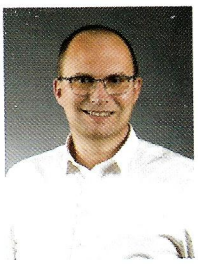
Data science w piekarni

Grupowanie klientów sklepu wielobranżowego techniką klastra aglomeracyjnego



Nadchodzące ciężkie czasy zmuszają przedsiębiorców do „oglądania złotówki z każdej stron". Wydatki muszą być racjonalne i przynosić właściwe efekty. Spadek siły nabywczej konsumentów niewątpliwie spowoduje spadek obrotów. W zestawieniu z rosnącymi cenami surowców i energii oznaczać to może poważnej kłopoty dla firm. Intuicyjnie wiemy, że spadek wolumenu sprzedaży może być częściowo ograniczony przez bardzo dobrą, trafną ofertę marketingową. Wszyscy staramy się osiągnąć przewagę konkurencyjną w obszarze cenowym, technologicznym oraz poprzez budowanie zaufania klientów. Gdy wreszcie uda nam się stworzyć doskonały produkt lub usługę, nasza oferta musi jakoś zostać zakomunikowana, żeby klienci mieli szansę z niej skorzystać. I tu powstaje problem rynku doskonale konkurencyjnego. Podaż informacji jest przeogromna. Klienci zasypywani są ofertami, promocjami, kuszeni są konkursami i nagrodami. Przedsiębiorcy starają się dotrzeć do jak największej grupy klientów. Informacje reklamowe się kumulują. W obliczu przytłaczającej ilości informacji, ofert i promocji klienci stają się obojętni. Bronią się przed szumem i reklamowym zgiełkiem, pozbywają się zbędnych informacji, czyniąc tym samym wysiłki przedsiębiorców bezowocnymi.

Segmentacja przez filtrowanie

Podstawowym sposobem poprawy efektywności reklamy jest segmentacja rynku. To prosty i bardzo stary sposób dzielenia populacji klientów na grupy. Klientów można dzielić na kobiety i mężczyzn, na ludzi starych, młodych i w średnim wieku. Wydzielamy grupy konsumentów z uwagi na poziom dochodów, wykształcenie lub tworzymy grupy w oparciu o miejsce zamieszkania, region. Taka forma segmentacji nazywana jest filtrowaniem. Może mieć ono wiele poziomów, na przykład możemy stworzyć pojedynczy segment konsumentów płci żeńskiej, w młodym wieku i mieszkających na wsi. Kobiety w młodym wieku zamieszkałe na terenach wiejskich powinny otrzymywać reklamy w określonej kolorystyce i zawierające określony komunikat, który akurat będzie opracowany dla tej wybranej grupy konsumentów. Rozwiązanie to wydaje się bardzo rozsądne.

Niestety istnieją dwa poważne problemy.

1. Brak cech do filtrowania

W sprzedaży internetowej dzięki imieniu na zamówieniu możemy wyodrębnić płeć klienta oraz dzięki adresowi dostawy miejsce zamieszkania, natomiast bardzo ciężko będzie nam określić wiek, który jest bardzo ważny przy formułowaniu treści marketingowej.

Inaczej rozmawia się z nastolatkami, inaczej z osobami na emeryturze. Wiek można próbować szacować analizując zamówione towary. Niestety nie ma pewności czy zamówienie nie było robione dla osoby trzeciej.

2. Brak jednorodności segmentu

Drugim problemem z segmentacją opartą na filtrowaniu jest wątpliwe założenia, że wszystkie kobiety w młodym wieku mieszkające na wsi będą podatne reagować na określony komunikat marketingowy. Kobiety te przecież mogą mieć różne usposobienie, mogą mieć różne wykształcenie i zainteresowania. Wysyłając do określonej grupy odbiorców skomponowane specjalnie dla nich informacje marketingową możemy mieć poczucie, że już nie strzelamy kulą w płot. Poczucie to może być bardzo złudne, ponieważ powstały przez filtrowanie segment rynku będzie tylko pozornie jednolity.

Podsumujmy, jak mówi chat GPT: „Filtrowanie jest procesem wybierania określonych elementów lub informacji z dużego zbioru danych na podstawie określonych kryteriów. Celem filtrowania jest eliminacja niepożądanych elementów lub redukcja zbioru danych do mniejszego zestawu zawierającego tylko te elementy, które spełniają określone warunki. Filtrowanie jest szeroko stosowane w analizie danych, bazach danych, przetwarzaniu sygnałów, przetwarzaniu obrazów itp.".

Filtrowanie pozwala wyodrębnić określone grupy, w których teoretycznie osoby mogą mieć podobne preferencje. Przy filtrowaniu możemy nie mieć odpowiednich cech konsumentów do filtrowania. Ponadto osoby zgromadzone w grupie w oparciu o stałe cechy, takie jak wiek czy płeć, mogą się istotnie różnić między sobą. Dlatego możemy przyjąć założenie, że segmentacja tworzona przez filtrowanie może okazać się narzędziem nieefektywnym.

Segmentacja klientów za pomocą klastrów

Jak definiuje tę technikę chat GPT?

„Klastrowanie to technika analizy danych, która polega na dzieleniu zbioru danych na grupy (klastry). Celem jest umieszczenie podobnych obiektów w tym samym klastrze, a różnych obiektów w innych klastrach. Klastrowanie ma na celu znalezienie wewnętrznej struktury danych i odkrycie ukrytych wzorców. Jest używane w wielu dziedzinach, takich jak eksploracja danych, uczenie maszynowe, analiza

obrazu itp.". W mojej ocenie tę definicję należy głębiej omówić.

Zapytajmy więc chat GPT czym różni się filtrowanie od klastrowania?

„Klastrowanie jest techniką grupowania danych w klastry w celu znalezienia wzorców, podczas gdy filtrowanie polega na wybieraniu lub eliminowaniu określonych elementów z dużego zbioru danych na podstawie określonych kryteriów."

Mówiąc prościej filtrowanie polega na segregacji według określonych stałych cech, natomiast klastrowanie polega na grupowaniu opartym na cechach zmiennych, behawioralnych. Klastrowanie nie opiera się na cechach stałych, takich jak kolor oczu czy miejsce zamieszkania lecz na zachowaniach, charakterystycznych zwyczajach, pewnych cechach, które teoretycznie są zmienne lecz w określonym przedziale czasu są stałe. Nie jestem pewny czy moje tłumaczenie jest prostsze niż to, które zaprezentował czat GPT.

Jest bardzo trudno wyjaśnić jak działa mechanizm klastrowania, dlatego stworzyłem krótki przykład używając środowiska języka programowania Python.

Przykład klastrowania danych

Sklep spożywczy chce rozesłać do określonych grup odbiorców przeznaczone dla nich oferty marketingowe. Sklep nie ma żadnych informacji na temat klientów oprócz numeru ID klientów i adresu internetowego. Należy wyodrębnić pięć grup klientów na podstawie ich wydatków w poszczególnych kategoriach produktów.

Na bazie rejestru sprzedaży udało się utworzyć bazę 440 transakcji dokonanych przez 100 indywidualnych klientów. Tabela 1 przedstawia fragment tej bazy.

Następnie tworzymy zestawienie średnich wydatków, jakie dokonywał każdy ze 100 analizowanych klientów.

Tabela 1. Rejestr sprzedaży produktów według transakcji dokonanych przez klientów indywidualnych.

	ID_customer	Pieczywo	Mleko	Art_przeysłowe	Mrożonki	Chemia	Słodycze	Napoje
0	47	12.67	9.66	10.80	0.71	6.68	2.68	11.57
1	43	7.06	9.81	13.67	5.87	8.23	3.55	12.20
2	43	6.35	8.81	10.98	8.02	8.79	15.69	17.53
3	78	13.26	1.20	6.03	21.35	1.27	3.58	3.14
4	82	22.62	5.41	10.28	13.05	4.44	10.37	11.15
...	...	...	...	...	...	...	...	...
435	97	29.70	12.05	22.90	43.78	0.46	4.41	15.01
436	93	39.23	1.43	1.09	15.03	0.23	4.69	3.98
437	74	14.53	15.49	43.20	1.46	37.10	3.73	18.27
438	36	10.29	1.98	3.19	3.46	0.42	4.25	4.32
439	4	2.79	1.70	3.59	0.22	1.19	0.10	1.84

440 rows × 8 columns

Tabela 2: Średnie wydatki klientów w grupach asortymentowych.

In [27]:
```
PA7 = data7.pivot_table(index=['ID_customer'],
                        values=['Pieczywo', 'Mleko', 'Art_przeysłowe',
       'Mrożonki', 'Chemia', 'Słodycze', 'Napoje'],
                        aggfunc='mean').applymap('{:,.2f} zł'.format).reset_index()

PA7
```

Out[27]:

	ID_customer	Art_przeysłowe	Chemia	Mleko	Mrożonki	Napoje	Pieczywo	Słodycze
0	0	3.02 zł	0.45 zł	0.70 zł	0.72 zł	0.93 zł	0.42 zł	0.37 zł
1	1	2.65 zł	1.33 zł	2.47 zł	1.90 zł	2.66 zł	0.41 zł	0.11 zł
2	2	3.31 zł	1.98 zł	2.14 zł	1.70 zł	2.61 zł	0.84 zł	0.67 zł
3	3	6.60 zł	3.34 zł	2.31 zł	1.46 zł	3.08 zł	0.60 zł	1.24 zł
4	4	5.02 zł	3.41 zł	3.74 zł	1.34 zł	4.94 zł	1.22 zł	1.91 zł
...	...	...	...	...	...	...	...	...
95	95	36.84 zł	27.58 zł	13.37 zł	24.09 zł	17.08 zł	30.52 zł	5.72 zł
96	96	9.37 zł	3.20 zł	5.23 zł	18.88 zł	9.93 zł	47.79 zł	8.40 zł
97	97	25.17 zł	14.41 zł	15.78 zł	37.74 zł	19.05 zł	36.67 zł	4.63 zł
98	98	25.41 zł	15.76 zł	15.58 zł	51.24 zł	20.21 zł	45.22 zł	7.24 zł
99	99	19.42 zł	4.36 zł	19.47 zł	99.14 zł	34.29 zł	62.81 zł	26.21 zł

100 rows × 8 columns

Teraz należy zestandaryzować wartości średnie. Robi się tak dlatego, ponieważ algorytm szukający podobieństw działa wadliwie gdy wartości drastycznie się od siebie różnią. Teraz wszystkie liczby z tabeli 2 przyjmą wartości w przedziale od 0 do 1.

In [42]:
```
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler (feature_range = (0, 1))

PA8 = PA7

PA8[['Pieczywo', 'Mleko', 'Art_przeysłowe', 'Mrożonki',
     'Art_przeysłowe', 'Słodycze','Napoje']] = scaler.fit_transform (PA8[['Pieczywo',
                                                                          'Mleko',
                                                                          'Art_przeysłowe',
                                                                          'Mrożonki',
                                                                          'Art_przeysłowe',
                                                                          'Słodycze',
                                                                          'Napoje']])
PA8
```

Out[42]:

	ID_customer	Art_przeysłowe	Chemia	Mleko	Mrożonki	Napoje	Pieczywo	Słodycze
0	0	0.028206	0.4520	0.000000	0.000000	0.000000	0.000304	0.009962
1	1	0.017530	1.3300	0.081029	0.012004	0.051940	0.000000	0.000000
2	2	0.036540	1.9780	0.066029	0.009997	0.050366	0.006971	0.021609
3	3	0.130900	3.3375	0.073381	0.007584	0.064381	0.003165	0.043199
4	4	0.085566	3.4050	0.139021	0.006314	0.120218	0.012980	0.068774
...	...	...	...	...	...	...	...	...
95	95	1.000000	27.5800	0.578758	0.237422	0.484246	0.482509	0.214847
96	96	0.210647	3.2000	0.206831	0.184592	0.269667	0.759338	0.317720
97	97	0.664516	14.4080	0.688509	0.376112	0.543111	0.581142	0.173257
98	98	0.671528	15.7550	0.679559	0.513284	0.578007	0.718195	0.273276
99	99	0.499274	4.3580	0.857006	1.000000	1.000000	1.000000	1.000000

100 rows × 8 columns

Teraz możemy przystąpić do tworzenia klastrów. Zawsze istnieje optymalna ilość klastrów dla określonych cech. Ilość tę znajduje się za pomocą metod statystycznych. Tym razem ilość klastrów dla grupy 100 klientów zostanie określana arbitralnie na 5.

```
from sklearn.cluster import AgglomerativeClustering

F = PA8[['Pieczywo', 'Art_przeysłowe', 'Chemia', 'Mleko',
       'Mrożonki', 'Napoje', 'Słodycze']].values

cluster = AgglomerativeClustering(n_clusters=5, affinity='euclidean', linkage='ward')
PA8['cluster']= cluster.fit_predict(F)
```

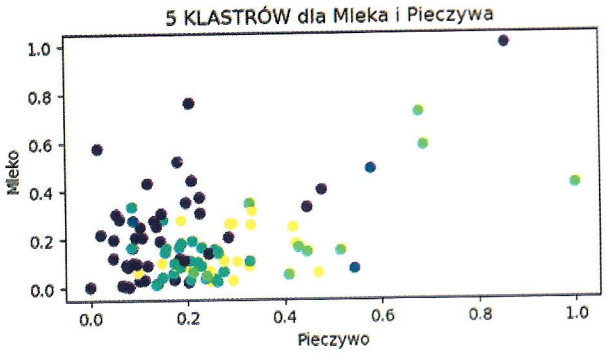
W wyniku tego każdy klient z tabeli 2 zostaje przyporządkowany do określonego klastra.

Wizualizacja klastrowania

Poniższy wykres przedstawia przyporządkowanie klientów do klastrów dla dwóch grup produktów, dla mleka i pieczywa. Każdy punkt na wykresie to jeden ze 100 klientów indywidualnych. Aby odzwierciedlić jakość klastrowania, należałoby stworzyć wykres siedmiowymiarowy. Oczywiście jest to niemożliwe, dlatego poniższy wykres z trudem odzwierciedla pięć „chmur" klastrów. Niestety algorytm musiał „pogodzić" aż siedem rodzajów produktów. Poniższy wykres przedstawia tylko dwa.

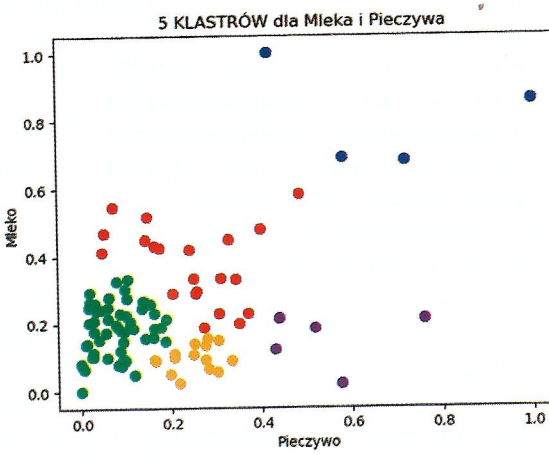
```
In [49]: plt.figure(figsize=(6, 3))
         plt.scatter(PA8['Mleko'], PA8['Pieczywo'], c=cluster.labels_)
         plt.ylabel('Mleko')
         plt.xlabel('Pieczywo')
         plt.title ('5 KLASTRÓW dla Mleka i Pieczywa')

Out[49]: Text(0.5, 1.0, '5 KLASTRÓW dla Mleka i Pieczywa')
```



Aby pokazać bardziej wiarygodny dowód na skuteczność tej metody tworzenia segmentów, przeprowadziłem analizę klastrowania wyłącznie dla dwóch produktów. Wcześniejsza analiza obejmowała aż 7 produktów.

Każdy punkt na wykresie to indywidualny klient spośród analizowanej grupy 100 klientów. Klienci otrzymali przyporządkowanie do jednej z 5 grup. Grupy różnią się między sobą średnią sumą wydatków na pieczywo i mleko. Grupa oznaczona kolorem zielonym wydaje mało pieniędzy zarówno na mleko jak i na chleb, podczas gdy grupa oznaczona kolorem fioletowym, składająca się z pięciu osób wydaje na pieczywo istotnie więcej niż na mleko



Podsumowanie

W dzisiejszych czasach szumu informacyjnego i spadającej siły nabywczej konsumentó skuteczna metoda segmentacji i będąca jej kon sekwencją precyzyjnie adresowana do klientó oferta może okazać się istotną przewagą konku rencyjną. Zaprezentowany dzisiaj prosty przykła segmentacji opartej na klastrowaniu pokazu wielki potencjał tego narzędzia.

Wojciech Moszczyńs

Wojciech Moszczyński – ekspert z zakresu opty malizacji matematycznej oraz modelowania pre dykcyjnego. Od lat zaangażowany w popularyz zację metod ekonometrycznych w środowiskach biznesowych. Specjalizuje się w optymalizac procesów sprzedażowych, produkcyjnych ora logistycznych. Przez 15 lat pracował jako eksper finansowy ze specjalizacją w obszarze kontrolin gu i rachunkowości zarządczej. Od 10 lat pracuj jako analityk danych (*data scientist*). Absolwen katedry Ekonometrii i Statystyki Uniwersytetu M kołaja Kopernika w Toruniu. Obecnie zatrudnion jako Senior Data Scientist w polskiej firmie Unit Group.