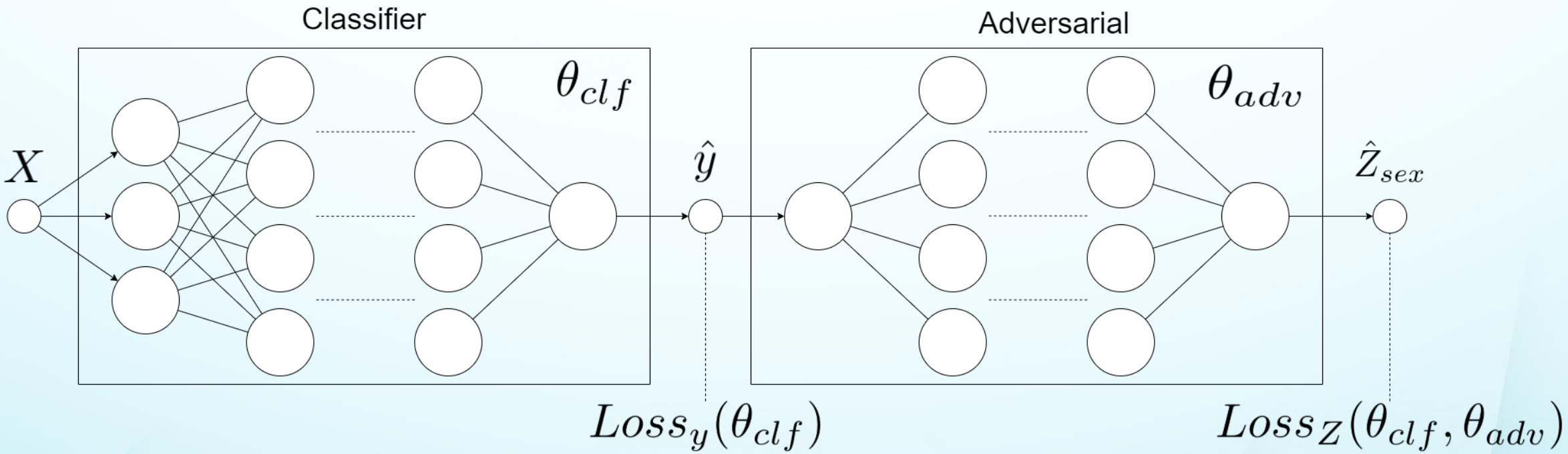
FairPANs - bringing fairness to neural networks

Author: Hubert Ruczyński

Mentor: Jakub Wiśniewski

Meet FairPANs – Fair Predictive Adversarial Networks



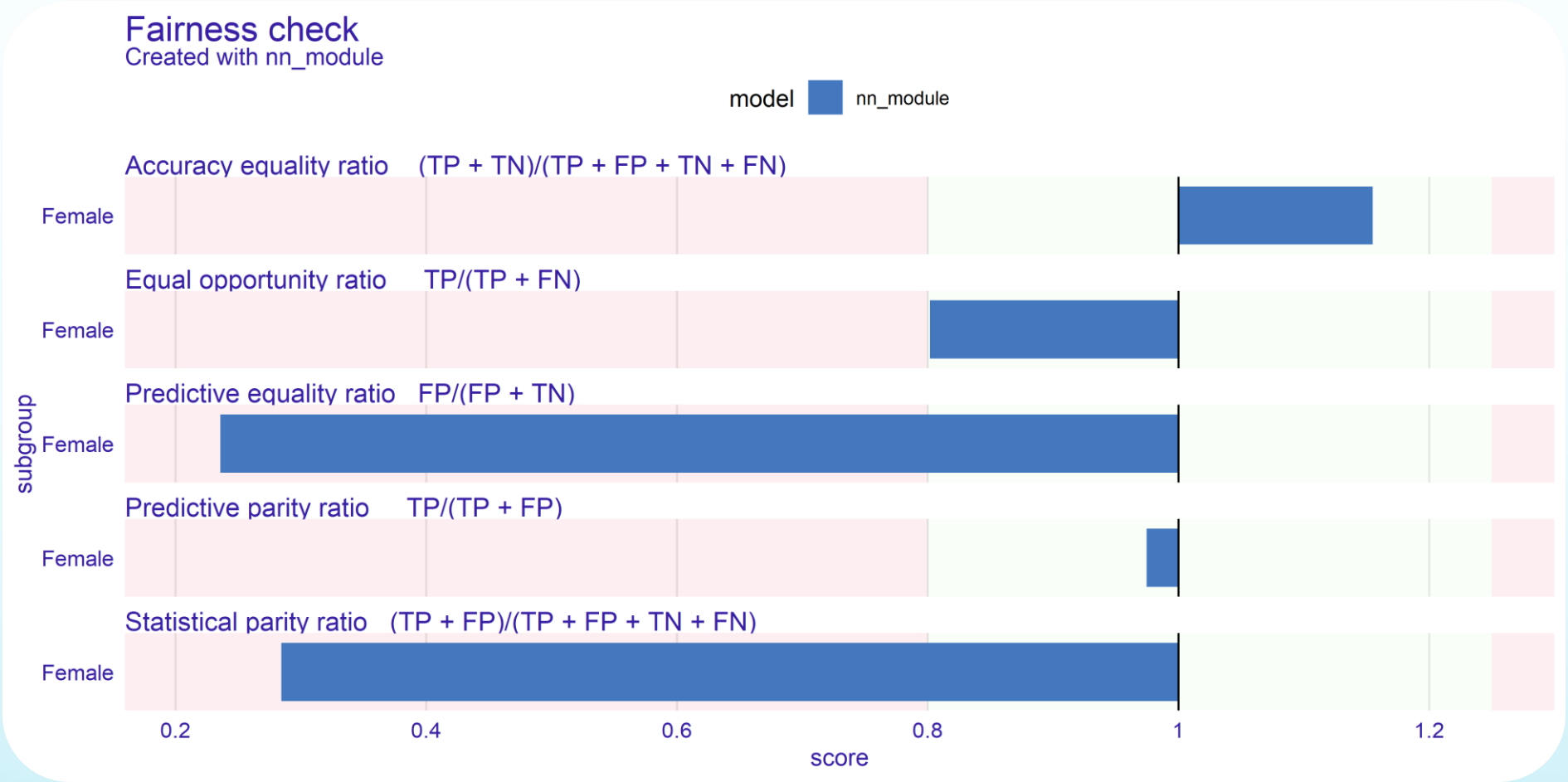
Custom Loss Function

$$min_{\theta_{clf}}[Loss_Z(\theta_{clf}, \theta_{adv})]$$

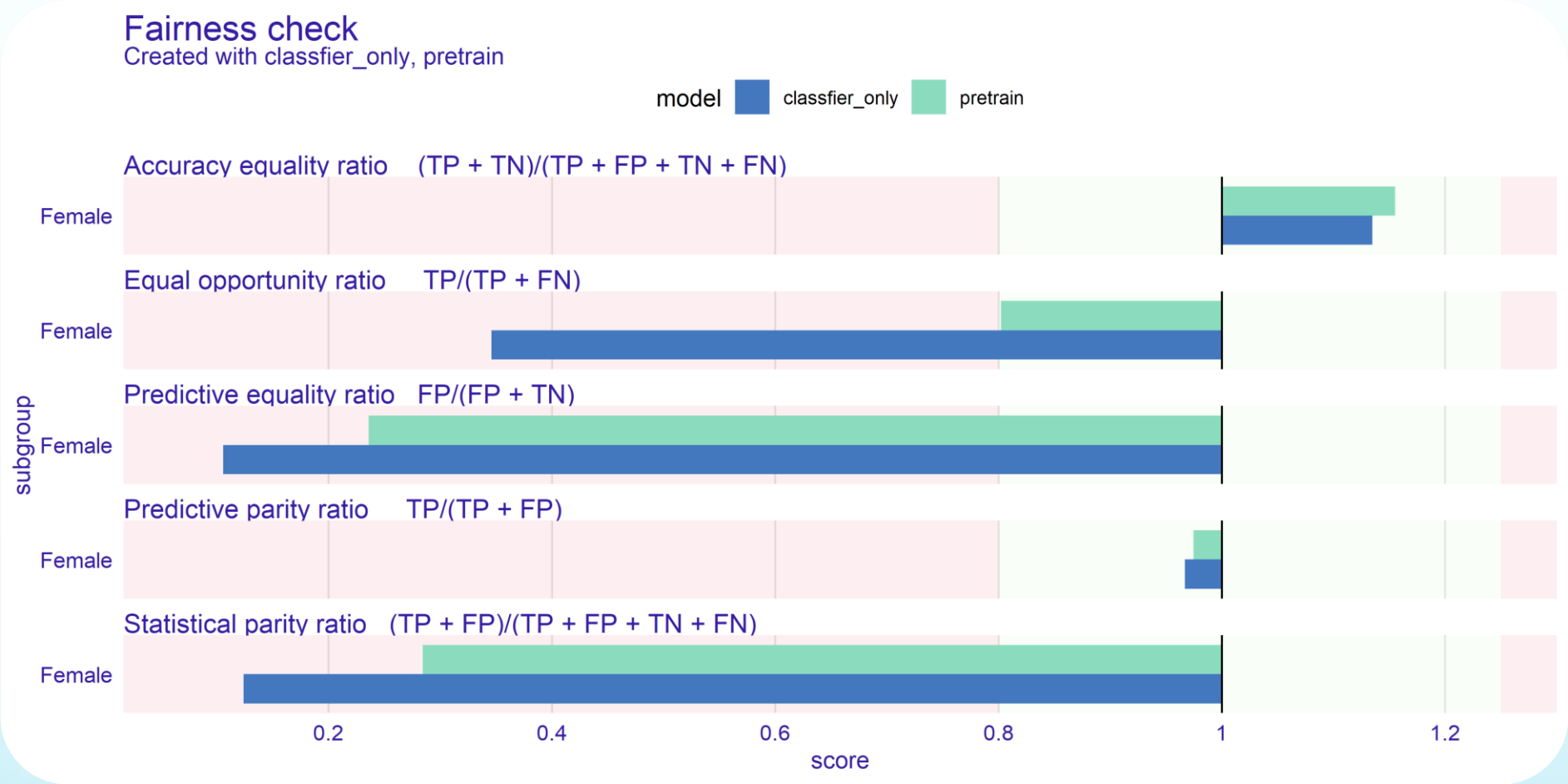
$$min_{\theta_{clf}}[Loss_y(\theta_{clf}) - \lambda Loss_Z(\theta_{clf}, \theta_{adv})]$$

FairWHY ?

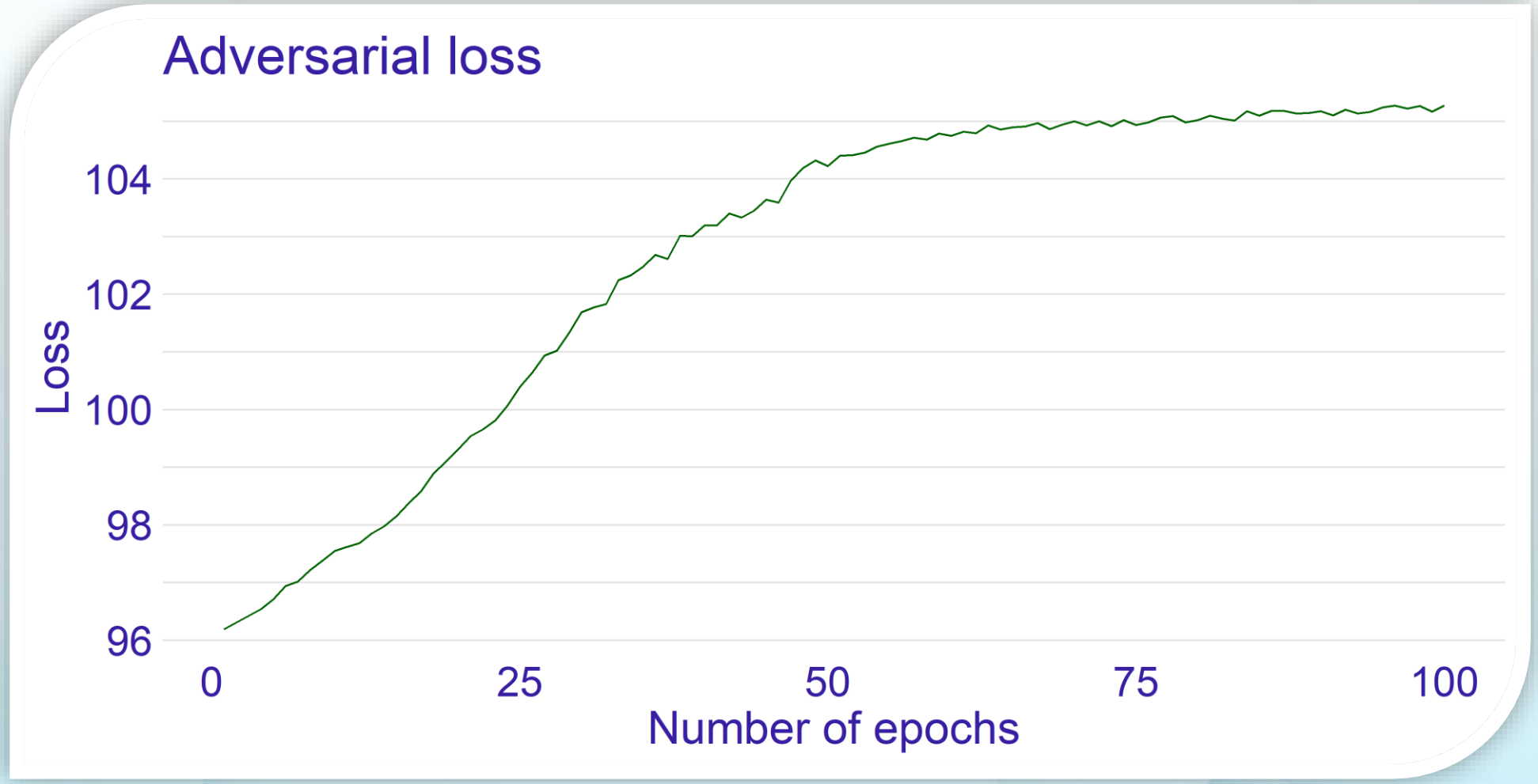
Pre-train


Fairness check
Created with nn_module
model
nn_module
Accuracy equality ratio (TP + TN)/(TP + FP + TN + FN)
Female
Equal opportunity ratio TP/(TP + FN)
Female
Predictive equality ratio FP/(FP + TN)
Female
Predictive parity ratio TP/(TP + FP)
Female
Statistical parity ratio (TP + FP)/(TP + FP + TN + FN)
Female
subgroup
0.2
0.4
0.6
0.8
1
1.2
score

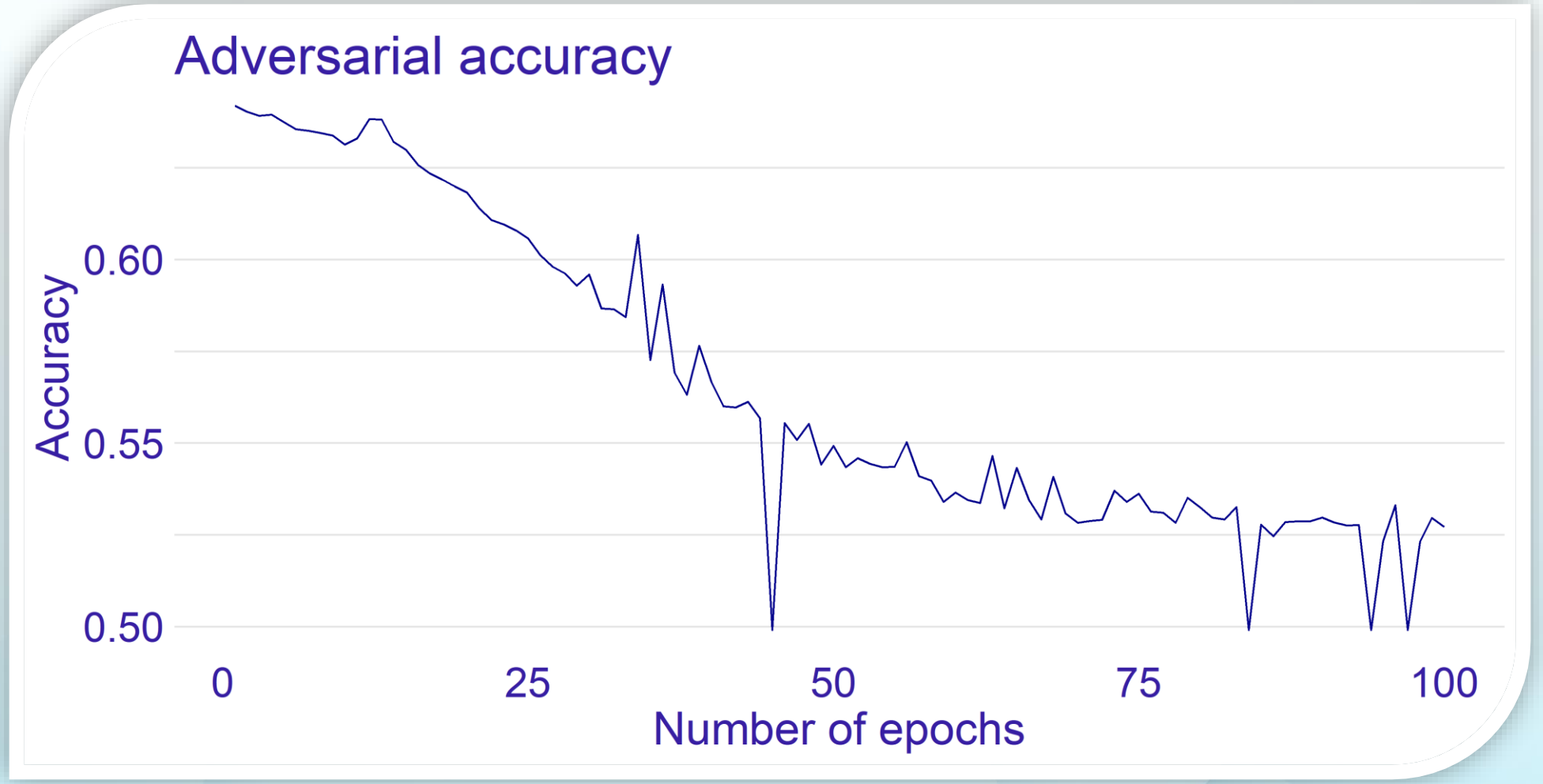
Classifier training



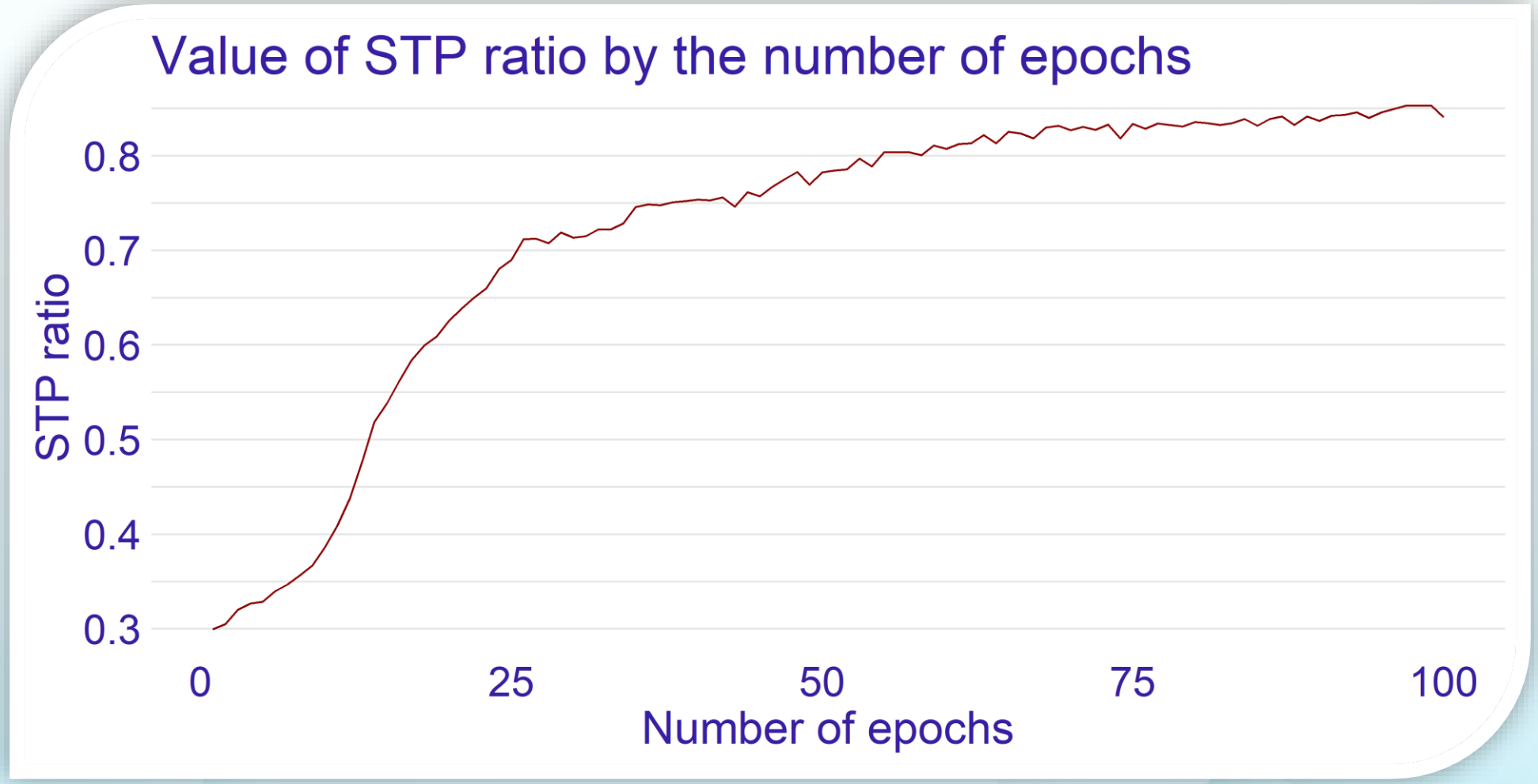
Adversarials loss



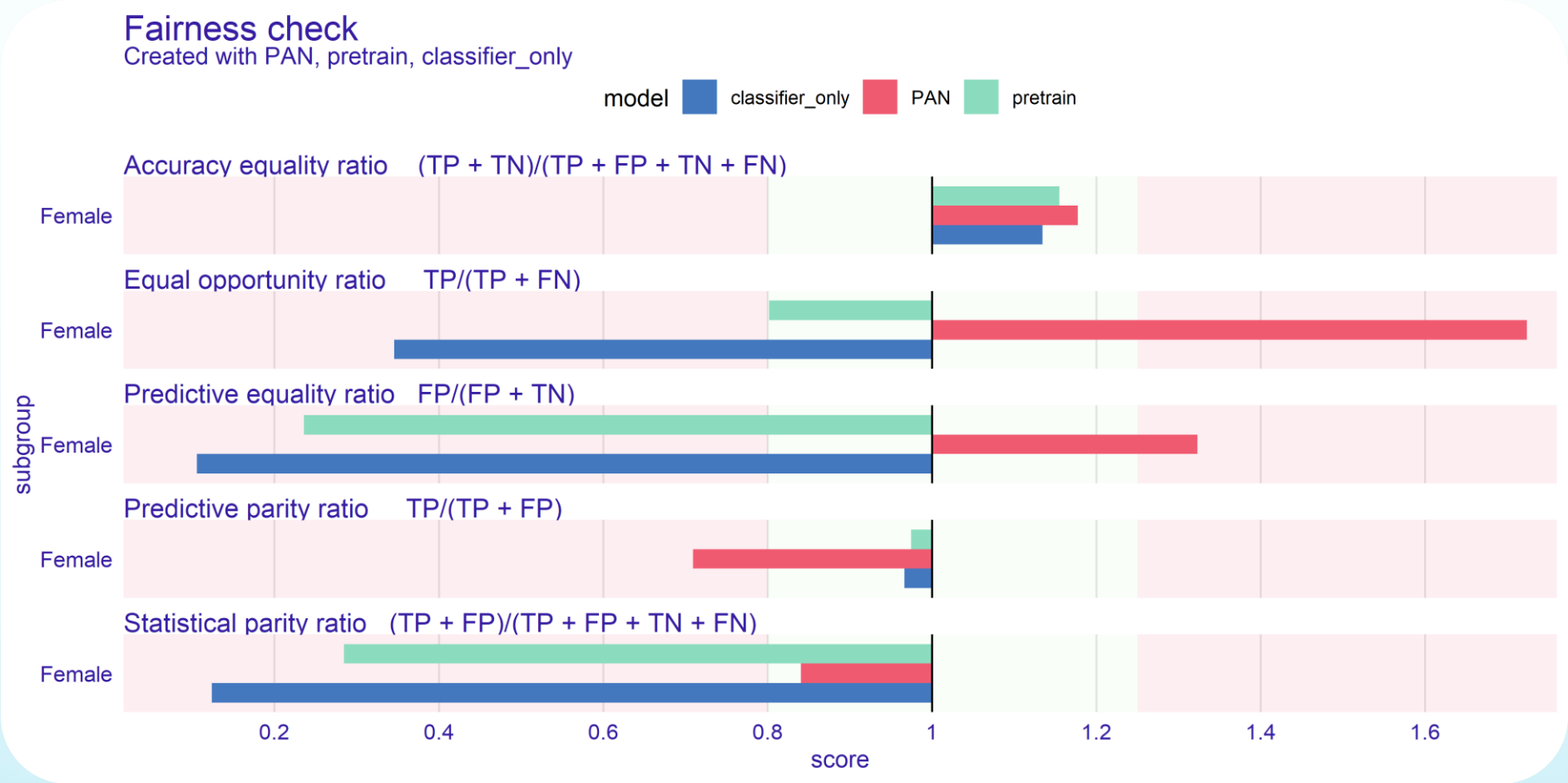
Adversiarials accuracy



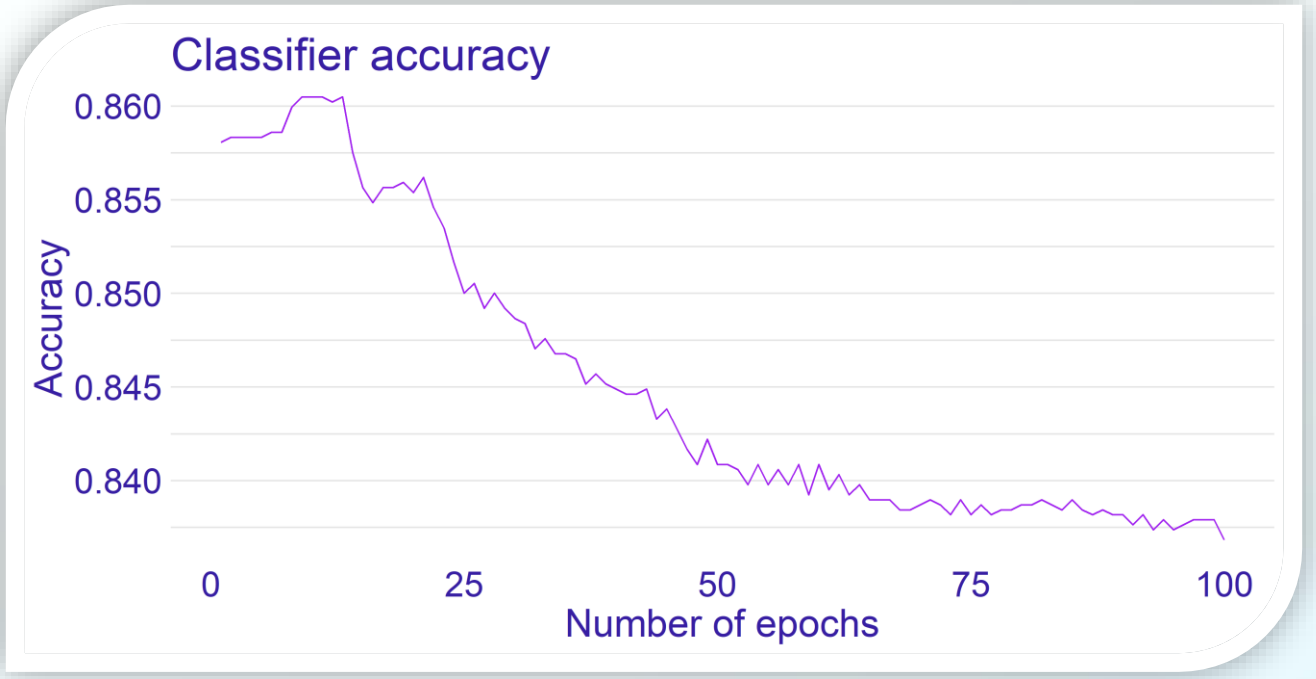
Statistical Parity Ratio



The final outcome



Metrics comparison



	Accuracy	AUC	F1	Precision	Recall
Classifier	0.850	0.871	0.541	0.773	0.417
PAN	0.837	0.857	0.536	0.680	0.443

Thank you for your attention