Raport

Wstęp do uczenia maszynowego - Projekt 1

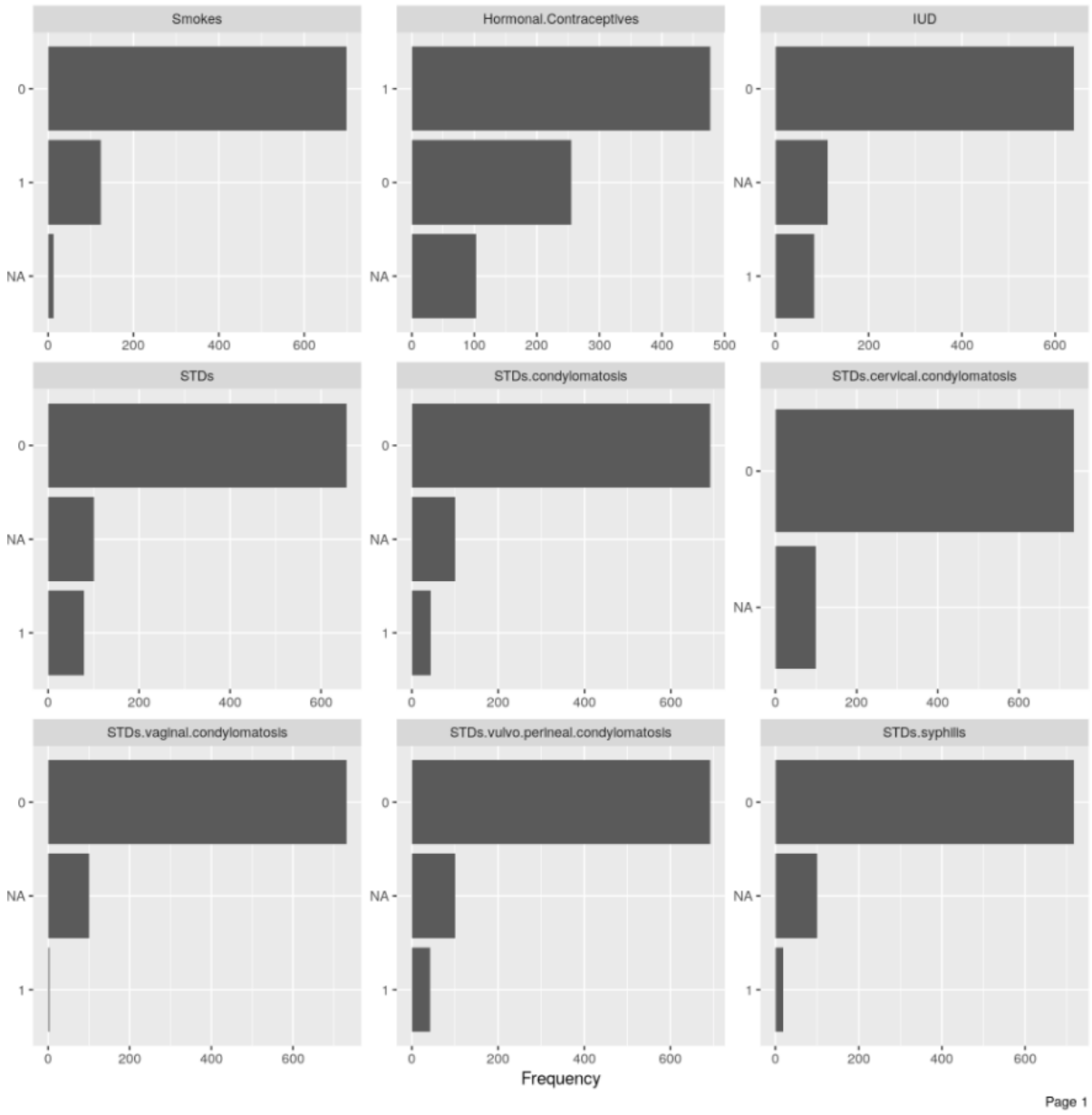
Ngoc Anh Nguyen, Piotr Piątyszek

Wstęp biznesowy

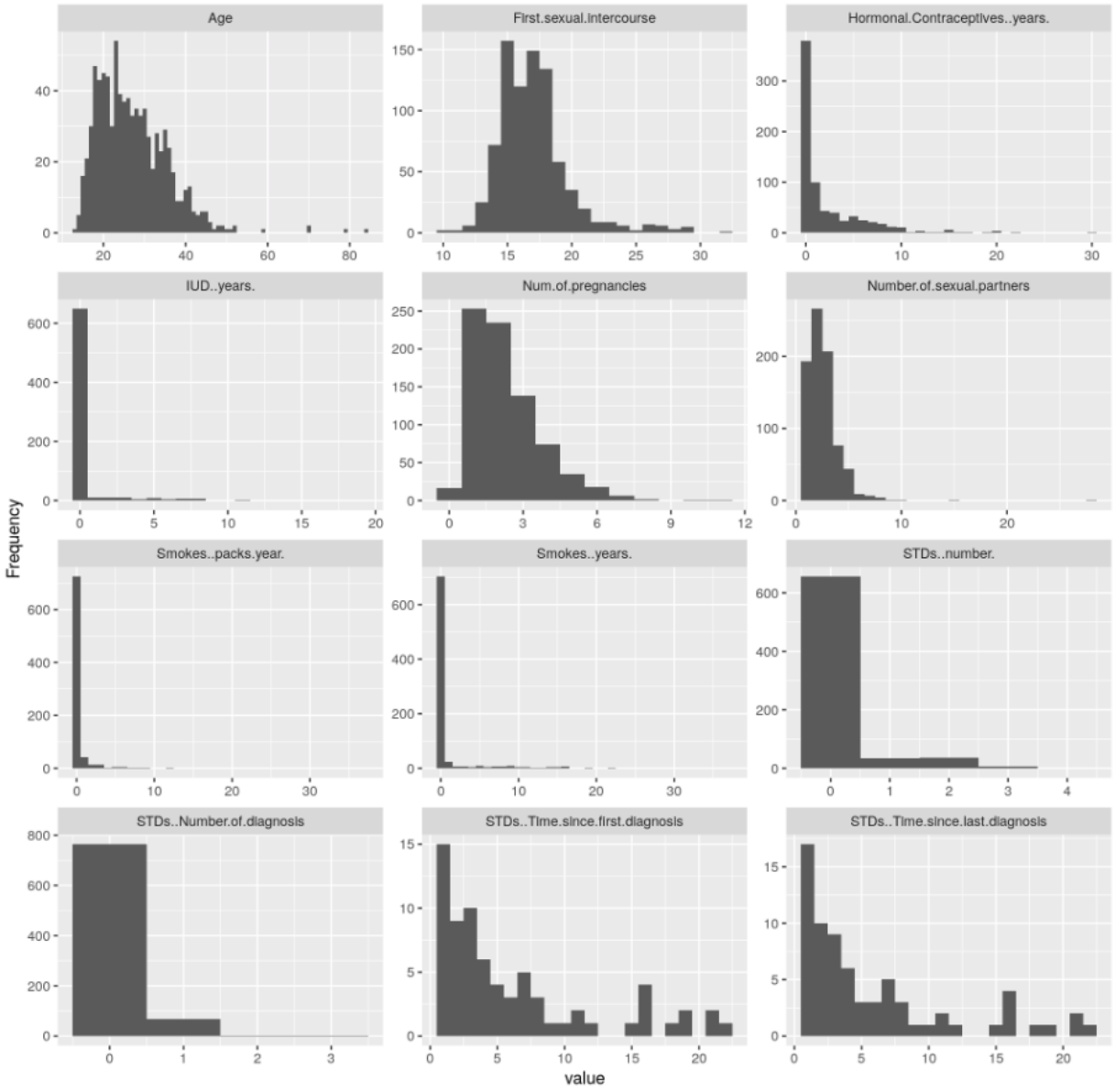
Istnieje wiele poważnych chorób, które wydają się być wyrokiem dla osoby zdiagnozowanej. W znakomitej większości przypadków zachorowań tak naprawdę czynnikiem decyzyjnym jest etap, na którym schorzenie zostało wykryte i zdiagnozowane. Z tego powodu powstało mnóstwo kampanii społecznych promujących profilaktyczne badania przeprowadzane wśród osób niemających objawów, ponieważ wczesne wykrycie oraz leczenie zapobiega poważnym następstwom zawczasu. Nasz model mógłby pomóc szczególnie w takich właśnie sytuacjach, dokładniej mówiąc: w badaniach przesiewowych. Jako, że praktycznie nigdy nie diagnozuje on osób zdrowych jako chore nie powodowałby marnowania czasu, surowców, pracy lekarzy, a także stresu pacjentów. Natomiast jeśli chodzi o chorych – wykrywa 10 % przypadków chorych. Mimo, że nie wykryłby oczywiście wszystkich osób potrzebujących leczenia to jego przydatność jest niezaprzeczalna. Diagnoza postawiona przez nasz model wysoce prawdopodobnie byłaby prawdziwa – rzadko zdarza się tzw. 'fałszywie pozytywny'.

Eksploracja danych

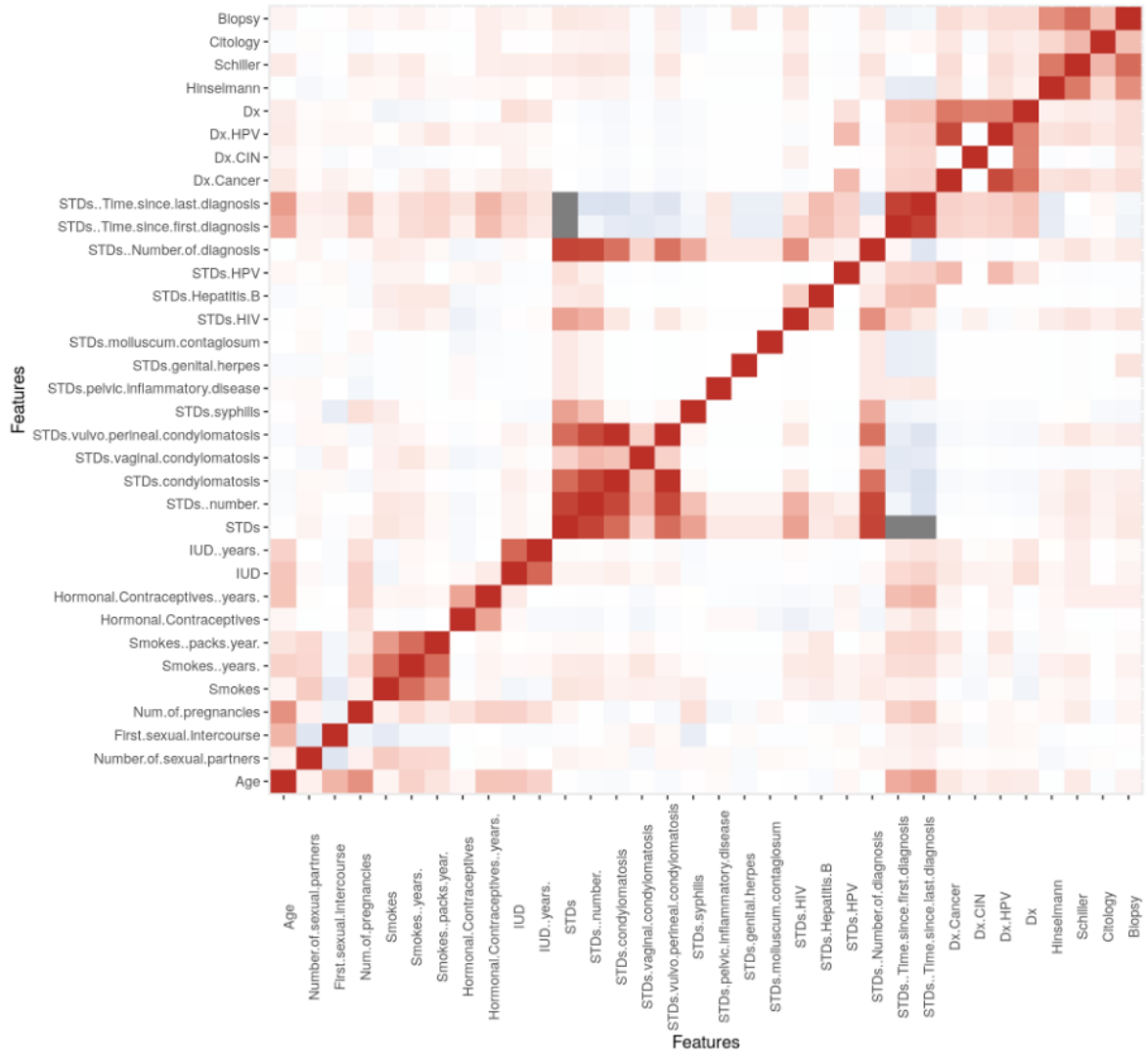
Pierwszym wnioskiem z eksploracji danych jest bardzo dużo zmiennych logicznych ze sporymi brakami danych. Poniżej znajduje się rozkład około połowy, pozostałe wyglądają podobnie. Większośc to pytania o konkretne choroby weneryczne.



Poza tym znajdują się zmienne numeryczne

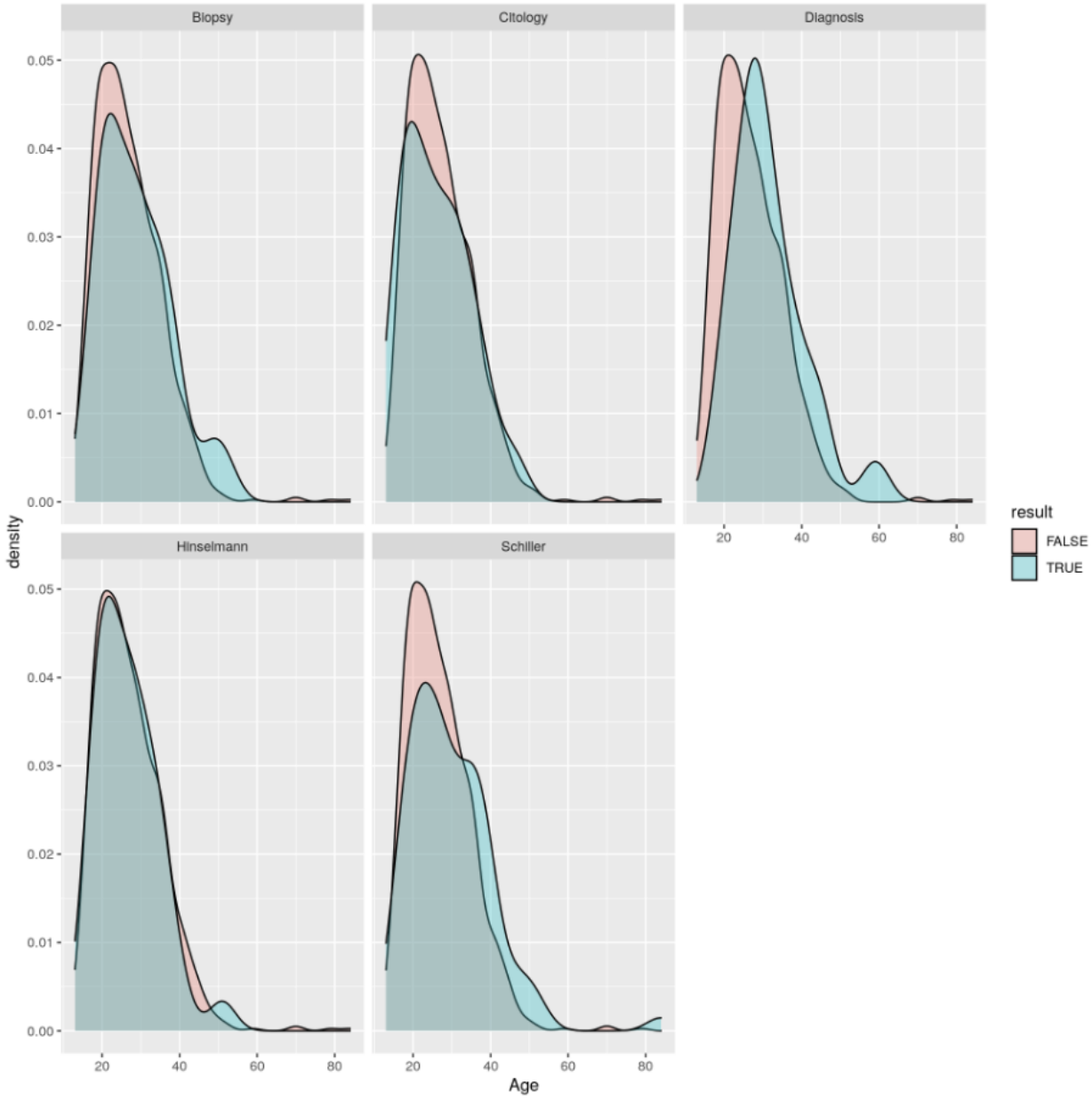


Zbadaliśmy korelację zmiennych i niestety nie znaleźliśmy nieoczywistych zależności. Jedynie między zmiennymi które zależą od siebie w oczywisty sposób, np. zmienne celu czy choroby weneryczne.



Zmienne celu

W zbiorze znajdują się 4 zmienne celu odpowiadające różnym metodą diagnostyki oraz jedna metoda `Dx` co oznacza Diagnoza. Porównaliśmy rozkład tych zmiennych względem wieku. Dość ciekawym wnioskiem jest, że lekarz podejmując diagnozę znacznie bardziej sugeruje się wiekiem, niż wynika to z testu.



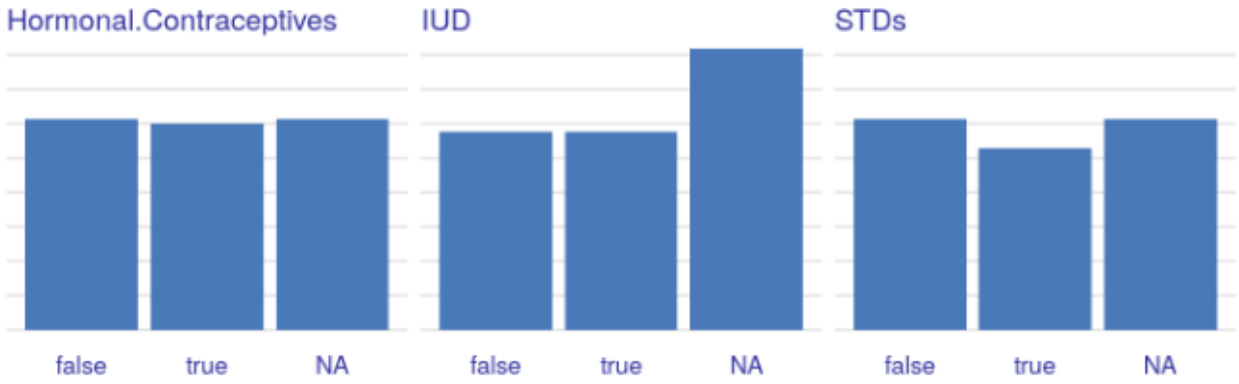
Feature Engineering

Wybór zmiennych

Ze zbioru usunęliśmy wszystkie kolumny celu inne niż nasza - wynik biopsji. Poza tym usunęliśmy kolumny zaczynające się od 'Dx' - diagnoza, ponieważ nie byliśmy pewni ich znaczenia i chcieliśmy uniknąc target leakage. Po drugim kamieniu milowym na podstawie wykresów PD określiliśmy też, które zmienne mają nieistotny wpływ i też je usunęliśmy dla uproszczenia modelu.

Braki danych

W zbiorze występuje bardzo dużo zmiennych logicznych z licznymi brakami danych. Pierwotnie próbowaliśmy te braki imputować, jednak zauważyliśmy, że nie odpowiedzenie na pytanie o choroby weneryczne czy antykoncepcję może być istotną informacją samą w sobie, na przykład o niskiej świadomości seksualnej pacjetntki, co może mieć wpływ na korzystanie z profilaktyki. Dlatego postanowiliśmy te zmienne kodować jako 3 poziomową zmienną kategoryczną. Profile partial dependencje jednego z wytrenowanych modeli pokazuję, że to słuszna droga.



Zmienne numeryczne imputowaliśmy przy pomocy pakietu `mice`.

Niezbalansowanie klas

Zmienna celu - wynik biopsji jest bardzo niezbalansowana, dlatego w procesie uczenia wypada zastosować jedną z technik balansowani. Down-sampling odrzuciliśmy z uwagi na bardzo niewielką liczbę danych, metody ROSE nie udało się użyć z uwagi na problemy techniczne. Przetestowaliśmy up-sampling i metodę SMOTE. Przy trenowaniu finalnego modelu używaliśmy już tylko up-samplingu, ponieważ w drugim kamieniu milowym dawał lepsze wyniki.

Model performance

Wytrenowaliśmy 3 modele ze strojeniem parametrów. Poniżej znajdują się metryki dla danych z cross validacji oraz dla danych testowych dla każdego rodzaju modelu w najlepszej wersji względem AUC.

Boosted GLM

	Accuracy	Kappa	ROC	Sens	Spec	Balanced Accuracy
Train	0.6268673	0.0301969	0.5619212	0.4219048	0.6410318	0.5314683
Test	0.8078078	0.0239077	0.5268620	0.1904762	0.8493590	0.5199176

Random Forest

	Accuracy	Kappa	ROC	Sens	Spec	Balanced Accuracy
Train	0.8214674	0.0048647	0.5299212	0.1352381	0.8696362	0.5024372
Test	0.8678679	0.1171367	0.5937882	0.2380952	0.9102564	0.5741758

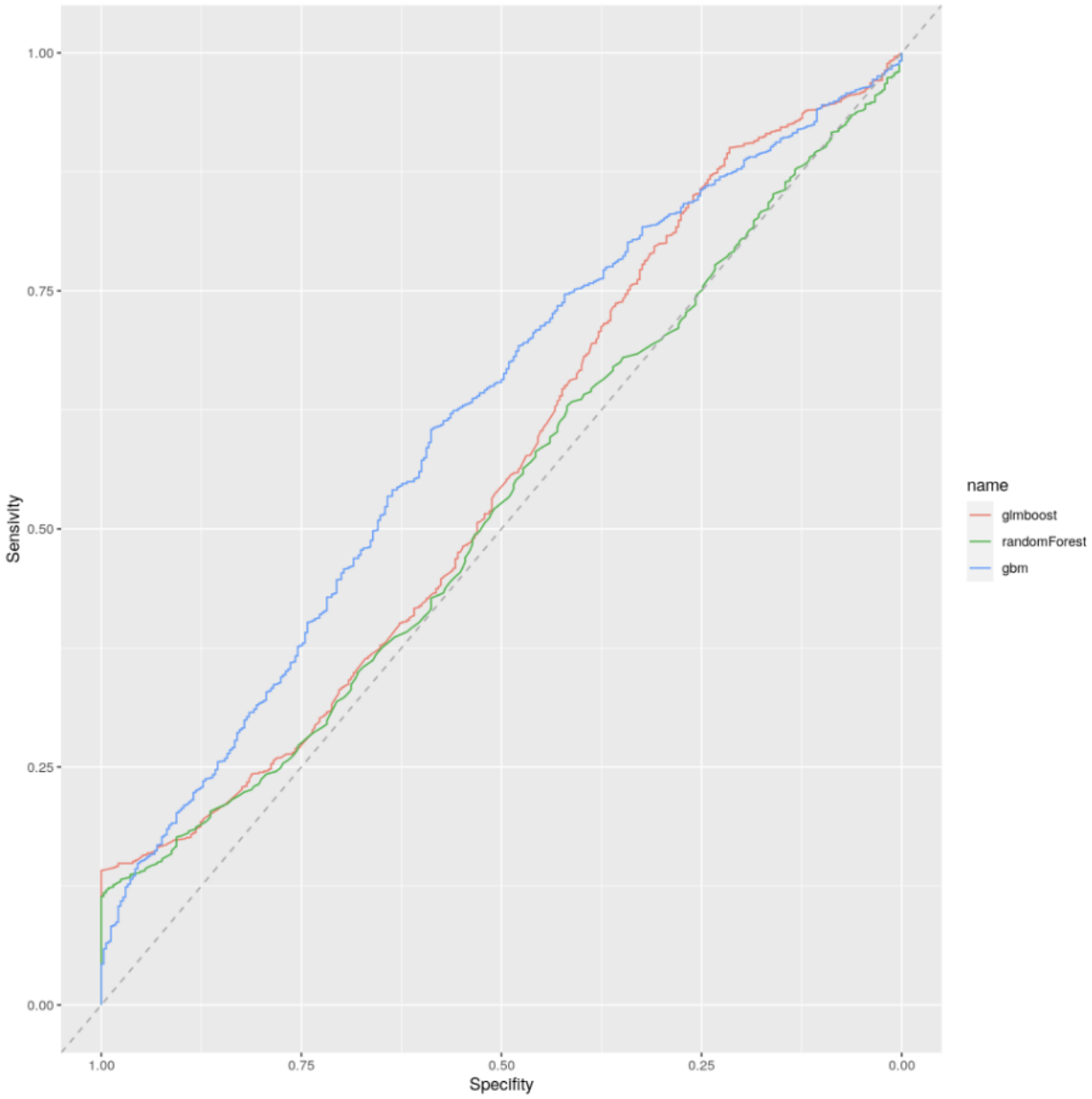
GBM

	Accuracy	Kappa	ROC	Sens	Spec	Balanced Accuracy
Train	0.7643586	0.0672419	0.6174235	0.3433333	0.7940105	0.5686719
Test	0.7597598	0.0938776	0.6088217	0.4285714	0.7820513	0.6053114

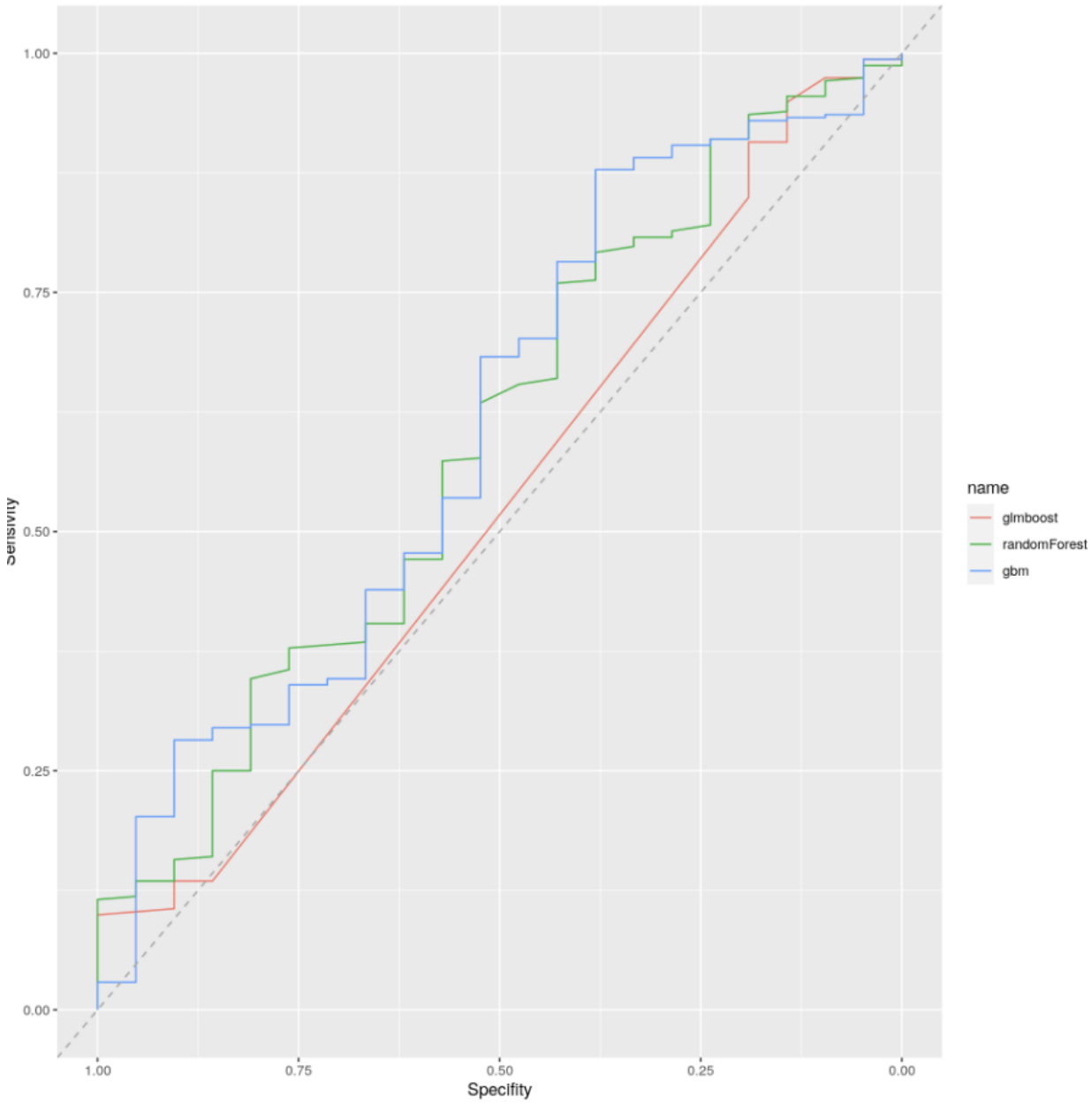
Jak można zauważyć, największą wartość AUC osiąga GBM, jednak to z uwagi na następny wykres zdecydowaliśmy się wybrać **Random Forest**. Zarówno Boosted GLM jak i Random Forest mają ogromną zaletę nad GBM, ponieważ dla bardzo wysokiego cutoff nie popełniamy praktycznie false positve'ów i nadal wykrywamy około 10% chorych. Taki model można wykorzystać do badań przesiewowych. Ponadto GLM ma przewagę na tym, że ma bardziej prostą strukturę i łatwiej go zrozumieć. Jednak Po przeanalizowaniu variable importance można zauważyć, że model GLM wykrył korelację z konkretną chorobą weneryczną. Jej posiadacze prawdopodobnie wiedzą, że są zagrożeni. Random forest za to opiera swój wynik o zmienne które można zadać w ankiecie i są mniej oczywiste.

Krzywe ROC

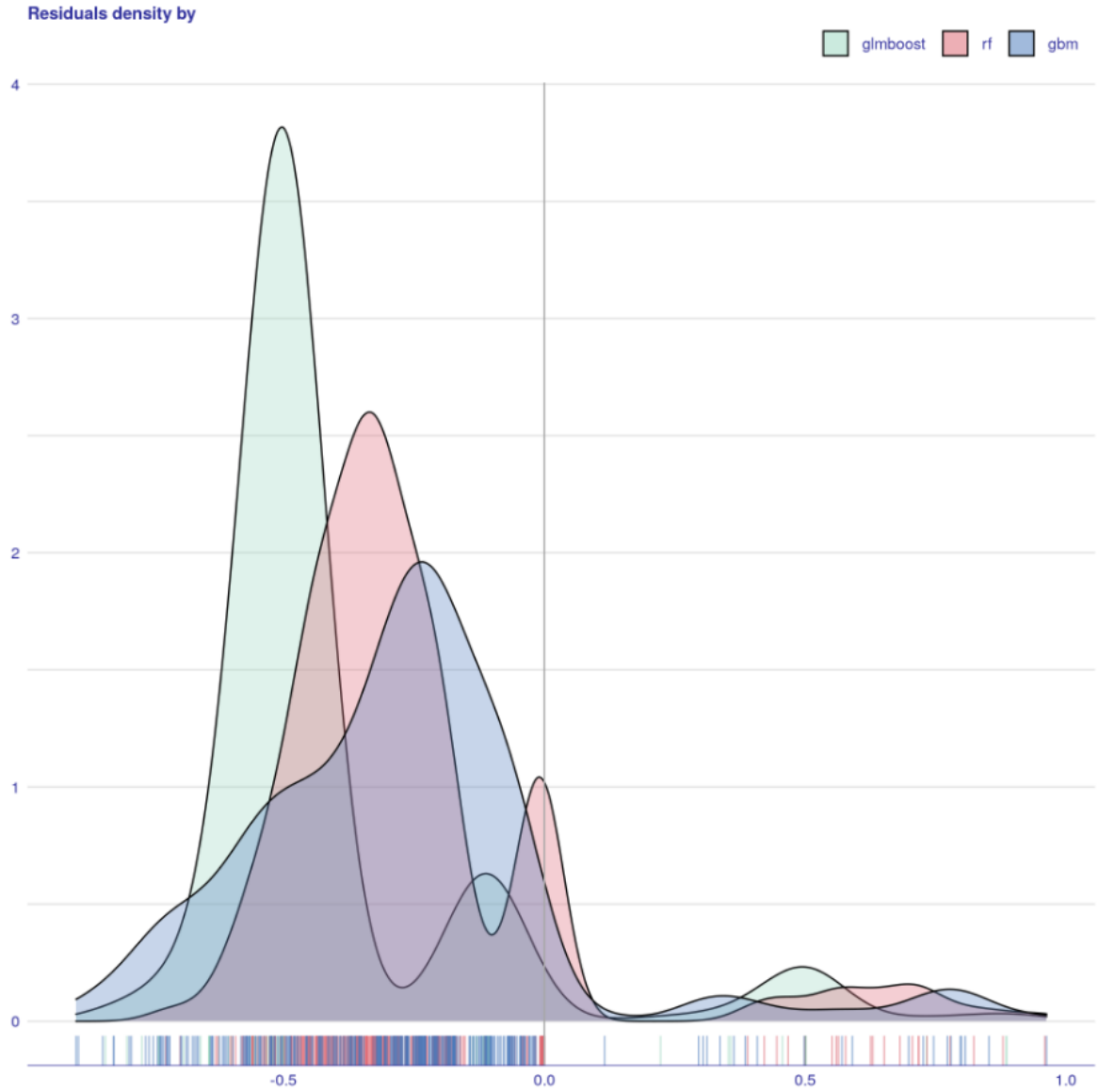
Dla danych z CV



Dla danych testowych


1.00
0.75
0.50
0.25
0.00
Sensivity
1.00
0.75
0.50
0.25
0.00
Specifity
name
glmboost
randomForest
gbm

Model residuals

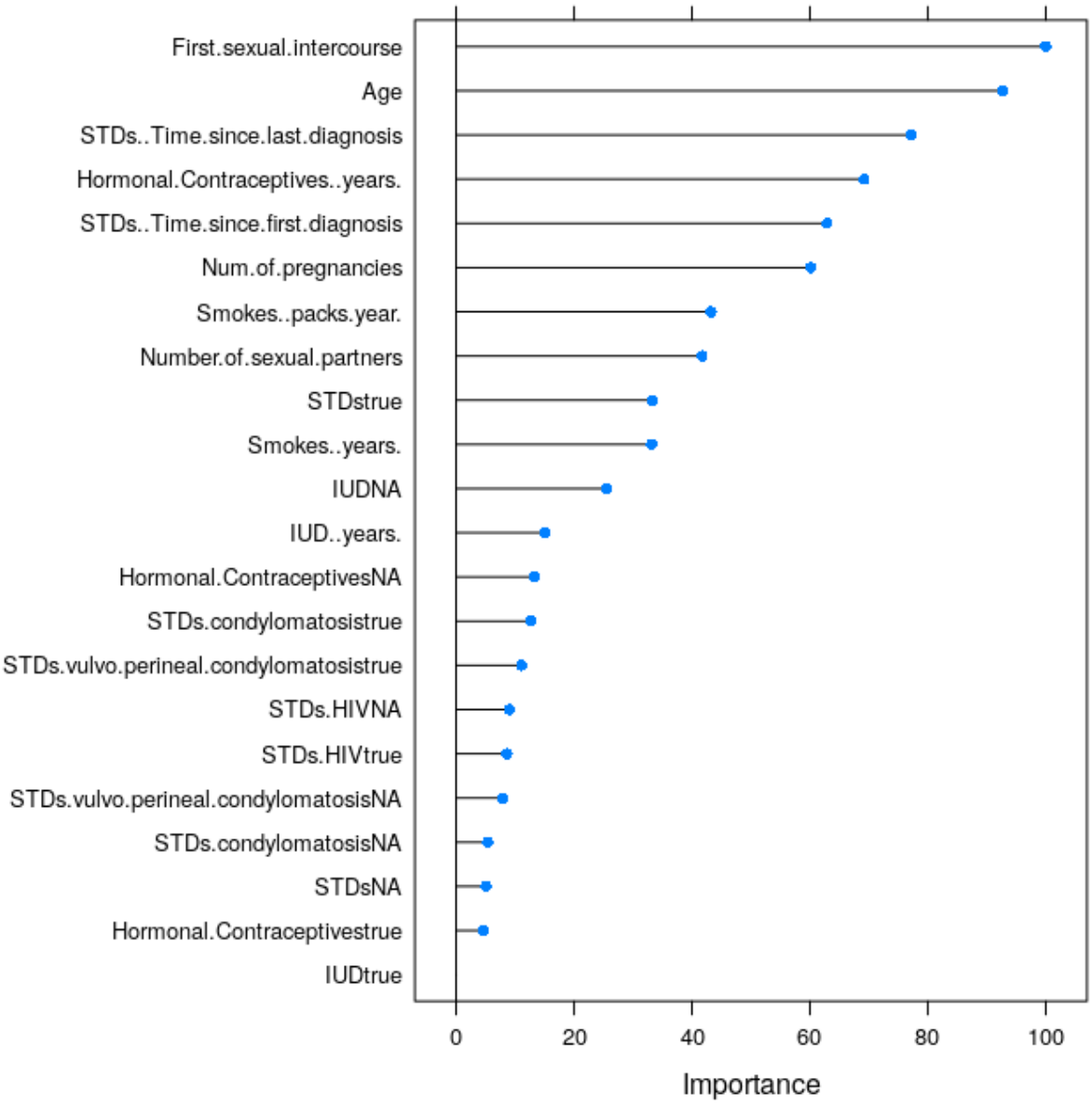


Na rozkładzie błędów widać fragment dla którego random forest popełnia mały błąd.

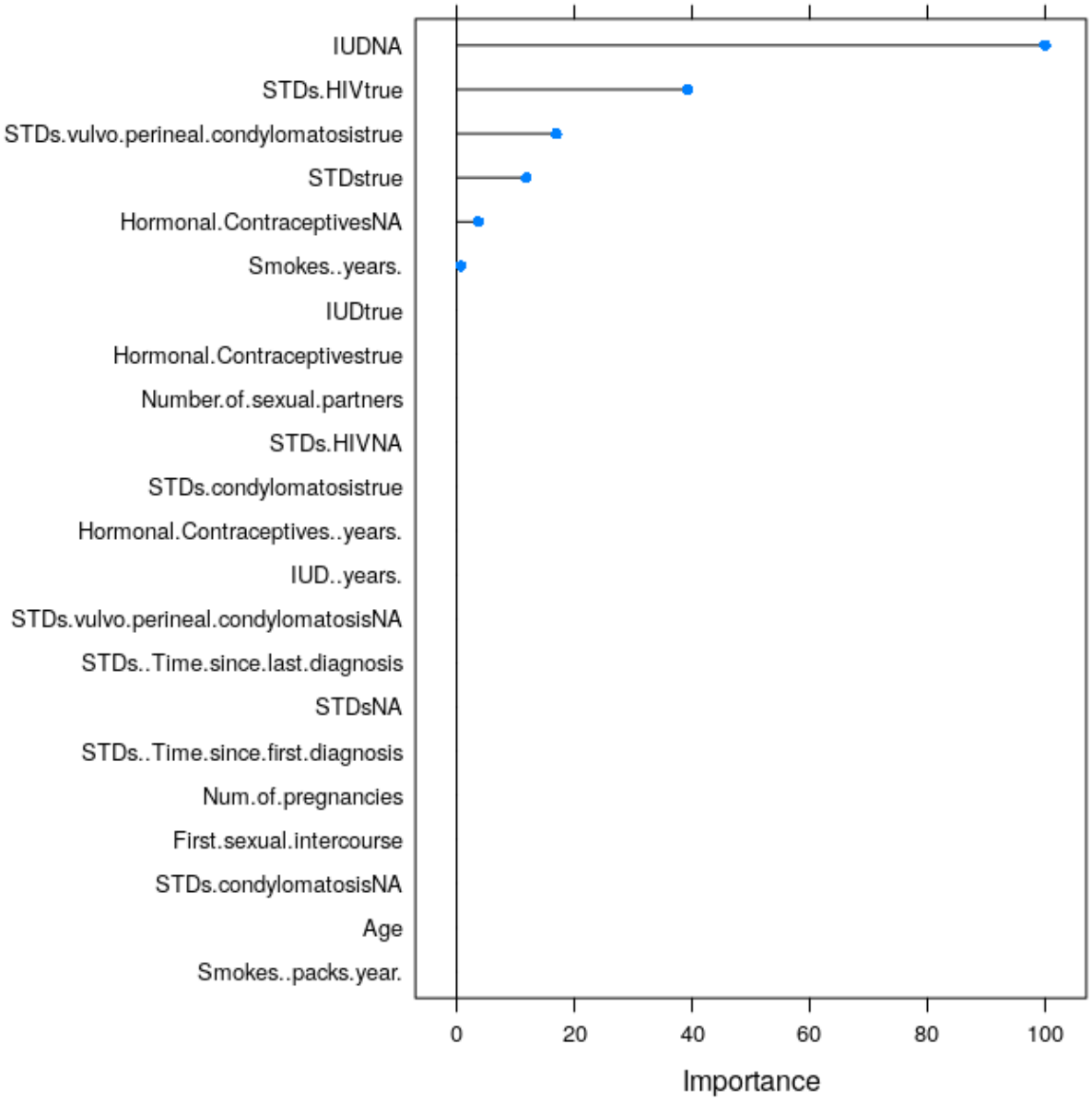
XAI

Variable importance

Random Forest


First.sexual.intercourse
Age
STDs..Time.since.last.diagnosis
Hormonal.Contraceptives..years.
STDs..Time.since.first.diagnosis
Num.of.pregnancies
Smokes..packs.year.
Number.of.sexual.partners
STDstrue
Smokes..years.
IUDNA
IUD..years.
Hormonal.ContraceptivesNA
STDs.condylomatosistrue
STDs.vulvo.perineal.condylomatosistrue
STDs.HIVNA
STDs.HIVtrue
STDs.vulvo.perineal.condylomatosisNA
STDs.condylomatosisNA
STDsNA
Hormonal.Contraceptivestrue
IUDtrue
0
20
40
60
80
100
Importance

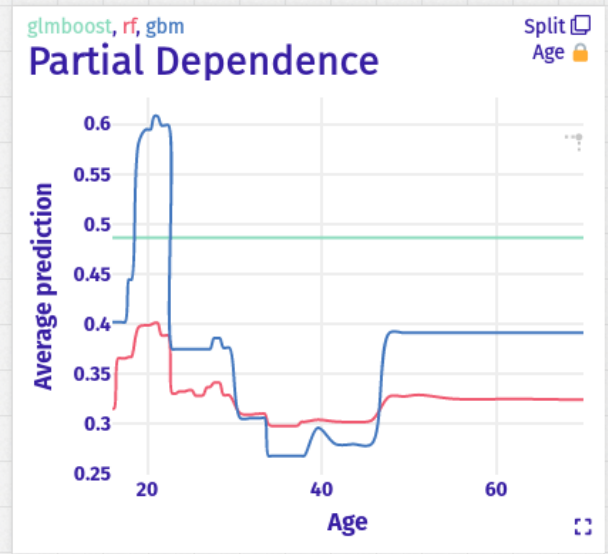
Boosted GLM

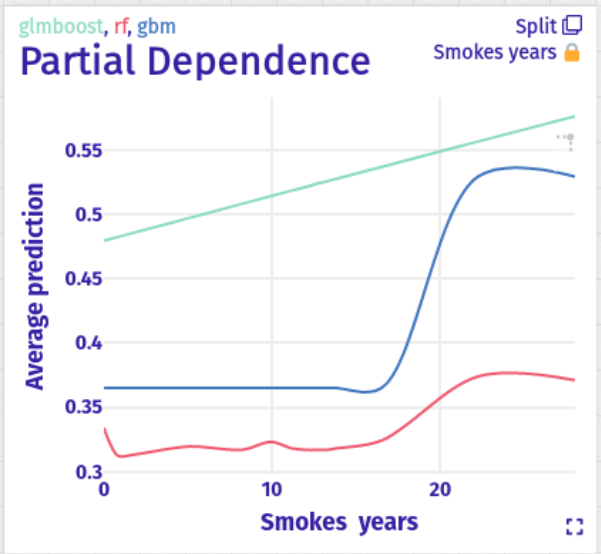


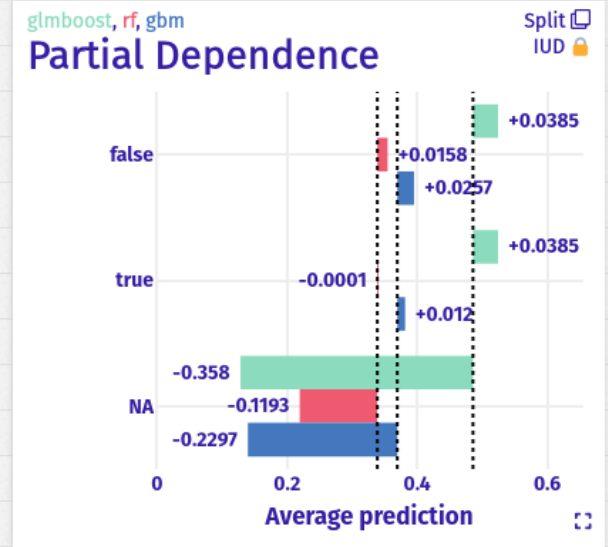
Arena

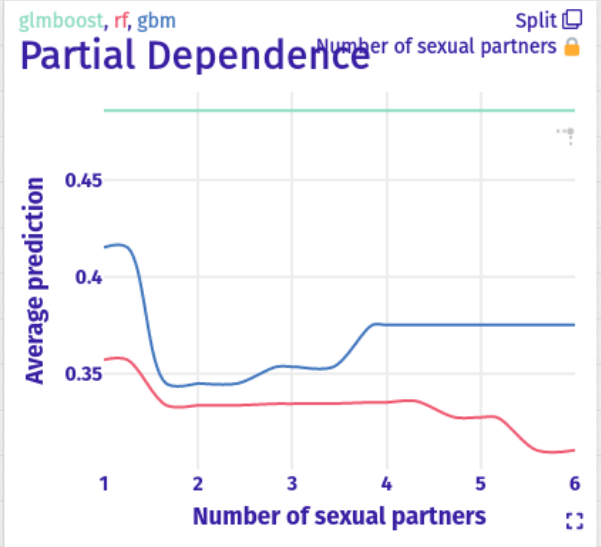
Pod poniższym adresem możliwa jest eksploracja wyjaśnień globalnych i lokalnych dla modeli. Poniżej zamieszczamy tylko wybrane. https://arena.drwhy.ai/branch/dev/?session=https://gist.githubusercontent.com/piotrpiatyszek/204135c535190bb4e6e585d636f216b5/raw/ac3dfd0be685136de1de9fbe696e246bd2466fdc/wum

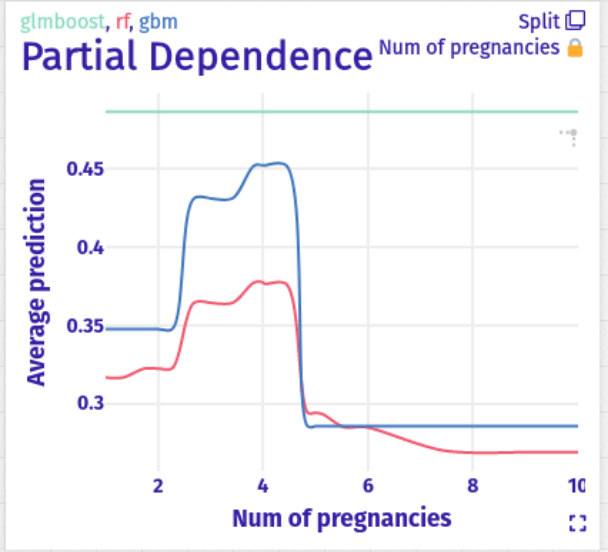
PDP

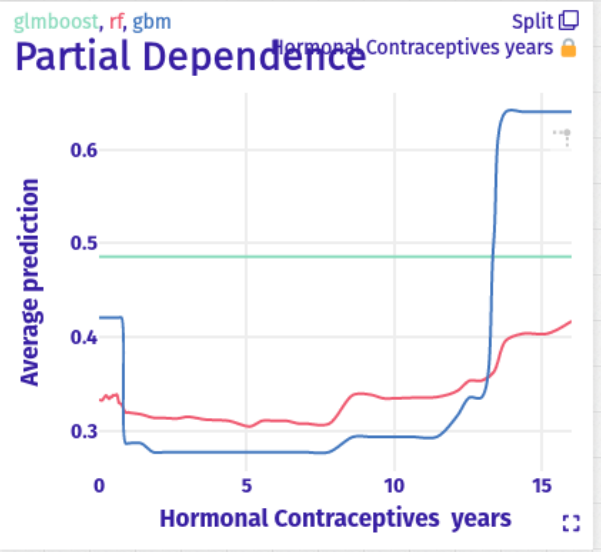

glmboost, rf, gbm
Split
Age
Partial Dependence
Average prediction
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
20
40
60
Age


glmboost, rf, gbm
Split
Smokes years
Partial Dependence
Average prediction
0.55
0.5
0.45
0.4
0.35
0.3
0
10
20
Smokes years


glmboost, rf, gbm
Split
IUD
Partial Dependence
false
true
NA
+0.0385
+0.0158
+0.0257
+0.0385
-0.0001
+0.012
-0.358
-0.1193
-0.2297
0
0.2
0.4
0.6
Average prediction


glmboost, rf, gbm
Split
Number of sexual partners
Partial Dependence
Average prediction
0.45
0.4
0.35
1
2
3
4
5
6
Number of sexual partners


glmboost, rf, gbm
Split
Num of pregnancies
Partial Dependence
Average prediction
0.45
0.4
0.35
0.3
2
4
6
8
10
Num of pregnancies


glmboost, rf, gbm
Split
Hormonal Contraceptives years
Partial Dependence
Average prediction
0.6
0.5
0.4
0.3
0
5
10
15
Hormonal Contraceptives years


glmboost, rf, gbm
Split
First sexual intercourse
Partial Dependence

Wyjaśnienia lokalne