrossValidator/TrainValidationSplit.
Ćwiczenia (40%) – pełny proces przygotowania danych, logowania pipeline’u w MLflow oraz rejestracji cech w Feature Store.
Główne narzędzia: Databricks, PySpark, Delta Lake, Spark MLlib, MLflow, Feature Store, Unity Catalog.
Uwaga:W trakcie wybranych szkoleń Akademia zapewnia Uczestnikom dostęp do płatnych narzędzi i usług niezbędnych do realizacji programu szkolenia (w szczególności usług Azure, Fabric oraz innych zasobów chmurowych). Uczestnik zobowiązany jest do korzystania z udostępnionych zasobów wyłącznie w zakresie wynikającym z agendy szkolenia oraz zgodnie z poleceniami trenera. Wykorzystywanie tych zasobów w sposób wykraczający poza zakres szkoleniowy, w szczególności do realizacji projektów prywatnych, testowania dodatkowych usług, modyfikowania konfiguracji środowiska lub podejmowania działań generujących koszty niezwiązane z realizacją ćwiczeń – skutkować będzie obciążeniem Uczestnika kosztami przekraczającymi przewidziany limit, na podstawie wydatków poniesionych przez Akademię (https://www.altkomakademia.pl/ogolne-warunki-uczestnictwa-w-szkoleniach/)

Język

Wykład: polski
Materiały: angielski

Metoda egzaminacyjna

Szkolenie nie zawiera formalnego egzaminu. Po ukończeniu kursu zaleca się podejście do egzaminu Databricks Certified Machine Learning Associate jako kolejny krok rozwojowy.

Czas trwania

1 dni / 8 godzin

Opis egzaminu

Szkolenie nie zawiera formalnego egzaminu. Po ukończeniu kursu zaleca się podejście do egzaminu Databricks Certified Machine Learning Associate jako kolejny krok rozwojowy.