

Uczenie maszynowe w Azure Databricks: od zasad, przez proces tworzenia modeli, po optymalizację

Jest to autorskie szkolenie, które prowadzi uczestników przez pełną ścieżkę uczenia maszynowego na platformie Azure Databricks – od zrozumienia, czym jest uczenie z danych, przez przygotowanie danych i budowę pierwszego modelu, po zarządzany, wdrożony i zoptymalizowany model produkcyjny. Program łączy solidne podstawy teoretyczne z aktualnym stanem platformy: Databricks Runtime ML, Spark MLlib, MLflow 3, Models in Unity Catalog, Model Serving, Optuna i AutoML.

Szkolenie kładzie nacisk na decyzje, które w praktyce przesądzają o jakości modelu: sformułowanie problemu, unikanie wycieku danych, dobór metryki do kosztu biznesowego błędu, diagnozę niedouczenia i przeuczenia (bias-variance) oraz świadomy wybór między strojeniem ręcznym, Optuna i AutoML. Materiały uwzględniają najnowsze zmiany platformy – m.in. zastąpienie Hyperopt przez Optuna, aliasy modeli zamiast etapów w rejestrze oraz wymagania AutoML dotyczące wersji Databricks Runtime ML.

Szkolenie składa się z 3 modułów, 4 dem na rzeczywistych danych telekomunikacyjnych oraz 5 laboratoriów „krok po kroku” realizowanych w jednym środowisku Azure Databricks tworzonym w pierwszym module 1.



Odbiorcy szkolenia

Szkolenie przeznaczone jest dla:

- Analityków danych i początkujących data scientistów, którzy znają Pythona i SQL, a chcą przejść od analizy danych do budowy, oceny i wdrażania modeli uczenia maszynowego w skali.
- Inżynierów danych pracujących z Azure Databricks, którzy przygotowują dane i cechy dla zespołów ML i chcą rozumieć cały cykl życia modelu – od architektury medalionowej po serwowanie predykcji.
- Programistów Python, którzy chcą wejść w uczenie maszynowe w sposób uporządkowany: z teorią tam, gdzie jest potrzebna, i z kodem, który działa na rozproszonej platformie Spark.
- Architektów i liderów zespołów danych, odpowiedzialnych za standardy pracy z modelami – śledzenie eksperymentów, ład w Unity Catalog, wersjonowanie i wdrażanie modeli.
- Specjalistów migrujących z innych narzędzi ML, którzy chcą poznać aktualne, zalecane przez Databricks praktyki (MLflow 3, Models in Unity Catalog, Optuna, AutoML).



Korzyści

Po ukończeniu szkolenia uczestnik osiągnie następujące korzyści:

- Solidne podstawy uczenia maszynowego – zrozumiesz, dlaczego uczenie na próbie w ogóle działa, jak wygląda pętla treningowa i jak rozpoznać niedouczenie, przeuczenie oraz pułapki ekstrapolacji.
- Dane gotowe do modelowania – zaprojektujesz architekturę medalionową w Unity Catalog, zautomatyzujesz walidację jakości danych (Great Expectations, Lakeflow pipeline expectations) i zbudujesz cechy bez wycieku danych.
- Cechy jako zarządzane produkty danych – utworzysz tabele cech w Unity Catalog, wykorzystasz FeatureLookup i wygenerujesz embeddingi tekstu modelami hostowanymi w Databricks.
- Potoki Spark MLlib – zbudujesz powtarzalne pipeline'y przygotowania danych i trenowania modeli regresji oraz klasyfikacji, które działają identycznie podczas treningu i wnioskowania.
- Świadoma ocena modeli – dobierzesz metrykę do kosztu biznesowego błędu (RMSE, MAE, R^2 , precision, recall, F1, AUC) i porównasz modele uczciwie, na tych samych danych.
- Pełny cykl życia modelu – będziesz śledzić eksperymenty w MLflow, rejestrować modele w Unity Catalog z aliasami champion/challenger i udostępniać je przez Model Serving.
- Systematyczna optymalizacja – zdiagnozujesz model przed strojeniem, zastosujesz Optuna (TPE) do przeszukiwania hiperparametrów i skalowania prób oraz wykorzystasz AutoML jako szybki, „szklany” punkt odniesienia.

W praktyce oznacza to, że wyjdiesz ze szkolenia ze zdolnością samodzielnego przeprowadzenia projektu ML na Azure Databricks – od surowych danych, przez zarządzane cechy i śledzone eksperymenty, po zarejestrowany, wdrożony i zoptymalizowany model.



Program szkolenia

Moduł 1 – Zasady uczenia maszynowego (Principles of machine learning)

- Czym jest uczenie maszynowe: definicja Mitchella (zadanie, doświadczenie, miara jakości), wykonalność uczenia na próbie, pętla treningowa, model jako funkcja
- Niedouczenie i przeuczenie, dopasowanie złożoności modelu do danych, granice modelu i pułapki ekstrapolacji
- Cykl życia modelu ML i architektura referencyjna na Azure Databricks; architektura medalionowa (bronze, silver, gold) w Unity Catalog
- Jakość danych: najczęstsze problemy, Great Expectations (GX Core) i Lakeflow pipeline expectations
- Inżynieria cech: imputacja braków, normalizacja, dyskretyzacja, one-hot i target encoding, feature hashing, TF-IDF, embeddingi; balansowanie zbioru (SMOTE); podział train/validation/test bez wycieku danych
- Tabele cech w Unity Catalog (FeatureEngineeringClient, FeatureLookup) i modele embeddingów (Foundation Model APIs, ai_query)
- Typy zadań ML; regresja liniowa, funkcja kosztu i metryki MAE, MSE, RMSE, R^2 ; trenowanie ze śledzeniem w MLflow
- Azure Databricks dla ML: architektura, Databricks Runtime ML, serverless i AI Runtime, notatniki i Genie Code, biblioteki (scikit-learn, Spark MLlib, XGBoost, PyTorch)
- Dema: Data architecture (raw → bronze → silver), Feature store & embeddings, Regression model – na danych telekomunikacyjnych
- Laboratorium: Machine learning in Azure Databricks – Regression

Moduł 2 – Proces tworzenia modeli w Azure Databricks (Model development workflow)

- Spark MLlib (pyspark.ml): transformers, estimators, Pipeline i PipelineModel – brak rozbieżności między trenowaniem a wnioskowaniem
- Przygotowanie cech w MLlib: StringIndexer, OneHotEncoder, VectorAssembler, MinMaxScaler i StandardScaler, Imputer
- Algorytmy klasyfikacji: regresja logistyczna i próg decyzyjny, drzewa decyzyjne, bootstrapping, lasy losowe, boosting (GBT, XGBoost, LightGBM), sieci neuronowe
- Ocena klasyfikatorów: macierz pomyłek, accuracy, precision, recall, F1, krzywa ROC i AUC; ewaluatory MLlib, CrossValidator i ParamGridBuilder
- MLflow 3: eksperymenty i przebiegi, autologging i logowanie jawne, Logged Models
- Models in Unity Catalog: wersje, aliasy (champion, challenger), uprawnienia; Model Serving – endpointy czasu rzeczywistego, wnioskowanie wsadowe (ai_query), AI Gateway
- Laboratoria: Machine learning in Azure Databricks – Multiclass classification oraz Binary classification with MLlib pipelines

Moduł 3 – Optymalizacja wydajności modeli (Optimizing model performance)

- Diagnoza przed optymalizacją: dźwignie jakości modelu, kompromis bias-variance, odczytywanie błędów treningowego i walidacyjnego
- Wiarygodna ocena: k-krotna walidacja krzyżowa i jej koszt, wybór metryki optymalizacji, uczciwe

użycie zbioru testowego

- Hiperparametry a parametry modelu; problem CASH; strategie przeszukiwania: grid search, random search, optymalizacja bayesowska (TPE)
- Optuna na Azure Databricks: funkcja celu, przestrzeń przeszukiwania, Study i Trial, integracja z MLflow; skalowanie prób (MLflowSparkStudy, Ray Tune); Hyperopt jako rozwiązanie wycofywane
- AutoML: zakres automatyzacji, wymagania (Databricks Runtime ML, Dedicated access mode), konfiguracja w interfejsie i przez Python API (classify, regress, forecast), generowane notatniki
- Wybór podejścia: strojenie ręczne, Optuna czy AutoML; uczciwe porównanie champion/challenger
- Demo: Optimize the telco model with the AutoML wizard
- Laboratoria: Optimize hyperparameters in Azure Databricks oraz Train a model with AutoML



Oczekiwane przygotowanie uczestnika

- Podstawowa znajomość języka Python (zmienne, funkcje, praca z bibliotekami; mile widziane pandas).
- Podstawowa znajomość SQL (zapytania SELECT, filtrowanie, agregacje).
- Podstawowe pojęcia statystyki opisowej (średnia, odchylenie standardowe, rozkład).
- Ogólna znajomość platformy Microsoft Azure (portal, subskrypcja, grupy zasobów).
- Mile widziana wcześniejsza styczność z Apache Spark lub notatnikami Jupyter/Databricks.
- Nie jest wymagane wcześniejsze doświadczenie z uczeniem maszynowym – zasady ML omawiane są od podstaw w module 1.
- Mile widziane ukończenie AI-901 lub DP-900 albo równoważna wiedza.

Dla zwiększenia komfortu pracy oraz efektywności szkolenia zalecamy skorzystanie z dodatkowego ekranu. Brak dodatkowego ekranu nie jest przeciwwskazaniem do udziału w szkoleniu, ale w znaczący sposób wpływa na komfort pracy podczas zajęć.

Informacje oraz wymagania dotyczące uczestniczenia w szkoleniach w formule zdalnej dostępne na: <https://www.altkomakademia.pl/jak-dzialamy/o-szkoleniach/?distance-learning>



Szkolenie obejmuje

* dostęp do portalu słuchacza AltKom Akademii

Metoda szkolenia:

- Wykład (30%)
- Warsztaty (20%)
- Ćwiczenia (50%)

Główne narzędzia dydaktyczne: prezentacje PowerPoint, praktyczne środowisko laboratoryjne Azure Databricks oraz zasoby Microsoft Learn i dokumentacji Azure Databricks.

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 kończy się egzaminem. Przygotowuje jednak do egzaminu Databricks Certified Machine Learning Professional.

Czas trwania

1 dni / 7 godzin

Opis egzaminu

Szkolenie nie kończy się egzaminem. Przygotowuje jednak do egzaminu Databricks Certified Machine Learning Professional.

Opis egzaminu: <https://www.databricks.com/learn/certification/machine-learning-professional>