Rozpoznawanie płci na podstawie własności akustycznych głosu

Marcin Wilk, Hubert Drążkowski

Wydział Matematyki i Nauk Informacyjnych, Politechnika Warszawska

Wstęp do uczenia maszynowego
20.04.2021

Plan prezentacji

1. Opis problemu i zbioru danych
2. Analiza eksploracyjna
3. Podstawowe modele
4. Zaawansowane modele
5. Wnioski

Opis problemu

Nasz zbiór danych powstał poprzez nagrywanie głosu mężczyzn i kobiet, a następnie przepuszczenie nagrań przez program R-owy, który generował na podstawie nagrania różne statystyki dotyczące na przykład średniej częstotliwości osoby mówiącej. Naszym zadaniem było na podstawie tych danych oraz ich etykiet nauczyć nasz model rozpoznawać kiedy mówiącym jest kobieta, a kiedy mężczyzna.

Opis zbioru danych

Zbiór ten był stosunkowo nieduży, liczył sobie początkowo 3168 obserwacji, 20 zmiennych objaśniających oraz jedną zmienną z etykietami. Nie było żadnych braków danych, więc nie musieliśmy na szczęście nic uzupełniać. Wszystkie zmienne objaśniające były również numeryczne. Proporcje między mężczyznami i kobietami były dokładnie równe, więc nie mieliśmy problemu z niezbilansowanymi klasami.

Analiza eksloracyjna

Podczas wstępnej analizy odkryliśmy, że:

1. Większość zmiennych ma wartości z przedziału (0,1), zatem nie trzeba zastanawiać się nad ich skalowaniem. Było jednak kilka zmiennych o stosunkowo dużych wartościach, z tego powodu w dalszej części spróbowaliśmy je przetransformować tak, by miały podobne zakresy jak pozostałe.
2. Jest kilka par zmiennych bardzo silnie skorelowanych ze sobą. Również wartości współczynników VIF przy wielu zmiennych były bardzo wysokie, co świadczyło o silnej współliniowości zmiennych. Z tego powodu rozważyliśmy usunięcie kilku z takich zmiennych, aby to poprawić.
3. Zmienne meanfun oraz Q25 wydawały się najlepiej separować obie klasy. Ich histogramy z podziałem na płeć były niemalże rozłączne.

Przygotowanie zbiorów danych do konstruowania modeli

Chcieliśmy sprawdzić jak transformacje czy usunięcie niektórych zmiennych lub obserwacji odstających wpłyną na jakość naszych modeli. Poniżej zamieszczamy krótki opis jaką metodyką były przygotowywane te zbiory danych.

1. df1 - wejściowy zbiór danych z usuniętymi jedynie zmiennymi o korelacjach bardzo bliskich jedynce.
2. df2 - df1 bez kolejnych 4 zmiennych, które uznaliśmy za zbyt współliniowe z innymi na podstawie macierzy korelacji lub współczynników VIF
3. df3 - df2 z zastosowanymi transformacjami na kilku zmiennych, które były dzięki temu lepiej skalowane lub ich rozkład stawał się symetryczny
4. df4 - df3 z usuniętymi obserwacjami odstającymi na podstawie metod bardziej intuicyjnych jak odstawanie od rozkładu, nierówność Markowa
5. df5 - df3 z usuniętymi obserwacjami odstającymi zidentyfikowanymi przez algorytm DBSCAN

Podstawowe modele

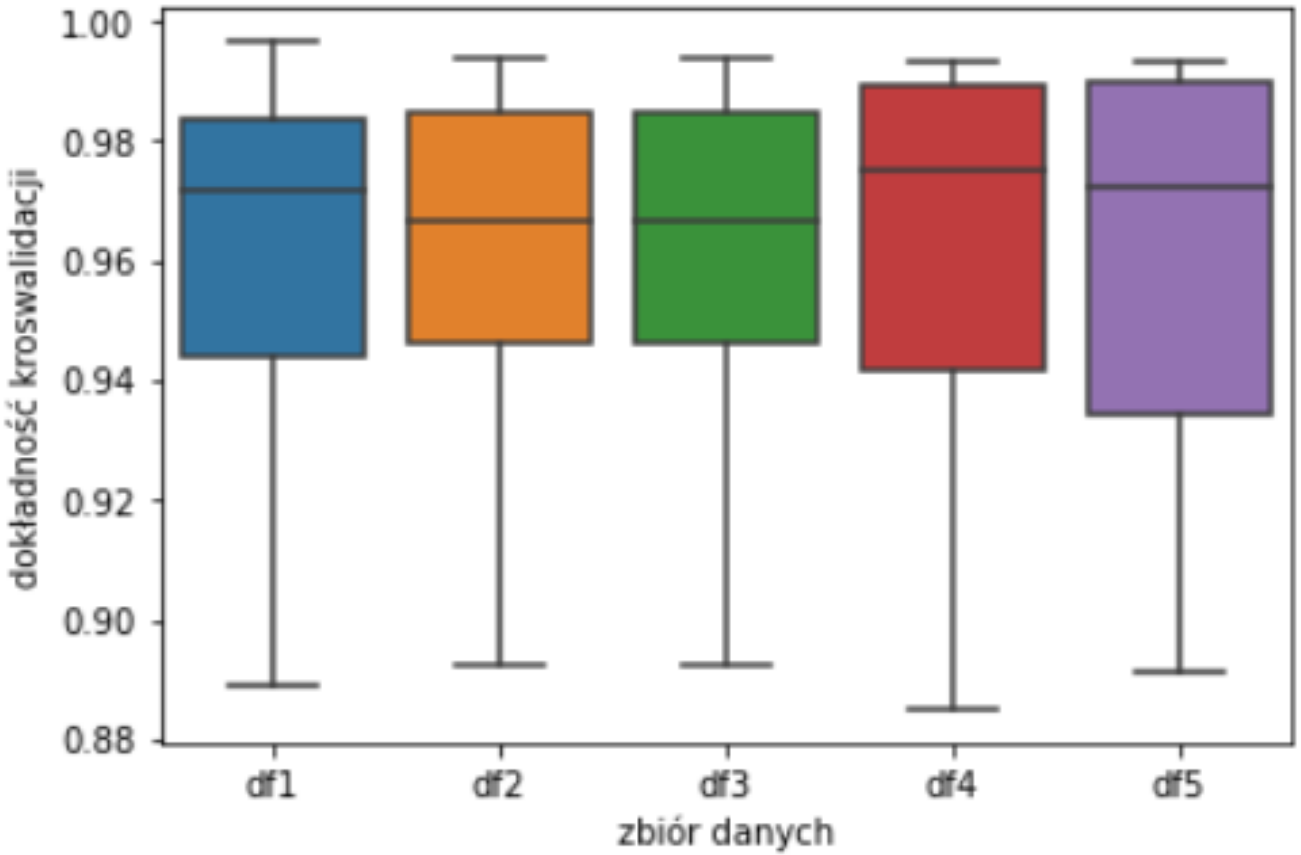
Rozważyliśmy 2 podstawowe modele do porównania z pozostałymi. Pierwszym z nich była zwykła regresja logistyczna jedynie dla zmiennej meanfun, która powinna najlepiej separować nam klasy. Okazuje się, że dokładność jaką daje model wytrenowany jedynie na tej zmiennej jest zaskakująco dobra, oscyluje ona około 0.96. Co ciekawe dla zbiorów z usuniętymi obserwacjami odstającymi jest ona trochę lepsza, ale ma większą wariancję. Drugim modelem była regresja logistyczna z regularyzacją grzbietową. Okazała się ona gorsza od poprzedniego modelu, jej dokładność obniżyła się do około 0.9.

Las losowy

Pierwszym z zaawansowanych modeli, które chcieliśmy wytestować był las losowy. Wzięliśmy do niego 1000 drzew. Aby zapobiec ewentualnemu przetrenowaniu ograniczyliśmy ich głębokość do 4 oraz minimalną liczbę obserwacji do podziału węzła na 10. Okazało się, że taki model minimalnie poprawia dokładność dla każdego ze zbiorów danych do około 0.97. Ponownie jednak ramki danych 4 i 5 okazały się mieć minimalnie lepszą dokładność kosztem wariancji.

Las losowy

Zobaczmy jeszcze jak dokładność na zbiorach kroswalidacyjnych prezentuje się na boxplocie.



Komitet 3 modeli

Kolejną z naszych prób było zastosowanie komitetu modeli QDA, naiwnego Bayesa oraz regresji logstycznej. Tym razem zastosowaliśmy go tylko do zbioru df3. Wagi stroiliśmy jako hiperparametry. Nie mogliśmy użyć jednak metody siatki lub też losowej siatki, gdyż nasze hiperparametry muszą sumować się do jedynki. Wagi te pochodziły z 'hipersiatki' zdefiniowanej na brzegu trójwymiarowego symplexu. Braliśmy wielkości będące wielokrotnościami 0.05.

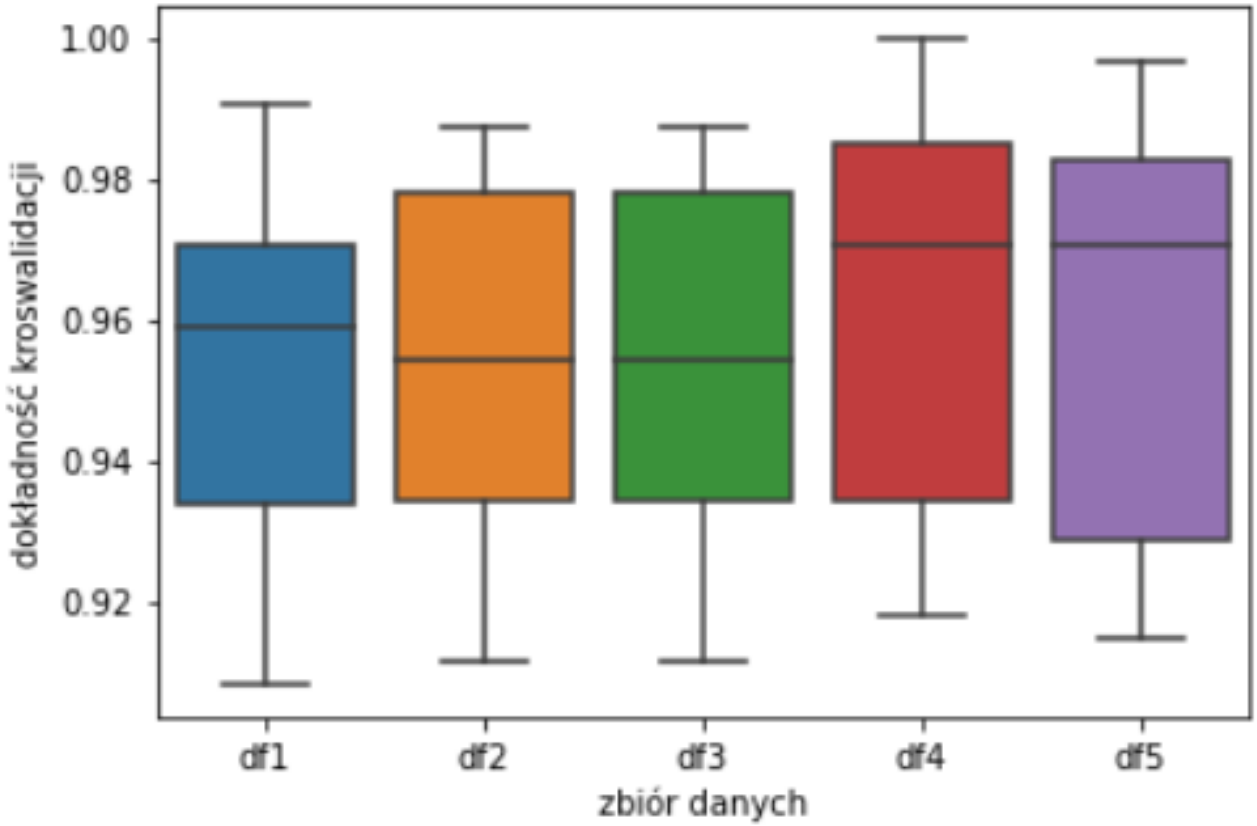
Komitet 3 modeli

Na wybranym zbiorze testowym najlepsze okazało się 5 komitetów z dokładnością 0.957413 oraz wagami przedstawionymi w tabelce.

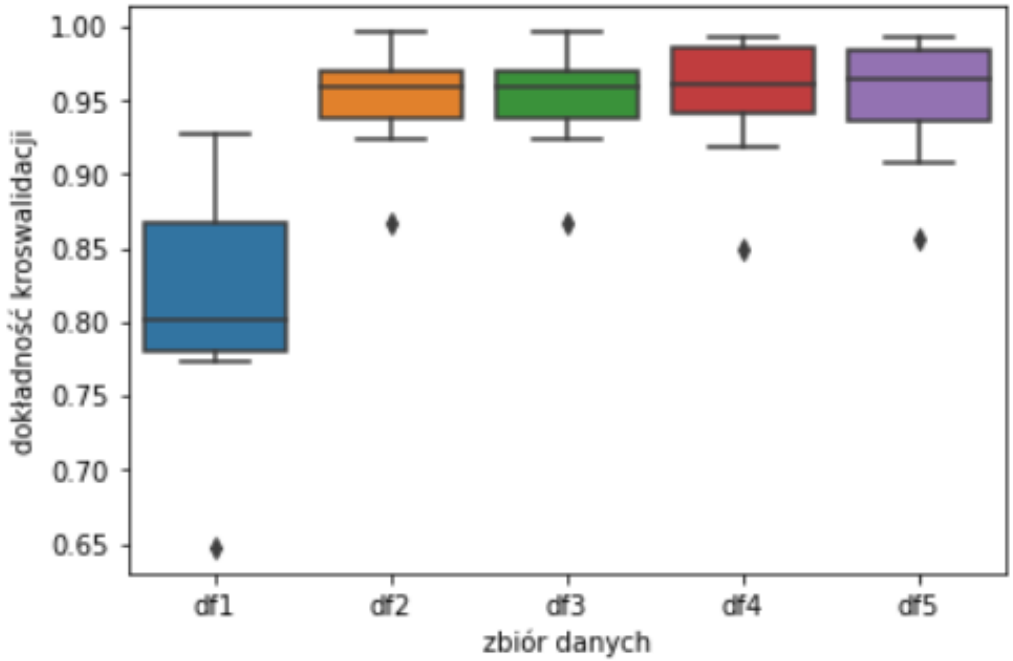
QDA	$Bayes$	RL
0.35	0.1	0.55
0.4	0.2	0.4
0.4	0.25	0.35
0.45	0.3	0.25
0.45	0.35	0.2

Komitet 3 modeli

Wzięliśmy nastepnie jeden z najlepszych modeli i sprawdziliśmy go kroswalidacyjnie na naszych zbiorach danych. Rezultaty były ciekawe, co możemy zobaczyć na wykresie.

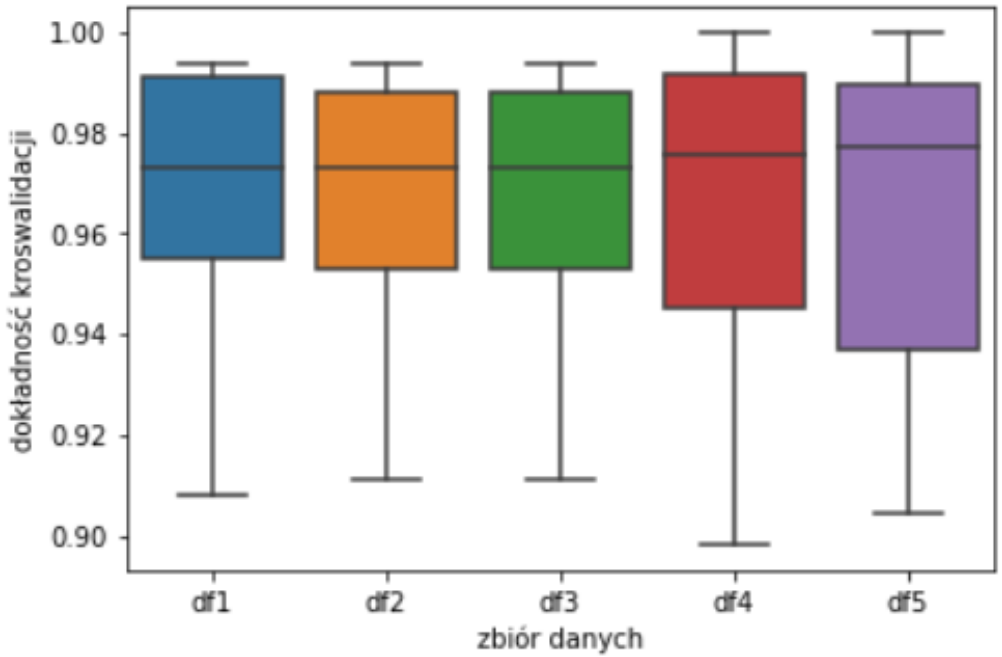


Sprawdziliśmy również jak spisze się tuningowany na siatce SVM na naszych zbiorach danych. Tutaj widać znaczną różnicę pomiędzy pierwszym zbiorem danych, gdzie mieliśmy najwięcej zmiennych współliniowych, a pozostałymi zbiorami.



XGBoost

Ostatnią próbą poprawienia dokładności był XGBoost tuningowany ze względu na liczbę estymatorów. Optymalna okazała się liczba 360 i tak prezentują się wyniki kroswalidacji:



Standaryzacja zmiennych

W kilku źródłach znaleźliśmy informację, że czasem standaryzacja zmiennych poprawia dokładność algorytmu. Postanowiliśmy więc to sprawdzić dla każdego z modeli i porównać wyniki czy dała ona pozytywną czy negatywną zmianę.

1. Regresja grzbietowa uległa znacznej poprawie, dokładność wzrosła o dobre kilka punktów procentowych.
2. Las losowy pozostał praktycznie na tym samym poziomie dokładności
3. Komitet i XGBoost nieznacznie zyskały na standaryzacji
4. SVM zdecydowanie poprawił dokładność po standaryzacji, szczególnie dla pierwszej ramki danych

Wnioski

- Bardzo poprawny model dało się wytrenować już na jednej ze zmiennych jaką była meanfun.
- Komplikacja modeli poprawiała dokładność przeciętnie o 1-3 punkty procentowe
- Odrzucenie obserwacji odstających zazwyczej powodowało niewielki wzrost dokładności kosztem wariancji
- Standaryzacja była w niektórych przypadkach bardzo pożyteczna, a w innych napewno nie szkodziła