WUM - Raport z projektu 1 - Klasyfikacja

Przemysław Chojecki, Michał Wdowski

28 04 2020

1 Opis

W ramach projektu zdecydowaliśmy sie na analize zbioru danych medycznych zawierajacego dane o kobietach ze szpitala z Wenezueli w różym wieku. U niektórych z nich zidentyfikowano raka szyjki macicy. Dane zawieraja informacje z przeprowadzonych ankiet oraz badań medycznych.

Celem projektu było stworzenie modelu przewidujacedo wynik testu na którykolwiek z raków na podstawie takich informacji jak: Wiek, ilość partnerów seksualnych w przeszłości, wiek inicjacji seksualnej, intensywność palenia papierosów, czas korzystania z antykoncepcji hormonalnej oraz wyników przeprowadzonych na pacjentkach badań na choroby AIDS, syfilis(kiła), Hepatitis B, kłykcina i tym podobne.

Projekt skupia sie 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. Jako najważniejsza uznaliśmy miare "Weighted TPR-TNR Measure" opisana w załaczonym artykule (pod nazwa W_R.pdf), gdyż jest ona tam opisana jako najleprza do oceny danych niezbalansowanych.

2 Analiza zbioru

Oryginalny zbiór zawiera wyniki ankiet 858 pacjentek ze szpitala w Caracas w Wenezueli. Informacje dotyczace każdej z kobiet podzieliliśmy na 3 typy:

1. numeryczne, czyli takie, których wartości sa różnymi liczbami dodatnimi. W oryginalnym zbiorze jest takich 12.
2. kategoryczne, których wartościami jest prawda lub fałsz. W oryginalnym zbierze jest takich 20.
3. celu, wynik danego badania na chorobe szyjki macicy. W oryginalnym zbierze jest takich 4.

Do analizy dla każdej z tych grup podchodziliśmy troche inaczej.

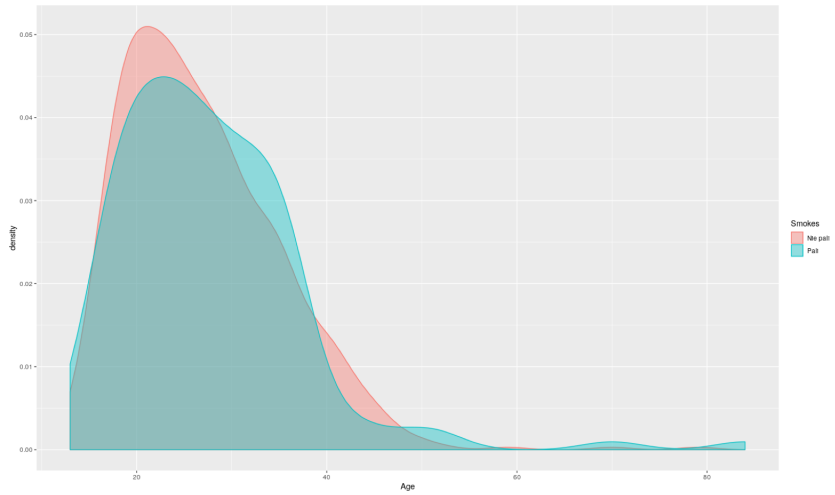

Figure 1: Rozkład wieku w zależności od palenia

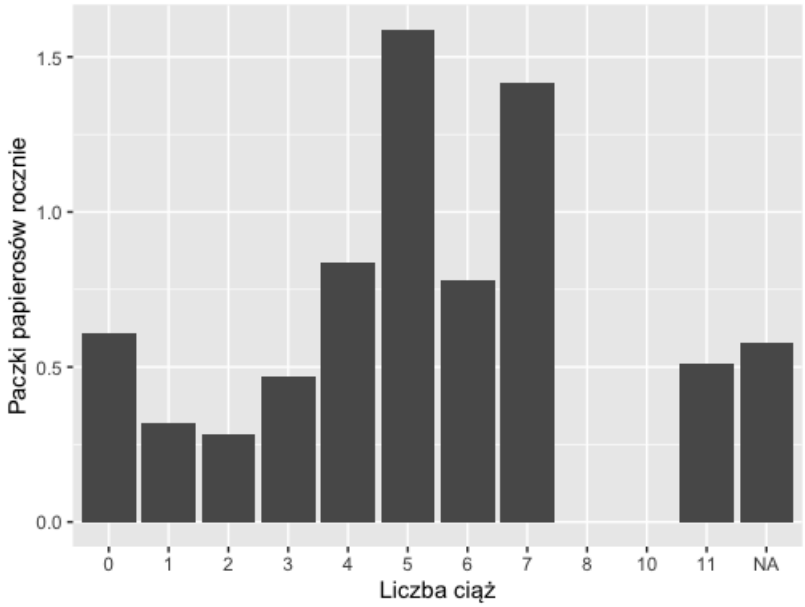

Figure 2: Rozkład palenia w zależności od liczby ciaż

Na pierwszy wykresie widzimy rozkład wieku w zależności od tego, czy dana kobieta pali, czy nie. Spodziewaliśmy sie zobaczyć dużo starsze kobiety wśród palaczek, jednakże tak niema.

Na drugim wykresie widizmy jak dużo średnio pala kobiety w zależności jak

dużo razu były w ciaży. Wyglada na to że ludzie bez dzieci pala troche, jak maja mało dzieci to pala mniej, a potem stresu jest za dużo i pala wiecej. Uwaga - dla osób z ciażami 8 i wiecej sa tylko 1-2 osoby w każdej grupie.

Jesteśmy bardzo zdziwieni niska ilościa palonych przez kobiety paczek papierosów. Wnioskujemy, że być może w Wenezueli standardy palenia sa niższe niż w Polsce i stad ta dysproporcja.

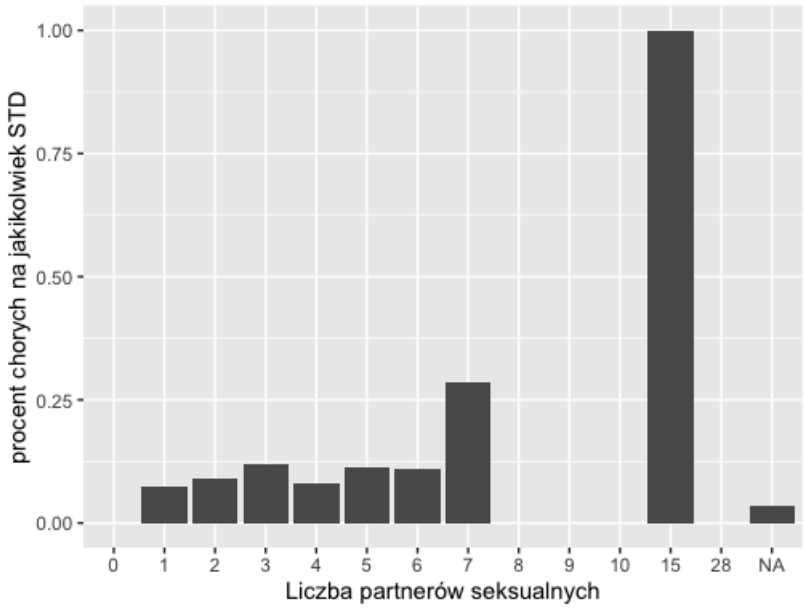

Figure 3: Rozkład wieku w zależności od palenia

Na trzecim wykresie widizmy jaka cześć ludzi ma choroby STD w zależności od liczby partnerów seksulanych. Uwaga - dla osób z liczba partnerów seksualnych 8 i wiecej robi sie po jednej osobie na dana liczbe partnerów

Wiele z danych sa brakujace(kwestja ta bedzie dokładnie omówiona w nastepnym paragrafie). W celach późniejszej analizy i modelowania zamieniliśmy je na wartości za pomoca algorytmu z pakietu "mice".

Na koniec ciekawostka. Jeden z wierszy przedstawia sie nastepujaco: 16-sto latka posiadajaca 28 byłych partnerów seksualnych, z czego pierwszy z kontaktów seksualnych odbyła w wieku 10-ciu lat. Co jeszcze ciekawsze, nigdy nie brała antykoncepcji, ale tylko raz była w ciaży. Od 5-ciu lat pali. Testy nie wykryły choroby na żadna z chorób.

3 Inżynieria cech

Najważniejsza kwestja w tej cześci było poradzenie sobie z danymi brakujacymi. Jak sie okazało, najważniejsze dla modelowania cechy, czyli te zawierajace

wyniki testów, sa wybrakowane. Dokładnie 100 kobiet w posiadanych przez nas danych nie posiada wyników testów na żadne z chorób, co oznacza, że jedyne informacje o tych 100 kobietach jakie posiadamy, to wyniki ich ankiet. Zdecydowaliśmy sie usunać te dane, co było trudna decyzja, gdyż jest to aż 12% posiadanych przez nas danych.

Zauważyliśmy istnienie kilku danych niepoprawnych, pozbyliśmy sie ich. Polegało to na przykład na tym, że pacjetka w ankiecie twierdziła, że miała inicjacje seksualna w wieku wiekszym niż jest jej obecny.

Postanowiliśmy równierz pozbyć sie kolumn, które powielały posiadane już informacje, lub były jej bardzo zbliżone.

Dwie z kolumn posiadały aż 91% danych brakujacych, zastanawialiśmy sie wiec nad jej usunieciem. Jadnakże doszliśmy do wniosku, że maja one jednak sens. Kolumny te nosiły nazwy: "czas od ostatniej diagnozy" oraz "czas od pierwszej diagnozy". Brak dancyn w nich zinterpretowaliśmy jako niedbycie przez pacjetke żadnej diagnozy, wiec zastapiliśmy te braki wartościami "0". Poza tym pozbyliśmy siekolumny "czas od ostatniej diagnozy", gdyż była praktycznie taka zama jak "czas od pierwszej diagnozy". Pozbyliśmy sie równierz kolumn z informacja o chorobie AIDS, gdyż żadna z pacjetek nie były chore, oraz kilku innych niewnoszacych wiele do analizy kolumn.

W tym miejscu zdecydowaliśmy sie wiele ze zmiennych ciagłych(numerycznych) zamienić na zmienne kategoryczne, tworzac tak zwane "kubełki". Decyzje o punktach takzwanych "cieć" podejmowaliśmy później na podstawie wyników dostosowanych modeli.

Dopiero po zredukowaniu wymiaru danych, na sam koniec etapu inżynieria cech, pozostałe brakujace dane wypełniliśmy za pomoca funkcji z pakietu "mice".

4 Modelowanie

Wieksza cześć tego procesu przeprowadzaliśmy w załaczonej, utworzonej przez nas aplikacji shiny. Umożliwiła ona nam łatwe i szybkie analizowanie wyników naszych modeli i dostosowywanie parametrów i kubełków z etapu inżynierji cech.

TODO(MOZE, wkleic zdjecia a appki?) Działanie aplikacji jest nastepujace:

1. Aplikacja wczytuje dane i wykonuje obróbke opisana w poprzednich dwóch paragrafach na podstawie wybranych przez użytkownika parametrów

2. Aplikacja dzieli zbiór danych na treningowy(70%) walidacyjny(15%) i testowy(15%). Wykonuje standaryzacje danych treningowych, po czym za pomoga średnich i wariancji ze zbioru tredningowego, stara sie ustandatyzować dane ze zbiorów treningowego i walidacyjnego.

3. Jeśli użytkownik tego sobie życzy (widnieje to pod nazwa "Create more positive data") aplikacja sztucznie zmultiplikuje i troszke zaburzy nowe dane kobiet ze zdiagnozowanym rakiem. My w naszych modelach decydowaliśmy sie na ta opcje, gdyż oryginalna kolumna celu w danych

jest bardzo niezbalansowana, a to utrudnia modelowmi proces uczenia. Dzieki takiemu małemu oszustwu model miał łatwiej w zrozumieniu, że ważniejszym dla nas było, żeby model wychwytywał kobiey chore, a mniej, żeby udawało mu sie ze zdrowymi. TODO(przetłumaczyć to zdanie na Polski)

4. Użytkownik decyduje sie który z modeli chciałby nauczyć. Do wyboru ma algorytmy:
   (a) kknn
   (b) las losowy
   (c) regresje logistyczna TODO(Sprawdz, czy tak to sie na pewno nazywa)
   (d) drzewo decyzyjne
5. Nastepnie aplikacja wykonuje tyle modeli, o ile użytkownik poprosił w opcji "Find best of" i porównuje je ze soba na podstawie wyników miary Weighted TPR-TNR liczonej na zbiorze walidacyjnym. Każdy z modeli jest liczony w systemie kroswalidacji takiej o jaka prosił użytkownik. TODO(Przetłumacz końcówke na Polski)
6. Na koniec wyświetlane sa użytkownikowi wyniki testów modelu wykonane na zbiorze testowym, czyli takiem, którego model nigdy nie widział i nie dostosowywał sie do niego. Wylistowane sa wyniki miar:
   (a) AUC
   (b) FBeta score
   (c) Weighted TPR-TNR
   (d) AUPRC
   (e) tabela liczby poprawnych/niepoprawnych klasyfikacji

Z ważnych rzeczy: każdy z dostepnych modeli zwracał jako wynik ciag prawdopodobieństw, że dany rekord bedzie chory. Zdecydowaliśmy sie na "uciecie" chorych w punkcie, który jest m-tym od góry prawdopodobieństwem, gdzie $m = n * procentowy\ udzial\ targetu\ w\ zbiorze\ treningowym$, gdzie n to liczba zmiennych w zbiorze testowym, lub walidacyjnym(wartości te sa takie same). TODO(spytać Michała, czy to jest dobrze wytłumaczone) Dzieki temu symulujemy w testowanym zbiorze taka ilość chorych, jaka była w danych uczacych. Taki rodzaj interpretacji sprawdza sie przy sprawozdaniach tego typu, jednakże jest bezużyteczny "na produkcji", gdzie dane napływaja w sposób ciagły, a nie w paczkach, co jest wymagana w naszym sposobie. Aktualny wkład procentowy chorych można podejrzeć z lewej storony aplikacji, a aktualnie wybrane prawdopodobieństwo graniczne w cześci środkowej nad wynikami modelu.

Na koniec ostrzeżenie dla użytkownika. Aplikacja czasem sie buguje, dlatego zamieszczony został łatwodostepny przycisk "RESET", którego naciśniecie zazwyczaj naprawia problem.

Za pomoca aplikacji ustaliliśmy najbardziej optymalny układ "kubełków" i innych parametrów, oraz uznaliśmy, że najlepiej sprawdzajacymi sie modelami sa knn i las losowy.

Ostatecznie dostrajaliśmy parametry modelów za pomoca randomsearch i wyniki sa nastepujace: TODO(Wkleić zdjecia krzywej ROC i cyferki miar)

5 Wnioski

TODO(Wymyślić wnioski, najlepiej nie kompletnie negatywne xd)