Rekrutacja: Intern - Data Science - Zadanie rekrutacyjne

Wstępne założenia:

Chcemy jak najbardziej dopasować rekomendacje produktów dla użytkowników, bazując na:

- ich trzech ulubionych kategoriach produktu
- ich geolokacji: stanu oraz miasta
- nie znamy id użytkownika, więc nie możemy bazować na ich poprzednich aktywnościach (zakupach)

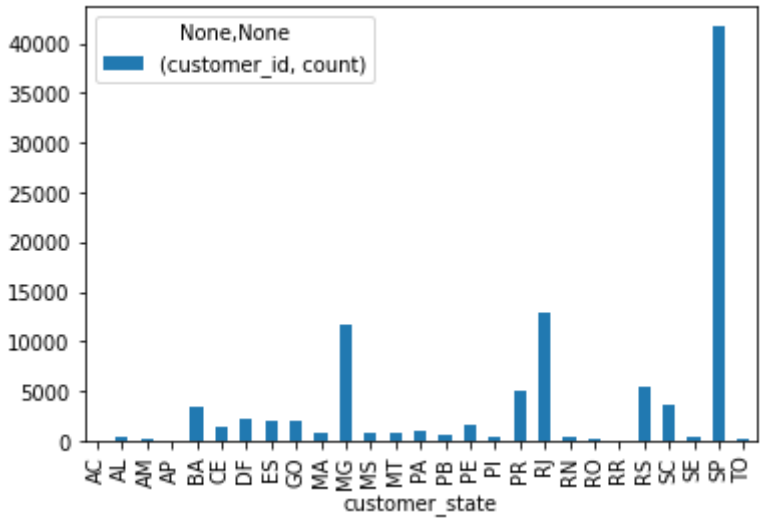
Jako dane wyjściowe oczekujemy rekomendację 10 id produktów.

Wstępna analiza danych:

- większość użytkowników z bazy zakupili tylko raz produktu
- zaledwie 3300 użytkowników dokonało ponownych tranzakcji
- najwięcej zamówień zrobionych przez jednego użytkownika to 17

customer_unique_id
99441
96096
8d50f5eadf50201ccdcedfb9e2ac8455
17

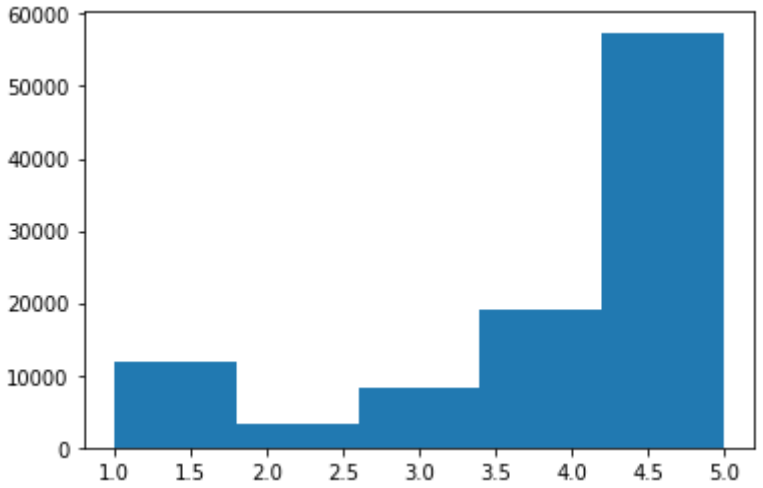
- występuje silna dominacja kilku stanów w aspekcie aktywności użytkowników



-

- większość zamówień opiewa na 1 produkt, maksymalna ilość kupionych produktów przy jednej transakcji to 21

- rozkład ocen zamówień jest nierównomierny, większość użytkowników ocenia zamówienia pozytywnie



Proponowane rozwiązanie:

Podjęto próbę zaimplementowania rekomendacji na podstawie algorytmu k nearest neighbours. Na pierwszym etapie szukamy podobnych użytkowników na podstawie lokalizacji oraz opracowanego współczynnika podobieństwa koszyka zakupowego. Dane dotyczące miasta i stanu użytkownika zamieniona na współrzędne geograficzne. Natomiast współczynnik podobieństwa został obliczony następująco:

- za 50% wartości współczynnika odpowiada to czy zawartość koszyka danego zamówienia mieści się w zainteresowaniu naszego użytkownika
- drugie 50% to różnorodność koszyka i pokrycie wszystkich kategorii zainteresowania, wynik maksymalny uzyskamy jeżeli koszyk zawiera wszystkie trzy kategorie
- współczynnik jest skalowany o ocenę zamówienia (szukamy zadowolonych użytkowników)

Tak zdefiniowany trójwymiarowy wektor posłuży algorytmowi KNN do znalezienia sąsiedztwa. Nie bierzemy pod uwagę unikalnego id użytkownika, ponieważ jeżeli dany użytkownik (zdecydowana mniejszość) dokonał kilkukrotnie zakupu w obrębie naszego zainteresowania, to powinien się pojawić kilkukrotnie w poszukiwanym sąsiedztwie.
Następnie sprawdzane są bestsellery w obrębie dopasowanych użytkowników.
Proponowane będą (**implementacja nie została dokończona ze względu na koniec czasu**):

- losowe 3 z 10 najbardziej popularny produkt w ulubionej kategorii nr 1 wśród podobnych użytkowników
- losowe 3 z 10 najbardziej popularny produkt w ulubionej kategorii nr 2 wśród podobnych użytkowników
- losowe 3 z 10 najbardziej popularny produkt w ulubionej kategorii nr 3 wśród podobnych użytkowników
- losowy 1 z 10 najbardziej popularny produkt w pozostałych kategoriach (ekspozycja użytkownika również na inne kategorie poza ulubionymi)

Użytkownik anonimowy:

Dla anonimowego użytkownika sugerowanym rozwiązaniem byłoby:

- obliczyć średnią dzienną sprzedaż dla każdego produktu
- obliczyć dzienną sprzedasz za ostatni tydzień
- zaproponować użytkownikowi te produktu, które w ostatnim czasie zyskały na popularności w stosunku do historycznych danych

Tego rozwiązania również nie zaimplementowano w pełni - zaproponowano top 10 najbardziej popularnych produktów na dzień 2018-08-27.

Rozwiązania:

Użytkownik 1: (cama_mesa_banho, papelaria, fashion_calcados), (sao paulo, SP)

99a4788cb24856965c36a24e339b6058
f1c7f353075ce59d8a6f3cf58f419c9c
06edb72f1e0c64b14c5b79353f7abea3
ec2d43cc59763ec91694573b31f1c29a
777d2e438a1b645f3aec9bd57e92672c
84f456958365164420cfc80fbe4c7fab
5411e9269501a870cabf632f05655131
fb55982be901439613a95940feefd9ee
363218ba55c610b750224f90bdd34be1
64fb265487de2238627ce43fe8a67efc

Użytkownik 2: (esporte_lazer, moveis_decoracao, telefonia), (rio de janeiro, RJ)

aca2eb7d00ea1a7b8ebd4e68314663af
b532349fe46b38fbc7bb3914c1bdae07
9ecadb84c81da840dbf3564378b586e9
78efe838c04bbc568be034082200ac20
c6336fa91fbd87c359e44f5dca5a90ed
e44f675b60b3a3a2453ec36421e06f0f
11875b30b49585209e608f40e8082e65
d3c044bd42d84a79e3b0c42662806a48
eb8c629f70275fd1c4f809116cce1efc
c7fd13b5e515bffdab855d0812842edb

Użytkownik niezalogowany: (), ()

aca2eb7d00ea1a7b8ebd4e68314663af
99a4788cb24856965c36a24e339b6058
422879e10f46682990de24d770e7f83d
389d119b48cf3043d311335e499d9c6b
368c6c730842d78016ad823897a372db
53759a2ecddad2bb87a079a1f1519f73
d1c427060a0f73f6b889a5c7c61f2ac4
53b36df67ebb7c41585e8d54d6772e08
154e7e31ebfa092203795c972e5804a6
3dd2a17168ec895c781a9191c1e95ad7

Pomysł usprawnień:

1) Gdybyśmy mieli dostęp do id użytkownika, a sama baza miałaby więcej użytkowników, którzy dokonywali częściej zakupów, moglibyśmy lepiej dopasować podobnych użytkowników. Na podstawie poprzednich zakupów moglibyśmy ocenić gust bazowego użytkownika i następnie poszukać innych o podobnych cechach.
2) Dla anonimowego użytkownika prawdopodobnie na podstawie jego IP moglibyśmy określić jego lokalizacje (wyjątek VPN etc.)