Wstęp do uczenia maszynowego - projekt 2

Karol Degórski, Piotr Marciniak i Paweł Niewiadowski

1 Opis problemu

Problem postawiony w pracy projektowej nr 2 polega na stworzeniu klasteryzacji klientów danego sklepu internetowego na podstawie danych dotyczących podstawowych informacji o danej sesji użytkownika, danych pochodzących z Google Analytics i innych takich jak np. miesiąc.

2 Opis zbioru danych

W pracy projektowej wykorzystaliśmy zbiór danych **Online Shoppers Purchasing Intention**, pochodzący ze strony archive.ics.uci.edu.

Dane zawarte w tym zbiorze dotyczą zakupów dokonywanych w sklepie internetowym. Podstawowym podziałem klientów, o których mamy informację jest podział na tych, którzy dokonali zakupu, czyli przynieśli zysk dla sklepu, oraz tych, którzy nie dokonali zakupu (kolumna Revenue).

Zbiór danych zawiera też informacje ogólne takie jak miesiąc, czy jest weekend oraz informację o special day (naturalnie jeśli jest to jakiś dzień wyprzedaży, promocji, czy też np. dzień matki albo ojca to sprzedaż sklepu rośnie). Jest też informacja o tym, czy dany użytkownik jest użytkownikiem powracającym. Tacy użytkownicy mogą charakteryzować się sprecyzowanymi wymaganiami i dlatego też może być to pomocna informacja przy klasteryzacji.

Ponadto jest też informacja o sprzęcie z jakiego użytkownik nawiązywał połączenie ze sklepem internetowym np. rodzaj przeglądarki, system operacyjny, czy też rodzaj ruchu. Takie informacje również mogą być pomocne przy klasteryzacji, ponieważ można przypuszczać, że osoby używające systemu operacyjnego macOS, będą dokonywali wyższych zakupów z uwagi na ceny produktów marki Apple.

Co więcej, zbiór danych zawiera szereg informacji pochodzących z usługi Google Analytics, a dokładniej są to: BounceRates, ExitRates i PageValues, czyli średnie po odpowiednio Bounce Rate, Exit Rate i Page Value dla danej sesji. Te trzy cechy wydają się bardzo istotne dla naszego problemu. Bounce Rate jest stosunkiem użytkowników nieaktywnych do wszystkich użytkowników danej strony. Exit Rate jest stosunkiem wyświetleń strony jako ostatniej w danej sesji do wszystkich wyświetleń. Z kolei Page Value to po prostu liczba wizyt użytkownika danej strony, która poprzedziła dokonanie transakcji.

Ostatnimi informacjami zawartymi w zbiorze danych to Administrative, Informational i ProductRelated oraz odpowiadające im ich czasy trwania. Są to informacje o liczbie odwiedzin stron odpowiednio administracyjnej, informacyjnej oraz związanej z produktem przez użytkownika w danej sesji. Tak jak już napisaliśmy zbiór ten zawiera również informacje o całkowitym czasie spędzonym przez użytkownika na oglądaniu tych 3 rodzajów stron.

Dane składają się z 12330 rekordów bez żadnych braków. Podsumowanie zawierające opis oraz typ wszystkich zmiennych umieściliśmy w Tabeli 1 w załącznikach.

3 Opis EDA i preprocesingu zastosowanego w modelu

W celu zapoznania się z danymi oraz znalezienia między nimi zależności przeprowadziliśmy eksploracyjną analizę danych. Okazało się, że wiele zmiennych zawartych w zbiorze danych ma rozkład wykładniczy. Ponadto dużo zmiennych ma rozkład prawostronnie skośny oraz dużo obserwacji bliskich 0. Dlatego też, stwierdziliśmy, że być może warto będzie transformować zmienne logarytmicznie.

Zauważyliśmy również, że mamy wiele zmiennych kategorycznych takich jak np. Browser, czy OperatingSystems, które są zakodowane w postaci liczb naturalnych. Uznaliśmy, że w tym przypadku nie ma możliwości porównywania przeglądarek tak jak liczb naturalnych, np. że jedna jest lepsza od drugiej o 2. Dlatego też, aby lepiej oddać różnice pomiędzy przeglądarkami będziemy używali one hot encodera.

Ciekawą obserwację zauważyliśmy dla zmiennej month. Okazało się, że nie ma żadnych obserwacji ze stycznia ani z kwietnia, natomiast w marcu, maju, listopadzie i grudniu zaobserwowano znaczny wzrost obserwacji. Możliwe, że dane te zostały przesunięte w czasie. Wykres przedstawiono na Rysunku 1.

Eksploracyjna analiza danych pozwoliła nam też na odkrycie korelacji między zmiennymi w zbiorze danych. Zauważyliśmy silną korelację pomiędzy ExitRates, a BounceRates, czyli pomiędzy zmiennymi pochodzącymi z usługi Google Analytics. Ponadto zauważyliśmy zależność pomiędzy ProductRelated_Duration a ProductRelated, która nie jest dziwna, bo to po prostu zależność między liczbą odwiedzin strony powiązanej z danym produktem, a czasem przebywania na tej stronie. Podobne, choć nie tak silne korelacje zauważyliśmy pomiędzy Administrative a Administrative_Duration, oraz Informational a Informational_Duration. Analogicznie są to zależności pomiędzy liczbą odwiedzin, a czasem spędzonym na danym typie strony.

Następnie przeszliśmy do wykonania preprocessingu danych. Stworzyliśmy dwa datasety: dane przetransformowane logarytmicznie oraz bez wykonania takiej transformacji. Tak jak wcześniej wspomnieliśmy: system operacyjny, przeglądarkę i region zakodowaliśmy używając one hot encodera, bo nie ma między nimi porządku, natomiast miesiące zakodowaliśmy ordinal encoderem, ponieważ występuje dla nich porządek.

Ponadto usunęliśmy outliery używając Isolation Forest. Jako, że mamy do czynienia z klasteryzacją, to przeskalowaliśmy nasze dane używając Standard Scalera. Z tak przygotowanymi dwoma zbiorami danych przeszliśmy do przeprowadzenia modelowania, a dokładniej klasteryzacji.

4 Opis wyboru modelu oraz sposobu doboru hiperparametrów

Do ewaluacji naszych modeli wybraliśmy 3 metryki: miarę silhouette, Calinski-Harabasz oraz Davies-Bouldin. Miara silhouette mówi nam o tym, czy nasze klastry są odpowiednio odseparowane i dobrze upakowane. Miara Calinski-Harabasz pozwala nam na sprawdzenie, czy nasze klastry są odpowiednio gęste i dobrze odseparowane. Natomiast Davies-Bouldin pozwala sprawdzić odseparowanie klastrów. Mimo, że metryki te oceniają podobne cechy klasteryzacji to użycie tych 3 pozwala na lepszy wybór optymalnego modelu.

Sprawdziliśmy klasteryzację dla następujących modeli: KMeans, Affinity Propagation, Agglomerative Clustering z linkage single, complete, average oraz ward, DBSCAN, Gaussian Mixture i OPTICS. Dla tych modeli, które wymagają podania liczby klastrów sprawdzaliśmy wartości metryk dla liczb od 2 do 20 (Patrz Rysunek 2). Dla każdego z tych modeli wybraliśmy na podstawie metryk kilka najlepszych i przy użyciu PCA oraz t-SNE wizualizowaliśmy ich działanie, na odpowiednio płaszczyznę 3D i 2D, aby móc też ocenić ich działanie na tej podstawie.

5 Podsumowanie

Po przeanalizowaniu danych i ewaluacji potencjalnych modeli doszliśmy do wniosku, że wybór odpowiedniego modelu jest niejednoznaczny. Różne metryki faworyzowały różne modele, dlatego temat potraktowaliśmy bardzo subiektywnie. Dla zdefiniowanych klastrowań przypisaliśmy naszym zdaniem najlepsze hiperparametry. Po czym zwróciliśmy uwagę na wyniki przedstawione w tabeli 2 oraz na ich wizualizację. Naszym zdaniem najlepiej spisał się KMeans, który jako preprocessing brał dane zlogarytmowane i dzielił je na 4 klastry. Po przeanalizowaniu rozkładów wartości dla danych klastrów zauważyliśmy kilka ciekawych obserwacji:

- Osoby należące do klastra 0
  - mają **PageValues** równy 0.
  - rzadko odwiedzają strony i nie spędzają tam dużo czasu.
  - mają wyższy wskaźnik **BounceRate** i **ExitRate** od pozostałych.
  - nie przynoszą zysku.
- Osoby należące do klastra 1
  - w całości są klientami powracającymi.
  - używają głównie systemu nr 2 lub 3
  - używają głównie przeglądarki nr 2
- Osoby należące do klastra 2
  - mają prawie w całości **Revenue** równy 0, czyli nie przynoszą dochodu,
  - korzystają z przeglądarki nr 2.
  - korzystają z systemu nr 2.
- Osoby z klastra 3
  - korzystają w znacznej większości z przeglądarki nr 1
- **Weekend** oprócz outlierów zawiera dla prawie każdego klastra podobny stosunek dni zakupów w weekend, a w tygodniu.
- Osoby z klastrów 1, 2, 3 mają niższe wartości **BounceRates**.
- Osoby z klastrów 1, 2, 3 mają niższe wartości **ExitRates**.

6 Załączniki

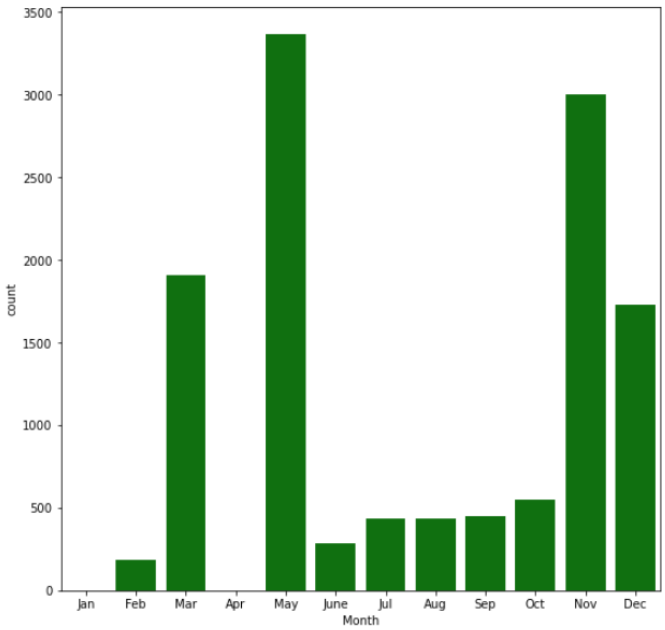
Tablica 1: Opis zmiennych

nazwa	typ	opis
Administrative	integer	liczba odwiedzin strony administracyjnej przez użytkownika w danej sesji
Administrative_Duration	float	całkowity spędzony przez użytkownika czas na oglądaniu stron administracyjnych
Informational	integer	liczba odwiedzin strony informacyjnej przez użytkownika w danej sesji
Informational_Duration	float	całkowity spędzony przez użytkownika czas na oglądaniu stron informacyjnych
ProductRelated	integer	liczba odwiedzin strony powiązanej z produktem przez użytkownika w danej sesji
ProductRelated_Duration	float	całkowity spędzony przez użytkownika czas na oglądaniu stron o produkcie
BounceRates	float	procent odwiedzających stronę i ją opuścili, bez wykonania żadnych interakcji z serwerem analizującym daną sesję
ExitRates	float	procent odsłon strony w ostatniej sesji
PageValues	float	średnia liczba stron odwiedzonych przed wykonaniem transakcji e-commerce
SpecialDay	float	cecha reprezentuje, bliskość czasu odwiedzenia strony, do specjalnego dnia np. dnia mamy, walentynek, kiedy to sfinalizowanie transakcji jest bardziej prawdopodobne
Month	object (string)	miesiąc odwiedzenia strony
OperatingSystems	integer	system operacyjny użytkownika
Browser	integer	przeglądarka użytkownika
Region	integer	region z którego użytkownik ogląda stronę
TrafficType	integer	rodzaj ruchu na stronie
VisitorType	object (string)	typ odwiedzającego stronę
Weekend	boolean	informacja czy użytkownik odwiedził stronę w weekend
Revenue	boolean	informacja o przychodzie

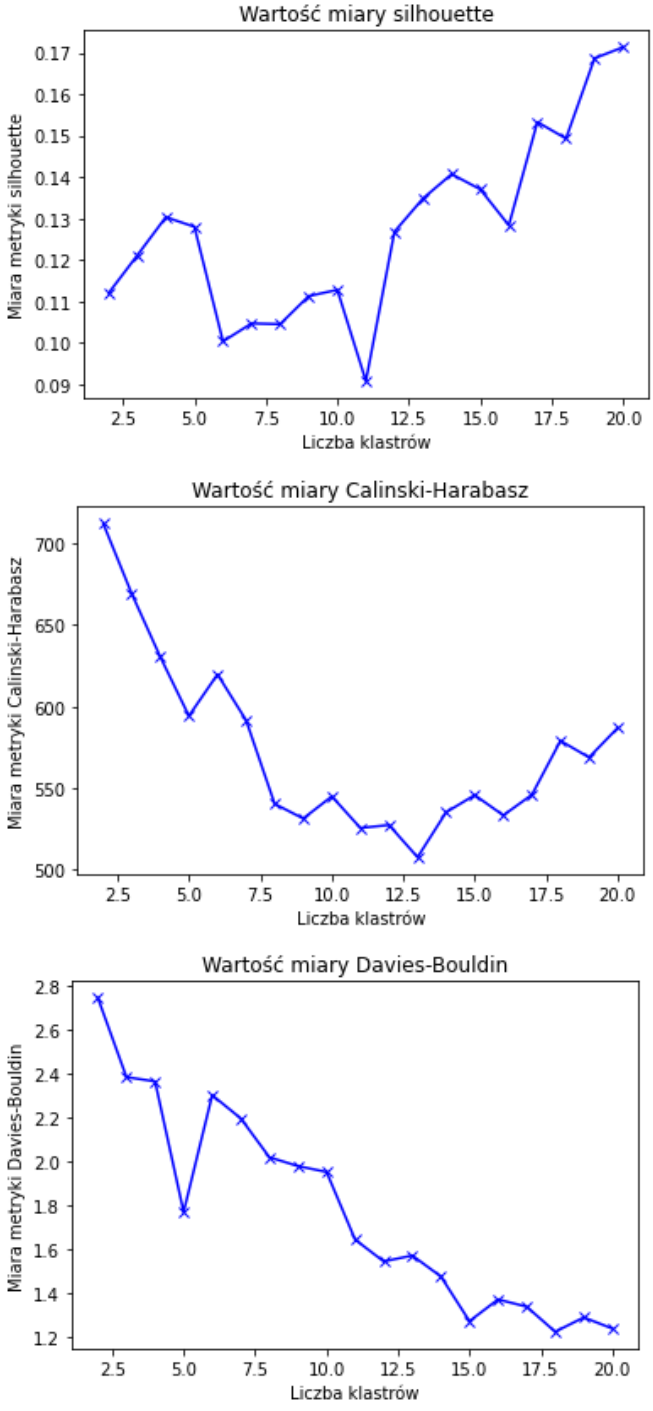
Tablica 2: Podsumowanie wybranych modeli

	Silhouette	Calinski-Harabasz	Davies-Bouldin
K-Means	0.130302	630.447644	2.363371
Agglomerative ward	0.143914	532.000408	1.316800
DBSCAN	0.299328	275.802153	2.919611
Gaussian	0.128698	495.283169	1.640207
Optics	0.281913	271.441550	1.182212

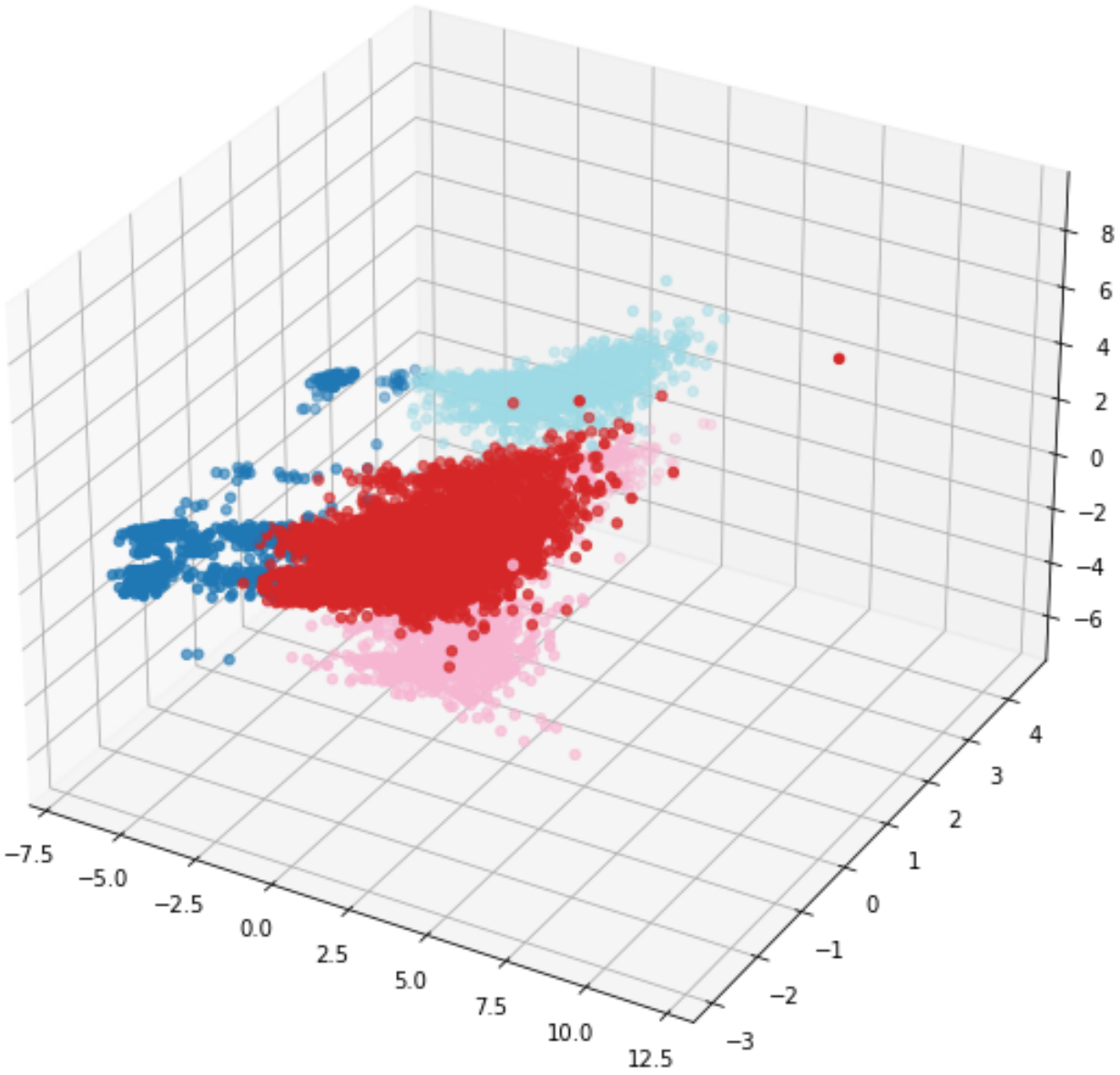
Rysunek 1: Rozkład miesięcy



Rysunek 2: Wykresy metryk dla K-Means



Rysunek 3: Podział na klastry K-means



Rysunek 4: Podział na klastry z t-SNE