Gender voice recognition

AUTORZY: MICHAŁ PIASECKI, BARTOSZ SIŃSKI

Cel projektu

Naszym zadaniem było wytrenowanie modelu, który będzie prawidłowo klasyfikował płeć osoby, która wypowiadała się w danym nagraniu.

Dane, które posiadaliśmy to około 3000 rekordów. Każdy z nich posiadał 20 zmiennych, które były różnymi parametrami statystycznymi danego nagrania oraz kolumna, która wskazywała płeć osoby.

Jak widzimy po prawej mamy zmienne takie jak: średnia częstotliwość, odchylenie standardowe, pierwszy kwartyl i wiele wiele innych.

Nazwy kolumn w zbiorze danych

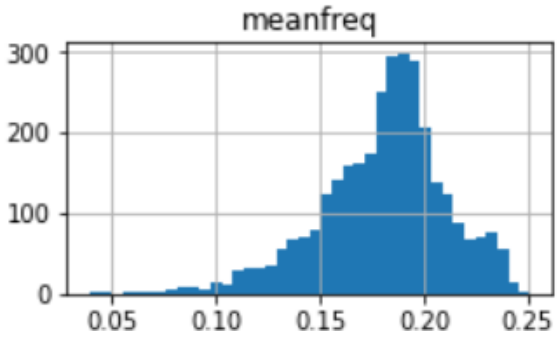
meanfreq
sd
median
Q25
Q75
IQR
skew
kurt
sp.ent
sfm
mode
centroid
meanfun
minfun
maxfun
meandom
mindom
maxdom
dfrange
modindx
label

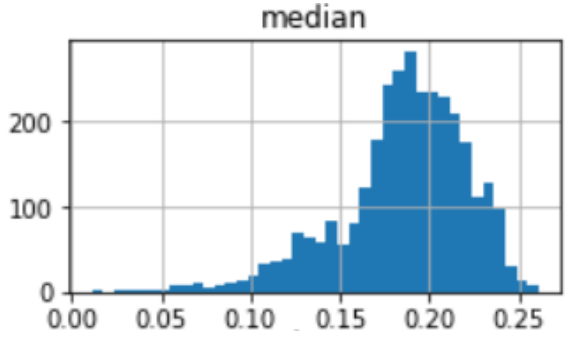
Eksploracja danych

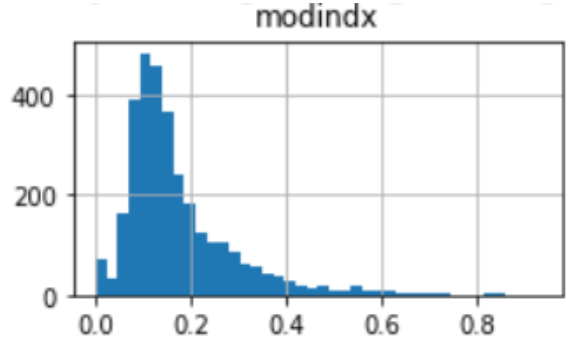
Wszystkie zmienne, które opisują naszą kategoryczny target są zmiennymi ilościowymi. Dodatkowo nie mamy w naszym zbiorze żadnych wartości brakujących. Ilość rekordów zaklasyfikowanych jako męskie jest dokładnie taka sama jak ilość rekordów damskich (po 1584 w każdym)

Poniżej możemy zobaczyć przykładowe histogramy zmiennych opisujących nasz target

(od lewej: średnia częstotliwość nagrania, mediana częstotliwości oraz indeks modulacji)







Korelacje zmiennych ze sobą

Jako że posiadamy 20 zmiennych statystycznych opisujących dane nagrania wideo, możemy domyślać się, że wiele z nich będzie bardzo mocno skorelowanych. Po szybkiej analizie okazuje się to być prawdą

Jak możemy zobaczyć w tabeli po prawej stronie dla wielu zmiennych współczynnik korelacji jest większy niż 0.8 ! Daje nam to jasno do zrozumienia, że w celu uproszczenia naszego modelu będziemy mogli pozbyć się wielu z tych zmiennych. Widzimy, że będziemy mogli się pozbyć jednej z elementów z dwójek (dfrange, maxdom), (meanfreq,median) czy (centroid, Q26)

dfrange	maxdom	0.999
maxdom	dfrange	0.999
meanfreq	median	0.920
median	meanfreq	0.920
	centroid	0.920
centroid	median	0.920
meanfreq	Q25	0.906
Q25	meanfreq	0.906
	centroid	0.906
centroid	Q25	0.906
sp.ent	sfm	0.894
sfm	sp.ent	0.894
skew	kurt	0.888
kurt	skew	0.888
sd	sfm	0.879
sfm	sd	0.879
Q25	IQR	0.870
IQR	Q25	0.870
sp.ent	sd	0.861
sd	sp.ent	0.861
	IQR	0.859
IQR	sd	0.859
sd	Q25	0.831
Q25	sd	0.831
sfm	centroid	0.824

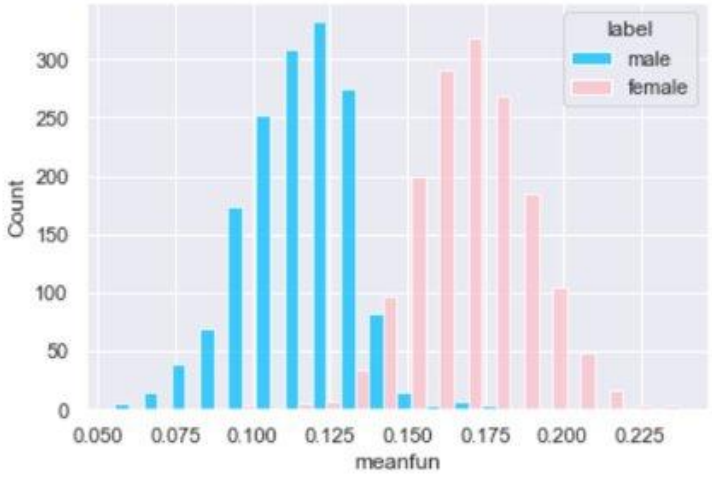
Korelacje zmiennych z targetem

Kolejnym elementem naszej eksploracji jest przyjrzenie się, które zmienne są najbardziej skorelowane z naszym targetem. Pozwoli nam to przewidzieć, które zmienne będą najważniejsze przy budowie modelu predykcyjnego.

	label_male	label_female
label_male	1.000000	-1.000000
label_female	-1.000000	1.000000
meanfun	-0.833921	0.833921
IQR	0.618916	-0.618916
Q25	-0.511455	0.511455
sp.ent	0.490552	-0.490552
sd	0.479539	-0.479539
sfm	0.357499	-0.357499
centroid	-0.337415	0.337415
meanfreq	-0.337415	0.337415
median	-0.283919	0.283919

Wykresy wzorów dla poszczególnych zmiennych

Poniższe wykresy pokazują, jak bardzo nasze niektóre pojedyńcze zmienne klasyfikują nasz target. Po lewej stronie widzimy histogram zmiennej meanfun przy podziale na głosy damskie i męskie. Możemy zobaczyć, że praktycznie wszystkie głosy męskie są poniżej 0.14 meanfreq, a damskie powyżej. Po prawej stronie widzimy jak zmienne median "ładnie" dzielą nam zbiór głosów na damskie i męskie.



Wybrane kolumny:

meanfreq
sd
median
Q25
IQR
skew
sp.ent
sfm
centroid
meanfun

Feature engineering

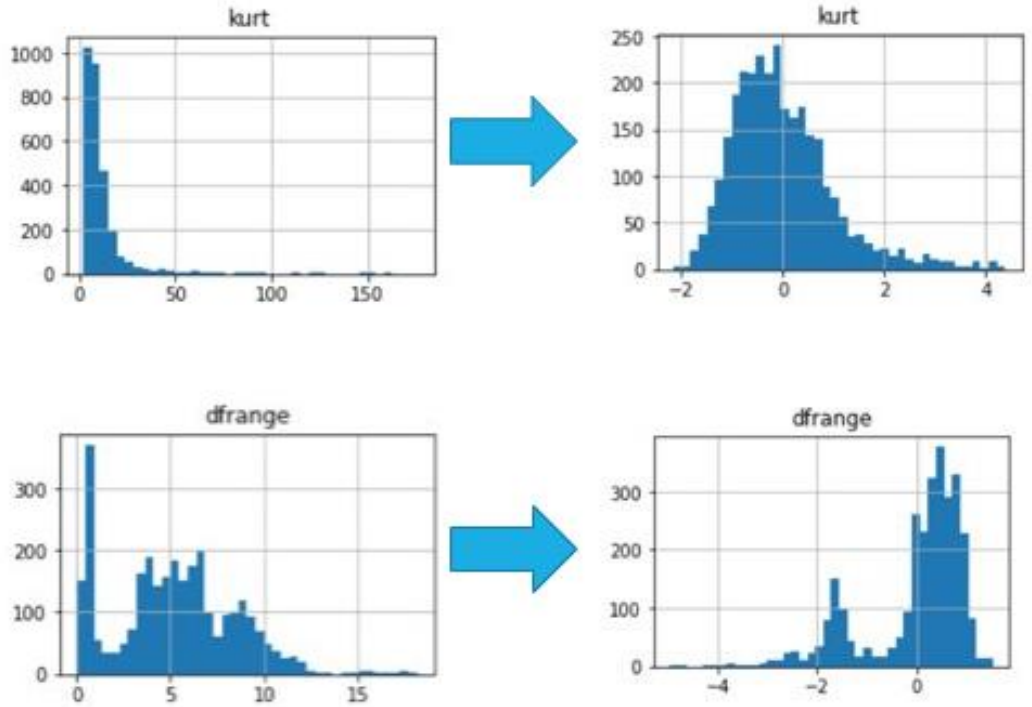
W ramach feature engineering zdecydowaliśmy się na następujące kroki:

1. Usunęliśmy 219 outlierów z naszego zbioru, które zakrzywiałyby działanie naszego modelu

2. Za pomocą SelectKBest wybraliśmy 10 zmiennych opisujących najlepiej nasz target. Dzięki temu uprościliśmy nasz wstępny model. Wybrane kolumny znajdują się w prawym górnym rogu.

3. Ustandaryzowaliśmy zmienne które pozostały . Dzięki temu wszystkie zmienne mają wartości z tego samego, małego przedziału liczbowego.

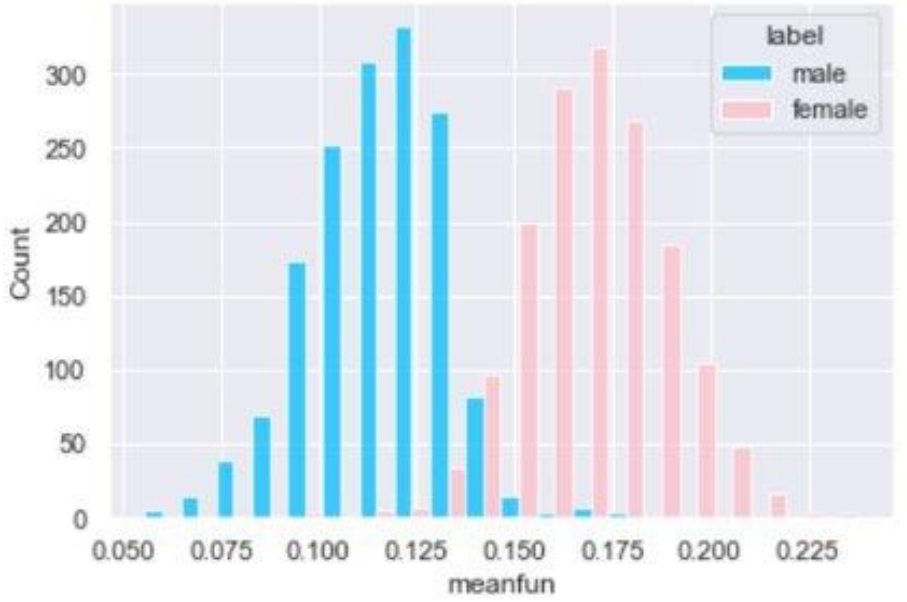
(po prawej stronie możemy zobaczyć histogramy zmiennych kurtosis oraz dfrange przed oraz po standaryzacji, przed: wartości bardzo rozproszone, po: gęsto obok siebie)



Pierwszy model baselinowy

Przed budowaniem naszego modelu postanowiliśmy zbudować dwa proste modele baselinowe, które będą wyznaczać nam standard, który chcemy na pewno przebić. Pierwszy model:

1. Inspirując się rozkładem zmiennej meanfun (patrz rysunek poniżej), zbudowaliśmy najprostszy model baselinowy, będący po prostu jedną **instrukcją warunkową.** Jeśli rekord ma meanfun poniżej 0.14 klasyfikujemy jako mężczyzna, jeśli powyżej jako kobieta. Otrzymana accuracy wynosi 95 %.



Drugi model baselinowy

Drugim modelem , który zastosowaliśmy była prosta regresja logistyczna bez regularyzacji. Tym sposobem osiągneliśmy accuracy na poziomie 98 %.

Model końcowy

Model, który postanowiliśmy użyć to klasyfikator SVM. Bez tuningu parametrów osiągneliśmy accuracy 0.991, natomiast po tuningu parametrów: C, kernel oraz gamma nasz model poprawił sie o 0.002 punkty procentowe I osiągnął accuracy 0.993. Próba zastoswania Polynomial Features pogorszyła rezultaty. Najlepsze parametry dla modelu są podane poniżej:

```
{'classifier_C': 8, 'classifier_gamma': 0.3, 'classifier_kernel': 'rbf'}
```

Pełne wyniki modelu

Wyniki naszego modelu są następujące:

1. Accuracy : **99.3** %
2. Precision: **99.7**%
3. Recall : **99.8%**
4. ROC AUC: **99.3%**
5. Średnie accuracy na CV : **96.8%**

Poniżej z lewej strony możemy zobaczyć tabelkę jak klasyfikował nasz model na zbiorach testowych. Po prawej stronie mamy wykres ROC-AUC.

	Actual positives	Actual negatives
Positive predictions	354	4
Negative predictions	1	379