Projekt dot. klasyfikacji - raport końcowy

Paulina Przybyłek
Ada Gąssowska
Paweł Koźmiński

21 kwiecień 2020

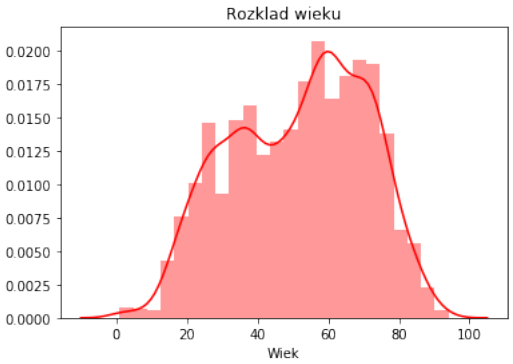
Abstrakt. Projekt dotyczy klasyfikacji na zbiorze o pacjentach i przewidzeniu, czy osoba o podanych cechach ma chorobę tarczycy czy też nie. Skupia się on na trzech obszarach: analizie zbioru i danych o pacjentach, inżynierii cech i przekształcenie zbioru jak najlepiej dla modelowania oraz modelowanie i wybór najlepszego klasyfikatora na podstawie kilku różnych miar. Dobre przewidzenie choroby u pacjenta jest ważne, dlatego należy znaleźć najlepszy klasyfikator. Dodatkowo w projekcie duży nacisk będzie kładziony na strojenie hiperparametrów - wykorzystano randomsearch i gridsearch do znalezienia najlepszych parametrów a jako miarę główną klasyfikatorów wybrano średnią geometryczna, gdyż zbiór jest niezbalansowany (geometric mean została uznana za jedną z najlepszych z dostępnych miar do oceny niezbalansowanych danych w artykule dot. miar dla niezbalansowanych zbiorów: https://wum-2020l.slack.com/files/U010B26JFJ8/F011PJTFB2Q/10.1016_j.eswa.2020.113391.pdf).

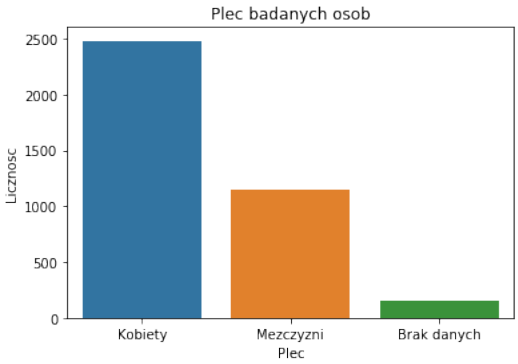
1 Wstęp

Cała analiza w projekcie oparta jest na zbiorze **sick**, który należy do OpenML100 (link: https://www.openml.org/d/38). Zbiór ten zawiera rekordy z badaniami pacjentów zebrane w roku 1987 w Australii a za jego autora uznawany jest Ross Quinlan. Dane są pogrupowane w 30 atrybutów (kolumn), oznaczających informacje o danym pacjencie. Większość atrybutów określana jest poprzez *prawda/fałsz*, a dane o takich wartościach przedstawiają informacje medyczne m.in. o ciąży, możliwej chorobie (np. przeziębienie) podczas badań, podawanego litu czy różnych hormonów i leków, a także występujących wałów tarczycy, niedoczynności przysadki czy guzów. Są także wartości numeryczne (ciągłe) określające wiek oraz zmierzone wartości hormonów. Dodatkowo dana jest płeć pacjenta, szpital w jakim przebywał i określenie czy ma on chorą tarczycę czy nie (zmienna celu wykorzystana przy modelowaniu).
Na podstawie tego zbioru wykonano model do przewidywania czy pacjent ma chorą tarczycę czy też nie na podstawie reszty atrybutów. Oczywiście zbiór został wcześniej do tego specjalnie przygotowany, a modelowanie opierało się na sprawdzeniu kilku możliwości inżynieri cech danych, aby określić najlepszy do tego sposób.

2 EDA - wstępna analiza zbioru

EDA polega głównie na zapoznaniu się z danymi i wyciągnięciu wniosków, które mogą pomóc przy późniejszych etapach. Zbiór zawiera informacje o 3772 pacjentach badanych na obecność choroby tarczycy. Jednak są to dane niezbalansowane - **chorobę tarczycy stwierdzono jedynie u około 6 procent pacjentów**. Dodatkowo spośród badanych to kobiet jest więcej i to ponad dwukrotnie. Wiek pacjentów natomiast zawiera się w przedziale 1-90 kilka lat (oraz u jednego pacjenta 455, co jest błędem i na wykresie jest usunięte), a przewagę mają osoby dobiegające 60-70 lat.





Jak można zauważyć na rozkładzie płci, niektórzy pacjenci nie mają podanych wszystkich informacji, co zdarza się w kartach pacjentów, gdyż nie każdy musi mieć wszystkie atrybuty zmierzone i sprawdzone, jeśli jego stan zdrowia tego nie wymaga. W szczególności to poziom hormonów nie jest badany u wszystkich pacjentów, jednak nie wszyscy mieli podany też wiek (1 osoba) i płeć (150 osób).
Do analizy danych numerycznych zastosowano także sumaryczne podsumowanie wartości - średnia, mediana, min, max i kwantyle. Jednak, aby określić poprawność danych, należy wiedzieć czym są hormony określone jedynie skrótem:

- TSH (tyreotropina), która pobudza tarczycę do produkcji T4 i T3
- T3 (trijodotyronina) - wyróżniamy dwie: całkowitą i wolną. w sick przedstawiona jest prawdopodobnie T3 wolna
- TT4 (tyroksyna), czyli T4 całkowita
- T4U, oznaczające wykorzystanie tyroksyny przez organizm człowieka (eksploatacja, pobór)
- FTI, czyli wolny testosteron, wiarygodny indeks do oceniania stanu tarczycy określony wzorem FTI = Thyroxine (T4)/Thyroid Binding Capacity
- TBG, czyli stężenie globuliny wiążącej tyroksynę

Wiedząc z czym mamy do czynienia, można było sprawdzić, że wszelkie podane dane u pacjentów są możliwe do osiągnięcia oraz korzystając z danych o hormonach, że TSH wpływa na T4 i T3, a FTI zależy od T4, sprawdzić korelacje między atrybutami, które potwierdziły prawdziwość postawionej tezy.

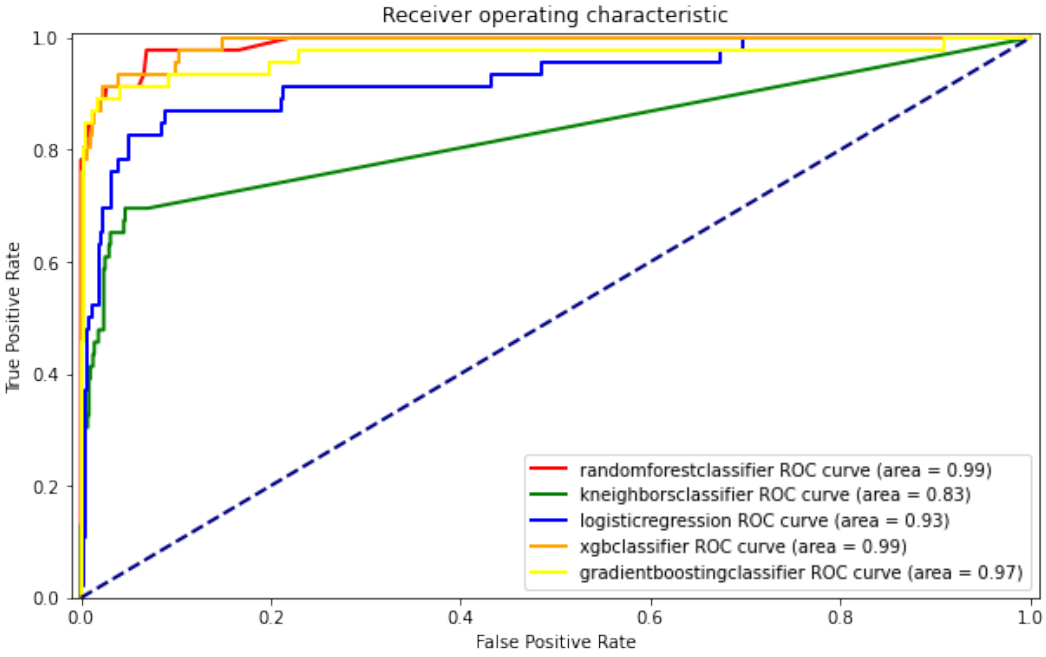
3 Inżyniera cech

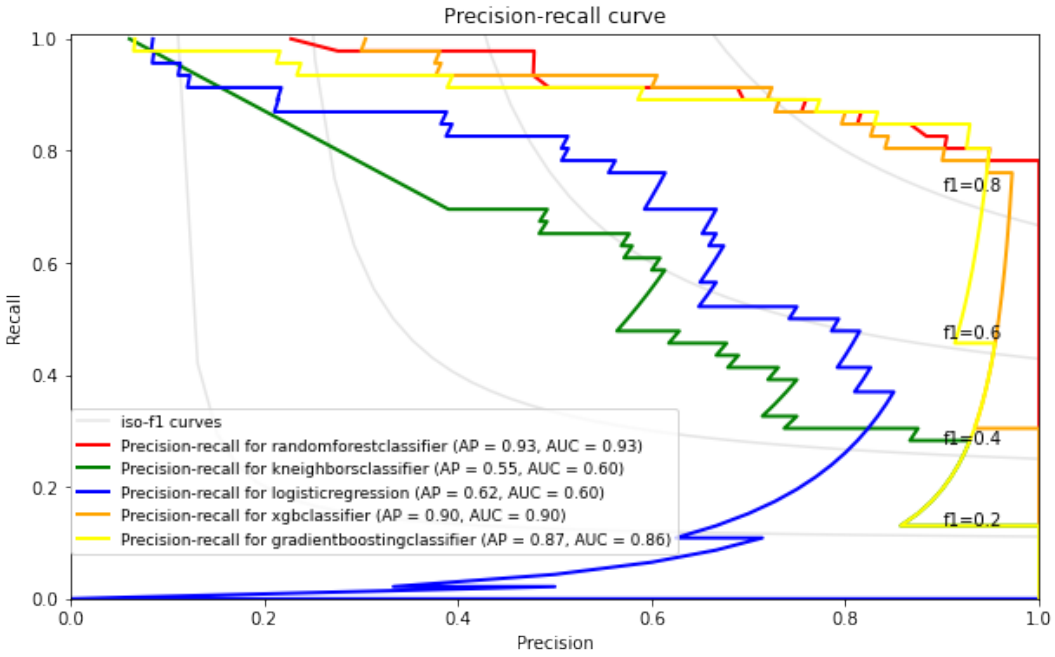
Krótko przedstawiając ten punkt: usunięto kolumnę TBG (brak danych), wartości o typach *prawda/fałsz* zostały zamienione na 0/1, zmienne kategoryczne zakodowano a brakujące wartości poddano imputacji. Zanim zaczęto kodowania i imputację zbiór danych został podzielony na treningowy 80% i testowy 20%. Kodowanie zmiennych kategorycznych dotyczyły: sex (płci) i referral_source (szpitala, z którego uzyskano dane o pacjencie). Pierwszy atrybut miał zastosowaną zamianę na 0,1 i NaN, a także Target Encoding i Oridinal Encoding traktując Missing jako oddzielną kategorię. Przy imputowaniu NaN wykorzystano losowy wybór z dobranym prawdopodobieństwem zgodnym z proporcjami płci w zbiorze. Drugi atrybut otrzymał kodowania: Target Encoding oraz One Hot Encoding. Jeśli chodzi o imputację braków to atrybuty age, TSH, T3, TT4, T4U, FTI zaimputowano średnią, medianą, modą, Knn Imputer'em oraz Iterative Imputer'em. Przed imputacją w kolumnie age na dwa sposoby zmieniono dziwną wartość 455 - zamiana na NaN i na 45. W ten sposób powstało wiele zbiorów, będącymi kombinacjami wszystkich zastosowanych metod, tak aby sprawdzić i wybrać najlepszy możliwy sposób inżynierii cech.

4 Modelowanie

Wstępne modelowanie opierało się na wyborze najlepszego zbioru przy wykorzystaniu modeli XGBoost oraz Random Forest z domyślnymi parametrami. Na podstawie score, geometric mean (w późniejszym modelowaniu stosowana jako główna miara) oraz macierzy konfuzji/pomyłek, wybrano zbiór do dalszego modelowania o cechach: **sex zakodowana na 0,1 z wartościami imputowanymi losowo, refferal_source**

zakodowana target encodingiem, age równy 455 jako NaN, a zmienne numeryczne zaimputowane medianą.
W modelowaniu na wybranym zbiorze zaczęto od sprawdzenia wpływu standaryzacji i normalizacji na dwóch modelach już wykorzystywanych oraz Regresji logistycznej, a po otrzymanych wynikach uznano standaryzację danych za wartą dalszego wykorzystywania.
Następnie stworzono funkcję tuner, która dla pięciu modeli (Random Forest, K Neighbors, Regresji logistycznej, XGBoost oraz GradienBoostingClassifier) sprawdza czy standaryzacja dobrze wpływa na działanie modelu oraz dobiera najlepsze hiperparametry (sprawdzając zarówno grid search jak i random search). Przy danym wywołaniu zebrano następujące wyniki: standaryzacja jest lepsza dla Regresji logistycznej i K Neighbors oraz dla random search Random Forest, w większości przypadków to grid search wybrał parametry, przy których model osiągał lepszy wynik, a najlepszym modelem okazał się XGBoost. Powtórzono ten zabieg jeszcze dwukrotnie: dla danych bez artybutów określających zmierzenie poziomu hormonów oraz przy usunięciu refferal_source i najbardziej skorelowanego atrybutu TT4. Jednak wnioski okazały się podobne, wyniki miary się minimalnie pogorszyły, jedynie dla Regresji Logistycznej w pierwszym przypadku i K Neighbors były minimalnie większe, jednak one i tak miały gorsze modele od reszty.
Wracając do pierwszego wywołania funkcji tuner, poniżej przedstawiono miary graficzne dla uzyskanych najlepiej dostrojonych modeli na zbiorach testowych:





Do wyboru wykorzystano miarę GM, krzywą ROC oraz krzywą Precission-Recall. W graficznych miarach najlepiej prezentuje się Random Forest (kolor czerwony), choć pod względem średniej geometrycznej plasuje się na drugiej pozycji, za Gradient Boosting Classifierem.
Dla trzech najlepszych zadecydowano wykonać feature importance, gdzie w każdym przypadku wybrane zostały te same atrybuty. Porównując feature importance z macierzą korelacji - hormony nie są ukazane jako mające wpływ kolumny, natomiast występują inne dane medyczne np. ciąża, choroba czy przyjmowana tyroksyna, a najbardziej znaczący jest wiek, który w macierzy korelacji nie miał wcale dużej korelacji ze zmienną celu.
Dla tych samych modeli postanowiono kontynuować wybór k najlepszych kolumn i sprawdzić kiedy model osiąga najlepsze wartości. Okazało się, iż to Gradient Boosting przy 7 i 9 kolumnach jest najlepszym modelem i osiągnął najwyższy dotychczasowy wynik - 0.925.
Na zakończenie pracy sprawdzono jeszcze model automatyczny i wybrany klasyfikator miał wynik miary GM bardzo dobry, jednak nieco niższy od Gradient Boostingu dla 7 najlepszych kolumn.

5 Wyniki i konkluzje

Krótko podsumowując, najlepsza inżynieria cech dla zbioru sick to: sex zakodowana na 0,1 z wartościami imputowanymi losowo, refferal_source zakodowana target encodingiem, age równy 455 jako NaN, a zmienne numeryczne zaimputowane medianą. A wedle wykorzystanych miar ogólnie najlepszym klasyfikatorem okazał się Gradient Boosting. Natomiast najsłabszymi modelami okazały się Regresja logistyczna oraz K Neighbors.