Aspects éthique

Véronique Tremblay

Confidentialité

- Anonymisation
- Limiter les détails
- Regrouper les catégories de faibles effectifs
- Ajouter du bruit
- Confidentialité différentielle

Les implications sociales

L'équité algorithmique

Les mythes

«La machine apprend toute seule.»

«C'est objectif, c'est basé sur des données.»

Les sources de discrimination

- Échantillon biaisé
- Cibles biaisées
- Validité variable
- Classes mal représentées

Un autre mythe

«Mon modèle ne peut pas être sexiste car je n'ai pas utilisé le genre pour le construire.»

Qu'est-ce qu'on peut faire?

- Modifier l'échantillon avant l'ajustement du modèle (*pre-processing*)
- Modifier le modèle *a posteriori* (*post-processing*)
- Modifier la fonction de perte

Mesurer l'équité d'un modèle

Notation

$\hat{Y}$: Valeur prédite par le modèle

Y: Vraie valeur

S: est la variable sensible et $S = s$ correspond à la modalité habituellement discriminée

Pour simplifier les explications, nous allons supposer que Y et S ont deux modalités mais ces mesures se généralisent.

Parité démographique

On parle parfois de *equal parity*, *statistical parity* ou d'indépendance.

$$\hat{Y} \perp S$$

$$\mathbb{P}(\hat{Y} = 1) = \mathbb{P}(\hat{Y} = 1|S = s) = \mathbb{P}(\hat{Y} = 1|S \neq s)$$

Avec la librairie `fairness`, utilisez `dem_parity`.

Égalité de l'exactitude

$$\mathbb{P}(\hat{Y} = Y | S = s) = \mathbb{P}(\hat{Y} = Y | S \neq s)$$

Avec la librairie `fairness`, c'est la fonction `acc_parity`.

Il est préférable d'utiliser le taux de faux négatifs et de faux positifs en complément, avec les fonctions `fnr_parity` et `fpr_parity`.

Égalité des chances

On parle parfois de *Positive Rate Parity* ou de séparation.

$$(\hat{Y} = 1 | Y = 1) \perp S$$

$$\mathbb{P}(\hat{Y} = 1 | Y = 1, S = s) = \mathbb{P}(\hat{Y} = 1 | Y = 1, S \neq s)$$

Avec la librairie `fairness`, utilisez `equal_odds`.

Égalité de la précision

$$(Y = 1|\hat{Y} = 1) \perp S$$

$$\mathbb{P}(Y = 1|\hat{Y} = 1, S = s) = \mathbb{P}(Y = 1|\hat{Y} = 1, S \neq s)$$

Avec la librairie `fairness`, utilisez `pred_rate_parity`.

En pratique - exemple avec les données de COMPAS

Librairies

```
#install.packages("fairness")
library(fairness)
```

```
data("compas")
data("germancredit")
```

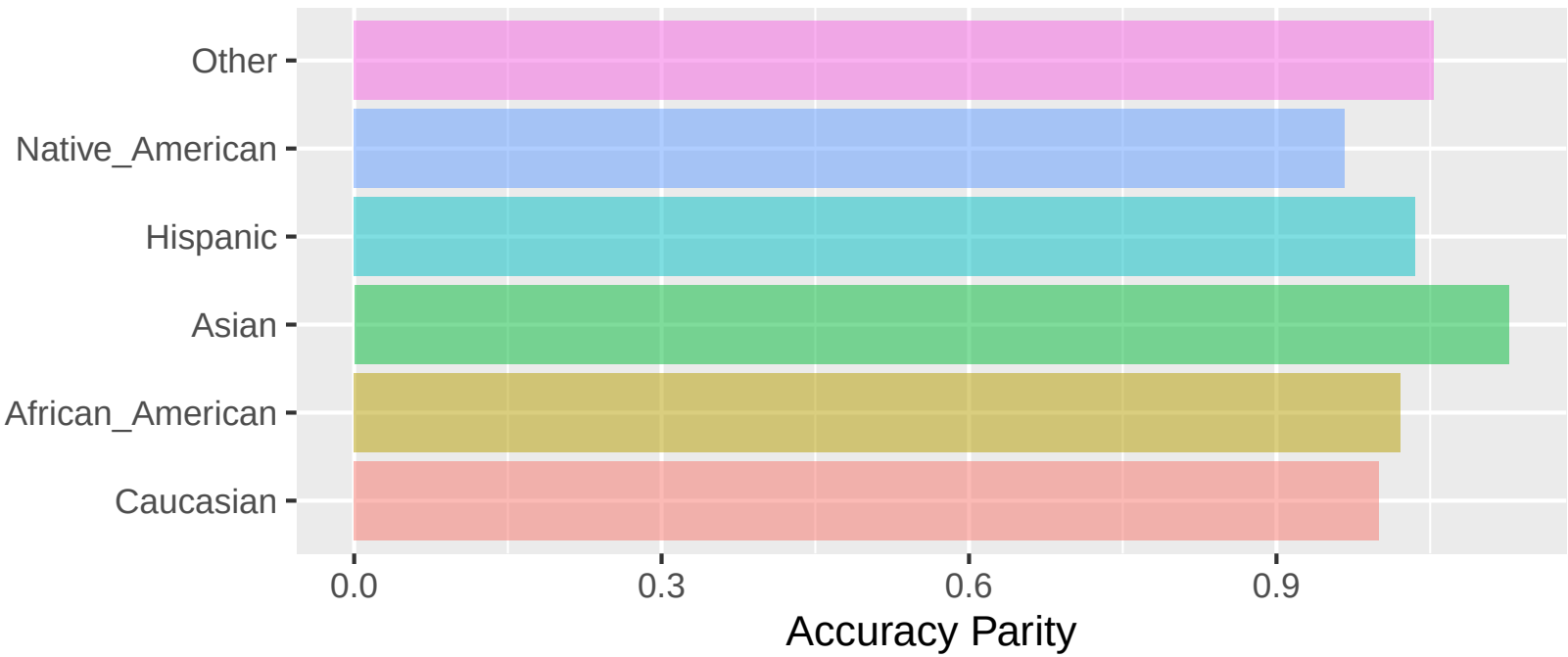
Égalité de l'exactitude

```
egal_exact <- acc_parity(data    = compas,
            outcome = "Two_yr_Recidivism",
            group   = "ethnicity",
            probs   = "probability",
            #preds   = NULL,
            cutoff = 0.5,
            base   = "Caucasian")
```

Égalité de l'exactitude

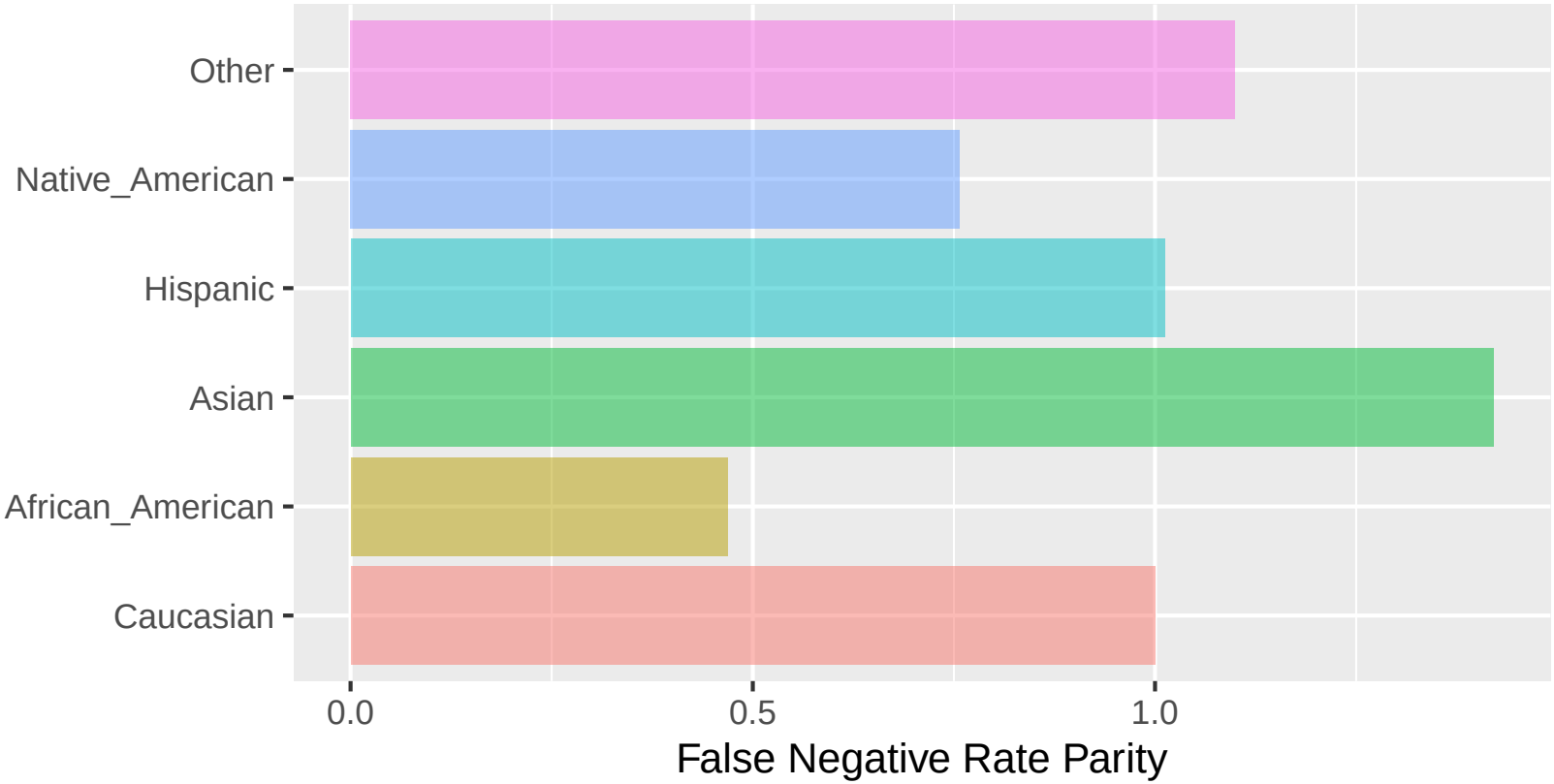
	Caucasian	African_American	Asian	Hispanic	Native_American
Accuracy	0.66	0.67	0.74	0.68	0.64
Accuracy Parity	1.00	1.02	1.13	1.04	0.97

	Other
Accuracy	0.69
Accuracy Parity	1.05



Taux de faux négatif

	Caucasian	African_American	Asian	Hispanic	Native_American	Other
FNR	0.53	0.25	0.75	0.53	0.40	0.58
FNR Parity	1.00	0.47	1.42	1.01	0.76	1.10



Taux de faux positifs

	Caucasian	African_American	Asian	Hispanic	Native_American	Other
FPR	0.22	0.42	0.09	0.19	0.33	0.15
FPR Parity	1.00	1.87	0.39	0.86	1.50	0.68