Klasteryzacja - Raport

Konrad Welkier, Piotr Sieńko, Jacek Wiśniewski

1. Wstęp

Badany zbiór danych pochodzi ze strony https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset#, opisuje on 12,330 sesji użytkowników przeglądarek internetowych. Posiada on 10 zmiennych numerycznych oraz 8 kategorycznych, które zawierają różnego rodzaju dane dotyczące sposobu i czasu odwiedzania stron internetowych, stosowanego oprogramowania oraz informację, czy klient dokonał transakcji zakupu w czasie trwania danej sesji. Jest to naturalny sposób rozdziału klientów na tych, którzy przynoszą i nie przynoszą zysków. Postanowiliśmy więc sprawdzić w jakim stopniu podział uzyskany za pomocą metod klasteryzacji jest podobny do oryginalnego.

W tym celu przetestowaliśmy następujące algorytmy klasteryzacyjne:

- K-średnich → w czasie każdej iteracji przyporządkowuje obserwacje do najbliższych skupień, następnie oblicza centroidy uzyskanych grup, które stają się nowymi środkami. Algorytm należy kilkakrotnie powtórzyć w celu eliminacji błędu początkowego wyboru punktów skupień.
- Gaussian mixture model → za pomocą własności rozkładu normalnego, zwraca prawdopodobieństwo przyporządkowania punktów do określonego skupienia
- Algorytm aglomeracyjny → każdy punkt jest początkowo oddzielnym klastrem. Stopniowo łączy najbliższe klastry, aż do uzyskania zadanej ich liczby.

Do oceny wyników klasteryzacji użyliśmy następujących metryk:

- Indeks Silhouette → średnia miara tego jak dany punkt pasuje do przydzielonego klastra, w porównaniu z drugim najlepszym wyborem. Im wyższy wynik tym lepsza klasteryzacja.
- Indeks Daviesa-Bouldina → opisuje średnie podobieństwo pomiędzy każdym klastrem. Im niższy wynik tym lepsza klasteryzacja
- Skorygowany indeks Randa → jedyna używana przez nas miara, która porównuje podział uzyskany z oryginalnym. Oblicza podobieństwo pomiędzy obydwoma podziałami. Jeśli są identyczne, indeks przyjmuje wartość 1.

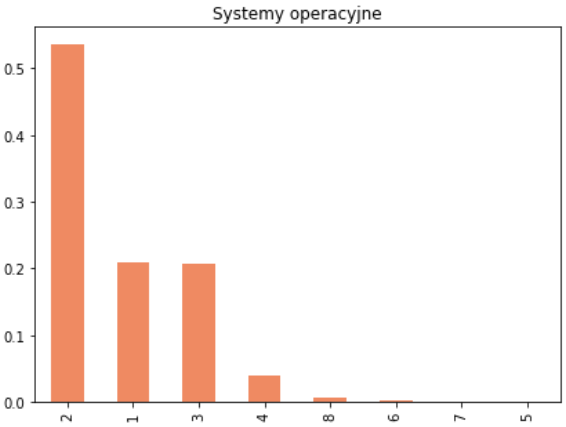
Dodatkowo, do wizualizacji wyników użyliśmy analizy głównych składowych.

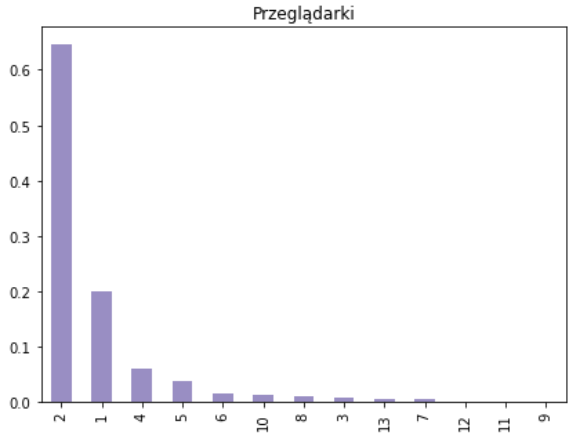
2. Pytanie Badawcze

Czy podział zbioru uzyskany metodami klasteryzacji odzwierciedla rzeczywiste grupy klientów?

3. Przygotowanie danych

W celu polepszenia działania algorytmów, ustandaryzowaliśmy zmienne numeryczne oraz zgrupowaliśmy rzadziej występujące przeglądarki i systemy operacyjne. Dotychczas rozkład przeglądarek oraz systemów operacyjnych w zbiorze przedstawiał się następująco:

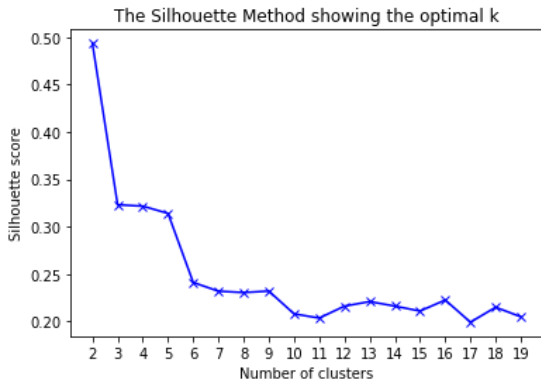


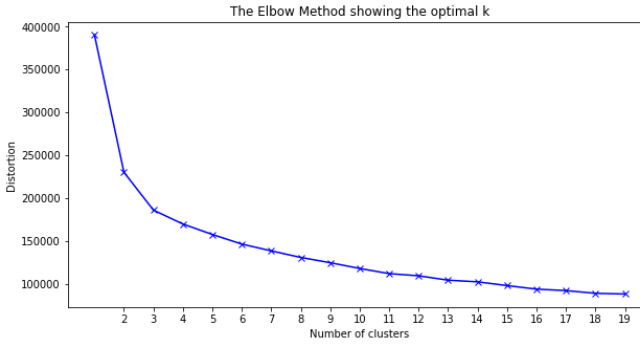


Natomiast po grupowaniu, razem przydzielone zostały systemy z numerami innymi niż 1, 2 lub 3, a w przypadku przeglądarek z numerami innymi niż 1 lub 2. Usunięta została kolumna celu – Revenue, a także kolumny Month, Weekend, VisitorType, Informational, Administrative i ProductRelated, które albo były mocno skorelowane z innymi kolumnami albo nie zawierały interesującej informacji.

4. Dobór liczby klastrów

Wszystkie używane przez nas algorytmy wymagają podania liczby klastrów. Musieliśmy więc w jak najbardziej obiektywny sposób wybrać na ile podzbiorów chcemy podzielić nasze dane. Do tego celu użyliśmy miary silhouette oraz metody "łokciowej".





Obie metody wskazują k = 2, jako najlepszą liczba klastrów. Stwierdziliśmy w tym momencie, że możliwe jest, iż nasz zbiór danych uda się podzielić klientów na kupujących i tylko odwiedzających.

5. Pierwsze modele

Po przygotowaniu danych oraz znalezieniu optymalnej liczby klastrów, stworzyliśmy dwa wstępne modele, wykorzystujące metodę K-średnich oraz GMM:

Indeks	K-średnich	GMM
Adjusted Rand score	0.002914198259632032	0.24542104703463424
Silhoute score	0.49335301746555515	0.038170524727392574
Davies-Bouldin score	0.8896926379870409	5.268457629172635

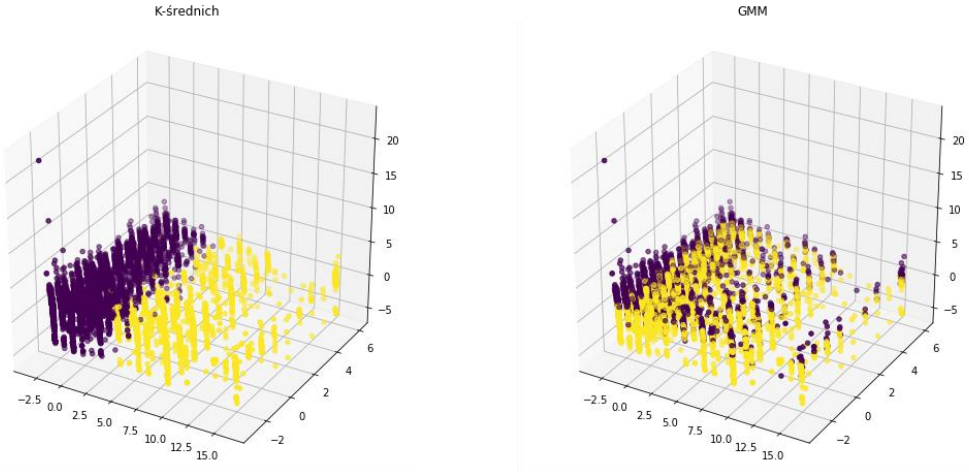
Porównanie K-średnich z rzeczywistym podziałem:

nasz/prawdziwy	0	1
0	8654	1575
1	1768	333

Porównanie GMM z rzeczywistym podziałem:

nasz/prawdziwy	0	1
0	2669	1659
1	7753	249

Wyniki były bardzo słabe, oba algorytmy nie zdołały podzielić zbioru na wyróżniające się, oddzielone od siebie klastry. Metoda K-średnich miała o wiele wyższy indeks Silhouette, lecz niestety indeks Randa był bliski zeru. Odwrotnie wyglądały wyniki Gaussian Mixture Model. Po uzyskanych wartościach widać, że oba algorytmy zadziałały zupełnie inaczej. Aby sprawdzić różnice między nimi, zdecydowaliśmy się na użycie analizy głównych składowych i graficzne przedstawienie ich działania.



Oba podziały są od siebie kompletnie różne. GMM posiadający wyższy indeks Randa, podzielił zbiór według wartości wektora pionowego. Zmienną, która w dużym stopniu warunkuje jego wartość jest *PageValues*. Postanowiliśmy zwiększyć jej znaczenie i sprawdzić, czy wpłynie to na lepszy podział zbioru. Po wielu próbach, do tej części eksperymentu wybraliśmy algorytm aglomeracyjny oraz używany wcześniej K-średnich.

6. Modele ze zwiększonym *PageValues*

Zdecydowaliśmy się na przemnożenie wartości z kolumny *PageValues* przez 20, a następnie na tak zmodyfikowym zbiorze zastosowaliśmy wybrane chwilę wcześniej algorytmy:

Indeks	K-średnich	Klasteryzacja aglomeracyjna
Adjusted Rand score	0.3090556140115334	0.22196899478319693
Silhoute score	0.8226984845752094	0.8388445443280641
Davies-Bouldin score	0.508916392470719	0.4675237227827504

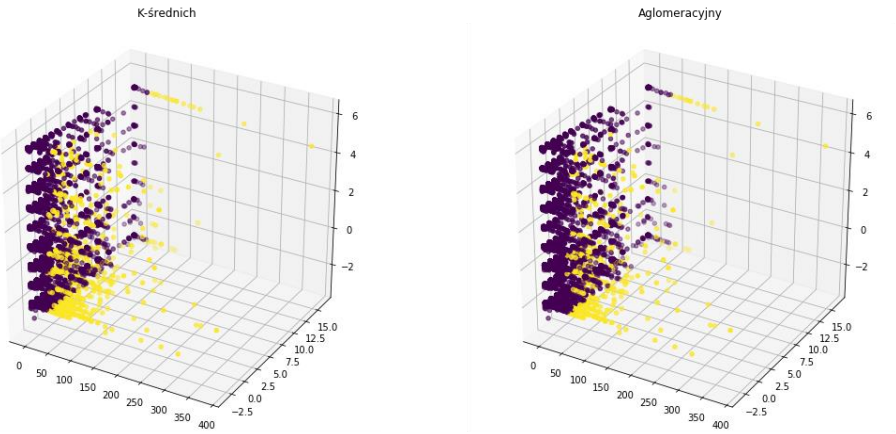
Porównanie K-średnich z rzeczywistym podziałem:

nasz/prawdziwy	0	1
0	10270	1365
1	152	543

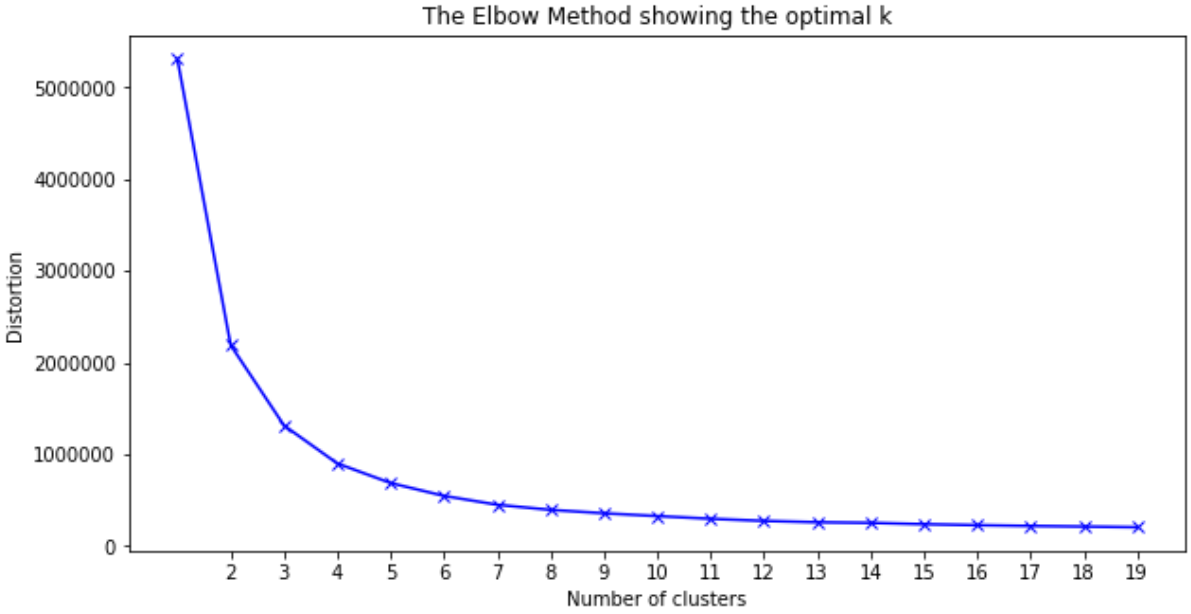
Porównanie klasteryzacji aglomeracyjnej z rzeczywistym podziałem:

nasz/prawdziwy	0	1
0	10336	1545
1	86	363

Okazało się, że tak drobna zmiana miała ogromny wpływ na działanie. Indeks Silhouette osiągnął ponad 0.8, natomiast indeks Daviesa - Bouldina zmalał w okolice 0.5. Teraz wizualizacja z wykorzystaniem głównych składowych prezentowała się następująco:



Oba algorytmy w podobny sposób rozbiły punkty na klastry. Widać, że obserwacje są podzielone wzdłuż zmodyfikowanego wektora. Upewnijmy się jeszcze, że dla tak zmienionych danych, optymalną liczbą klastrów jest 2.



Metoda 'łokciowa' znów wskazała k = 2. Przyjmijmy więc, że jest to optymalna liczba klastrów.

7. Weryfikacja hipotezy badawczej i wnioski

Pozostało nam sprawdzenie, czy rzeczywiste rozbicie na klientów, którzy dokonali zakupu i tych, którzy nie zakończyli transakcji jest możliwe do odwzorowania przy użyciu klasteryzacji. Do porównania z oryginalnym podziałem wybraliśmy algorytm K-średnich.

Kmeans

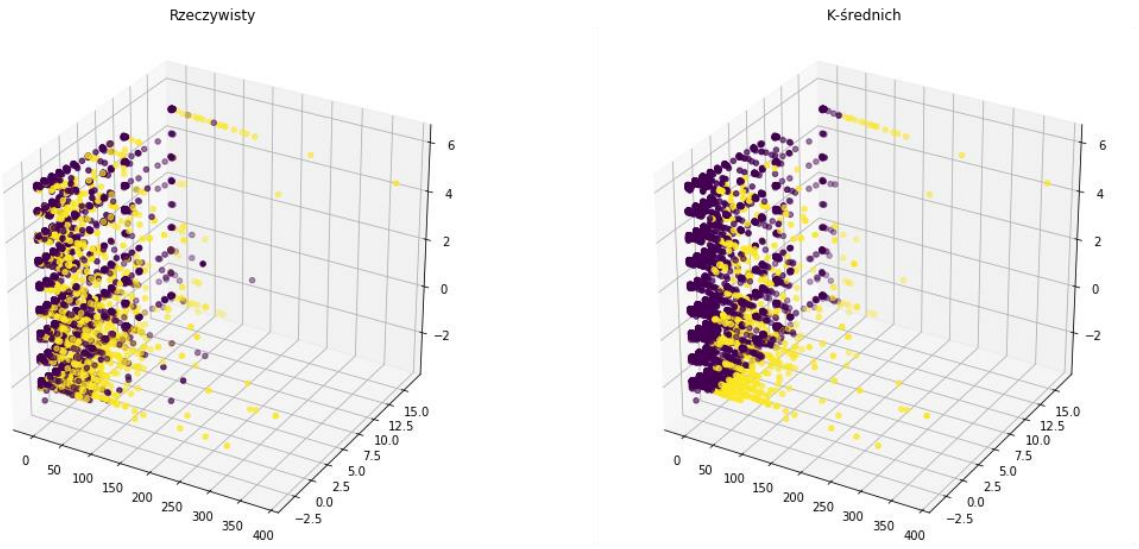
Indeks	Wartość
Adjusted Rand score	0.3090556140115334
Silhoute score	0.8226984845752094
Davies-Bouldin score	0.508916392470719

W końcowym porównaniu użyliśmy również miar znanych nam z zadań klasyfikacji:

Accuracy	0.8769667477696674
Precision	0.781294964028777
Recall	0.28459119496855345

nasz/prawdziwy	0	1
0	10270	1365
1	152	543

Podział rzeczywisty oraz ostateczny podział uzyskany za pomocą klasteryzacji prezentują się wizualnie w następujący sposób:



Wyniki nie są jednoznaczne. Na podstawie przeprowadzonych eksperymentów można stwierdzić, że zwiększenie wagi zmiennej *PageValues* poprawia działanie algorytmów klasteryzujących. Osiągnęliśmy podział, który wydaje się być na pierwszy rzut oka podobny do oryginalnego. Niestety,

rzeczywiste rozdzielenie podzbiorów na dany typ klienta jest znacznie bardziej skomplikowane. Pamiętajmy również, że dane zostały zmodyfikowane na podstawie indeksu Randa, który wykorzystuje wiedzę o prawdziwym podziale zbioru. Niestety w praktyce rzadko kiedy dysponujemy takimi danymi. Podsumowując, możemy przyjąć, iż wyniki działania algorytmów klasteryzujących nie odzwierciedlają rzeczywistego podziału zbioru względem typu klienta.

8. Oświadczenie

Oświadczamy, że niniejsza praca stanowiąca podstawę do uznania osiągnięcia efektów uczenia się z przedmiotu Wstęp do Uczenia Maszynowego została wykonana przeze nas samodzielnie.

Konrad Welkier, Piotr Sieńko, Jacek Wiśniewski