Gender Recognition by Voice

Self-explanatory

Zespół:

Mateusze Sperkowski

Szymon Rećko

Wstęp

- Zadaniem, którym będziemy się zajmować w ramach pierwszego projektu będzie stworzenie modelu uczenia maszynowego, który po cechach statystycznych ekstrahowanych z nagrań ludzkich głosów będzie klasyfikował czy należy on do kobiety czy mężczyzny.

Zbiór danych

- Oryginalne nagrania głosu zostały przetworzone za pomocą pakietu tuneR w R.
- Dostępne w zbiorze dane:
- Cel Predykcji: Podpis płci (ilość danych 1:1).
- Częstotliwość głosu (kHz) oraz dane statystyczne (średnia, odchylenie standardowe, moda, kwartyle, kurtoza, skośność itd.).
- Ton podstawowy i dominujący (fundamental and dominant frequency), oraz ich dane statystyczne.
- Modulacja głosu.
- Nie ma w kolumnach żadnych brakujących danych.
- Wszystkie zmienne objaśniające są zmiennymi ciągłymi.

- Dane pochodzą z: https://www.apispreadsheets.com/datasets/119

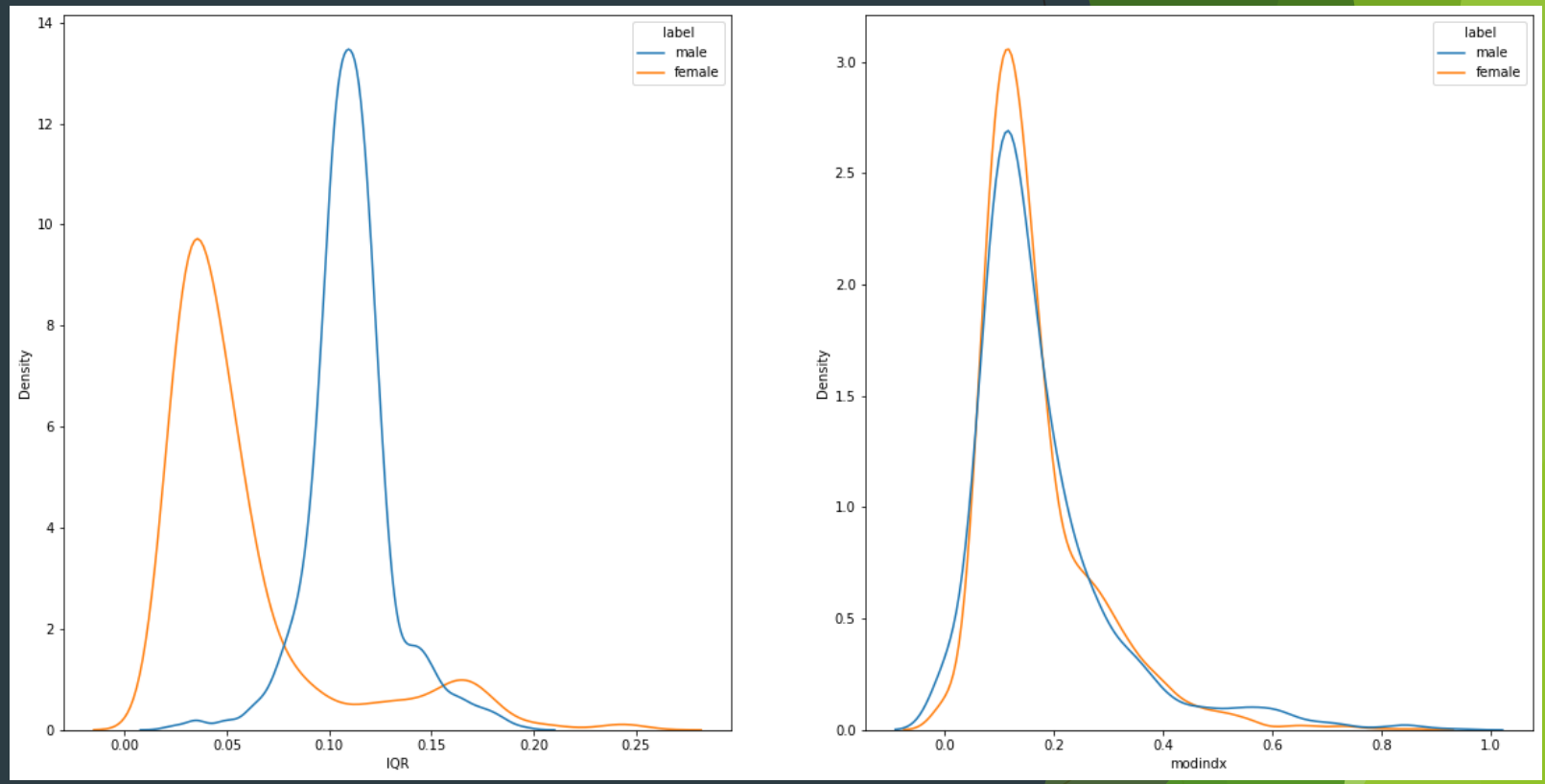
```
Data columns (total 21 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   meanfreq  3168 non-null   float64
 1   sd        3168 non-null   float64
 2   median    3168 non-null   float64
 3   Q25       3168 non-null   float64
 4   Q75       3168 non-null   float64
 5   IQR       3168 non-null   float64
 6   skew      3168 non-null   float64
 7   kurt      3168 non-null   float64
 8   sp.ent    3168 non-null   float64
 9   sfm       3168 non-null   float64
 10  mode      3168 non-null   float64
 11  centroid  3168 non-null   float64
 12  meanfun   3168 non-null   float64
 13  minfun    3168 non-null   float64
 14  maxfun    3168 non-null   float64
 15  meandom   3168 non-null   float64
 16  mindom    3168 non-null   float64
 17  maxdom    3168 non-null   float64
 18  dfrange   3168 non-null   float64
 19  modindx   3168 non-null   float64
 20  label     3168 non-null   object
```

Nasze wstępne Hipotezy

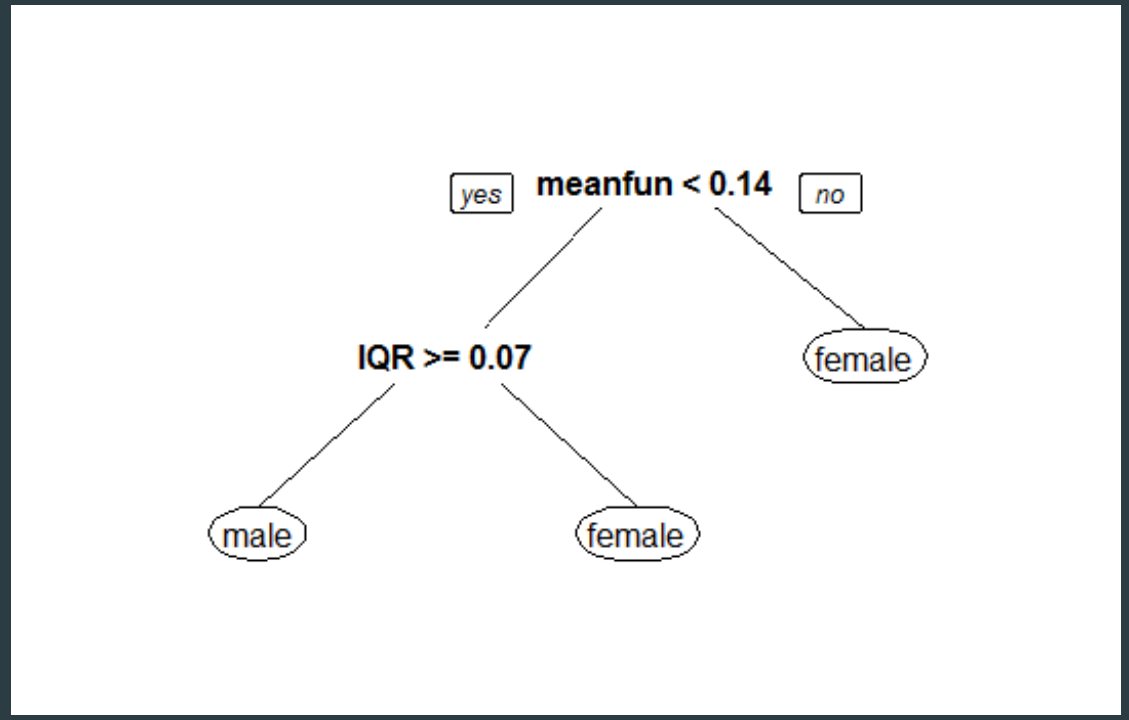
- Po doświadczeniach z życia codziennego, wiemy że rozpoznawanie płci danej osoby po głosie zazwyczaj jest zadaniem prostym. Jednak istnieją takie osoby których głos jest trudny do poznania płci, lub wręcz ich głos byśmy przydzielili do drugiej płci. Zapewne większość z nas popełniła w życiu taką pomyłkę.
- Założyliśmy że podobnie będzie w tym zbiorze danych. To znaczy, rozkład części danych dobrze rozgranicza płcie, ale są istnieją przypadki odstające, które wydają się być w drugiej z płci. Właśnie te będą wykorzystane do uczenia naszego modelu. Są zapewne też cechy danych które dla obu grup się pokrywają i nie będą użyteczne dla zadania rozpoznawania płci.

Eksploracyjna Analiza Danych

Dzięki stworzonej przez nas macierzy KDEplotów byliśmy w stanie sprawdzić czy rozkład cech znacząco się różni w zależności od płci (label'a). Na rysunku obok widzimy rozkłady zmiennej IQR oraz zmiennej modindx. W pierwszym przypadku obserwujemy wyraźną różnicę w rozkładach, co będzie później podstawą do tworzenia naszego modelu. Podobne do drugiego przykładu zależności rozkładów występowały w większości cech i nie są one aż tak użyteczne w rozdzielaniu.



Model CART



Okazuje się że istnieją dwie cechy które znacznie rozgraniczają dwie grupy klasyfikacji. Korzystając z wcześniejszej macierzy rozkładów stworzyliśmy 'model' niskiego drzewa decyzyjnego, którego accuracy na całym zbiorze wynosi 95.64(39)%.

Oczywiście chcemy w tym projekcie otrzymać model poprawiający skuteczność tego prostego klasyfikatora.

Image Source: http://www.primaryobjects.com/2016/06/22/identifying-the-gender-of-a-voice-using-machine-learning/

Inżynieria Cech

- Poszukiwanie najistotniejszych cech zbioru danych zaczęliśmy od stworzenia klasyfikatora RandomForest, który inherentnie tworzy hierarchię najważniejszych cech. Ten wybór okazał się potem tym za pomocą którego wybieraliśmy które cechy zostawić.
- Następnie stworzyliśmy alternatywne ramki danych gdzie wykorzystane zostały cechy wielomianowe stopnia 2, czyli wymnożenie przez siebie wszystkich kolumn naszego zbioru danych. Również dla tych cech wykorzystaliśmy RandomForest do znalezienia nowych najważniejszych cech.

- Postanowiliśmy za pomocą Analizy Głównych Składowych porównać wcześniej nadaną hierarchię cech z nową stworzoną właśnie tym algorytmem. Okazuje się że obie analizy doprowadziły nas do tych samych wyników.
- Dodatkowo chcieliśmy wypróbować metodę Non-negative Matrix Factorization, jednak przeciwnie do naszych założeń, okazało się że w tym zbiorze istnieją wartości ujemne, co uniemożliwiło jej zastosowanie.
- Tak jak sądziliśmy, znormalizowanie danych w tym zbiorze było istotnym krokiem do polepszenia wyników części bazowych modeli predykcyjnych. (Regresja logistyczna i SVM, ponieważ reszta modeli nie jest zależna od znormalizowania danych)

Eksperymenty z bazowymi modelami

- Do wstępnego porównania modeli stworzyliśmy bazowe wersje SVM, Regresja Logistyczna, Drzew Klasyfikujące, Naiwny Klasyfikator Bayesowski, Las Losowy. Najwyższe wyniki danych osiągał Las losowy i Drzewo Klasyfikujące.
- W trakcie prac dowiedzieliśmy się, że standaryzacja danych nie wpływa na wyniki osiągane przez Naiwny Klasyfikator Bayesowski, Drzewa Klasyfikujące i Las Losowy.
- Do zbioru testowego zostało przydzielone 33% danych, wartość ta została wybrana na podstawie wielkości naszego zbioru.
- Za pomocą RandomSearchCV oraz GridSearchCV staraliśmy się dostroić hiperparametry aby osiągać lepsze wyniki modelu, co zostanie szerzej opisane w następnym slajdzie.

Strojenie hiperparametrów

- Hiperparametry, które wybraliśmy do strojenia to: n_estimators, min_samples_split, min_samples_leaf, max_feature, max_depth, criterion i bootstrap.
- Otrzymane za pomocą RandomSearchCV najlepsze zestawy hiperparametrów użyliśmy do wytrenowania nowych modeli aby wyliczyć wyniki zwykłych metryk na tych zestawach. Ich wyniki wychodziły jednak gorzej niż dla modelu z domyślnymi parametrami, natomiast cross-validation wychodziło lepiej. Jednak w porównaniu z wynikami które otrzymaliśmy w RandomSearchCV, cross-validation na tych samych parametrach otrzymuje mniejsze wyniki. Nie udało się nam odkryć co jest tego powodem.

Finalny model

- Modelem, którego ostatecznie użyliśmy był RandomForest z 50 drzewami.

- Cechy, które wybraliśmy jako najważniejsze w trakcie inżynierii cech to: sd, Q25, IQR, sp.ent, sfm, mode, centroid, meanfun.

Wyniki finalnego modelu

Otrzymaliśmy wynik istotnie lepszy od prostego modelu CART.

- Accuracy: 0.9856
- Recall: 0.9847
- Precision: 0.9865
- F1: 0.9856
- ROC-AUC: 0.9856
- Cross-val: 0.9646

Wnioski

- Głosy niektórych ludzi są na tyle specyficzne, że sami czasem mamy problem z rozpoznaniem czy ich właścicielem jest mężczyzna czy kobieta. Z tego powodu nawet w przypadku osiągania świetnych wyników na zbiorze który mamy, trzeba brać pod uwagę, że pojawi się taki, którego nie będziemy umieli poprawnie skategoryzować.
- Mimo to problem w ogólności problem ten jest bardzo prosty i nie wymaga bardzo złożonego modelu do jego rozwiązania.