2021 Rozwiazanie Data Science Intern

Maciej Chylak

Maj 2021

1 Wstęp

Zadanie polegało na wybraniu 10 produktów, którymi najbardziej zainteresowałby się odpowiednio zalogowany jak i niezalogowany użytkownik. W przypadku zalogowanego użytkownika otrzymujemy 3 preferowane przez niego kategorie oraz jego adres zamieszkania, natomiast w przypadku użytkownika niezalogowanego nie dostajemy żadnych informacji na wejściu.

2 Realizacja

2.1 Ogólny zamysł

Stworzenie dwóch tabel:

- po dwa najchętniej kupowane produkty dla danej kategorii
- jeden najchętniej kupowany produkt w regionie klienta

Skala oceny chęci miałby być jak najbardziej niezależna od osoby od której kupujemy produkt. Miałaby ona być zależna od łącznej liczby zakupów, ogólnego zadowolenia z zakupu produktu oraz stosunku kwoty dostawy tego produktu do jego ceny. Istotne było dla mnie również to, aby skala ta nie zależała od czasu realizacji dostawy.

2.2 Wybór ramek danych

W mojej pracy zdecydowałem się na korzystanie z następujących ramek danych:

- orders_review - ze względu na zmienną review_score
- product_category_name_translation
- orders - ze względu na możliwość obliczenia różnicy w czasie pomiędzy oczekiwanym czasem dostawy a rzeczywistym oraz ze względu na zmienną order_status

- customers - zawiera ona informacje na temat miejsca zamieszkania klientów
- order_items - zawiera informacje na temat ceny dostawy oraz zamówienia. Będziemy w stanie policzyć ich stosunek

2.3 Przetworzenie ramek

2.3.1 Tłumaczenie

Dodałem tłumaczenie kategorii ramek danych, aby była ona przystosowana dla osób nie posługujących się językiem portugalskim. Zdaje sobie sprawe, że nie odegrało to kluczowej roli w realizacji zadania, gdyż w argumentach przyjmujemy dane na temat kategorii w języku portugalskim, jednak wydaje mi się, że ten sposób realizacji jest czytelniejszy.

2.3.2 Status zamówienia

Postanowiłem również przebadać zmienną order_status. Zauważyłem, że nasza ramka danych zawiera również te zamówienia, które sa w trakcie realizacji. Postanowiłem więc stworzyć trzy kategorie dla tej zmiennej:

- realized - zawiera ona zarówno te zamówienia, które zostały już wysłane jak i te w trakcie przetwarzania
- unavailable - to te zamówienia, które nie zostały wysłane z powodu braku towaru
- canceled - zamówienia, które zostały anulowane

Postanowiłem jednak ostatecznie usunąć również te dostawy, które miały status unavailable, gdyż jest to wina sprzedawcy, więc będzie to zaburzało sposób obliczania wyżej wspomnianej oceny.

Zauważyłem również, że te zamówienia anulowane, które otrzymały niską ocenę, najprawdopodobniej również w dużej mierze zależą od sprzedającego produkt.

Natomiast tych zamówień, które zostały anulowane, a posiadają wysoką ocenę, jest bardzo mało, zatem jesteśmy w stanie określić ocenę produktu wcześniej pozbywając się tych wierszy.

Sumarycznie, pozbyłem się wszytkich wierszy niesklasyfikowanych jako realized.

2.3.3 Stosunek ceny dostawy do ceny produktu

Stowrzenie go pomogło w późniejszej ocenie pożądania produktu

2.3.4 Różnica w czasie realizacji dostawy a estymowanym czasem dostawy

Obliczenie tej róznicy będzie istotne w przypadku stworzenia skali ocen produktów.

Postanowiłem w tym wypadku pozbyć się wszystkich wierszy, które zawierają nienaturalnie długi czas (górne 10%) realizacji oraz odpowiednią wysoką lub niską ocenę (w przypadku przedłużenia się czasu niską, a w przypadku szybszego czasu - wysoką), gdyż czas realizacji najprawdopodobniej wpłynał na ocene zamówienia.

2.4 Ocena pożądania produktu

Ostateczną ocena pożądania produktu została obliczona według następującego wzoru:

$$\sum_{i=1}^{n} \frac{1}{\frac{CenaDostawy_i}{CenaProduktu_i}} * OcenaProduktu_i$$

Produkty z wysoką ceną dostawy nie będą oczywiście tak pożądane jak te, które mają darmową dostawę. Również warto zwrócić uwagę, na ogólne zadowolenia ze zrealizowanego zamówienia. Dzięki sumie jesteśmy w stanie uwzględnić również całkowitą liczbę zamówień danego produktu.

2.5 Wybór produktów w przypadku zalogowanego klienta

W przypadku gdy chcemy wyświetlić polecane produktu zalogowanemu klientowi, robimy to według poniższego wzoru:

- po dwa najchętniej wybierane produkty z wybranych przez klienta kategorii
- jeden najchętniej wybierany produkt wśród klientów z tego samego regionu
- po jednym produkcie z kategorii podobnej do tej, którą użytkownik określił. Aby to zrobić, postanowiłem pogrupować produkty na siedem części: home, work and study, decoration, beauty, devices, free time oraz others

2.6 Wybór produktów w przypadku niezalogowanego klienta

W przypadku niezalogowanego klienta wybieramy po maksymalnie dwa najchętniej wybierane produkty z wyżej wspomnianych grup.

3 Rezultaty

Poniżej załączam wyniki dla trzech klientów zawartych w poleceniu:

```
In [1059]: print(result(["cama_mesa_banho", "papelaria", "fashion_calcados"], ["sao paulo", "SP"]))

['99a4788cb24856965c36a24e339b6058', 'f1c7f353075ce59d8a6f3cf58f419c9c', 'dc36a7859b743d8610a2bbbaea26ece9', '5411e9269501a870c
abf632f05655131', 'fb55982be901439613a95940feefd9ee', 'eb883e95b710f252cb15d0fb41d8bbe9', 'a2c75a23c2f838881dd4275c0cec519f',
'6fa96cbd1d917acb1ef4bc8040116105', 'f71973c922ccaab05514a36a8bc741b8', 'aca2eb7d00ea1a7b8ebd4e68314663af', '53b36df67ebb7c4158
5e8d54d6772e08']

In [1060]: print(result(["esporte_lazer", "moveis_decoracao", "telefonia"], ["rio de janeiro", "RJ"]))

['e44f675b60b3a3a2453ec36421e06f0f', '781afe929e3016a667f5f439afd55fce', 'db5efde3ad0cc579b130d71c4b2db522', 'aca2eb7d00ea1a7b8
ebd4e68314663af', '78efe838c04bbc568be034082200ac20', '7620a27f1d6747511f1c6f0ddb63c0ef', '431d674f9a4fbd0957ecf6ba3fcb6899',
'e7cc48a9daff5436f63d3aad9426f28b', 'e5eadf9be70a4a9fa514019542fc330a', '53b36df67ebb7c41585e8d54d6772e08', '99a4788cb24856965c
36a24e339b6058']

In [1061]: print(result([], []))

['e44f675b60b3a3a2453ec36421e06f0f', 'cbf96c04205dc933b89e025748c2a057', '8db75af9aed3315374db44d7860e25da', 'cd48f265a63e13b76
2601f5f794c5fca', 'eb883e95b710f252cb15d0fb41d8bbe9', '8ac47b3ab13c68f49f10dde899674149', '82c51c3938503a4ddc096fbed86428d6',
'93c902b021a9e594f658ab1b0351602a', 'a662d99595c3b6054702480cb7382afe', '7620a27f1d6747511f1c6f0ddb63c0ef']
```