Projekt IUM – 2020Z – zadanie 2 – etap 1

Adam Steciuk (300263)
Wojciech Moczydłowski (296258)

Słownik dziedziny problemu:

1. *Użytkownik/Klient* – osoba odwiedzająca stronę sklepu internetowego „eSzoppping”
2. *Zleceniodawca* – osoba zlecająca rozwiązanie problemu, dostarczająca dane i oceniająca otrzymane wyniki (tu: prowadzący przedmiotu IUM)
3. *Produkty* – asortyment sklepu
4. *Profil użytkownika* – wektor informacji kontekstowych użytkownika
5. *Rekomendacja* – wyświetlenie użytkownikowi spersonalizowanej listy produktów, którymi potencjalnie może być zainteresowany
6. *Trafna rekomendacja* – sytuacja, w której użytkownik kliknie na wyświetloną mu rekomendacje

Obecna sytuacja:

Klienci wchodzący na stronę nie mogą się zdecydować, który produkt obejrzeć. Średnia dzienna liczba odsłon produktów wynosi 300.

Zadanie biznesowe:

Sugerowanie (interesujących dla danego klienta) produktów na stronie głównej sklepu.

Oczekiwania:

System docelowo ma zwiększyć przychód ze sprzedaży produktów poprzez zwiększenie liczby przeglądanych produktów

Biznesowe kryterium sukcesu:

System powinien generować średnio co najmniej 60 odsłon produktów dziennie więcej niż obecnie (20% więcej niż obecnie)

Analityczne kryterium sukcesu:

Osiągnięcie:

$$\frac{\text{liczba trafnych rekomendacji}}{\text{liczba wszystkich rekomendacji}} \geq \frac{60}{\text{dzienna liczba wejść na stronę sklepu}}$$

Właściwości/cechy docelowego rozwiązania:

- Nie jednorazowa rekomendacja a jako wdrożenie system działającego ciągle
- Maksymalnie 1/10 s czasu odpowiedzi

Ograniczenia:

- Model może zostać zbudowany jedynie w oparciu o dostarczone przez zleceniodawcę dane
- Zebranie dodatkowych danych o użytkownikach jest czasochłonne (nie jesteśmy w stanie uzyskać danych od eksperta domenowego)
- Czas potrzebny do wytrenowania modelu
- Zasoby sprzętowe: moc obliczeniowa potrzebna do trenowania modelu

Zadanie modelowania:

W ramach rozważanego problemu rozwiążemy jedno zadanie modelowania. Będzie ono polegało na utworzeniu spersonalizowanej listy produktów, tak, aby niezdecydowany użytkownik kliknął na jeden

z nich. Wejściem modelu będzie lista produktów przeglądanych oraz kupionych przez użytkownika. Wyjściem natomiast spersonalizowana rekomendacja produktów.

Typ zadania:

Zadanie rekomendacyjne

Dane wykorzystane w procesie modelowania

- Baza użytkowników
  - Id użytkownika
  - Miasto zamieszkania
- Katalog produktów
  - Id produktu
  - Nazwa produktu
  - Kategoria produktu
  - Cena
- Historia sesji
  - Id sesji
  - Data i godzina akcji
  - Id powiązanego użytkownika
  - Id powiązanego produktu
  - Typ akcji
  - Oferowana zniżka
  - Id zakupu

Wstępna analiza dostarczonych danych:

- **Baza użytkowników**: przeprowadzona analiza nie wykryła błędów w podanym zbiorze.
- **Katalog produktów**: występują pojedyncze błędy w cenach produktów takie jak wartości ujemne oraz wartości zauważalnie odbiegające od średniej (bez wątpienia będące błędami przy wprowadzaniu np. Gra komputerowa o cenie 89990000 zł). Jeden z produktów ma wadliwą nazwę zawierającą hasła reklamowe. Liczebność tego typu produktów nie jest znaczące, toteż zostaną usunięte ze zbioru uczącego, gdyż mogą mieć negatywny wpływ na proces uczenia.

all	319	
valid	282	88.40%

- **Historia sesji:** przeprowadzając analizę zwróciliśmy uwagę na braki w atrybutach: „user_id” oraz ”product_id”. Pierwszy atrybut jesteśmy w stanie odtworzyć za pomocą innych sesji, które posiadają identyczne id. Jeżeli, w chociaż jednej został poprawnie zapisany atrybut „user_id” jesteśmy w stanie przekopiować go do reszty sesji o takim samym id. Atrybutu ”product_id” nie jesteśmy w stanie odtworzyć. W związku z tym, sesje o pustym atrybucie ”product_id” zostaną usunięte. Istnieje mała liczba sesji z pustym atrybutem ”product_id”, wobec czego ilość danych nie powinna ulec zauważalnemu skurczeniu. W historii sesji nie występuje ani jedna błędna wartość „user_id” czy ”product_id” nienależąca do znajdujących się w bazach products i users informacji.

all	90159	
initialy valid	81461	90.35%
valid + reproducible	85700	95.05%

- **Dane dotyczące wysyłek**: tworzony model nie będzie korzystał z tych danych.

Występujące błędy jesteśmy również w stanie wyeliminować poprzez usunięcie błędnych danych oraz otrzymanie od zleceniodawcy poprawnych na ich miejsce.

Badanie gęstości dostarczonych danych

Na podstawie danych z zapisanych sesji użytkowników została zbudowana macierz interakcji użytkowników z produktami o strukturze:

	P1	P2	P3	…
U1	1	1	0	…
U2	0	0	1	…
U3	1	1	1	…
…	…	…	…	…

P1, P2, P3, … - kolejne produkty
U1, U2, U3, … - kolejni użytkownicy

Jedynki w macierzy oznaczają interakcje danego użytkownika z danym produktem (zakup lub wyświetlenie). Gęstość tak utworzonej macierzy wyniosła, około 65%, co przekłada się na dużą gęstość informacji dla problemu rekomendacyjnego.

Badanie współczynnika informacji wzajemnej

Dla dostarczonych danych przeprowadziliśmy badanie współczynnika informacji wzajemnej. Przyjrzeliśmy się wpływowi kupna produktu w chwili t na kupno produktu w chwili t+1 przez danego użytkownika. Obliczenia przeprowadziliśmy dla dostarczonych danych oraz takiej samej ilości danych wygenerowanych losowo. Otrzymane wyniki:

- Dostarczone dane: 0.7172515674911283
- Dane wygenerowane losowo: 0.32369363139038637
- Dane wygenerowane losowo: 0.32374928138208137
- Dane wygenerowane losowo: 0.3226033887492894
- Dane wygenerowane losowo: 0.3235100745176365
- Dane wygenerowane losowo: 0.32435079532269634
- Dane wygenerowane losowo: 0.32264671285861923
- Dane wygenerowane losowo: 0.32405624902493385
- Dane wygenerowane losowo: 0.32312335018792404
- Dane wygenerowane losowo: 0.3225205685152983
- Dane wygenerowane losowo: 0.322738333262667

Z otrzymanych obliczeń wynika, że współczynnik informacji wzajemnej dla dostarczonych danych jest ponad dwukrotnie większy niż dla danych wygenerowanych losowo. Wobec tego wnioskujemy, iż kolejne produkty kupowane przez danego użytkownika są ze sobą powiązane.

Repozytorium projektu

https://github.com/strachus/IUM