Toepassingen van machine learning op registerdata om meer te leren over mobiliteit en armoede

Joep Burger (j.burger@cbs.nl)
Centraal Bureau voor de Statistiek, Team Methodologie Heerlen

Workshop datascience en machine learning voor openbare statistieken
Dag van het netwerk Statistiek Vlaanderen, Brussel, 21 maart 2024

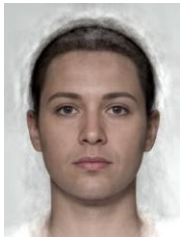
Inhoud

- Algemeen
  - statistische vs. algoritmische modellen
- Toepassing armoede
  - modelkwaliteit
  - uitlegbaarheid
  - deelpopulaties
- Overige toepassingen
- Centrale vragen

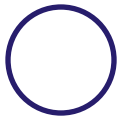
Algemeen

Model

- Vereenvoudiging van de werkelijkheid
  - Generaliseren
  - Signaal van ruis scheiden
  - Zo eenvoudig mogelijk maar niet eenvoudiger
  - Balans tussen fit en parsimonie



faceresearch.org



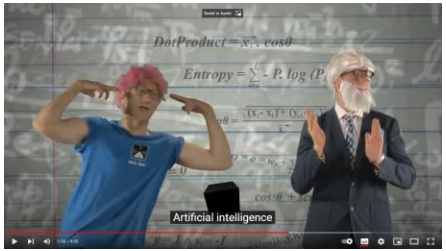
parsimonie →
← fit

Statistisch vs algoritmisch model

	Statistisch model	**Algoritmisch model**
Focus	Verklaren	Voorspellen
Aanname dataverdeling	Ja	Nee
Modelleer datagenererend mechanisme	Ja	Nee
Schaalbaar	Nee	Ja
Voorbeeld	$Y \sim \text{Binom}(n, p)$ $\text{logit}\, p = X\beta$	PageRank, DBSCAN, neural network

Breiman (2001)

[Rap battle data scientist vs statistician (2019)]



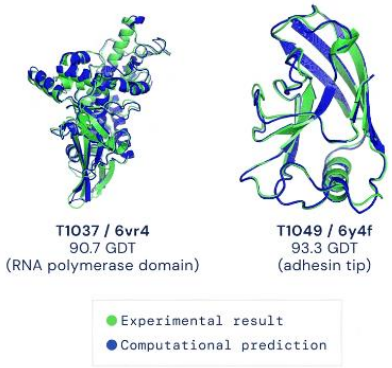
Toegenomen interesse in algoritmische modellen

- Nieuwe databronnen
  - websites
  - scanners
  - sensoren
- Niet-tabulaire data
  - tekst
  - afbeeldingen
- Open-source software
- Hardware
- Performance

caption → image
(DALL-E)

“a snail made of harp”

aminozuurvolgorde →
3D eiwitstructuur
(AlphaFold)



Toepassing machine learning officiële statistiek

- Informatie mijnen uit Stelsel van Sociaal-statistische Bestanden
  - Alle personen in Nederland
  - Demografie, sociaaleconomische status, criminaliteit, gezondheid, etc.

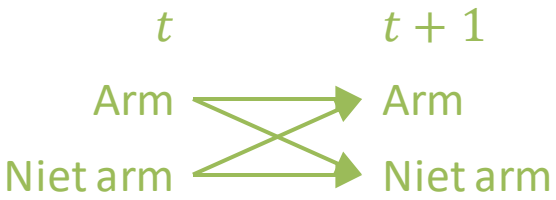
Toepassing armoede

Introductie

- UN Sustainable Development Goals (2015)
  - End poverty in all its forms everywhere by 2030
- Gemeentes: armoedepreventie
- Bekende risicofactoren voor transities in en uit armoede
  - Opleidingsniveau
  - Huishoudsamenstelling
  - Werkloosheidspercentage
  - Toegang tot voorzieningen
- Nieuwe inzichten door toepassen ML op registerdata?
  - Veel kenmerken tegelijk
  - Niet-lineaire verbanden en complexe interacties
- Verzoek ministerie Binnenlandse Zaken (BZK)

Data

- Doelpopulatie: volwassenen in particuliere huishoudens in NL op 31 dec met huishoudinkomen en –vermogen bekend in jaar t en $t+1$ (exclusief minderjarigen, institutionele huishoudens, studentenhuishoudens, migranten)
  - Arm in jaar t (0.6 mln)
  - Niet arm in jaar t (12 mln)
- Doelvariabele: arm in jaar $t+1$
- Definitie van arm
  - Gestandaardiseerd besteedbaar huishoudinkomen < lage-inkomensgrens
  - EN gestandaardiseerd vrij opneembaar vermogen < 0,5 lage-inkomensgrens
- $t = 2013,2017$



Data (vervolg)

- Stelsel van Sociaal-Statistische Bestanden (SSB)
- 500+ kenmerken
  - Persoon, huishouden~~, regio~~
  - Inkomen, uitgaven
  - Demografisch, socio-economisch, gezondheid, kinderwelzijn, criminaliteit
  - Huidig, afgelopen 3 jaar (life events)

Methode

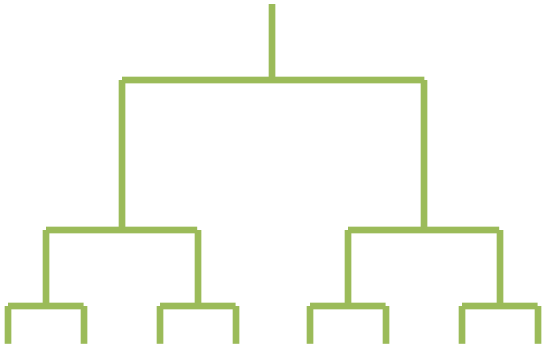
- Leer relatie tussen kenmerken en label
  - Arm: blijft arm (71 %)
  - Niet arm: wordt arm (1 %)
- eXtreme Gradient Boosting
  - Beslisboomachtig machine learning algorithme
  - Opeenvolgende bomen worden getraind op residuen van voorspellingen
  - Schaalbaar naar veel kenmerken
  - Niet-lineaire verbanden
  - Complexe interacties
  - Ontbrekende waardes
  - Multicollineariteit
- Geneste kruisvalidatie
  - Trainingset
  - Validatieset
  - Testset

Voorspelkwaliteit

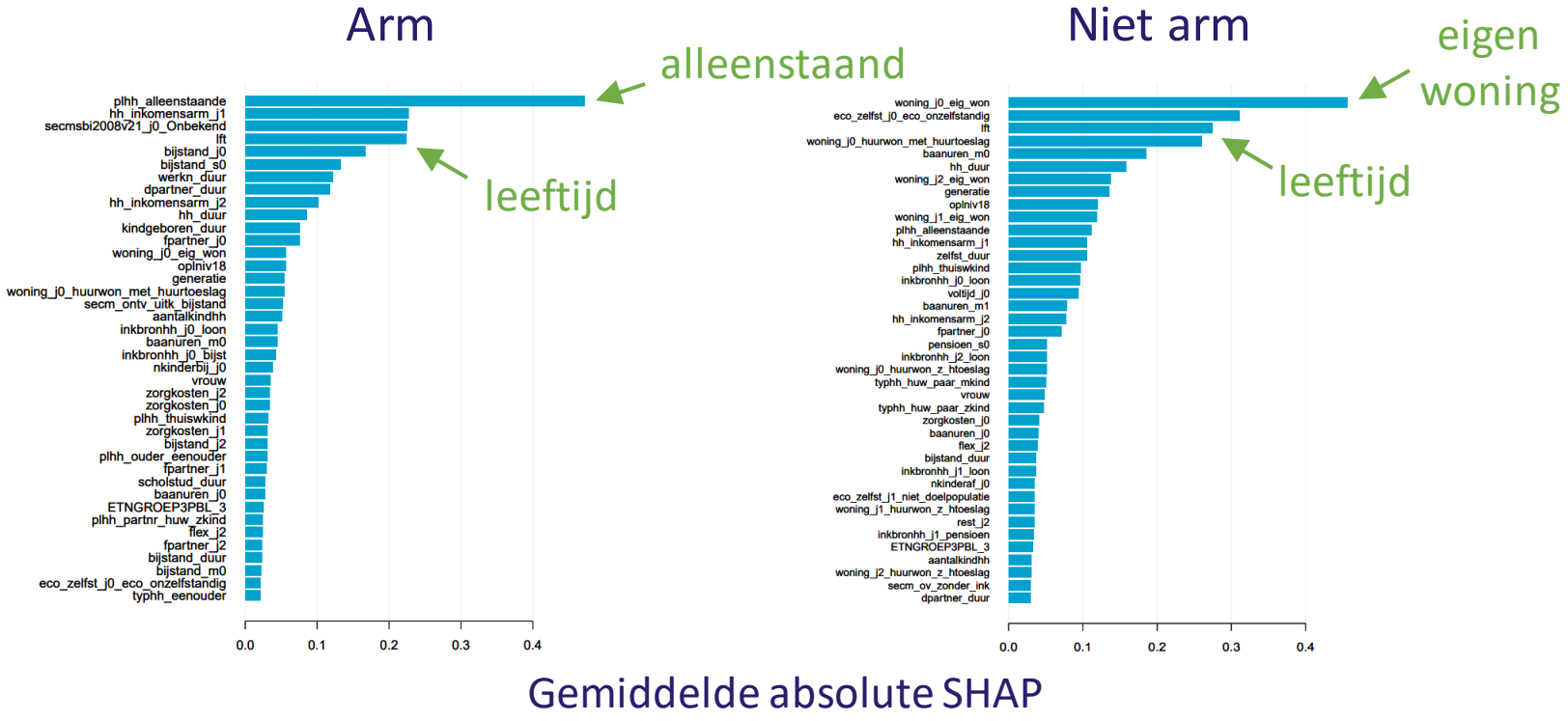
	Arm	Niet arm
AUC^{mmn}	0.67	0.90
$MCC(c^*)$	0.48	0.42
$F_1^{+,\text{mmn}}(c^*)$	0.53	0.42

Uitlegbaarheid

- SHAP-waarde
  - Relatieve belang van score van persoon i op kenmerk j voor geschatte kans om arm te blijven/worden
  - Positieve SHAP-waarde: potentiële risicofactor
  - Negatieve SHAP-waarde: potentieel beschermende factor
- Deelpopulaties: eenvoudige beslisboom op (logit van) geschatte kansen

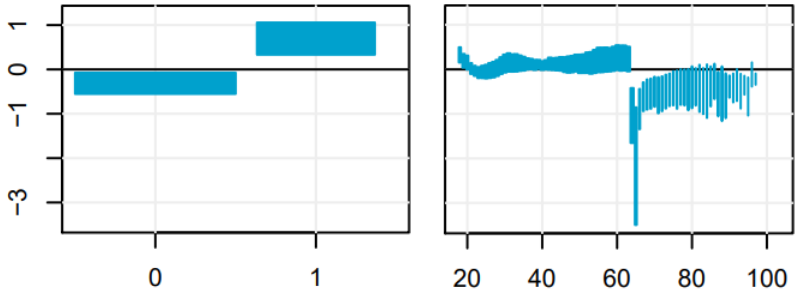


Belang van kenmerken



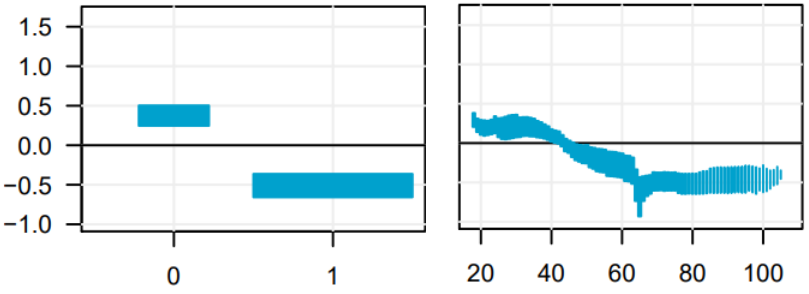
SHAP

Arm



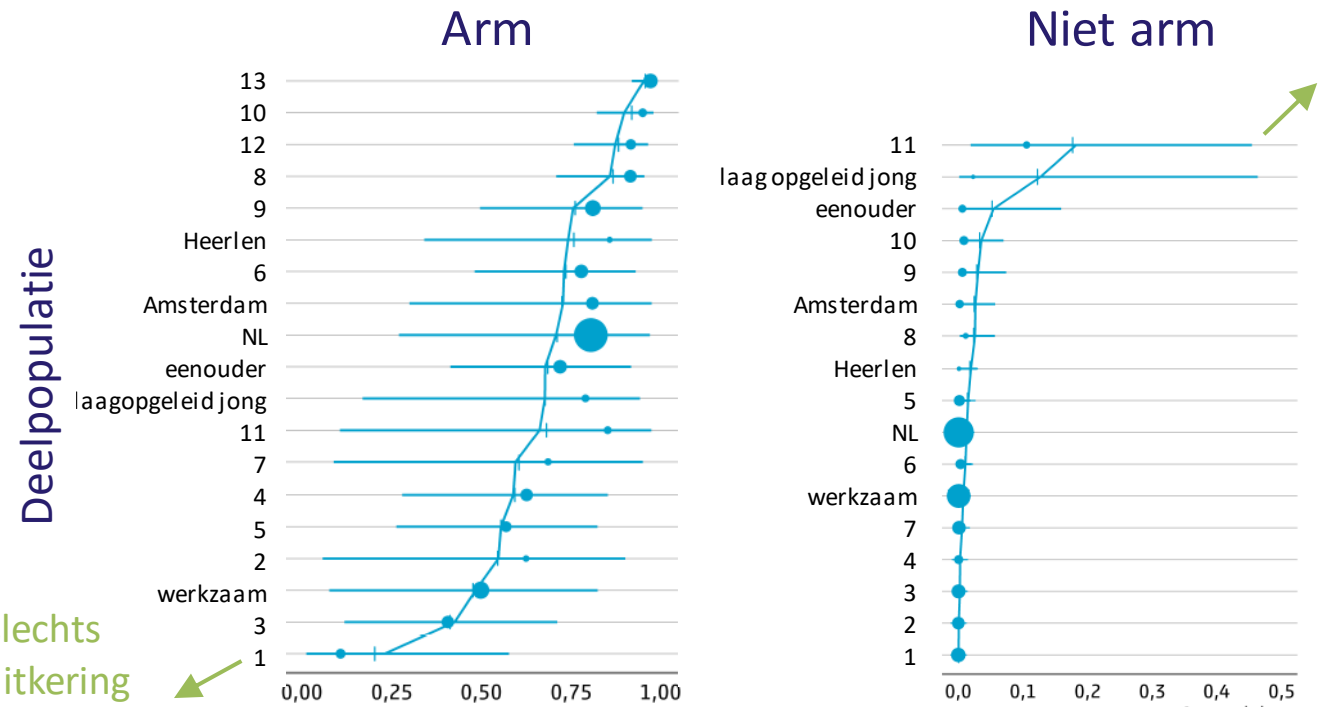
alleenstaand leeftijd

Niet arm



eigen woning leeftijd

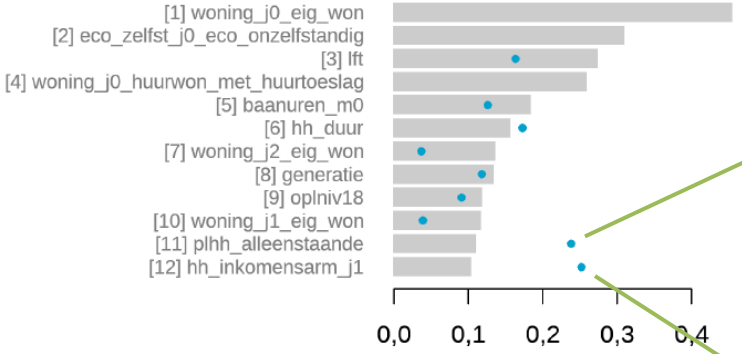
Deelpopulaties



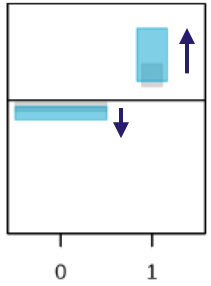
- ontvangt huurtoeslag
- economisch afhankelijk

- ontving slechts tijdelijk uitkering
- recente partner

SHAP decompositie

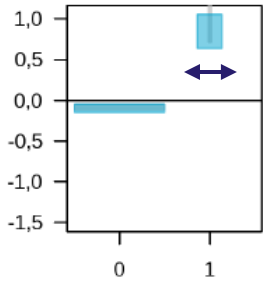


Gemiddelde absolute SHAP



alleenstaand

kenmerkeffect



laag inkomen $t-1$

frequentie-effect

Niet arm, deelpopulatie 11

Discussie

- Belangrijk kenmerk ≠ causaal effect (bv. herkomst, huurtoeslag)
- Belang van kenmerk gevoelig voor methode
- Geschikt om potentiele risicofactoren en beschermen factoren te identificeren, niet om individuen te profileren

Conclusies

- Modelvoorspellingen redelijk goed
- Bijdrage
  - Op belang rangschikken van potentiële risicofactoren
  - Gedetailleerde niet-lineaire verbanden
  - SHAP decompositie
  - Effect van *life events* in recente verleden
- Gebruik met verstand

Overige toepassingen

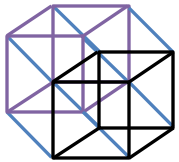
Vergelijkbare onderzoeken

- Mobiliteit
  - Verhuiskans (2018)
  - Vacantkans (lopend)
- Armoede
  - Kinderarmoede (lopend)
- In potentie vele andere thema's

Centrale vragen

Hoe inzetten voor bestaande en nieuwe statistieken?

- Automatiseren van taken
  - coderingen (ISCO, ISCED, NACE, COICOP, …)
  - modelselectie
- Verwerken tekst- en beelddata
  - tekstvelden, websites, sociale media
  - lucht-, satelliet-, straatfoto's
- Omgaan met hoogdimensionale ruimte



- Detecteren outliers
- Weging
- Imputatie
- Inference

Uitdagingen/valkuilen/opportuniteiten?

- *Fairness, accountability, safety, transparency (FAST) principles*
- Relatie gelabelde data en doelpopulatie (DGM)
- Onderscheid *prediction accuracy*, *distribution accuracy*, *estimation accuracy*, *imputation plausibility*
- Hiërarchische kenmerken
- Zeldzame categorieën
- Voorbij classificatie/regressie (ordinale labels, fracties)
- Causaliteit

- Kind niet met badwater weggooien

Gebruik van machine learning in officiële statistiek

- Aanvullend gereedschap
- Voor- en nadelen
- Statistische bril op

https://openai.com/dall-e-2

Meer info

Delden et al. (2023) Ten propositions on machine learning in official statistics. AStA, https://doi.org/10.1007/s11943-023-00330-0
https://www.cbs.nl/-/media/innovatie/rapport-armoede.pdf
https://www.cbs.nl/-/media/cbs/over-ons/innovatie/risicofactoren_armoede.pdf
https://dashboards.cbs.nl/v5/risicofactoren_transitie_armoede/
https://www.cbs.nl/nl-nl/over-ons/onderzoek-en-innovatie/project/verhuiskans-schatten-met-machine-learning

Kwaliteitsmaten

- Oppervlakte onder de ROC-curve (AUC)
- Matthews correlatiecoëfficiënt (MCC)
- Harmonisch gemiddelde van *recall* en *precision* van positieve categorie (F_1^+)

- Een persoon wordt geclassificeerd als ‘arm’ als geschatte kans boven drempelwaarde c ligt, welke geoptimaliseerd wordt (c^*) per maat

- Min-max genormaliseerd (mmn) met als minimum de score bij gokken met waargenomen armoedefractie

Quality metrics

Confusion matrix given cutoff c: $\hat{y}_i = \begin{cases} 1 & \text{if } \hat{p}_i > c \\ 0 & \text{otherwise} \end{cases}$

		Predicted			
		Poor	Not poor	Sum	
Observed	Poor	T^+	F^-	n^+	$R^+ = \frac{T^+}{n^+}$
	Not Poor	F^+	T^-	n^-	$R^- = \frac{T^-}{n^-}$
	Sum	m^+	m^-	n	
		$V^+ = \frac{T^+}{m^+}$	$V^- = \frac{T^-}{m^-}$	$\alpha = \frac{n^+}{n}$	

$$AUC = \int_{c=0}^{1} R_c^+ dR_c^-$$

$$ACC = \frac{T^+ + T^-}{n}$$

$$MCC = \frac{T^+T^- - F^+F^-}{\sqrt{n^+n^-m^+m^-}}$$

$$GEO = \sqrt{R^+R^-}$$

$$PSS = R^+ + R^- - 1$$

$$MRK = V^+ + V^- - 1$$

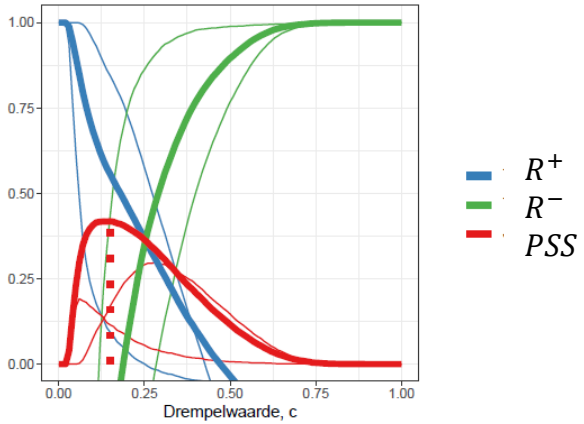
$$F_1^+ = \frac{2}{\frac{1}{R^+} + \frac{1}{V^+}}$$

$$F_1^- = \frac{2}{\frac{1}{R^-} + \frac{1}{V^-}}$$

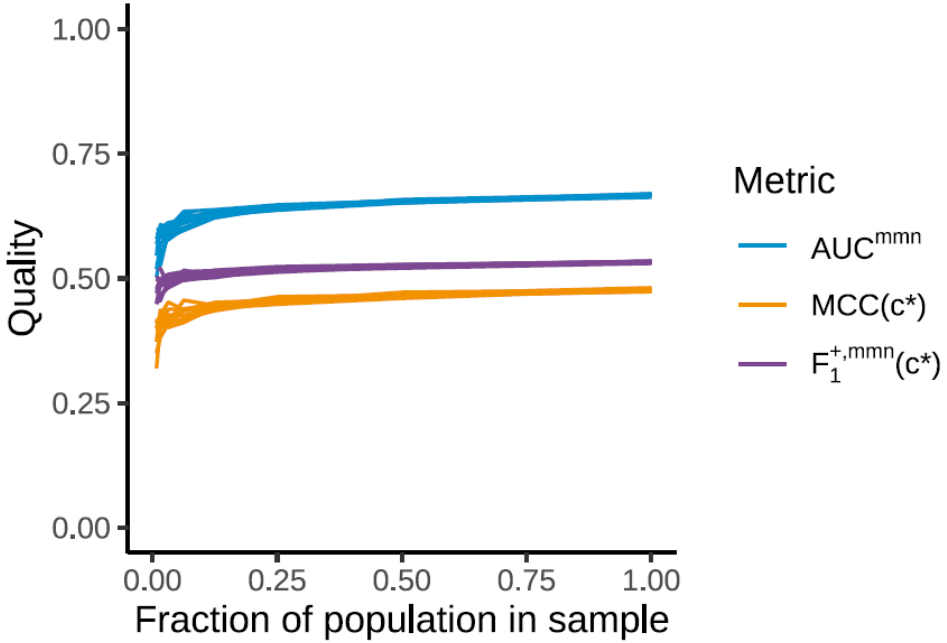
- orange: increases with c, blue: decreases with c
- Min-max normalization using as minimum the score when guessing with the observed fraction α

Trade-offs

- T^+T^- vs F^+F^- (MCC)
- R^+ vs R^- (balanced ACC, AUC, PSS, GEO)
- R^+ vs V^+ (F1, AP)

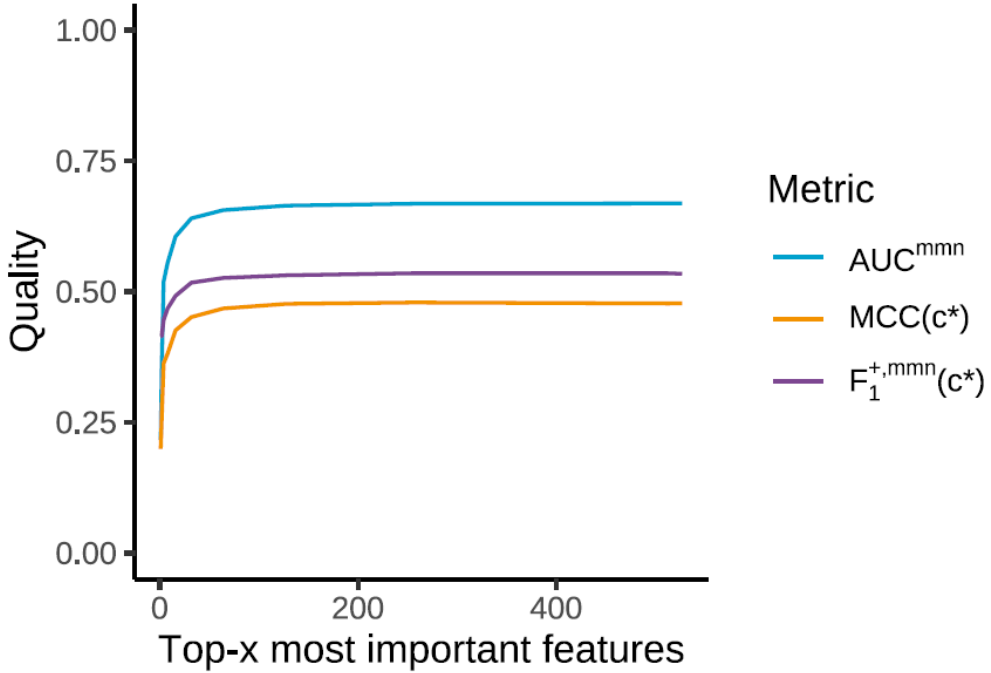


Effect of sample size on model performance



Poor population, $K = 10$ testsets

Adding features



Poor population, testset $k = 1$

SHAP

$$SHAP_{ij} = \sum_{s: j \in s} w_s C_{is}$$

$$C_{is} = f(\mathbf{x}_{i,s}) - f(\mathbf{x}_{i,s \setminus j})$$

$$w_s = \frac{1}{|s| \binom{p}{|s|}}$$

$$\mathbf{X} = \begin{pmatrix} \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & x_{i3} & x_{i4} \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

$|s| = 0$ features

$w_1 = \frac{1}{1 \cdot 4}$

$|s| = 1$ feature

$w_2 = \frac{1}{2 \cdot 6}$

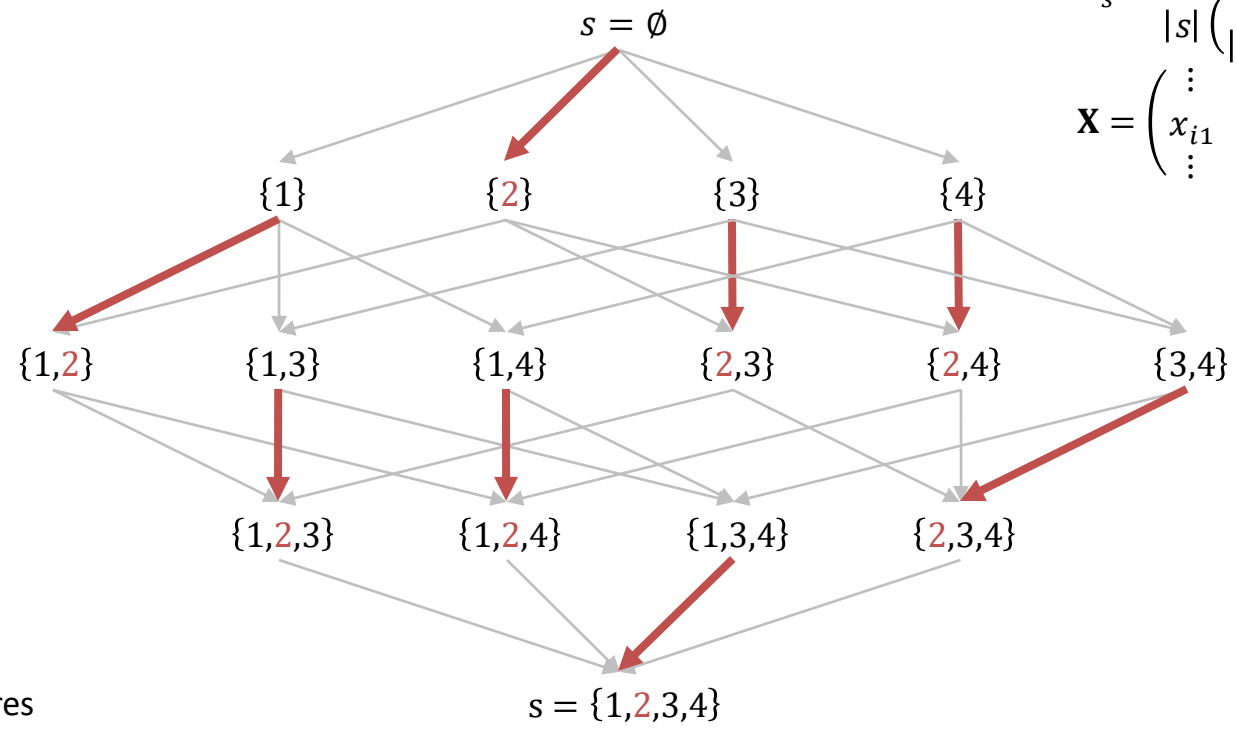
$|s| = 2$ features

$w_3 = \frac{1}{3 \cdot 4}$

$|s| = 3$ features

$w_4 = \frac{1}{4 \cdot 1}$

$|s| = p = 4$ features

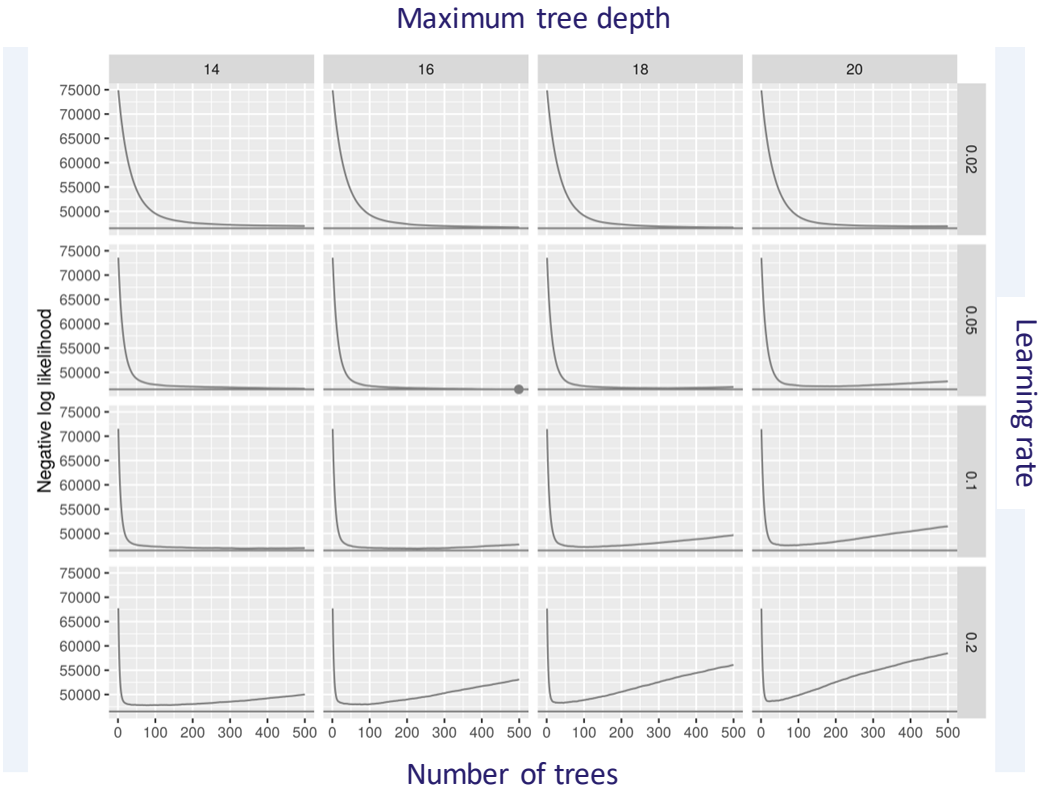


SHAP decomposition

$$
\begin{aligned}
\Delta I &= I^{(h)} - I \\
&= \sum_g f_g^{(h)} S_g^{(h)} - \sum_g f_g S_g \\
&= \sum_g f_g^{(h)} S_g^{(h)} - \sum_g f_g^{(h)} S_g + \sum_g f_g^{(h)} S_g - \sum_g f_g S_g \\
&= \sum_g f_g^{(h)} \left(S_g^{(h)} - S_g \right) + \sum_g S_g \left(f_g^{(h)} - f_g \right) \\
&= \Delta I^{\mathrm{s}} + \Delta I^{\mathrm{f}}
\end{aligned}
$$

$$
\begin{aligned}
I &= \sum_g f_g S_g \\
f_g &= \frac{n_g}{n} \\
S_g &= \frac{1}{n_g} \sum_{i:x==g} |SHAP_i|
\end{aligned}
$$

Hyperparameters



Poor population, testset $k = 1$