Product Recommendation Challenge

Team 3

Wybraliśmy ALS (Alternating Least Squares)

Sieci neuronowe (szczególnie głębokie), XGBoost, LightGBM okazały się zbyt ciężkie i kosztowne obliczeniowo dla naszego sprzętu albo osiągały zbyt słabe wyniki

ALS idealnie pasuje do danych w formacie użytkownik – produkt z ocenami od 1 do 5 oraz czy w ogóle ocenił
Skupiliśmy się wyłącznie na macierzy ocen użytkowników względem produktów używając tylko **train.csv** , pomijając metadane produktów co spowodowało redukcję wszystkich danych do przetworzenia z 1GB do 200mb i to wszystko w zapisie liczbowym bez tekstu i problemu z ogarnięciem niepełnych danych

Model szkolił się szybko i na **CPU** co było największym atutem. Czas treningu wynosi tylko **40s + 50s generacji submission.csv!**

Dodatkowo w naszych testach osiągnął lepsze wyniki niż inne podejścia, które próbowaliśmy co było niesamowitym wynikiem

Proces treningu

1. Wczytanie i przygotowanie danych – tworząc wymiary macierzy i ilości userów x produkty + ustawienie początkowych parametrów.

2. Tryb pojedynczego modelu – trening jednego modelu na wybranym zestawie parametrów, opcjonalnie bez walidacji.

3. Z Walidacją – podział danych w czasie, wybór użytkowników, tworzenie rzadkiej macierzy ocen, obliczanie skuteczności (MAP@10).

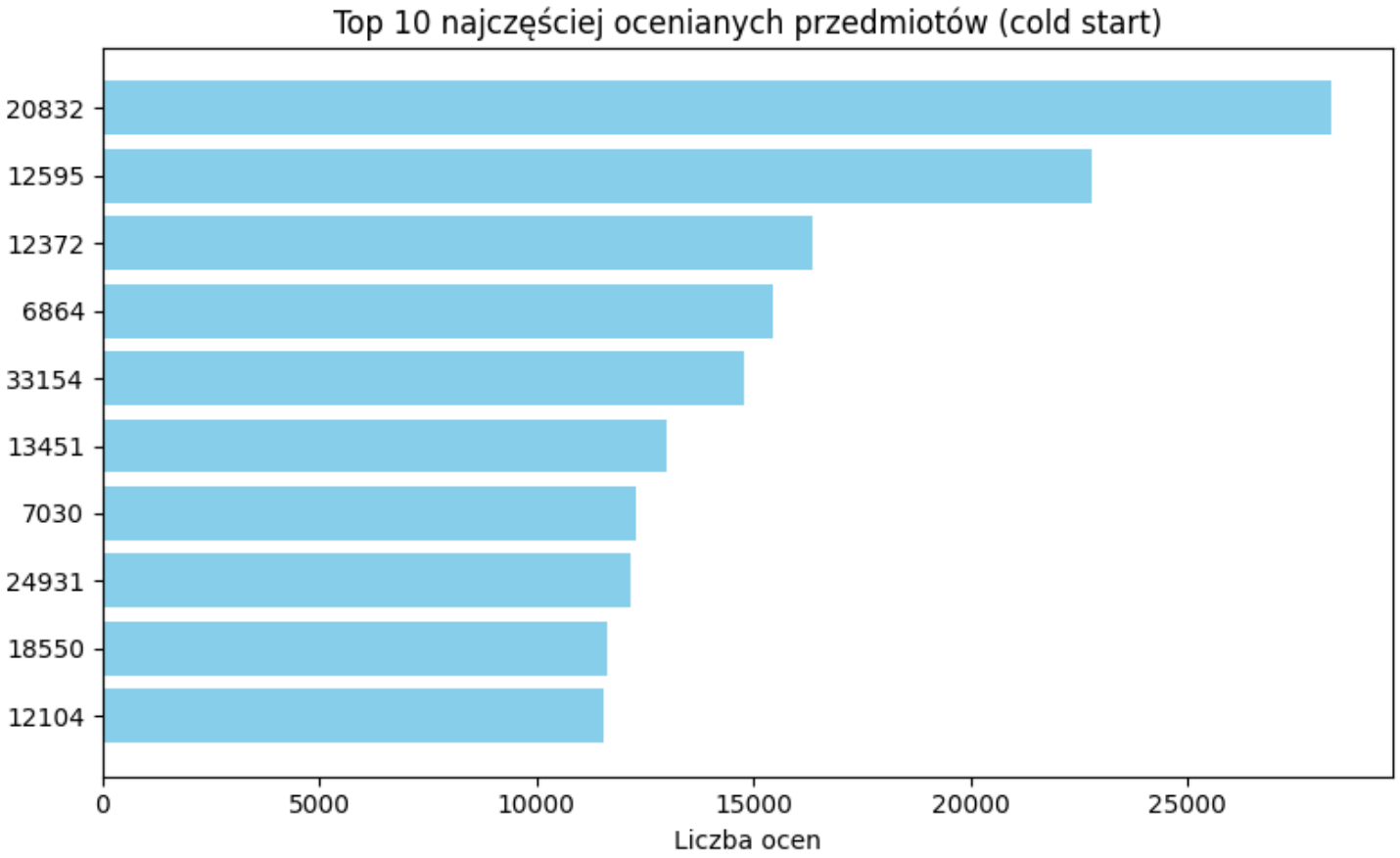
4. Tryb wielu konfiguracji – trenowanie wielu modeli z różnymi parametrami, porównywanie wyników walidacji, i wybór najlepszego.

5. Końcowy trening i predykcja – trening finalnego modelu na pełnych danych z użyciem najlepszych parametrów.

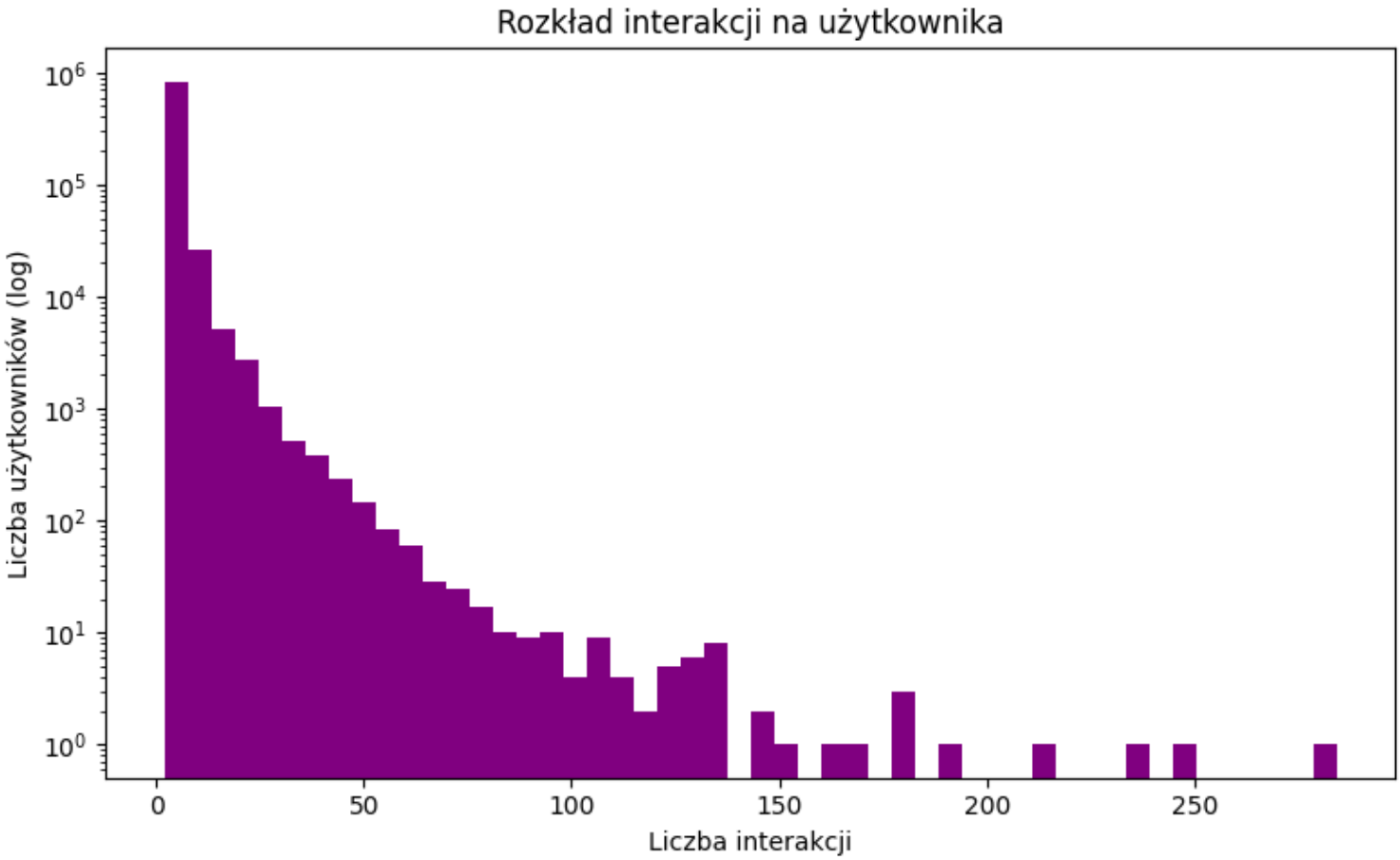
6. Generacja submission - model generuje top 10 dla każdego usera w test.csv

Wykresy

1. Przedstawia 10 najpopularniejszych produktów (najwięcej ocen). Wskazuje, że część produktów cieszy się dużym zainteresowaniem, co może być wykorzystywane w rekomendacjach, zwłaszcza dla nowych użytkowników.



2. Pokazuje, jak wielu użytkowników dokonało ile interakcji (ocen). Widzimy, że większość użytkowników jest mało aktywna – oceniła bardzo niewiele produktów. Oś Y jest logarytmiczna, więc różnice są ogromne – to znaczy, że tylko niewielka grupa użytkowników generuje większość danych.



Podsumowanie

Użycie ALS wyeliminowało problem z używaniem całego datasetu próbując znaleźć korelacje między użytkownikami produktami a ocenami jako wielka macierz

Te rozwiązanie przyspieszyło trening o wiele więcej niż wszystkie poprzednie opcje

Cold start - jest wyliczany na podstawie top 10 ocenianych produktów tak aby stworzyć możliwą predykcje dla usera który był użyty wcześniej w danych walidacyjnych albo nie pojawia się w test.csv

Optymalizacje jak czyszczenie za pomocą gc jak i wywalanie nieużywanych już macierzy do przygotowywania końcowej sprawiło możliwość używania na **CPU**

Możliwość tuningu - nasz model ma możliwość walki o najlepsze wyniki MAP@10 między sobą podając mu więcej **ALS_CONFIGU** w konfiguracji do sprawdzania różnych podejść i maksymalizowanie wyników

```
SINGLE_MODEL_TRAINING = True
SKIP_VALIDATION = False
SKIP_SUBMISSION = True

ALS_CONFIG = [{
'name': 'ALS: Najlepszy jak na razie',
'factors': 15,
'regularization': 0.001,
'iterations': 40
}]
```

Powoduje to możliwość lepszej optymalizacji najlepszych parametrów
Poprawiliśmy wynik końcowy z **0.03837**
na **0.05259** dzięki optymalizacji parametrów i kodu