Wakacje 2020

Jakub Wiśniewski

Plan prezentacji

- Fairness
- fairmodels
- Blog
- Moduł fairness w dalex

Czym jest fairness - podsumowanie

- Część Responsible ML
- Zajmuje się oceną sprawiedliwości decyzji podejmowanych przez modele
- W perspektywie fairness model to 3 wektory:
  - y - target (binary)
  - $\hat{y}$ - response
  - A - protected

Czy można jednoznacznie ocenić fairness? Na co uważać?

- Nie można, ale trzeba próbować!
- Zasada 4/5 (80%) w przypadku *statistical parity* [*UNIFORM GUIDELINES ON EMPLOYEE SELECTION PROCEDURES (1978)*]
- Możliwe sprawdzanie nie tylko binarnych *sensitive attributes* ale także tych niebinarnych oraz ich przecięć

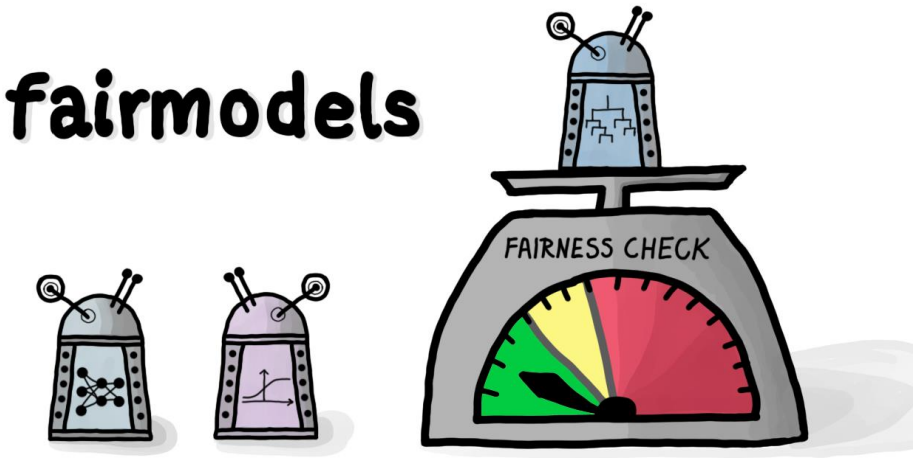
Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification
Joy Buolamwini , Timnit Gebru 2018

4.1. Key Findings on Evaluated Classifiers

- All classifiers perform better on male faces than female faces (8.1% – 20.6% difference in error rate)
- All classifiers perform better on lighter faces than darker faces (11.8% – 19.2% difference in error rate)
- All classifiers perform worst on darker female faces (20.8% – 34.7% error rate)
- Microsoft and IBM classifiers perform best on lighter male faces (error rates of 0.0% and 0.3% respectively)
- Face++ classifiers perform best on darker male faces (0.7% error rate)
- The maximum difference in error rate between the best and worst classified groups is 34.4%

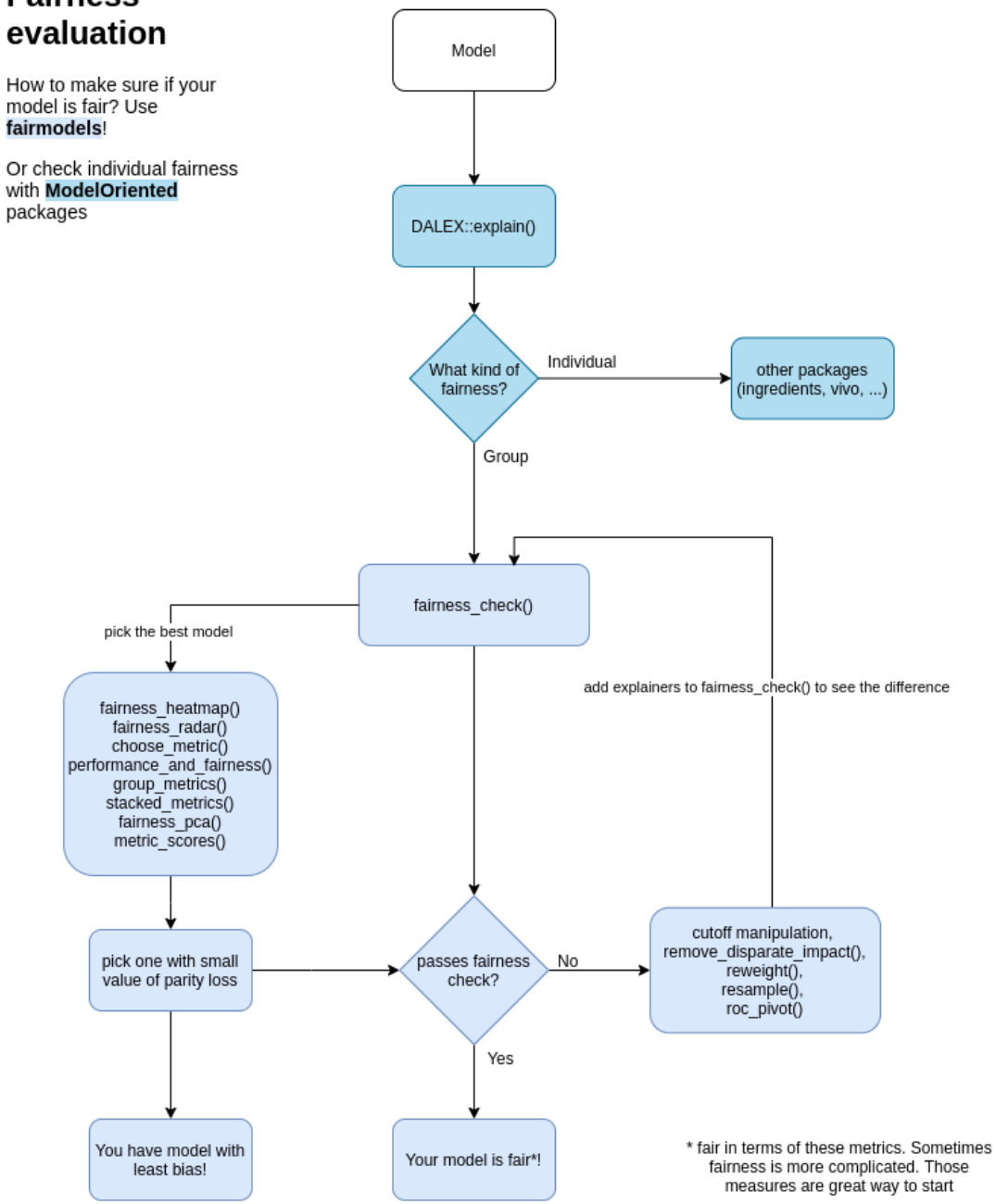
fairmodels

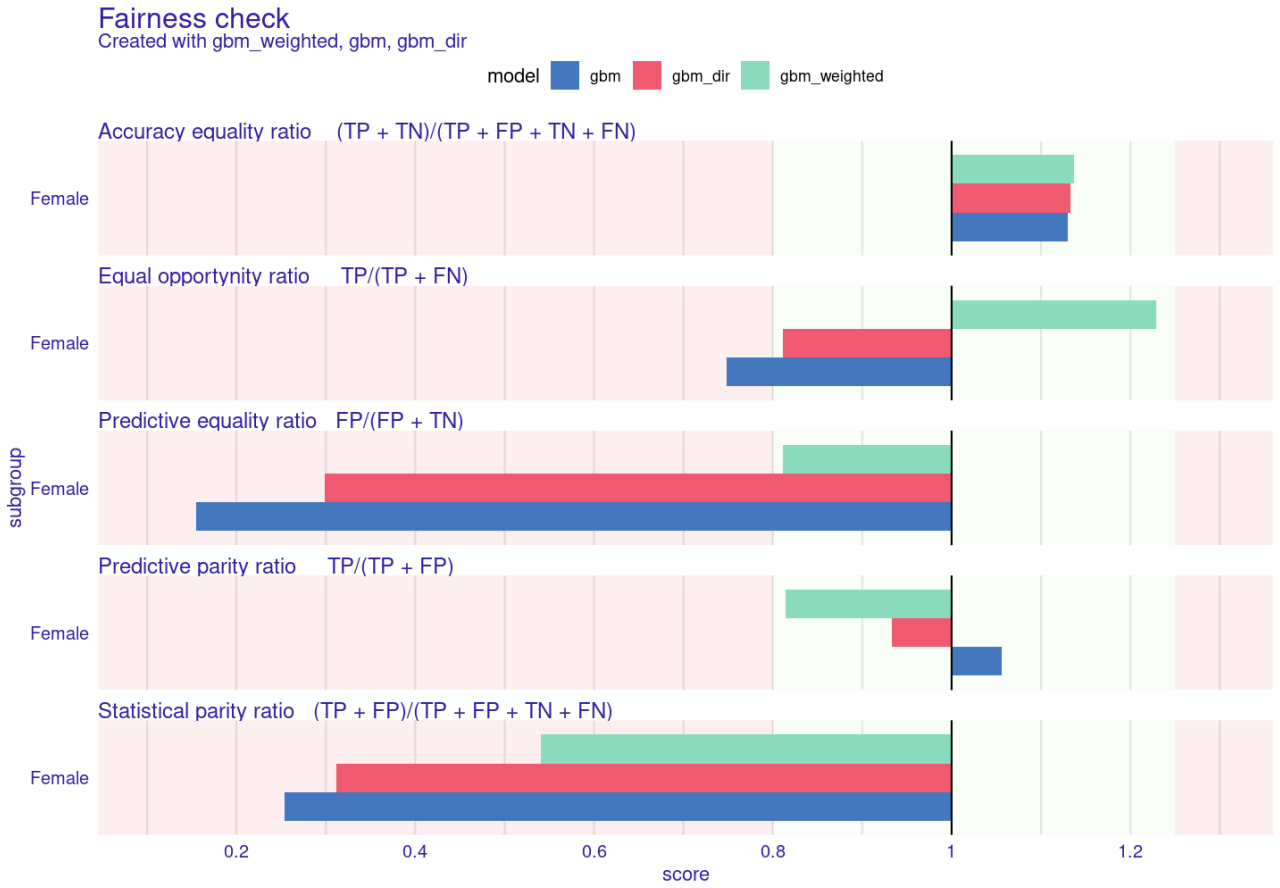
- Jedyne* takie narzędzie w R
- Na CRAN od 20 sierpnia
- Pojawia się w *top 40 CRAN new packages (August)*
- Wylistowane w “awesome-machine-learning-*interpretability*”



* które obsługuje modele oraz jest używalne

Jak używać fairmodels?





Metryki

- Metryki bazujące na macierzy pomyłek
- Pochodzące z nich metryki fairness
  - Equal opportunity (TPR)
  - Predictive parity (PPV)
  - Predictive equality (FPR)
  - Accuracy (ACC)
  - Statistical parity (STP = (TP+FP)/(TP+FP+TN+FN))

Parity loss – dlaczego zmiana?

- Kiedyś:

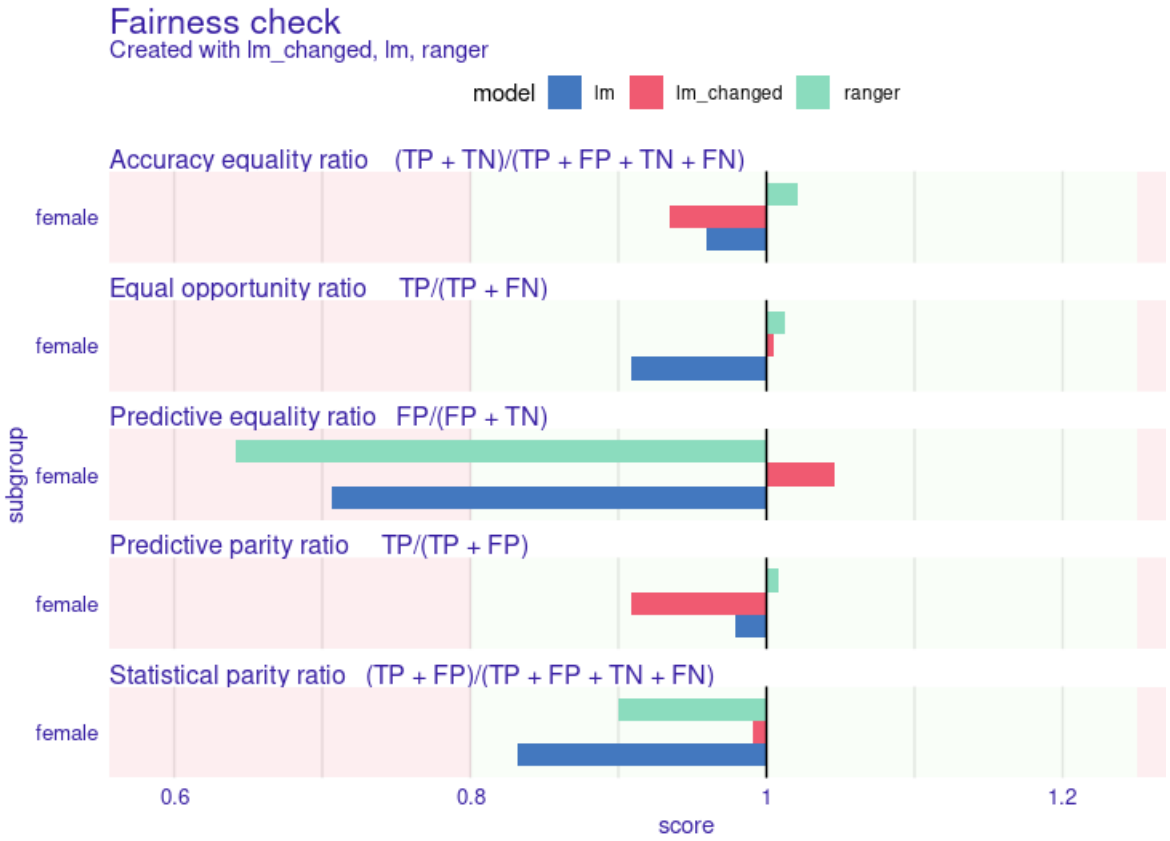
$$TPR_{parity\_loss} = \sum_i |TPR_{A=i} - TPR_{A=privileged}|$$

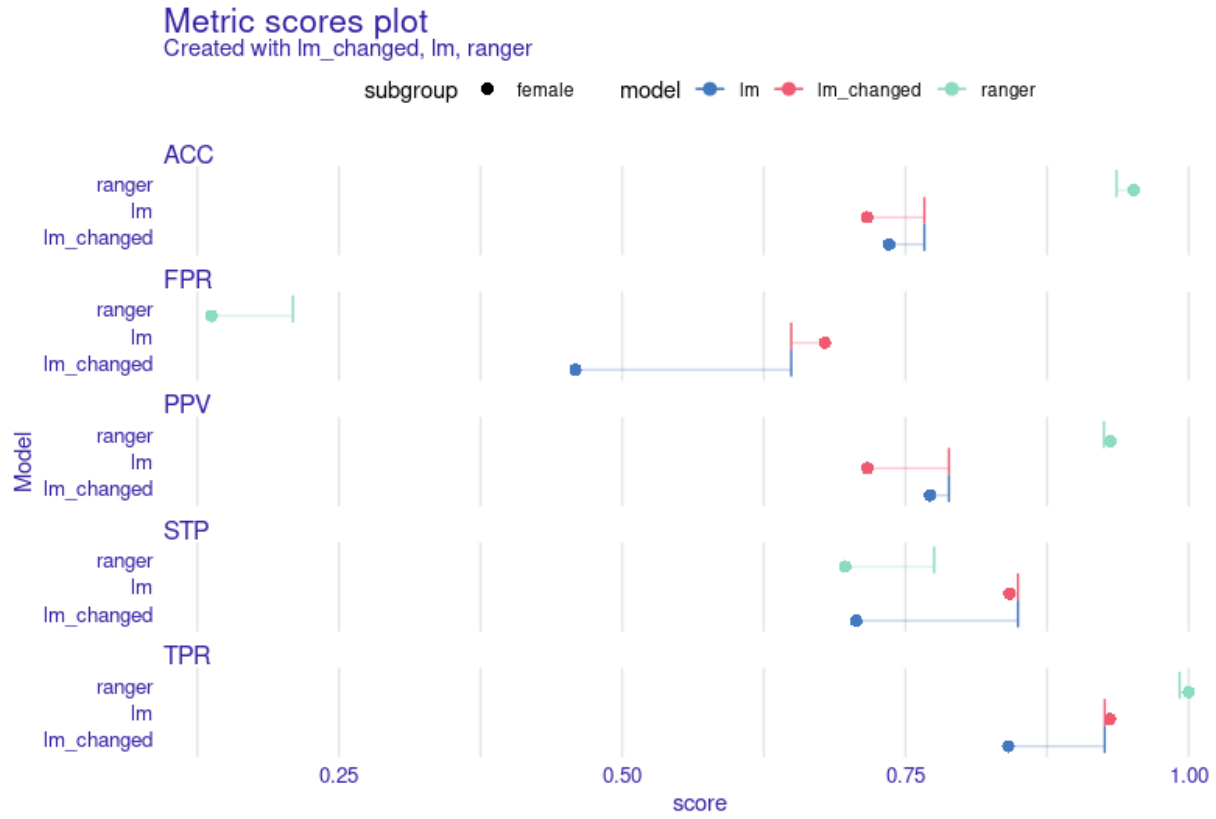
- Teraz:

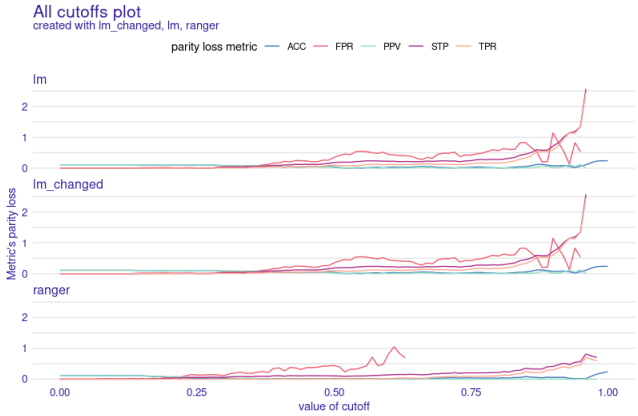
$$TPR_{parity\_loss} = \sum_i \left|\ln\left(\frac{TPR_{A=i}}{TPR_{A=privileged}}\right)\right|$$

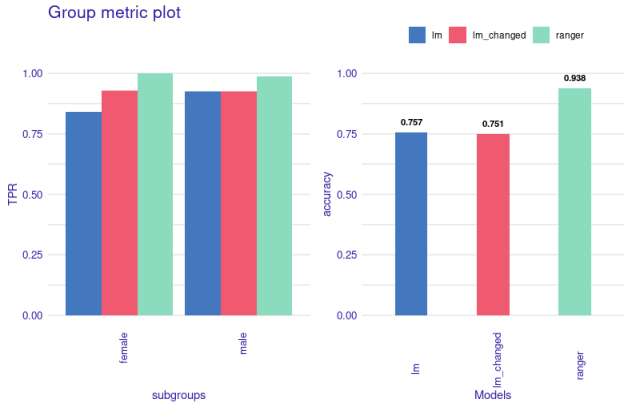
- Po co?
  - Aby dopasować się do Amerykańskiego prawa (four-fifths rule)
  - Aby dla dużych oraz małych wartości metryk interpretacja była ta sama.

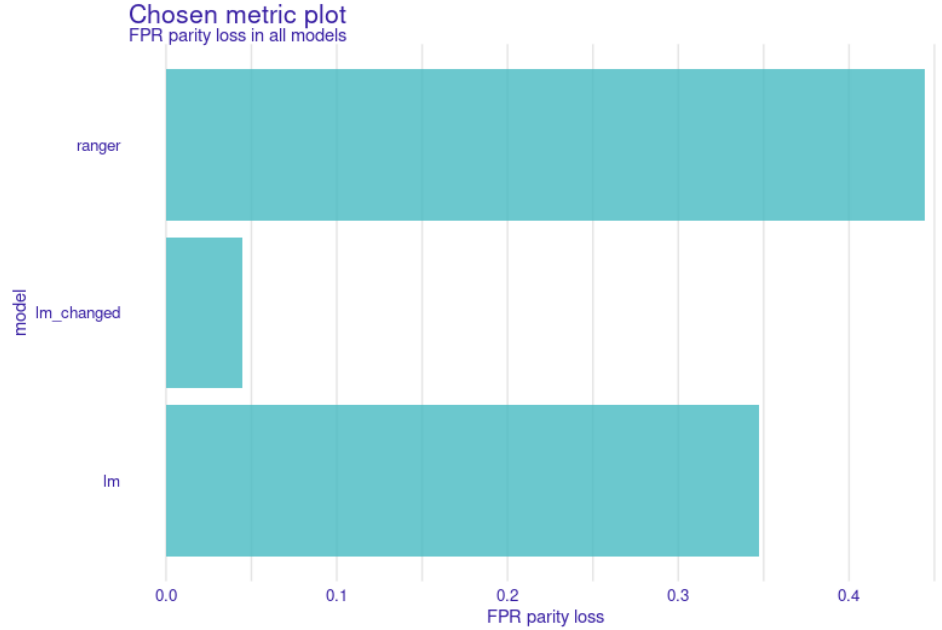
Wizualizacja fairness

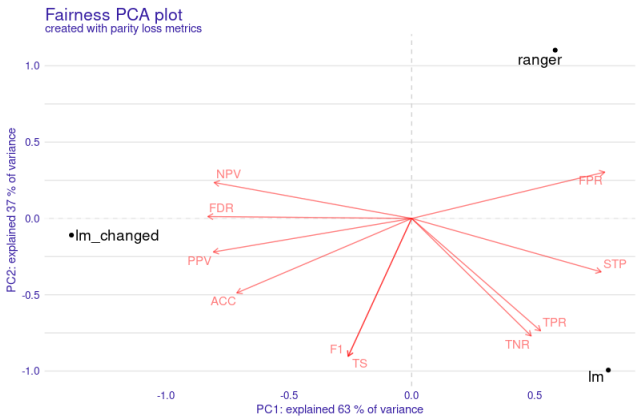


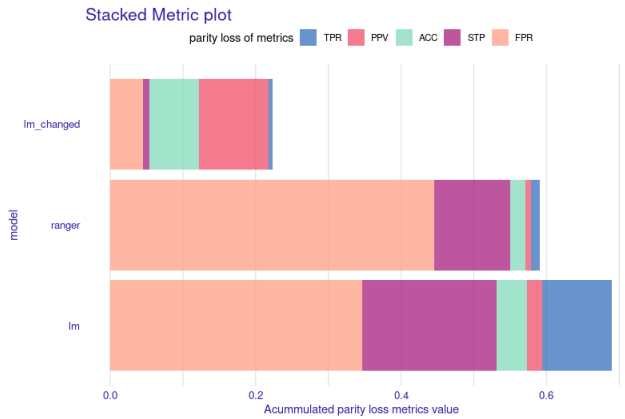


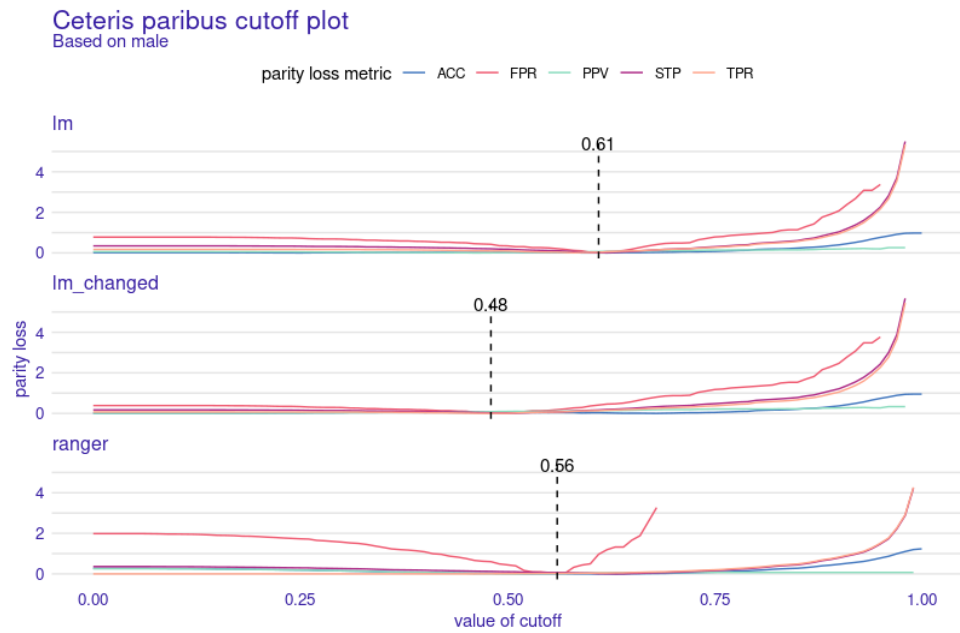

All cutoffs plot
created with lm_changed, lm, ranger
parity loss metric ACC FPR PPV STP TPR
lm
lm_changed
ranger
Metric's parity loss
value of cutoff

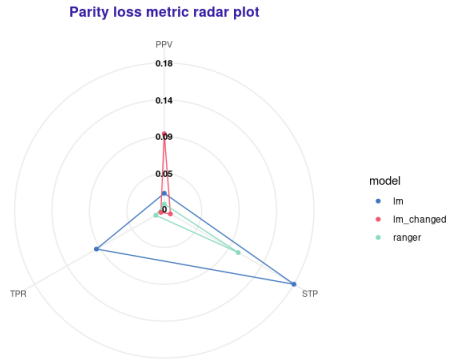

Group metric plot
lm lm_changed ranger
TPR
subgroups
female
male
accuracy
0.757
0.751
0.938
Models

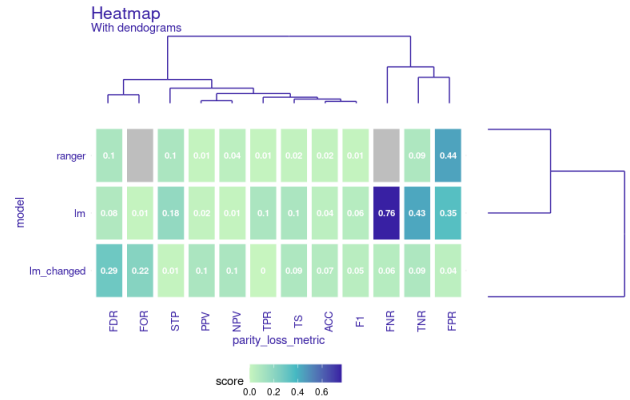

Chosen metric plot
FPR parity loss in all models
ranger
lm_changed
lm
model
FPR parity loss


Fairness PCA plot
created with parity loss metrics
ranger
lm_changed
lm
NPV
FDR
PPV
ACC
FPR
STP
TPR
TNR
F1
TS
PC2: explained 37 % of variance
PC1: explained 63 % of variance


Stacked Metric plot
parity loss of metrics TPR PPV ACC STP FPR
lm_changed
ranger
lm
model
Acummulated parity loss metrics value

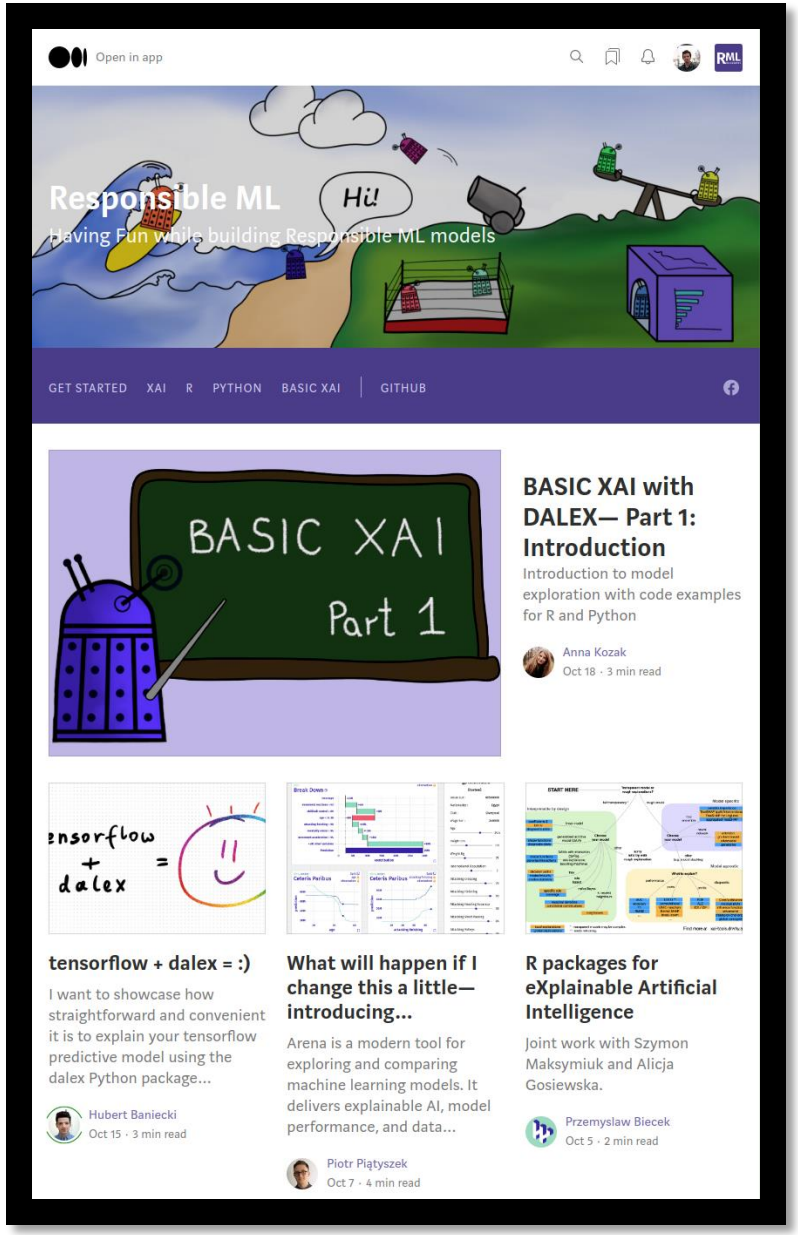

Ceteris paribus cutoff plot
Based on male
parity loss metric ACC FPR PPV STP TPR
lm
0.61
lm_changed
0.48
ranger
0.56
parity loss
value of cutoff


Parity loss metric radar plot
PPV
TPR
STP
model
lm
lm_changed
ranger


Heatmap
With dendograms
ranger
lm
lm_changed
model
FDR FOR STP PPV NPV TPR TS ACC F1 FNR TNR FPR
parity_loss_metric
score

Blog - miesięcznie

- 3027 wyświetleń
- Ponad 30 godzin spędzonych na czytaniu
- Średnio 63 unikalnych użytkowników dziennie
- Maksymalnie 294 użytkowników dziennie



Blog – statystyki

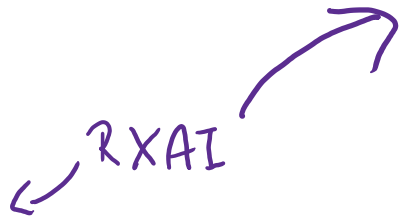
- Najwięcej zasięgów miały blogi Kasi oraz S&A&P
- Duży wpływ ma Linkedin Przemka: +230 wyświetleń

OCTOBER 2020

BASIC XAI with DALEX— Part 1: Introduction 3 min read · In ResponsibleML · View story · Details	176 +23K	115	65%	4
tensorflow + dalex = :) 3 min read · In ResponsibleML · View story · Details	162	95	59%	6
What will happen if I change this a little— i... 4 min read · In ResponsibleML · View story · Details	245 +34K	108	44%	5
R packages for eXplainable Artificial Intellig... 2 min read · In ResponsibleML · View story · Details	800 +28K	468	59%	10
Imputing missing data in mlr3 with EMMA 4 min read · In ResponsibleML · View story · Details	261 +24K	137	52%	6

SEPTEMBER 2020

What is new in fairmodels? 4 min read · In ResponsibleML · View story · Details	231 +35K	115	50%	6
What's new in DALEX and DALEXtra 7 min read · In ResponsibleML · View story · Details	435 +36K	155	36%	8
Explaining models with Triplot, part 1 4 min read · In ResponsibleML · View story · Details	658 +41K	325	49%	8
How to use the Arena for exploration of ML ... 3 min read · In ResponsibleML · View story · Details	119	82	69%	7
Introduction to ResponsibleML publication 3 min read · In ResponsibleML · View story · Details	151	87	58%	1



VIEWS BY TRAFFIC SOURCE	800
Internal	0%
External referrals	100%
RSS readers (full text)	28K
linkedin.com	266
email, IM, and direct	182
r-bloggers.com	152
Android device (not Medium app)	65
google.com	16
feedly.com	16
news.google.com	6
twitter.com	5
mail.google.com	4
All other external referrals	12

Moduł do dalex (in progress)

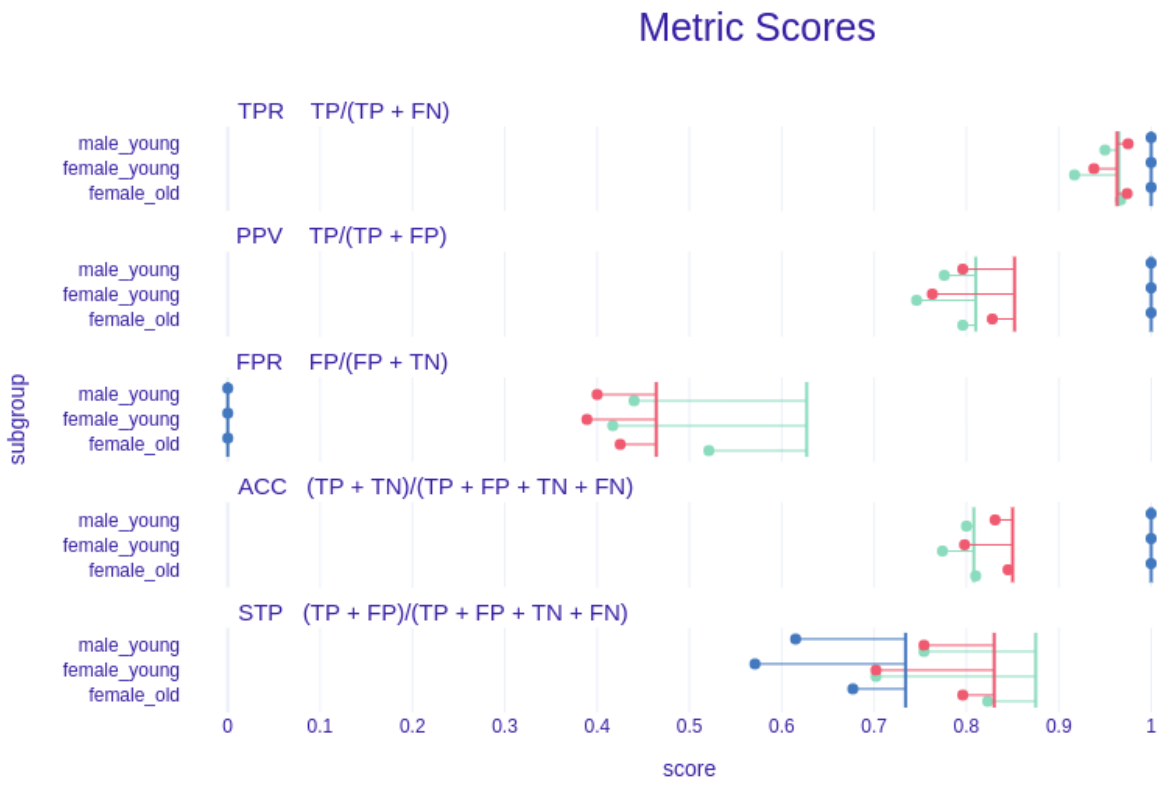
LEPIEJ ZAIMPLEMENTOWANY FAIRMODELS

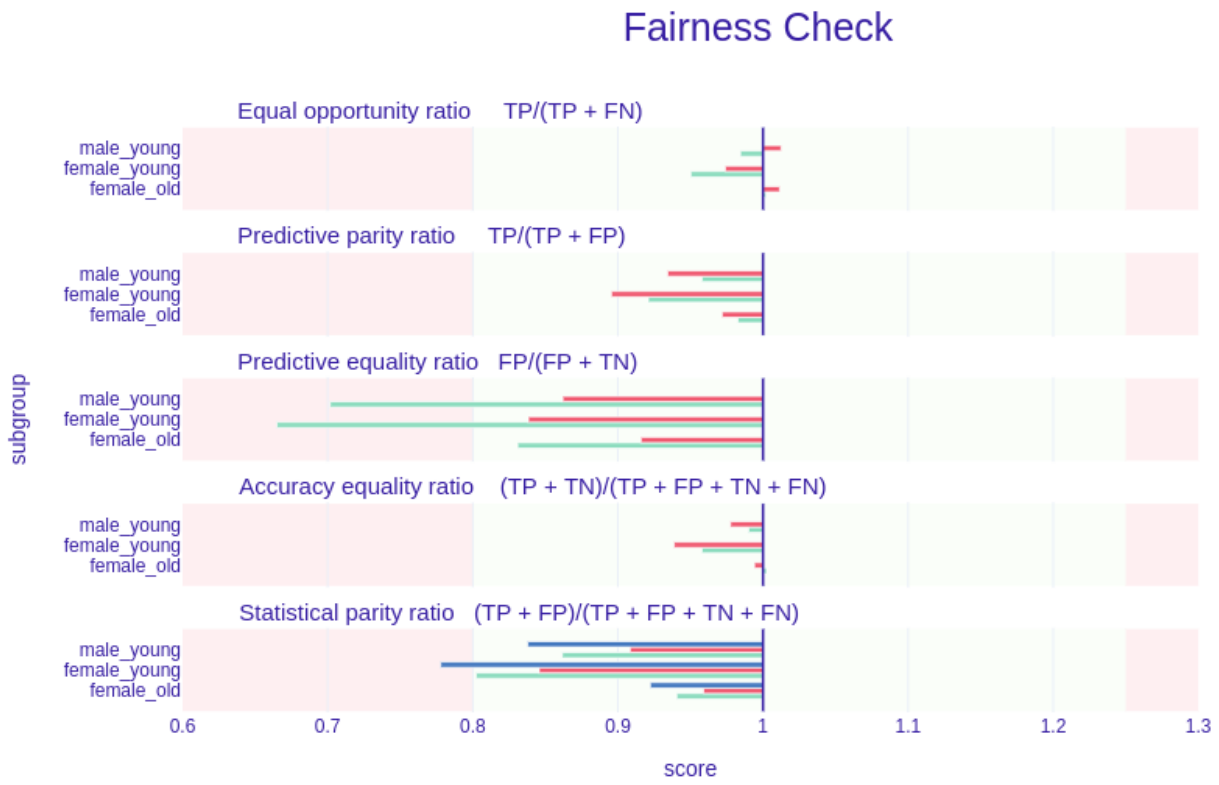
MNIEJ FUNKCJONALNOŚCI, SKUPIENIE SIĘ NA GŁÓWNYCH FEATURES

DOBRE PODSTAWY POD INDIVIDUAL FAIRNESS & FAIRNESS IN REGRESSION

ZAIMPLEMENTOWANY BAKCEND ORAZ 2 (MOIM ZDANIEM) NAJWAŻNIEJSZE WYKRESY

Interaktywne wykresy plotly





Koniec :)