Projekt 2 - klasteryzacja - raport końcowy

Martyna Majchrzak, Agata Makarewicz, Renata Rólkiewicz

11 06 2020

Wstęp

Klasteryzacja to metoda uczenia nienadzorowanego, której celem jest podział obserwacji na grupy obiektów podobnych do siebie, zwanych klastrami. Poniższy projekt ma na celu klasteryzację zbioru danych `Online Shoppers Purchasing Intention`

Zbiór danych

Każdy wiersz w zbiorze opisuję sesję innego użytkownika podczas zakupów internetowych. Dane zostały zebrane na przestrzeni jednego roku, aby uniknąć wpływów czynników takich jak kampanie reklamowe, specjalne oferty, profil użytkownika, pora roku itd. Zbiór danych nie zawiera żadnych brakujących wartości i składa się z 10 zmiennych numerycznych i 8 kategorycznych.

Odwiedzone podczas dokonywania zakupu witryny zostały podzielone na 3 typy: *Administrative* - administracyjne, *Informational* - informacyjne oraz *ProductRelated* - związane z produktem.

Zmienne związana z typami witryn:

- *Administrative* - liczba odwiedzonych stron typu "Administrative" (zarządzanie kontem)
- *Administrative_Duration* - sumaryczny czas (w sekundach) spędzony na stronach typu "Administrative"
- *Informational* - liczba odwiedzonych stron typu "Informational" (informacje o stronie, kontakt, adres)
- *Informational_Duration* - sumaryczny czas (w sekundach) spędzony na stronach typu "Informational"
- *ProductRelated* - liczba odwiedzonych stron typu "ProductRelated" (strony związane z produktami)
- *ProductRelated_Duration* - sumaryczny czas (w sekundach) spędzony na stronach typu "ProductRelated"

Zmienne związane z analizą Google Analytics (link do opisu miar) :

- *BounceRates* - procent odwiedzających stronę, którzy wychodzą z niej bez wykonania żadnego zapytania do serwera analitycznego poczas tej sesji
- *ExitRates* - procent wszystkich odwiedzin tej strony, które były ostanie w danej sesji
- *PageValues* - średnia wartość stron odwiedzonych przed dokonaniem transakcji

Informacje o użytkowniku:

- *VisitorType* - typ użytkownika (*Returning_Visitor/New_Visitor/Other*)
- *OperatingSystems* - system operacyjny użytkownika (zakodowane numerycznie liczbami od 1 do 8)
- *Browser* - przeglądarka użytkownika (zakodowane numerycznie liczbami od 1 do 13)

- *Region* - region, z którego pochodzi użytkownik (zakodowane numerycznie liczbami od 1 do 9)
- *TrafficType* - określa, w jaki sposób użytkownik dostał się na stronę, np. wpisując hasło w wyszukiwarkę/ klikając link (zakodowane numerycznie liczbami od 1 do 20)

Informacje o transakcji:

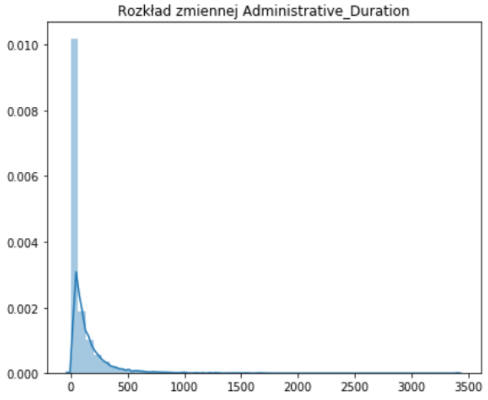
- *SpecialDay* - bliskość daty transakcji do szczególnych dni lub świąt (np. Dzień Matki) z uwzględnieniem np. czasu dostawy. Przykład: dla Walentynek - wartości niezerowe są od 2 do 12 lutego, a najwyższą wartość (równą 1) przyjmuje 8 lutego
- *Month* - miesiąc
- *Weekend* - informacja czy zakup został dokonany w trakcie weekendu (*False/True*)
- *Revenue* - informacja czy doszło do transakcji (*False/True*)

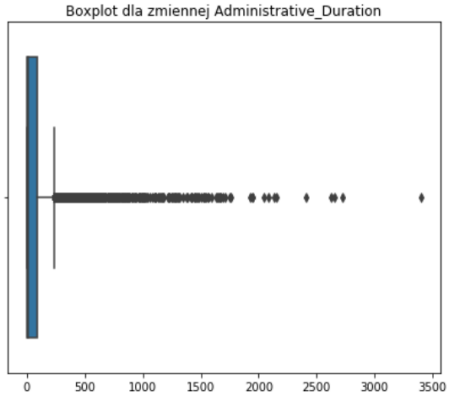
Przygotowanie zbioru danych

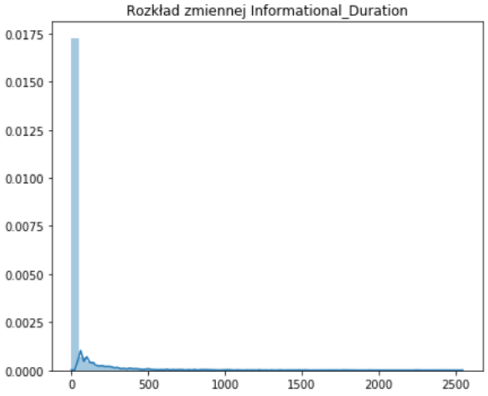
Na proces przygotowania zbioru danych składało się:

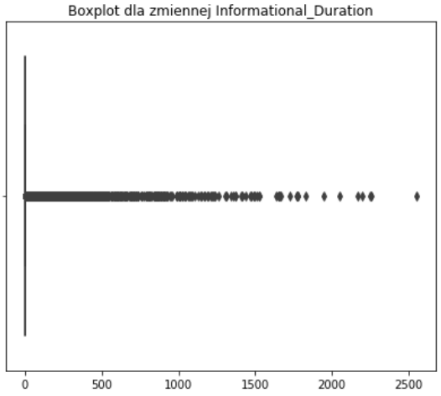
- **Usunięcie outlierów**

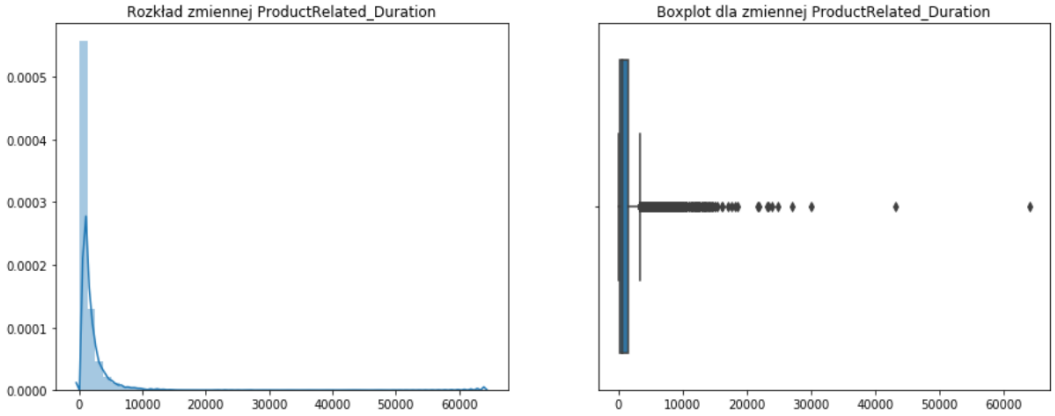
Cztery zmienne numeryczne zawierają bardzo dużą liczbę outlierów. Jest ich na tyle wiele, że zastosowanie tradycyjnych metod pozbywania się ich doprowadziłoby do znacznej redukcji liczby obserwacji.











Usunięta została więc tylko część obserwacji o wartościach większych od liczb wybranych na podstawie wykresów rozkładów zmiennych.

- **Zredukowanie liczby kategorii** dla zmiennych *OperatingSystems, Browser, Region* oraz *TrafficType*

Część kategorii tych zmiennych jest bardzo nieliczna. Zostanały one więc zagregowane do wspólnej kateorii *Other*.

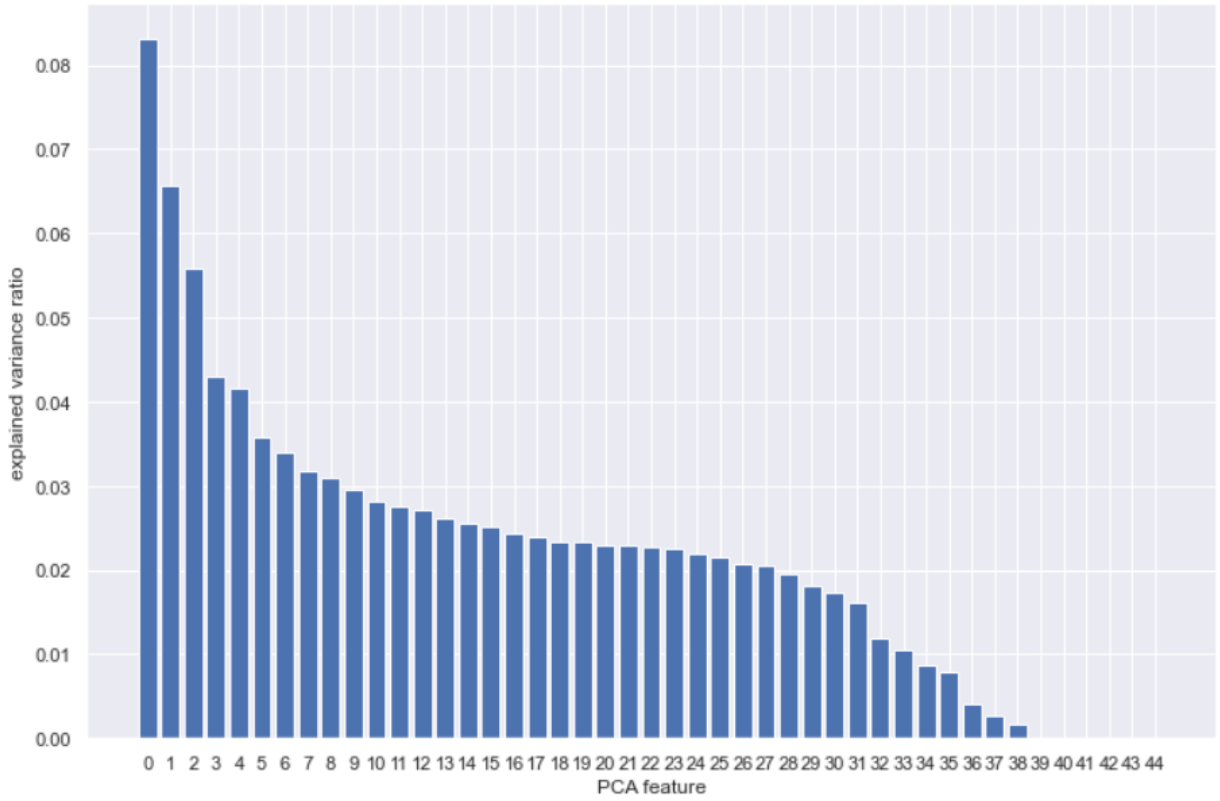
- **One-hot encoding**

Dla zmiennych kategorycznych zastosowano metodę One-hot encoding, polegającą na tym, że dla każdej z kategorii kodowanej zmiennej tworzona jest nowa kolumna, która przyjmuje wartość 0, jeśli rekord nie jest w danej kategorii lub 1, jeśli jest.

- **Standaryzacja zmiennych**
- **Redukcja wymiarów**

Po wykonaniu encodingu zbiór miał 47 kolumn. Usunięte zostały dwie zbędne kolumny (*Weekend_False*, *Revenue_False*). Ponieważ przed encodingiem były to zmienne logiczne, to informacje znajdują się już w analogicznych kolumnach `_True`.

Następnie przeprowadzona została PCA (Principal Component Analysis), czyli analizę składowych głównych.



Ponieważ powyżej 38 komponentu wyjaśniają one już bardzo niewielki ułamek procenta wariancji, dopasowany został do danych PCA o 38 komponentach i tymi danymi posłużono się w dalszej analizie.

Teza badawcza

Celem badań było znalezienie rodzajów klientów robiących zakupy internetowe na podstwie zgromadzonych danych. Posłużono się wszystkimi dostępnych zmiennych w zbiorze danymi uprzednio je przygotowywując według powyższego opisu. Dodatkowo przy eksperymencie sprawdzono dwa zbiory danych - przed PCA (tylko przeskalowane) i po PCA. Na każdym z nich wypróbowano 10 metod klasteryzacji.

Opis eksperymentu

Metody klasteryzacji

Na zbiorze przeprowadzono klasteryzację z użyciem 4 algorytmów: KMeans, Agglomerative z 4 wartościami parametru ‘linkage’, GMM z 4 wartościami parametru ‘covariance’ oraz DBSCAN.

- Kmeans
- Agglomerative average linkage
- Agglomerative single linkage
- Agglomerative ward linkage
- Agglomerative complete linkage
- GMM covariance full
- GMM covariance tied

- GMM covariance diag
- GMM covariance spherical
- DBSCAN

Dla każdego modelu sprawdzono liczbę klastrów od 2 do 9. Ponieważ DBSCAN nie przyjmuje liczby klastrów jako argumentu, jego działanie sprawdzono dla argumentu epsilon, dla wartości od 0.1 do 1.

Metryki

Do porównania modeli posłużono się trzema metrykami:

1. The Silhouette Coefficient:

Dla każdej obserwacji wyliczane są 2 wartości:

```
a: Średnia odległość od wszystkich punktów z tej samego klastra co obserwacja
b: Średnia odległość obserwacji od wszystkich punktów z kolejnego najbliższego klastra.
```

Silhouette Coefficient dla pojedynczej obserwacji to (a-b)/max(a,b), a dla zbioru średnia z wyników wszystkich obserwacji.

Zakres wartości:

```
-1 (dla niepoprawnej klasteryzacji)
1 (dla bardzo gęstych klastrów)
0 oznacza nakładające się klastry
```

2. Davies-Bouldin Index:

Oznacza średnie 'podobieństwo' między klastrami, gdzie miara podobieństwa porównuje odległość między klastrami z ich wielkością.

Wartość: im bliższa 0, tym podział lepszy

3. Calinski-Harabasz Index:

Oznacza stosunek pomiędzy sumą kwadratów odległości pomiędzy klastrami oraz sumą kwadratów odległości wewnątrz klastrów.

Wartość: im wyższa tym klastry gęstsze i lepiej oddzielone od siebie

Warto zauważyć, że według dokumentacji wszystkie powyższe metryki faworyzują wypukłe klastry ponad inne typy (np. oparte na gęstości, takie, jak zwraca DBSCAN i GMM).

Ocena wyników

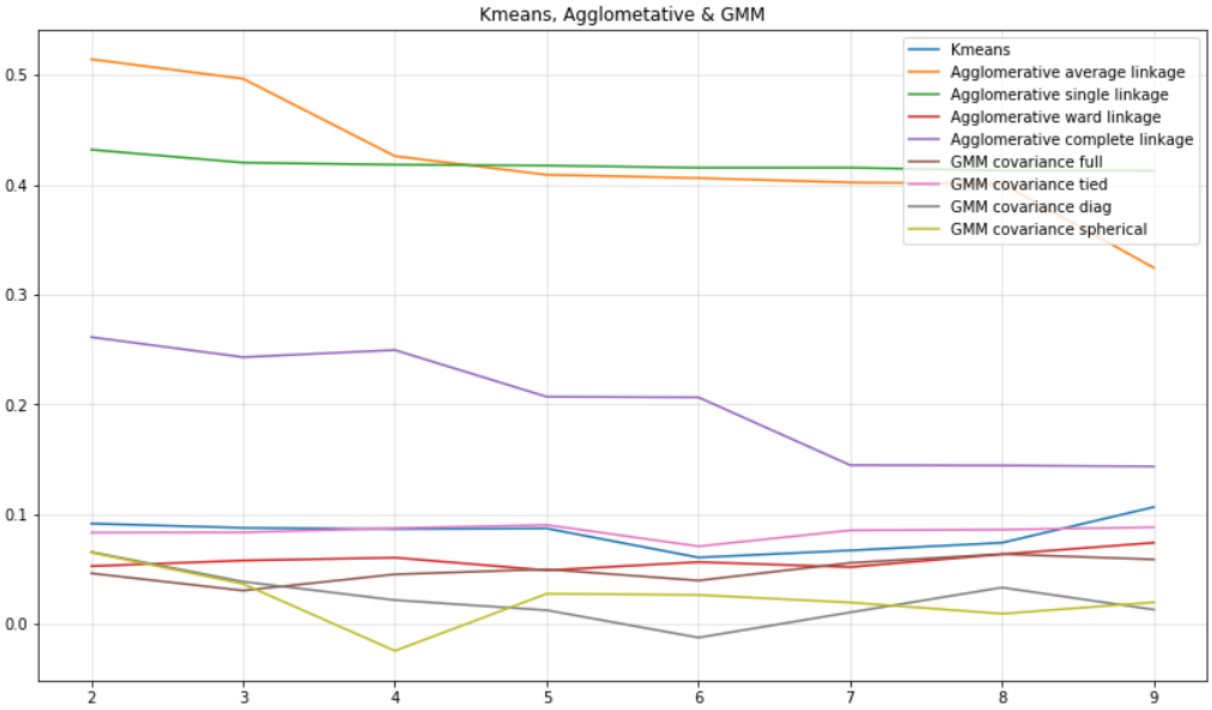
Po przetestowaniu modeli na danych ustandaryzowanych oraz po zastosowaniu PCA okazało się, że redukcja wymiarów nie przyniosła znacznej poprawy wyników modeli. Mimo to, dalsza analiza dotyczyć będzie tych drugich wyników, chociażby ze względu na to, że dane po PCA miały mniejszy wymiar. Dwie z rozważanych metryk (Silhouette oraz Davies-Bouldin Score) wskazały ten sam algorytm jako najlepszy, w odróżnieniu od Calinski-Harabasz Score.

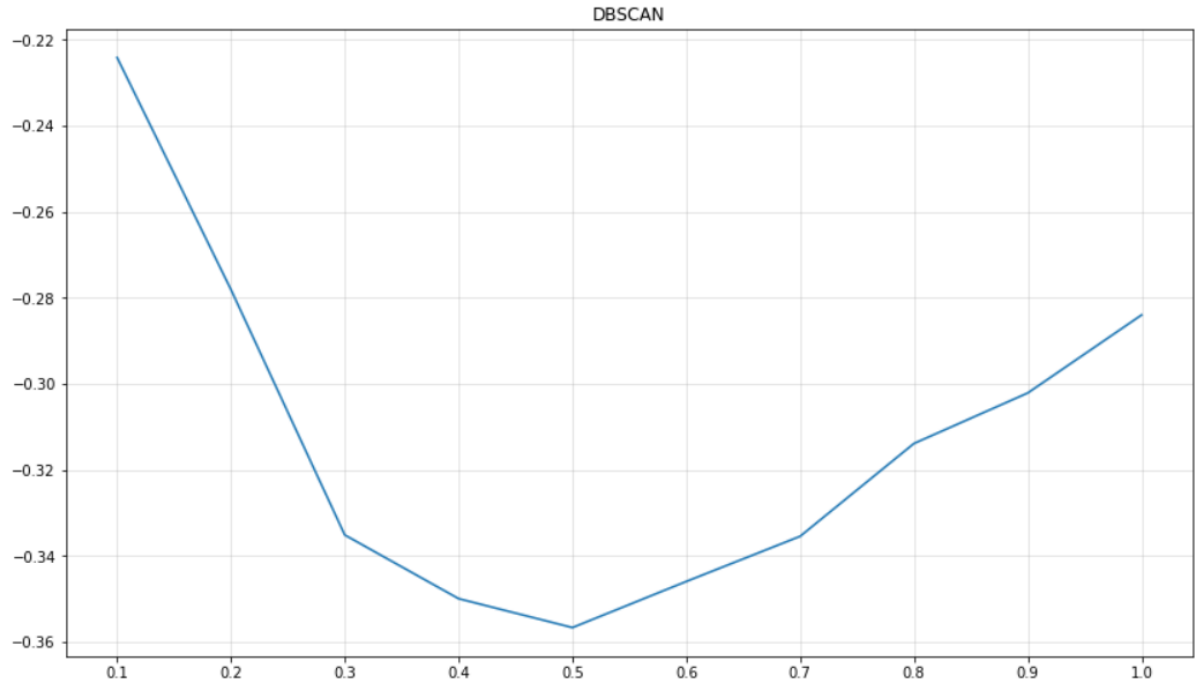
- **Silhouette Score** oraz **Davies-Bouldin Score** - *Agglomerative average linkage, n_clusters = 2*

- **Calinski-Harabasz Score** - *KMeans, n_clusters = 2*

Silhouette score

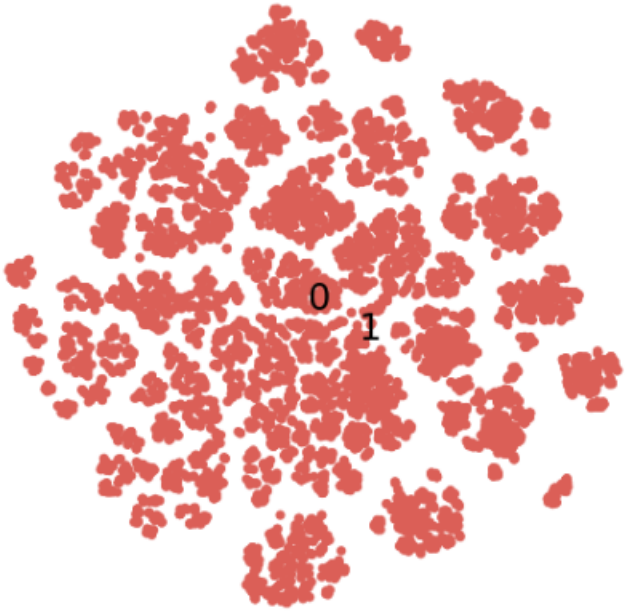
Według tej metryki, dla większości modeli zwiększenie liczby klastrów wiązało się z pogorszeniem wyniku. Najlepszy wynik to Agglomerative average linkage dla dwóch klastrów z wynikiem 0.51 Najgorzej poradził sobie GMM z wartościami parametru covariance: diag i spherical oraz DBSCAN, który jako jedyny osiągnął ujemne wartości metryki.





Z powyższych wykresów wynika, że pod względem *Silhouette score* najlepszy okazał się model AgglomerativeClustering z parametrem linkage='average', dla liczby klastrów równej 2. Wizualizacja za pomocą tSNE jednak szybko weryfikuje, że przydzielił on prawie wszystkie obserwacje do jednej kategorii.

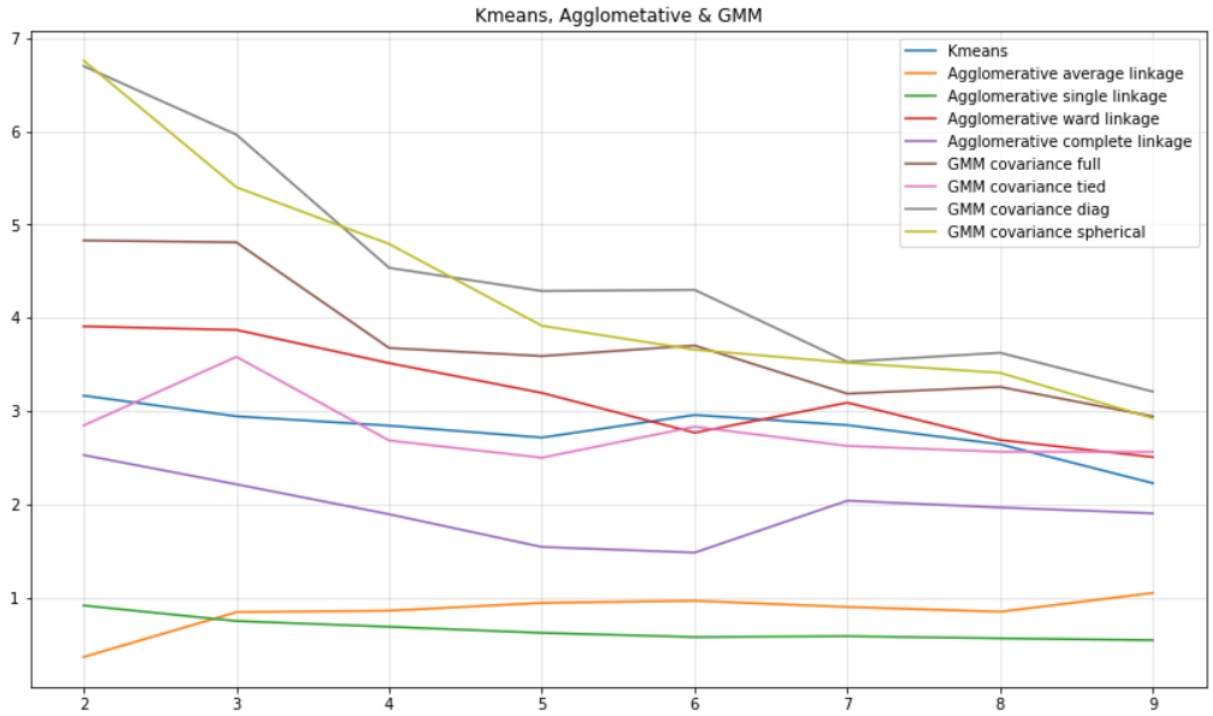
tSNE dla predykcji modelu Agglomerative, linkage 'average', 2 klastry:

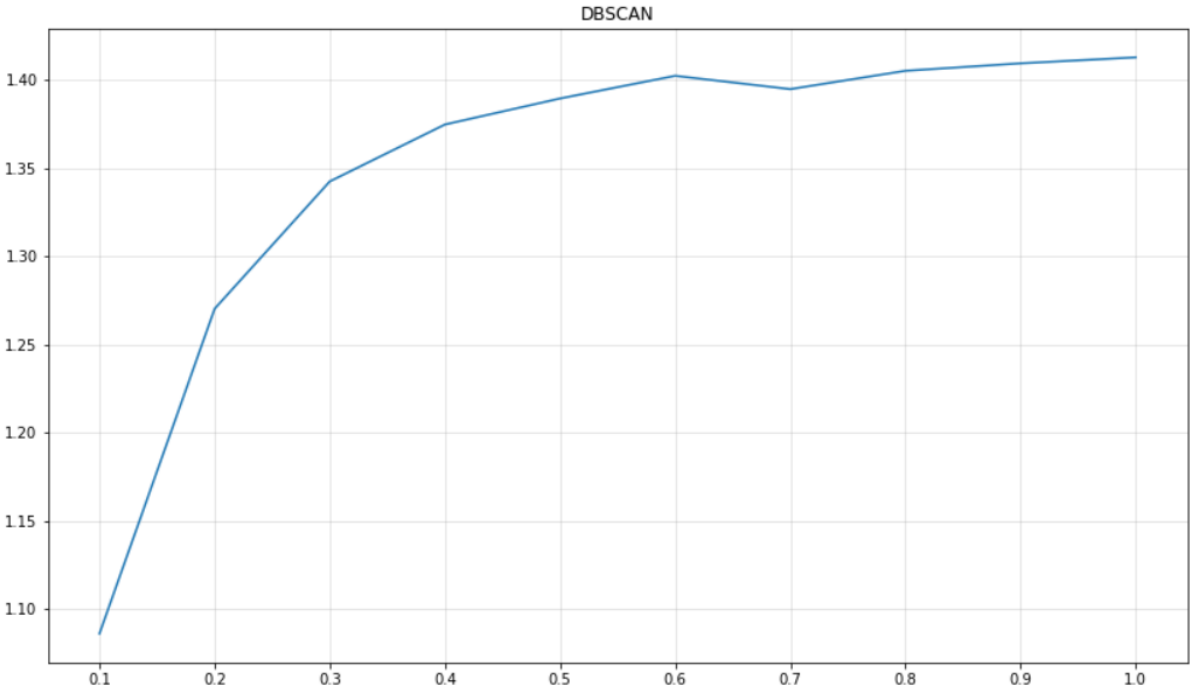


Davies-Bouldin Score

Index Daviesa-Bouldina sugeruje coś wprost przeciwnego do Silhouette - ocena większości modeli poprawia się wraz z wzrostem liczby klastrów. Jednakże, te same dwa modele: *Agglomerative average i single linkage*, uzyskują najlepsze wyniki.

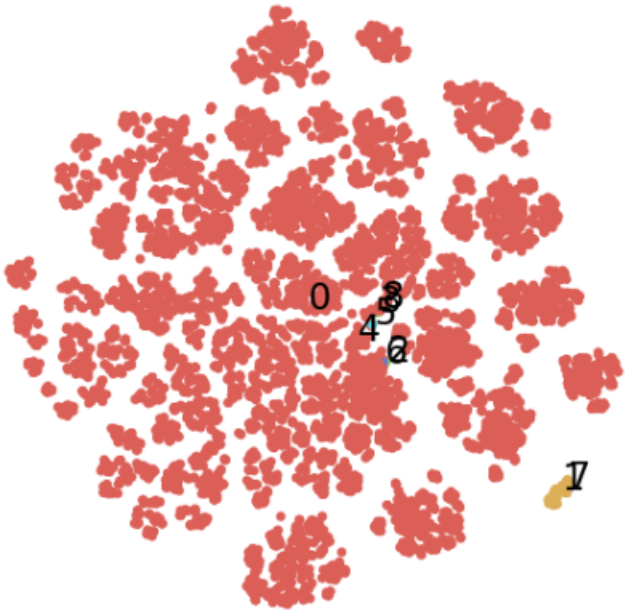
DBSCAN dla małych epsilonów daje daje dobre wyniki, ale średnio plasuje się na 3 miejscu za wspomnianymi modelami.





Z powyższych wykresów wynika, że pod względem *Davies-Bouldin score* również najlepszy okazał się model *AgglomerativeClustering* z parametrem *linkage='average'* i dla 2 klastrów (tak jak przy *Silhouette score*). Dla większych liczb klastrów jednak model z parametrem 'single' osiągał od niego zawsze lepsze wyniki. Najniższą wartość metryki osiąga on dla 9 klastrów.

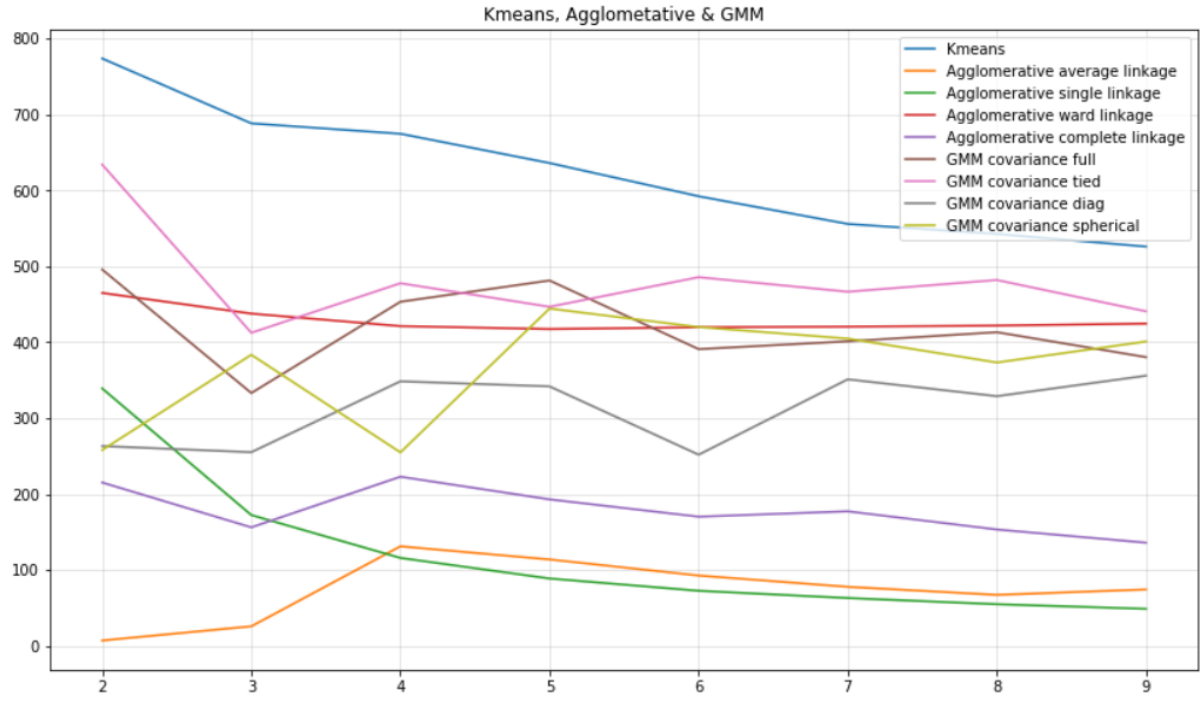
tSNE dla predykcji modelu Agglomerative, linkage 'single', 9 klastrów:

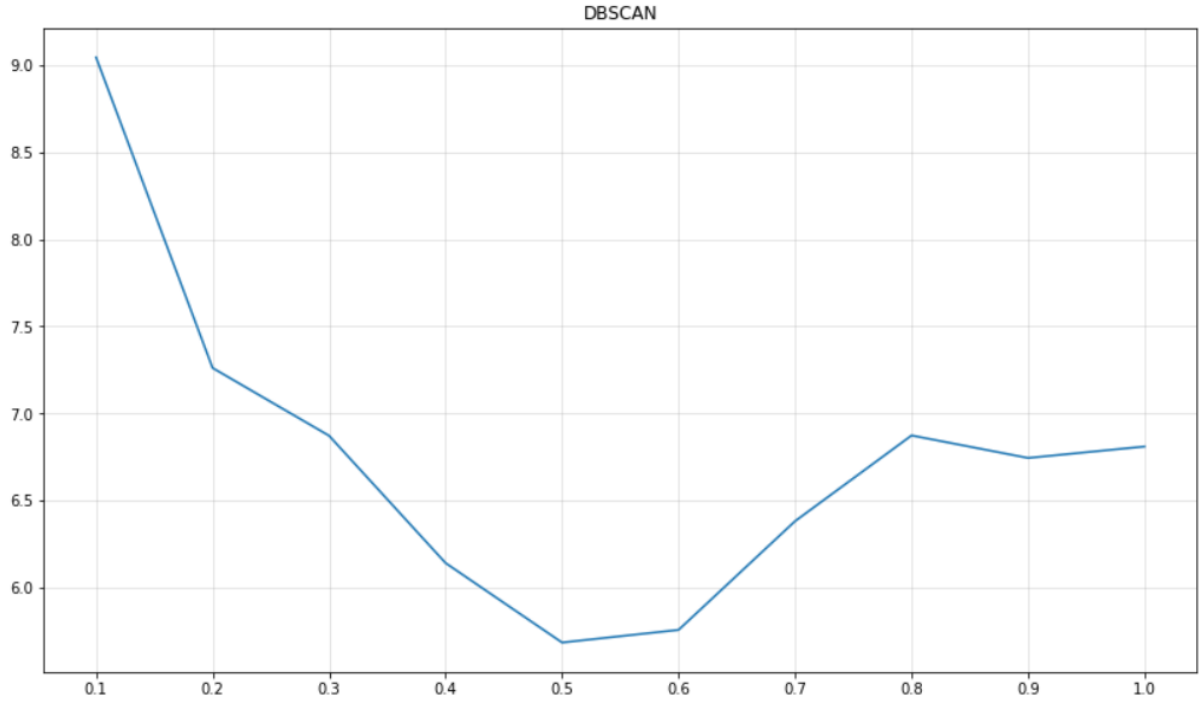


Powyższy wykres pokazuje jednak, że dla większej liczby klastrów algorytm Agglomerative również przydziela zdecydowaną większość obserwacji do jednej klasy.

Calinski-Harabasz Score

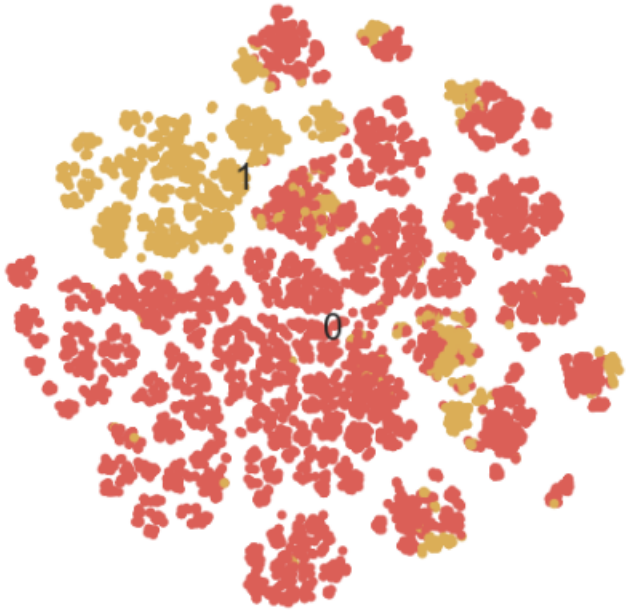
Index Calińskiego-Harabasza dla części modeli pogarsza się wraz z wzrostem liczby klastrów, chociaż jest tu też sporo wyjątków, np. wszystkie warianty modelu GMM. Jednak modele, które w poprzednich dwóch metrykach uzyskały najlepszy wynik, według tej metryki wypadają najgorzej. Zdecydowanie najlepsze wyniki osiąga z kolei model KMeans, który według poprzednich metryk wypadał średnio.





Z powyższych wykresów wynika, że pod względem *Calinski-Harabasz score* najlepszy okazał się model *KMeans* dla liczby klastrów równej 2.

tSNE dla predykcji modelu KMeans, 2 klastry:



Jest to pierwszy algorytm, dla którego klastry są dość zbilansowane co do wielkości (a przynajmniej żaden z nich nie jest pojedynczą obserwacją). W grupie oznaczonej nr 0 znajduje się 78% obserwacji.

Poniżej znajduje się tabela ze średnimi wartościami cech numerycznych dla podziału oryginalnego zbioru na te 2 grupy.

labels	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues	SpecialDay	Weekend	Revenue	labels
0	2.105322	78.497009	0.449741	26.267677	23.206948	849.938142	0.027104	0.047631	5.723428	0.050259	0.286401	0.147820	0.0
1	2.361527	79.948096	0.511652	35.749480	33.631398	1261.664208	0.020831	0.041822	5.864420	0.064648	0.217645	0.156471	1.0

Jedyna znacząca różnica widoczna jest w zmiennych Infromational Duration, ProductRelated oraz ProductRelated Duration.

Wnioski

Wszystkie zastosowane metryki sugerują, że dla tego zbioru najnaturalniejszym wyborem jest podział na dwa klastry. Grupy te nie są jednak dobrze wyodrębione, przez co wyniki żadnego z przebadanych modeli nie są satysfakcjonujące. Po rozkładach zmiennych widać, że wiele z nich przyjmuje dla znacznej większości obserwacji jedną wartość, co mogło spowodować, że wiele algorytmów przydzielało prawie wszystkie obserwacje do jednej klasy. Najlepszym w naszej opinii okazał się algorytm KMeans, który otrzymał najwyższy wynik według indeksu Calinskiego-Harabasza i w odróżnieniu od modeli, które uzyskały wysokie wyniki według innych metryk, zwracał sensowny podział zbioru. Podział ten nie pozwala jednak wyciągnąć zbyt wielu wniosków na temat tego, co odróżnia od siebie wskazane grupy nabywców.

Otrzymane wyniki mogą sugerować źle dobrane cechy zbioru, co bardzo utrudnia lub nawet uniemożliwia otrzymanie klasteryzacji, z której możnaby wyciągnąć wnioski na temat typów użytkowników, którzy robią zakupy internetowe.

Bibliografia

1. https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset
2. https://support.google.com/analytics/answer/2525491?hl=en&ref_topic=6156780
3. https://scikit-learn.org/stable/modules/clustering.html#silhouette-coefficient
4. https://scikit-learn.org/stable/modules/clustering.html#davies-bouldin-index
5. https://scikit-learn.org/stable/modules/clustering.html#calinski-harabasz-index

Oświadczenie

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

Martyna Majchrzak (298826), Agata Makarewicz (298827), Renata Rólkiewicz (298840)