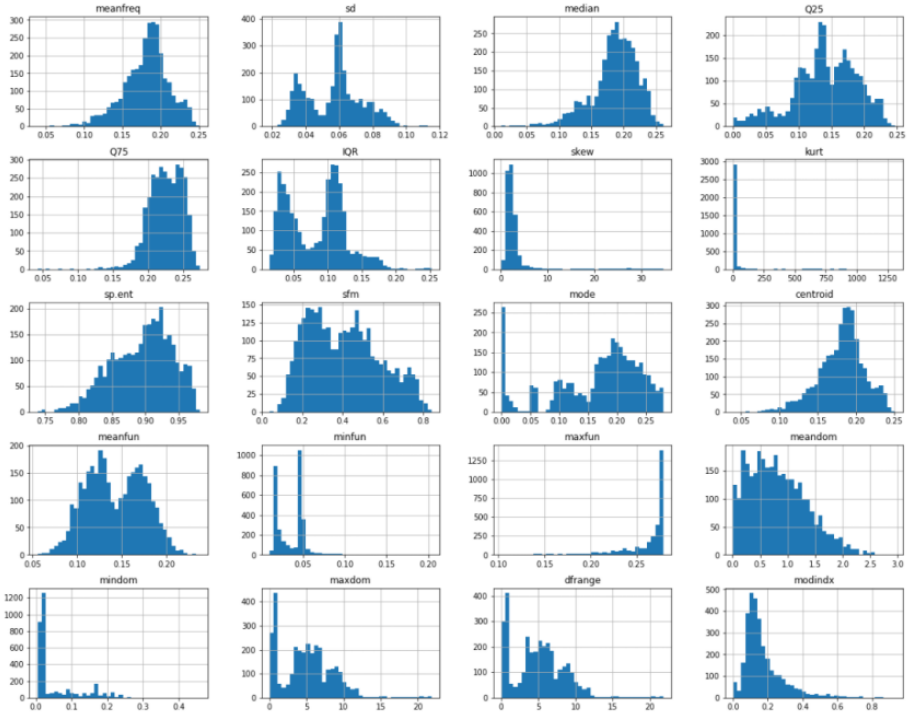
WUM - Projekt 1

Patryk Tomaszewski, Mateusz Ziemła

April 2021

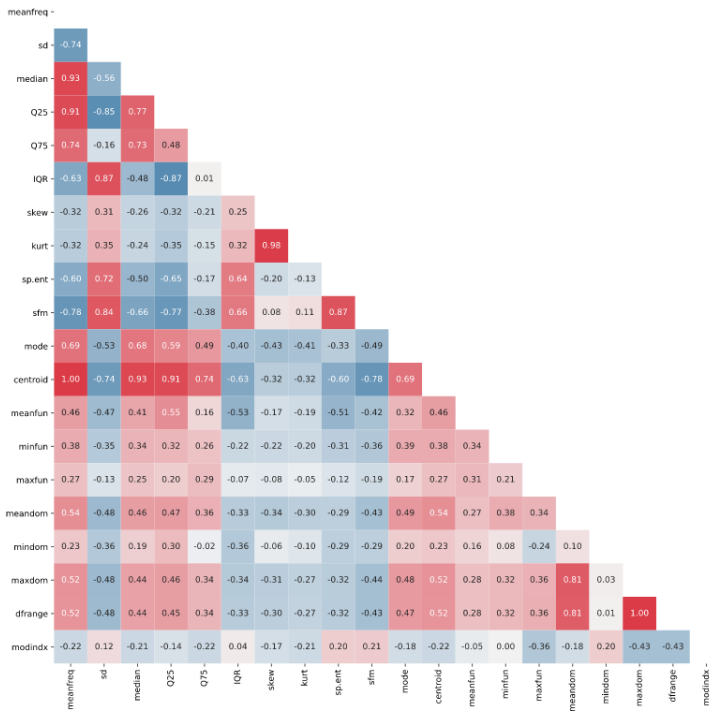
Eksploracja danych

Na początku skupiliśmy się na eksploracji danych, na których trenowaliśmy nasz model. Dostępnymi danymi były podstawowe zmienne opisujące dźwięk, oraz dane pochodzące z wywołania funkcji specprop w R. Są to między innymi takie zmienne jak średnia, mediana, oraz IQR częstotliwości. Wszystkie dane były odpowiednio podpisane oznaczeniami płci osoby do której należał głos na próbce, z której pochodziły dane.



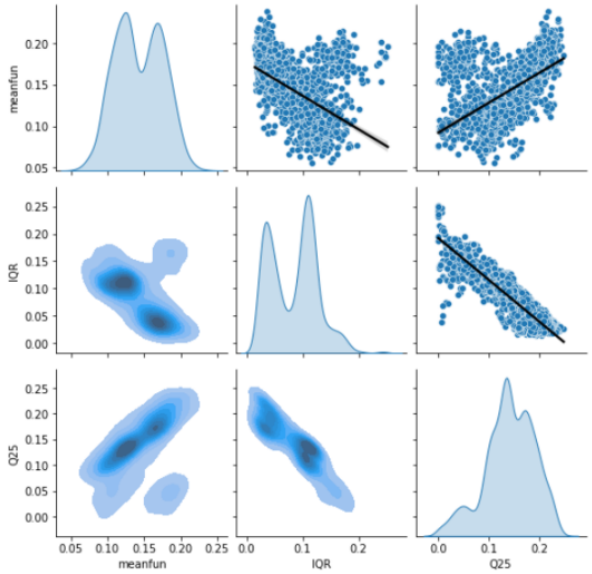
Rysunek 1: Rozkłady poszczególnych zmiennych

Sam dataset był bardzo przyjazny do pracy na nim - nie brakowało żadnych zmiennych, a wszystkie zmienne były w formacie numerycznym, bez konieczności konwersji. Także rozkłady najbardziej znaczących zmiennych były dogodne do pracy nad nimi i nie wymagały dodatkowych przekształceń.



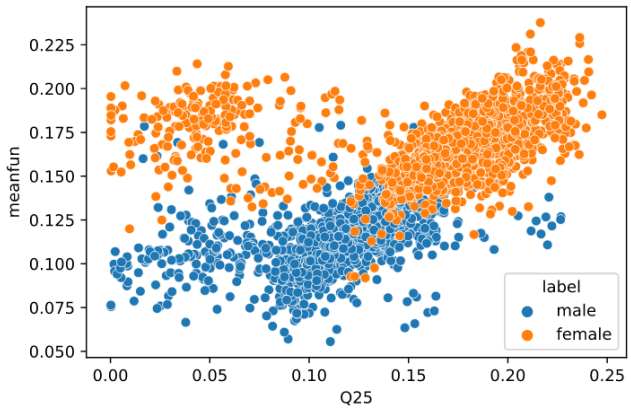
Rysunek 2: Korelacje poszczególnych zmiennych

Przy badaniu danych zauważyliśmy dużą korelację pomiędzy zmiennymi. W szczególności wyróżniły się zmienne centroid, kurt, oraz maxdom, które były praktycznie równoważne z innymi zmiennymi.



Rysunek 3: Korelacje oraz rozkłady trzech najważniejszych zmiennych

Zauważyliśmy również, że najbardziej skorelowane z podpisami danych (wg. korelacji Phika) są zmienne meanfun, IQR oraz Q25, opisujące odpowiednio średnią częstotliwość tonu podstawowego, IQR rozkładu częstotliwości, oraz pierwszy kwantyl częstotliwości. Z tych trzech zmiennych, dwie (IQR oraz Q25) były ze sobą najbardziej skorelowane, więc w późniejszych testach próbowaliśmy uzyskiwać wyniki z wyłączeniem jednej z nich.



Rysunek 4: Rozdział podpisów płci na wykresie meanfun Q25

Na powyższym wykresie widać, że całkiem dobrym modelem byłoby same rozróżnianie płci po tym, czy wartość zmiennej meanfun jest większa, czy mniejsza od 0.145, więc takiego modelu używamy później jako baseline.

Pierwsze modele

Na początku, pomijając wymyślony już wcześniej liniowy model polegający na jednym ifie, (opisujący 96% przypadków), sprawdziliśmy prostą sieć neuronową. Najpierw sprawdziliśmy, czy przy użyciu sieci neuronowej konieczny Przy użyciu trzech najważniejszych zmiennych, dostaliśmy 97,95% dokładności. Usunięcie dowolnej z tych zmiennych znacząco zmniejszało dokładność całego modelu.

Następnie sprawdziliśmy modele drzew decyzyjnych. Przy użyciu trzech najważniejszych zmiennych uzyskaliśmy dokładność 96%, ale nie byliśmy w stanie uzyskać większej dokładności.

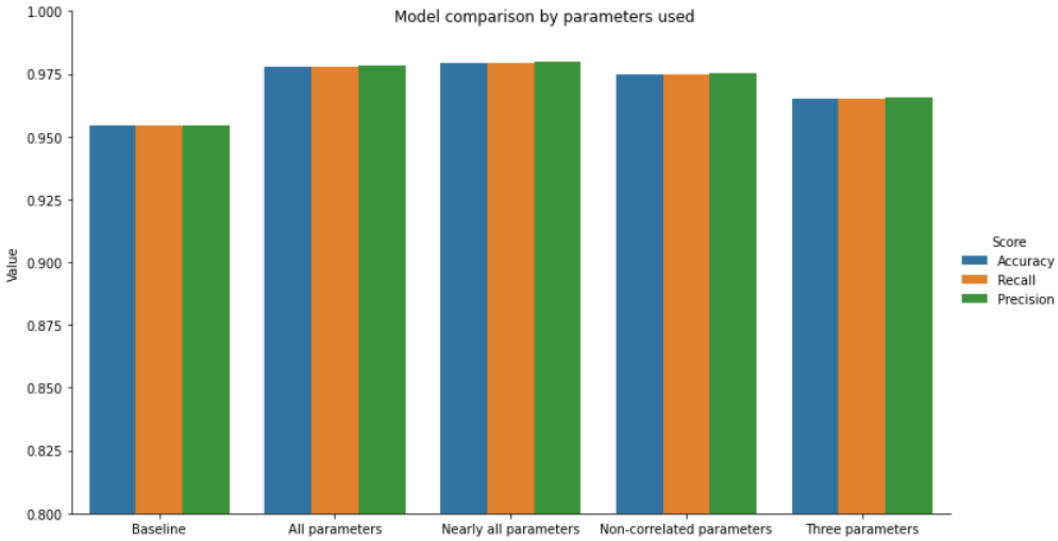
Przy użyciu algorytmu k najbliższych sąsiadów udało nam się uzyskać 97.3% dokładności, również korzystając z trzech najważniejszych zmiennych. Przy zwiększeniu ilości zmiennych, dokładność spadła.

Ostatecznie, sprawdziliśmy model gradient boosted trees, w którym przy użyciu wszystkich zmiennych, udało nam się uzyskać dokładność 98.2%, przy precyzji 99.3% oraz recallu 98.4%. Udało nam się uzyskać znacząco lepszą dokładność niż przy użyciu wszystkich zmiennych. Ze względu na znacząco największą dokładność modelu, postanowiliśmy użyć ten model jako finalny.

Ostateczny model

Po wybraniu modelu, chcieliśmy dokładnie sprawdzić jak zachowuje przy różnych warunkach trenowania.

Na początku sprawdziliśmy, czy może model nie trenuje się lepiej dla mniejszej ilości zmiennych. Wybraliśmy cztery potencjalne zbiory zmiennych - zawierających wszystkie zmienne, nie uwzględniający tych praktycznie równoważnych (korelacja większa niż 0.99), nie uwzględniający tych skorelowanych (korelacja większa niż 0.7) i zawierający tylko meanfun, IQR i Q25. W tym celu, porównaliśmy modele według ich celności, precyzji, oraz recallu.



Rysunek 5: Porównanie modeli przy treningu na róznej liczbie zmiennych

Model wytrenowany na prawie wszystkich zmiennych, z pominięciem zmiennych skorelowanych z innymi w bardzo wysokim stopniu, okazał się być najlepszy, dając minimalną przewagę nad tym zawierającym wszystkie zmienne.

Następnie postanowiliśmy użyć bayesowskiej optymalizacji dla jak najlepszego dobrania poszczególnych hiperparametrów. Niestety, przez sporą losowość boosted trees, użycie tego procesu dało nam wyniki niewiele lepsze niż te otrzymane wcześniej.

Ostatecznie zwiększamy czas treningu na dobranych hiperparametrach, kończąc z modelem o celności 98%, opartym na boosted trees wytrenowanych na zmiennych z wyłączeniem zmiennych kurt, centroid oraz dfrange.