Van de EBB naar integrale data

Baankenmerken voorspellen met machine learning

seo • economisch onderzoek

AMSTERDAM

April 24

Auteurs

Menno Pomp & William Luiten

In opdracht van

Bas ter Weel

DOEL VAN HET PROJECT

- Voor bepaalde (baan)kenmerken zijn we afhankelijk van enquêtes.
- Dit zorgt voor moeilijkheden als we specifieke subsamples onderzoeken.
  - De N wordt te laag.

- Is het mogelijk om de kenmerken te voorspellen met machine learning technieken?
- De kenmerken die we in deze proef achterhalen:
  - In ploegendienst werken
  - Avond- en of nachtdiensten werken
  - Thuiswerken

- In dit onderzoek achterhalen we hoe bruikbaar deze methode is, en welke factoren de kans op succes vergroten en welke niet.

VOOR WELKE ONDERZOEKEN IS DEZE TECHNIEK MOGELIJK INTERESSANT?

- Evaluatie leeftijdsverhoging AOW -> verschillen tussen leeftijd 67 en 68 bijvoorbeeld.
- Middellange termijn effecten COVID -> kwetsbare groepen
- Arbeidsmarktonderzoek naar arbeids/kennismigranten
- Arbeidsmarkteffecten van ziektes

METHODE

- We gebruiken de EBB om onze data te trainen
- We voorspellen de uitkomstkenmerken aan de hand van gegevens uit de SPOLISBUS, GBA en opleidingsgegevens
- De jaren 2021/2022/2023 zijn gebruikt om de modellen te trainen
  - We trainen het model op 64% van de data (2021/2022/2023)
  - Evalueren en hyperparameter tuning op 16% van de 2022 data
  - Uiteindelijk getest op 20% van de 2022 data.
- Verschillende modellen geschat om te kijken wat het beste werkt:
  - Logit (basis model)
  - Gradient Boost
  - Extreme Gradient Boost

HOE BEPALEN WE SUCCES?

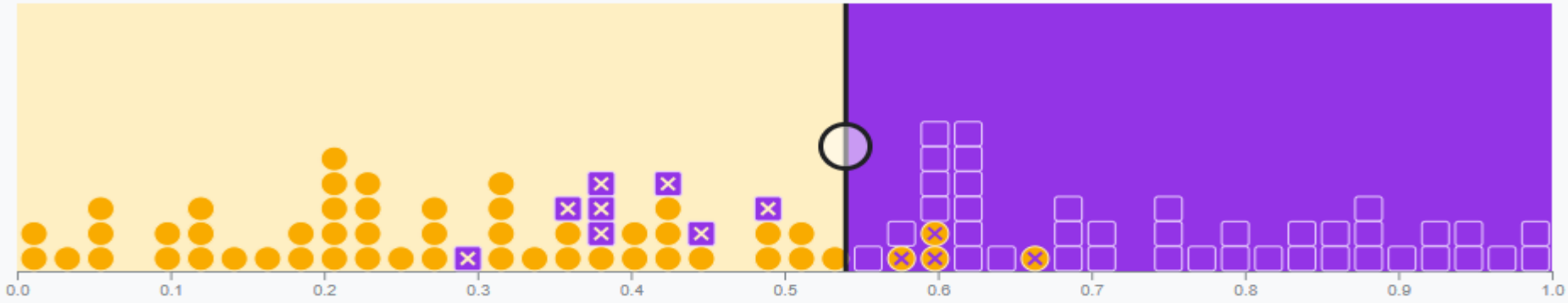
- Wanneer willen een groep identificeren die voornamelijk bestaat uit mensen die ook daadwerkelijk het kenmerk hebben.
  - Precision > 90%
  - Als het model zegt dat 10 mensen in ploegendienst werken, dat minstens 9 daarvan daadwerkelijk in ploegendienst werken -> 9/10 = 90%.

- De voorspelde groep moet minstens 10% bevatten van de degenen die het kenmerk werkelijk hebben
  - Recall > 10%
  - Als er in totaal 10 mensen in ploegendienst werken, en het model vindt er 1, dan is de recall 1/10 = 10%.
  - Balance table bekijken of de kenmerken overeenkomen.

- En de groep moet groot genoeg zijn om statistisch iets mee te kunnen
  - Minstens 2.000 mensen met het kenmerk in het analyse bestand

VOORBEELD: KIEZEN DREMPEL

Classification threshold



Threshold: 0.54

Confusion matrix

	Actually positive	Actually negative
Predicted positive	TP=40	FP=4
Predicted negative	FN=8	TN=47

Metrics

Accuracy	0.88
Precision	0.91
Recall	0.83

WANNEER IS DE METHODE EEN VERBETERING? EEN VOORBEELD

- Uit het model komen de volgende resultaten
  - De enquete bestaat uit 2.000 mensen, waarvan er 1.000 zwaar werk hebben

Threshold	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
Predicted Positives	2.000	1.552	1.212	972	779	617	476	345	222	108	31
True positives	1.000	900	800	700	600	500	400	300	200	100	30
Precision	50%	58%	66%	72%	77%	81%	84%	87%	90%	93%	96%
Recall	100%	90%	80%	70%	60%	50%	40%	30%	20%	10%	3%

- Stel je voor 20% in Nederland heeft een zwaar beroep.
  - Via de EBB moet de subgroep minimaal 17% van de Nederlandse bevolking zijn om 2.000 obs over te houden.
  - Onze methode werkt voor een subpopulatie van 1%

	Populatie	Aantal met kenmerk	Aantal true positives	Min subpopulatie
EBB	60.000	12,000	12.000	17%
Onze fictieve voorspelling	8.000.000	1.600.000	320.000	1%

BESCHRIJVENDE STATISTIEK

- De vragen in de EBB zijn:
  - Werkt u in ploegendienst of wisseldienst?
  - Werkt u weleens 's avonds/'s nachts?
  - Werkt u weleens thuis?

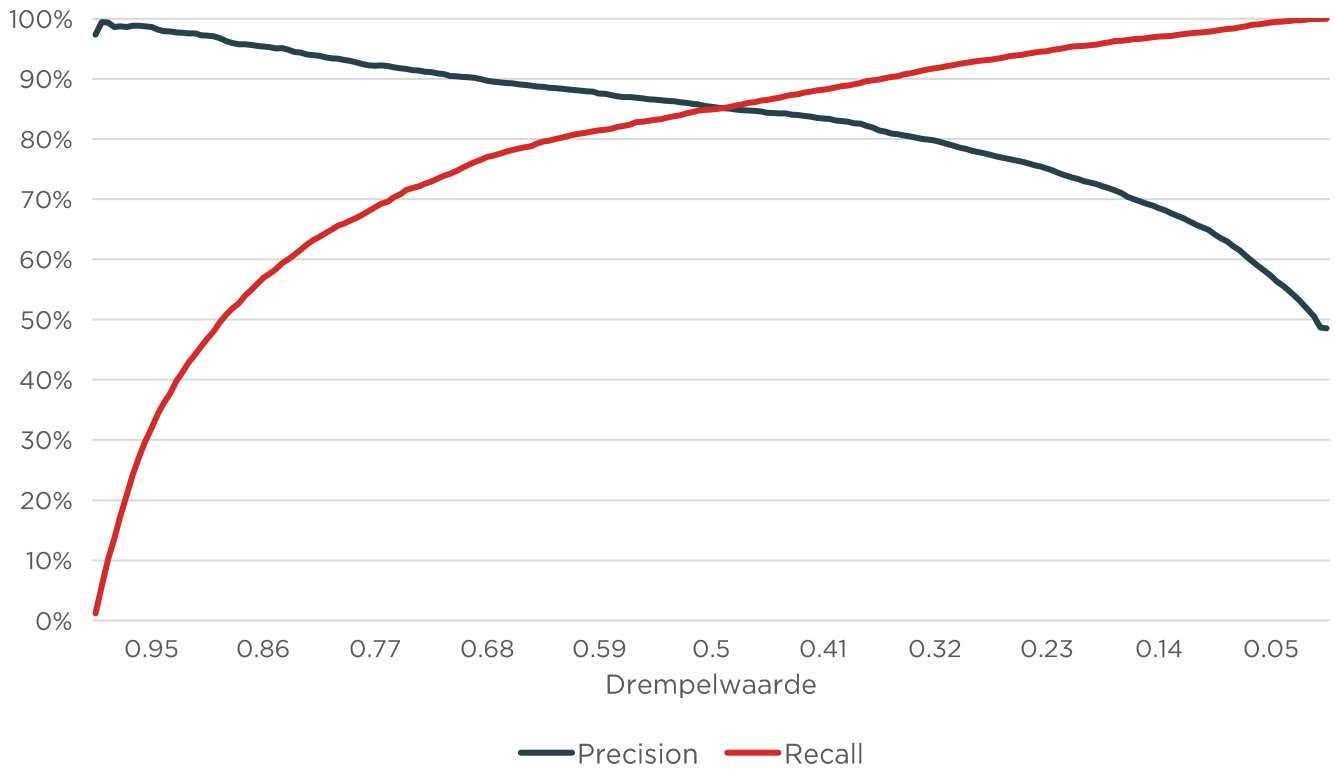
	Ploegendienst	Thuiswerken	Avond/nachtdiensten
Ja (soms/altijd)	16%	51%	63%
Nee	84%	49%	37%
N	54.368	63.900	63.348

- Ploegendienst komt weinig voor. Dit maakt voorspellen moeilijker.

RESULTATEN: THUISWERKEN

- Thuiswerken is zeer goed te voorspellen
- De Extreme Gradient Boost voorspelt thuiswerken het best
- Opleidingsgegevens verbeteren het model niet echt

	Precision	Recall
Zonder opleidingsinformatie		
Logit	90%	60%
GBM	90%	65%
XGB	90%	76%
Met opleidingsinformatie		
Logit	90%	66%
GBM	90%	70%
XGB	90%	77%



BALANCETABLE: THUISWERKEN

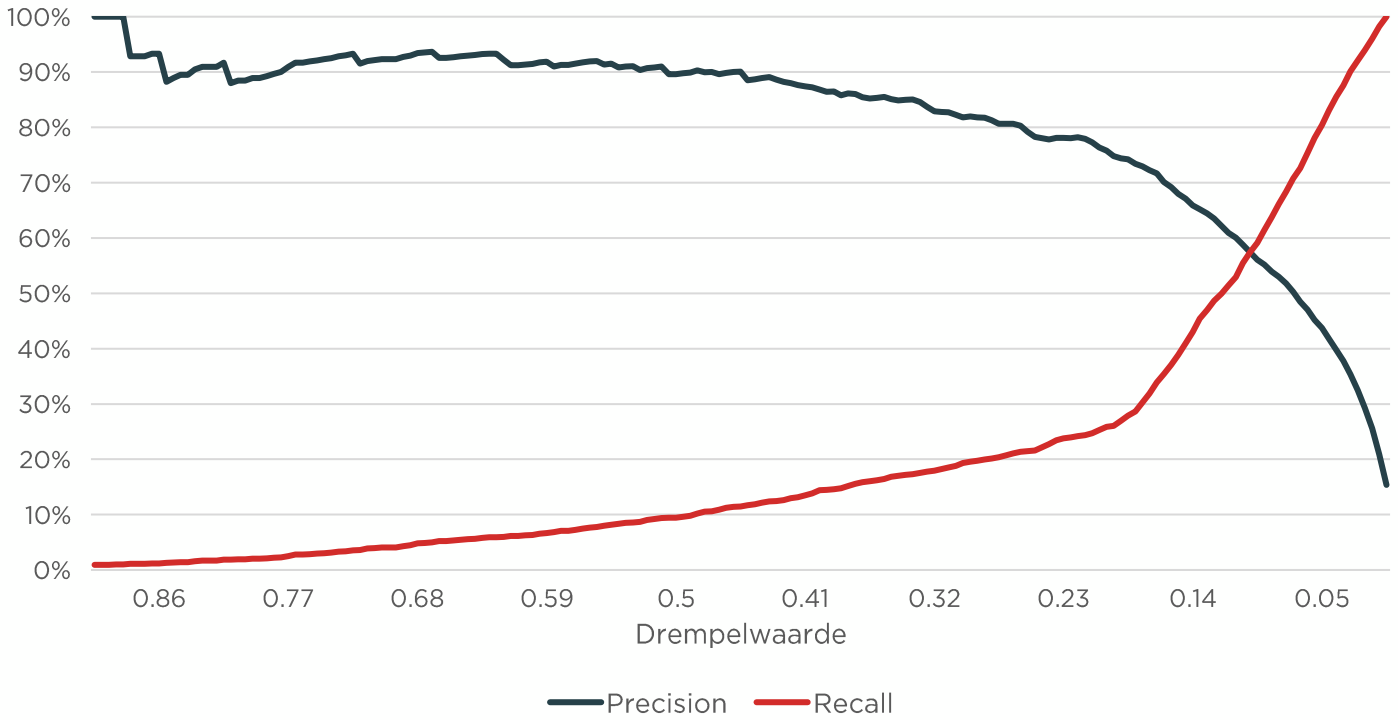
- Over het algemeen is onze voorspelde groep een goede afspiegeling van alle thuiswerkers
- Wel vaker hoog opgeleiden dan middelbaar opgeleiden
  - Komt waarschijnlijk door de sterke verklaringskracht van loon

	Voorspeld	EBB	Verschil
Man	52%	50%	2%
Vrouw	48%	50%	-2%
20 jaar of jonger	0%	1%	-1%
20 tot 25 jaar	4%	7%	-3%
25 tot 30 jaar	12%	12%	-1%
30 tot 35 jaar	15%	15%	0%
35 tot 40 jaar	15%	13%	1%
40 tot 45 jaar	15%	14%	1%
45 tot 50 jaar	11%	10%	1%
50 tot 55 jaar	12%	11%	1%
55 tot 60 jaar	11%	10%	0%
60 tot 65 jaar	6%	6%	0%
65 tot 70 jaar	1%	1%	0%
Nederland	83%	82%	1%
Europa (excl	4%	4%	0%
Buiten Europa	13%	14%	-1%
Laag	1%	3%	-2%
Midden	16%	27%	-11%
Hoog	82%	70%	12%

RESULTATEN: PLOEGENDIENST

- Ploegendienst is te voorspellen, maar wel voor een beperkte groep
- Extreme Gradient Boost weer het beste model

	Precision	Recall
Zonder opleidingsinformatie		
Logit	52%	15%
GBM	69%	10%
XGB	90%	11%
Met opleidingsinformatie		
Logit	58%	18%
GBM	71%	11%
XGB	86%	12%



BALANCETABLE: PLOEGENDIENST

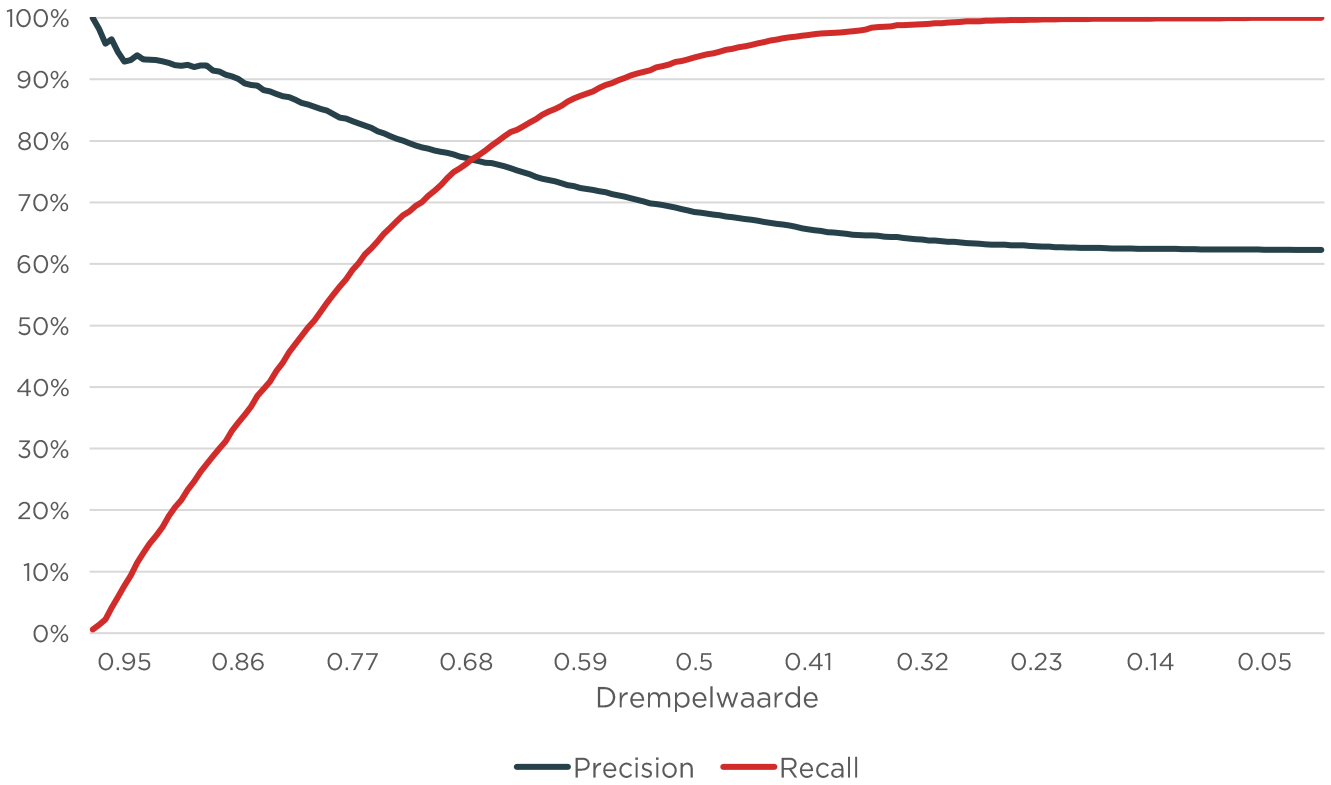
- Jongeren worden niet goed voorspeld door het model
- Ouderen juist te vaak aangezien voor mensen die werken in ploegendienst
- Middelbaar opgeleiden ook te vaak aangeduid als mensen met ploegendienst

	Voorspeld	EBB	Verschil
Man	50%	46%	4%
Vrouw	50%	54%	-4%
20 jaar of jonger	1%	24%	-23%
20 tot 25 jaar	14%	19%	-5%
25 tot 30 jaar	16%	11%	5%
30 tot 35 jaar	16%	9%	8%
35 tot 40 jaar	6%	7%	-1%
40 tot 45 jaar	11%	6%	5%
45 tot 50 jaar	10%	5%	5%
50 tot 55 jaar	7%	5%	2%
55 tot 60 jaar	12%	6%	6%
60 tot 65 jaar	8%	6%	2%
65 tot 70 jaar	0%	1%	-1%
Nederland	75%	77%	-2%
Europa (excl	4%	5%	-2%
Buiten Europa	21%	18%	4%
Laag	5%	24%	-19%
Midden	74%	58%	16%
Hoog	20%	18%	2%

RESULTATEN: AVOND/NACHTWERK

- Avond/nachtwerk goed te voorspellen
- Opleidingsniveau voegt iets toe, maar niet veel

	Precision	Recall
Zonder opleidingsinformatie		
Logit	87%	12%
GBM	89%	10%
XGB	90%	32%
Met opleidingsinformatie		
Logit	85%	15%
GBM	90%	11%
XGB	90%	34%



BALANCETABLE: AVOND/NACHTDIENSTEN

- Jongeren te vaak aangezien voor mensen die werken in ploegendienst
- Vrouwen ook minder vaak voorspeld als iemand die avond/nachtdienst werkt

	Voorspeld	EBB	Verschil
Man	57%	51%	6%
Vrouw	43%	49%	-6%
20 jaar of jonger	29%	14%	15%
20 tot 25 jaar	13%	13%	1%
25 tot 30 jaar	7%	11%	-4%
30 tot 35 jaar	8%	12%	-4%
35 tot 40 jaar	7%	10%	-3%
40 tot 45 jaar	9%	10%	-1%
45 tot 50 jaar	7%	8%	0%
50 tot 55 jaar	9%	8%	0%
55 tot 60 jaar	7%	8%	-1%
60 tot 65 jaar	4%	6%	-2%
65 tot 70 jaar	0%	1%	-1%
Nederland	82%	81%	1%
Europa (excl	5%	5%	0%
Buiten Europa	13%	14%	-1%
Laag	22%	16%	6%
Midden	39%	38%	1%
Hoog	39%	46%	-7%

CONCLUSIE

- Thuiswerken kunnen we heel goed voorspellen
- Avond/nachtdiensten voorspellen gaat ook goed
- Ploegendienst is goed te voorspellen, maar omdat het relatief weinig voorkomt is de voorspelde groep voor specifieke groepen niet representatief

- Opleidingsinformatie toevoegen heeft minder effect dan gehoopt
- Meerdere jaren mensen volgen lastig, omdat CBS constant de bestanden verbetert
  - Dus verandering in variabelen
  - Standaardset van POLIS variabelen gebruiken kan een oplossing zijn

AANBEVELINGEN

- We hebben het getest voor 3 kenmerken, maar in principe kan elke enquêtevraag gebruikt worden
- Wat verhoogt de succeskans?
  - De groep is een aanzienlijk aandeel van de Nederlandse bevolking?
  - Zijn er goede voorspellers te bedenken (zoals sector en inkomen in het geval van thuiswerken)?
  - Is enquete groot genoeg?
  - Naar welke jaren wil je kijken?
- Voor SEO echt nuttig als we het nog wat complexer maken en we beroepen kunnen voorspellen

ONZE EXPERTISE

SEO heeft onderstaande expertise gebieden

Our life is what our thoughts make it