WUM - Raport z projektu 1 - Klasyfikacja

Przemysław Chojecki, Michał Wdowski

28.04.2020

1 Wprowadzenie

W ramach projektu zdecydowaliśmy się na analizę zbioru danych medycznych zawierającego dane o pacjentkach z wenenezuelskiego szpitala. U niektórych z nich zidentyfikowano raka szyjki macicy. Dane zawierają informacje z przeprowadzonych ankiet oraz badań medycznych.

Celem projektu było stworzenie modelu przewidującego wynik testu na którykolwiek z testów na nowotwór na podstawie informacji takich jak: wiek, liczba partnerów seksualnych w przeszłości, wiek inicjacji seksualnej, liczba palonych papierosów, czas korzystania z antykoncepcji hormonalnej oraz wyników przeprowadzonych na pacjentkach badań na choroby takie jak: AIDS, syfilis (kiła), Hepatitis B, kłykcina i inne choroby transmitowane drogą płciową.

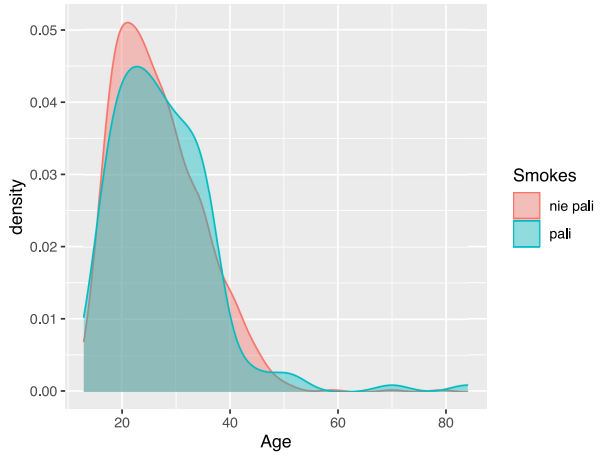
Projekt skupia się na trzech obszarach: analizie zbioru i danych o pacjentkach, inżynierii cech i przekształceniu zbioru jak najlepiej dla modelowania, oraz modelowaniu i wyborze najlepszego klasyfikatora na podstawie kilku różnych miar, co jest o tyle ważne, że zależy nam na prawidłowym przewidzeniu choroby u pacjenta. Jako najważniejszą uznaliśmy miarę "Weighted TPR-TNR Measure" opisaną w załączonym artykule (pod nazwą W_R.pdf), gdyż jest ona tam opisana jako najlepsza do oceny danych niezbalansowanych.

2 Analiza zbioru

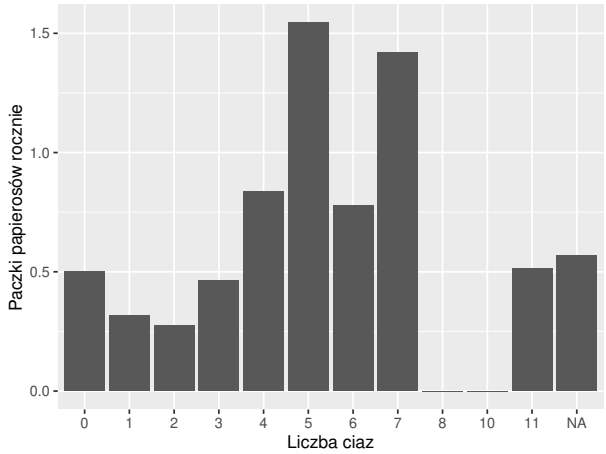
Oryginalny zbiór zawiera wyniki ankiet 858 pacjentek ze szpitala w Caracas w Wenezueli. Informacje dotyczące każdej z pacjentek podzieliliśmy na 3 typy:

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

Do analizy dla każdej z tych grup podchodziliśmy trochę inaczej.



Rysunek 1: Frakcja palących w zależności od wieku



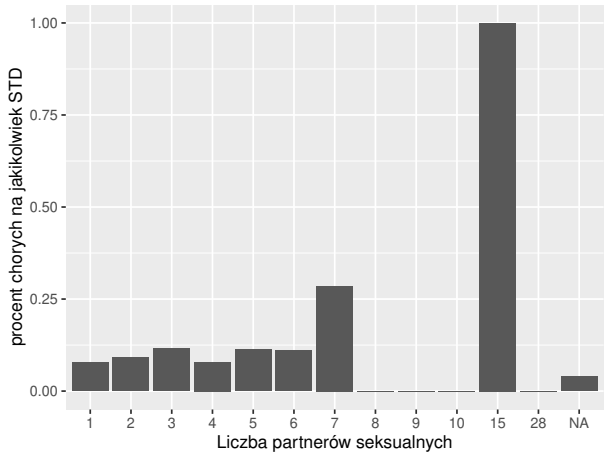
Rysunek 2: Średnia liczba wypalanych paczek papierosów w zależności od liczby ciąż

Na pierwszym wykresie widzimy rozkład palących kobiet w zależności od jej wieku. Spodziewaliśmy się zobaczyć dużo starsze kobiety wśród palaczek, jednakże okazało się, że tak nie jest.

Na drugim wykresie zostało przedstawione, jak dużo średnio palą kobiety w zależności od liczby przebytych ciąż. Wygląda na to, że kobiety bez dzieci palą trochę, gdy mają pierwsze dzieci to zaczynają ograniczac palenie, a potem stresu jest za dużo i palą więcej. Uwaga - dla kobiet z ciążami 8 i więcej są tylko 1-2 osoby w każdej grupie, więc wyniki moga nie być miarodajne.

Jesteśmy bardzo zdziwieni niską liczbą palonych przez kobiety paczek pa-

pierosów. Wnioskujemy, że być może w Wenezueli standardy palenia są niższe niż w Polsce i stąd ta dysproporcja. Możliwe jest też, że paczki papierosów w Wenezueli są większe niż w Polsce.



Rysunek 3: Frakcja chorych na choroby transmitowane drogą płciową w zależności od liczby partnerów seksualnych

Na trzecim wykresie widizmy jaka część ludzi ma choroby STD w zależności od liczby partnerów seksulanych. Uwaga - dla osób z liczbą partnerów seksualnych 8 i więcej robi się po jednej osobie na daną liczbę partnerów, więc wyniki znów mogą nie byc miarodajne.

Duża część danych jest wybrakowana (kwestia ta będzie dokładnie omówiona w następnym paragrafie). W celach późniejszej analizy i modelowania zamieniliśmy je na wartości za pomocą algorytmu z pakietu `mice`.

Na koniec pokażemy przykład ciekawgo rekordu. Jeden z wierszy wygląda następująco: szesnastolatka posiada 28 byłych partnerów seksualnych, z czego pierwszy kontakt seksualny odbyła w wieku dziesięciu lat. Co jeszcze ciekawsze, nigdy nie stosowała antykoncepcji, ale tylko raz była w ciąży. Od pięciu lat pali. Testy nie wykryły choroby na żadną z chorób. Zwróciliśmy uwagę na ten jeden wiersz, ponieważ w tych danych ta osoba, pomimo wieku, jest rekordzistką pod względem liczby partnerów seksualnych.

3 Inżynieria cech

Najważniejszą kwestią w tej części było poradzenie sobie z brakami w danych. Jak się okazało, duża część cech, dokładniej mówiąc - kolumny zawierające występowanie chorób układu rozrodczego, są wybrakowane. Dokładnie 100 kobiet w posiadanych przez nas danych nie podało wyników testów na żadną z tych chorób, co oznacza, że jedyne informacje o tych 100 kobietach jakie posiadamy,

to te nie dotyczące historii medycznej. Zdecydowaliśmy się usunąć te dane, co było trudną decyzją, gdyż jest to aż 12% rekordów.

Zauważyliśmy istnienie kilku danych niepoprawnych, pozbyliśmy się ich. Polegało to na przykład na tym, że pacjentka w ankiecie twierdziła, że miała inicjację seksualną w wieku większym, niż jej obecny wiek.

Postanowiliśmy również pozbyć się kolumn, które powielały posiadane już informacje, lub były bardzo zbliżone do siebie.

Dwie z kolumn posiadały aż 91% danych brakujących, zastanawialiśmy się więc nad jej usunięciem. Jednakże doszliśmy do wniosku, że akurat w tych brakach danych kryje się zawarta informacja. Kolumny te nosiły nazwy: "czas od ostatniej diagnozy" oraz "czas od pierwszej diagnozy". Te braki danych zinterpretowaliśmy jako nieprzebycie przez pacjentkę żadnej choroby, więc zastąpiliśmy te braki wartościami "0". Poza tym pozbyliśmy się kolumny "czas od ostatniej diagnozy", gdyż niosła praktycznie taką samą informację jak "czas od pierwszej diagnozy". Pozbyliśmy się również kolumn z informacją o chorobie AIDS, gdyż żadna z pacjentek nie była na nią chora, oraz kilku innych niewnoszących wiele do analizy kolumn.

W tym miejscu zdecydowaliśmy się wiele ze zmiennych ciągłych (numerycznych) zamienić na zmienne kategoryczne, tworząc tak zwane "kubełki". Decyzje o punktach tak zwanych "cięć" podejmowaliśmy później na podstawie wyników dostosowanych modeli.

Dopiero po zredukowaniu wymiaru danych, na sam koniec etapu inżynierii cech, pozostałe brakujące dane wypełniliśmy za pomocą funkcji z pakietu `mice`.

Próbowaliśmy wykorzystać metodę PCA z nadzieją, że poleprzy to modelowanie. Jak się jednak okazało, modele po zastosowaniu tej metody wcale nie były leprze niż bez niej, dlatego zdecydowaliśmy się z niej zrezygnować.

4 Modelowanie

Większą część tego procesu przeprowadzaliśmy w załączonej stworzonej przez nas aplikacji w `Shiny`. Umożliwiła nam ona łatwe i szybkie analizowanie wyników naszych modeli i dostosowywanie parametrów i kubełków z etapu inżynierii cech.

Działanie aplikacji jest następujące:

1. aplikacja wczytuje dane i wykonuje obróbkę opisaną w poprzednich dwóch paragrafach, na podstawie wybranych przez użytkownika parametrów;

2. aplikacja dzieli zbiór danych na treningowy (70% z całego zbioru), walidacyjny (15%) i testowy (15%). Następnie wykonuje standaryzację danych treningowych, po czym, za pomocą średnich i wariancji ze zbioru treningowego, stara się ustandaryzować dane ze zbiorów testowego i walidacyjnego;

3. jeśli użytkownik sobie tego życzy, aplikacja sztucznie zmultiplikuje i w lekkim stopniu zaburzy nowe dane tych kobiet, które uzyskały pozytywny

Rysunek 4: Przykładowy widok aplikacji

wynik testu (widnieje to pod nazwą "Create more positive data"). W naszych modelach zdecydowaliśmy się na tę opcję, gdyż oryginalne kolumny celu w danych są bardzo niezbalansowane, a to utrudnia modelowi proces uczenia. Dzięki takiemu "oszustwu" model mógł łatwiej zrozumieć, że ważniejszym dla nas było, żeby wychwytywał kobiey chore, a mniej, żeby udawało mu się ze zdrowymi.

4. użytkownik decyduje się który z modeli chciałby nauczyć. Do wyboru ma algorytmy:
   (a) kknn;
   (b) las losowy;
   (c) regresję logistyczną;
   (d) drzewo decyzyjne;
5. następnie aplikacja tworzy tyle modeli, o ile użytkownik poprosił w opcji "Find best of" i porównuje je ze sobą na podstawie wyników miary Weighted TPR-TNR liczonej na zbiorze walidacyjnym. Każdy z modeli jest liczony w systemie kroswalidacji o takiej liczbie zbiorów, o jaką prosił użytkownik;
6. na koniec wyświetlone zostają wyniki modelu wykonane na zbiorze testowym, czyli takie, z którym model nigdy nie miał styczności, i nie dostosowywał się do niego. Wylistowane są wyniki miar:
   (a) AUC;
   (b) FBeta score;

(c) Weighted TPR-TNR;

(d) AUPRC;

(e) tabela liczby poprawnych/niepoprawnych klasyfikacji - tzw. macierz konfuzji;

Wynikiem każdego z dostępnych modeli był ciąg prawdopodobieństw zdarzenia, że odpowiednia pacjentka jest chora. Zdecydowaliśmy się na "ucięcie" chorych w punkcie, który jest m-tym od góry prawdopodobieństwem, gdzie $m = n * (\text{frakcja chorych w zbiorze treningowym})$, gdzie n to liczba zmiennych w zbiorze testowym lub walidacyjnym (wartości te są takie same). Dzięki temu symulujemy w testowanym zbiorze taką część chorych, jaka była w danych uczących. Ten rodzaj interpretacji sprawdza się przy sprawozdaniach takiego rodzaju, gdy dane są podawane w zbiorach, a nie w sposób ciągły. Oznacza to, że nasz model nie powinien być używany do sprawdzania pojedyńczych rekordów - zamiast tego należy mu podawać zestaw danych. Aktualny wkład procentowy chorych można podejrzeć z lewej strony aplikacji, a aktualnie wybrane prawdopodobieństwo graniczne w części środkowej, nad wynikami modelu.

Na koniec ostrzeżenie dla użytkownika. W aplikacji mogą pojawiać się błędy, dlatego zamieszczony został przycisk "RESET", którego naciśnięcie przywraca pierwotny stan aplikacji, co eliminuje błędy.

Za pomocą aplikacji ustaliliśmy najbardziej optymalny układ "kubełków" i innych parametrów.

Uzyskane "kubełki" wyglądają następująco:

1. Number_of_sexual_partners_bins: $1, 2, 3, 4$;
2. Num_of_pregnancies_bins: $1, 2, 3, 4, 5$;
3. Smokes_packs_year_bins: $0.0001, 1.5$;
4. Hormonal_Contraceptives_years_bins: $0.0001, 2$;
5. IUD_years_bins: $1, 2, 3, 4, 5, 6, 7, 8$;
6. STDs_Time_since_first_diagnosis_bins: $1, 2, 3, 4, 5, 6, 7, 8$;

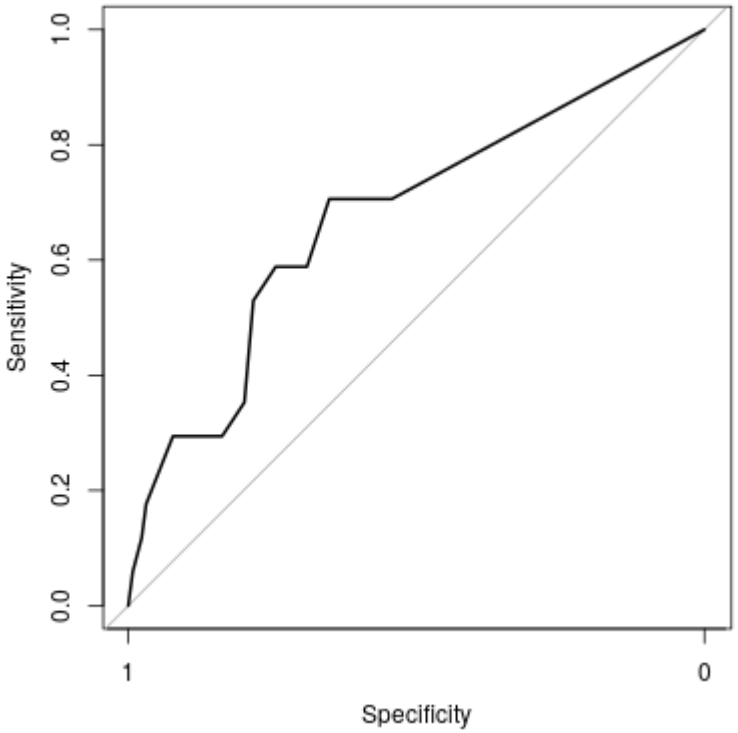
Podane liczby oznaczają punkty podziału zbioru. Przykładowo, w Number_of_sexual_partners_bins liczby $1, 2, 3$ i 4 oznaczają, że utworzone zostaną kategorie dla liczb:

- od 0 włącznie, mniejsze niż 1;
- od 1 włącznie, mniejsze niż 2;
- od 2 włącznie, mniejsze niż 3;
- od 3 włącznie, mniejsze niż 4;
- 4 i większe.

Ostateczne parametry modeli knn i lasu losowego zostały dostrojone za pomocą metody `randomsearch`. Wyniki dla poszczególnych modeli są następujące:

4.1 Model knn

Ustalone parametry to $k = 4$ i $distance = 2$.



Rysunek 5: Krzywa ROC dla modelu knn

Predicted/True	0	1
0	101	8
1	28	9

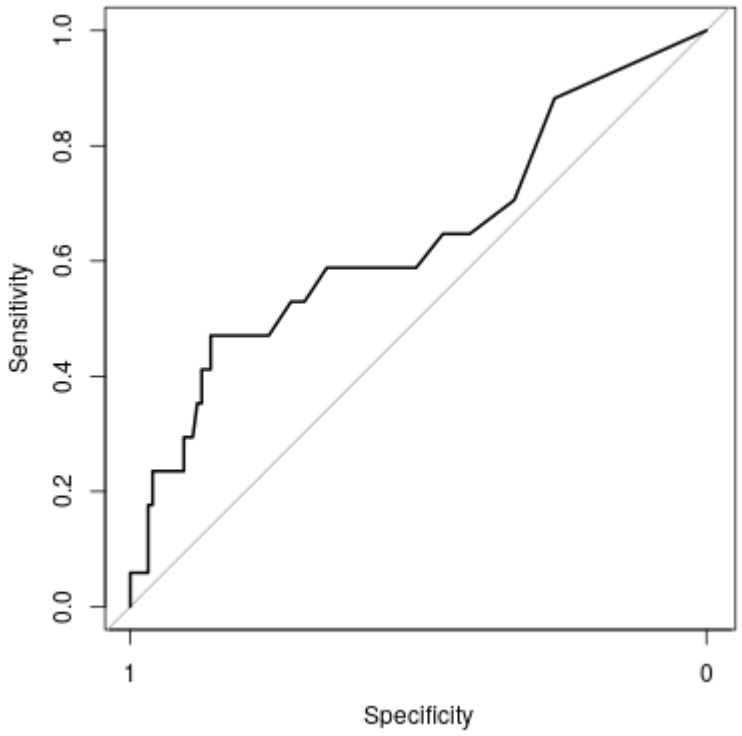
Tabela 1: Macierz konfuzji dla knn

Miara	Wynik
AUC	0.676
F-Beta	0.783
Weighted TPR-TNR	0.559

Tabela 2: Wyniki miar dla knn

4.2 Model lasu losowego

Ustalone parametry to $num.trees = 517$, $mtry = 10$ i $num.random.splits = 10$.



Rysunek 6: Krzywa ROC dla modelu lasu losowego

Predicted/True	0	1
0	101	9
1	28	8

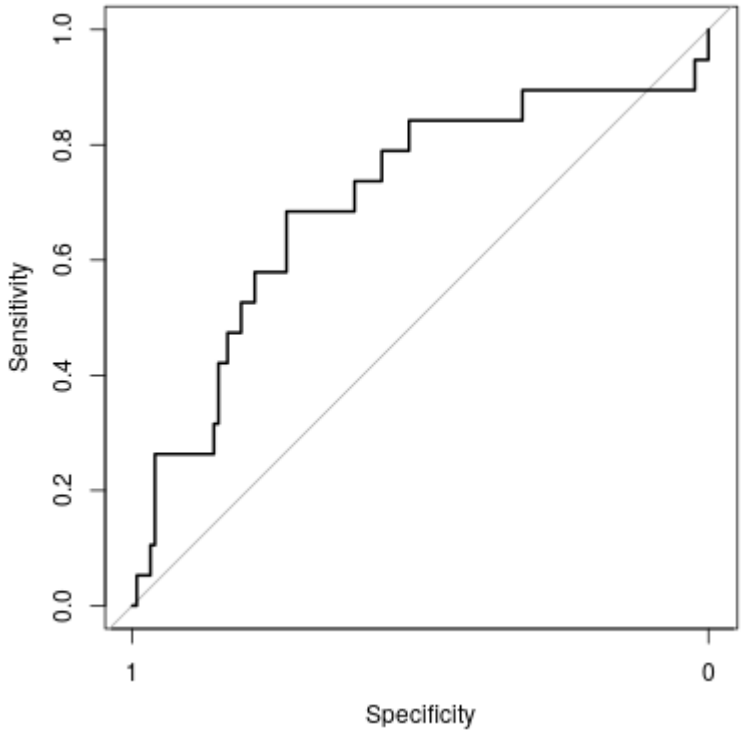
Tabela 3: Macierz konfuzji dla lasu losowego

Miara	Wynik
AUC	0.639
F-Beta	0.783
Weighted TPR-TNR	0.507

Tabela 4: Wyniki miar dla lasu losowego

4.3 Model regresji logistycznej

Model w używanym przez nas pakiecie `mlr` nie pozwala na strojenie parametrów.



Rysunek 7: Krzywa ROC dla modelu regresji logistycznej

Predicted/True	0	1
0	101	9
1	26	10

Tabela 5: Macierz konfuzji dla regresji logistycznej

Miara	Wynik
AUC	0.702
F-Beta	0.795
Weighted TPR-TNR	0.561

Tabela 6: Wyniki miar dla regresji logistycznej

5 Podsumowanie

Obserwując wyniki można zauważyć, że w tym wypadku model regresji logistycznej otrzymał najlepsze wyniki. Ponadto, w tym modelu oraz w knn, udało się osiągnąć większą liczbę TP niż FN w macierzy konfuzji, co można uznać za sukces. Należy wspomnieć, że przedstawione przez nas modele są tymi, które osiągnęły najlepszy wynik pośród 100 iteracji.

Nietypowym jest, że model lasu losowego poradził sobie niewiele lepiej niż model losowy. Wiemy z praktyki, że modele lasu losowego powinien w miarę dobrze wykrywać podstawowe, jak i bardziej złożone zależności. Jeśli tak nie jest, to najprawdopodobiej takich zależności nie ma, a to z koleji mogłoby sugerować słabą jakość zbioru badanych danych.