Warsztaty z technik uczenia maszynowego

Rekomendacje produktowe na podstawie danych z poprzednich transakcji

Mikołaj Kida
Yelyzaveta Liubonko
Agnieszka Orzechowska
Wojciech Trybulec

---

1 Opis tematu

HM Group to rodzina marek i firm z 53 rynkami internetowymi i około 4850 sklepami. Ich sklepy internetowe oferują kupującym szeroki wybór produktów. Przy zbyt wielu możliwościach klienci mogą mieć trudność ze znalezieniem tego, co ich interesuje lub czego szukają.Zagubiony klient może ostatecznie nie dokonać zakupu. Aby poprawić wrażenia z zakupów, kluczowe znaczenie mają rekomendacje produktów. Co ważniejsze, pomaganie klientom w dokonywaniu właściwych wyborów ma również pozytywny wpływ na zrównoważony rozwój, ponieważ np. zmniejsza ilość zwrotów.

Celem tego projektu jest opracowywanie narzędzia pozwalającego na znajdowanie rekomendacji produktowych dla danego klienta na podstawie danych z poprzednich transakcji, a także metadanych klientów i produktów. Do dyspozycji mamy proste dane jak rodzaj odzieży, czy wiek klienta, dane tekstowe z opisów produktów, a także dane obrazów ze zdjęciami odzieży.

2 Podział zadań

Nazwisko	Rola w zespole
Orzechowska Agnieszka	Obróbka danych
Liubonko Yelyzaveta	Testy
Kida Mikołaj	Implementacja w chmurze
Trybulec Wojciech	Budowa modelu

3 Założenia techniczne

3.1 Język programowania

Aplikacja zostanie napisana w języku Python. Wstępnie zakładamy wykorzystanie biblioteki TensorFlow oraz Keras do stworzenia modelu/modeli oraz skorzystanie z usług chmurowych AWS w celu przeprowadzania obliczeń. Wybór platformy AWS jest motywowany znacznym rozmiarem danych wejściowych, którymi dysponujemy.

3.2 Dane

Będziemy korzystać z trzech zbiorów danych: *articles*, *customers* i *transactions_train* oraz ze zbioru zdjęć ubrań.
W tabeli *articles* (z 25 kolumnami) mamy informacje o produktach, czyli na przykład: *article_id* - unikatowy identyfikator produktu, *product_code* - 6-cyfrowy kod produktu, *prod_name* - nazwa produktu, *product_group_name* - nazwa grupy produktu (łącznie 19 grup), *section_name* - nazwa sekcji, czy *detail_desc* - opis produktu.
Z kolei w tabeli *customers* mamy informacje o klientach: *customer_id* - unikalny identyfikator klienta, *FN* - mówi, czy klient otrzymuje Fashion News, dwie unikalne wartości (1, 0), *Active* - mówi, czy klient jest otwarty do komunikacji, dwie unikalne wartości (1, 0), *club_member_status* - status w klubie, trzy unikalne wartości (Active, Pre-create, Left club), *fashion_news_frequency* - częstotliwość wysyłania wiadomości do klienta, trzy unikalne wartości (NONE, Regularly, Monthly), *age* - wiek klienta, *postal_code* - kod pocztowy.
W zbiorze danych *transactions_train* mamy natomiast informacje o transakcjach: data transakcji w formie YYYY-MM-DD (string), identyfikator klienta, identyfikator produktu, kwota zamówienia oraz kanał sprzedaży: dwie unikalne wartości (1 - sklep stacjonarny, 2 - online).

3.3 Wybór modelu

W kwestii modelu planujemy wzorować się na architekturze głębokiej sieci neuronowych zaproponowanych w `https://dl.acm.org/doi/10.1145/2959100.2959190`. Powyższa architektura została użyta przez YouTube do implementacji ich systemu rekomendacji. Jako, że problem który staramy się rozwiązać konceptualnie jest bardzo podobny, mamy nadzieję, że to podejście będzie skuteczne.