



Rapportage

Risicofactoren voor transities in en uit armoede

Jan van der Laan
Joep Burger
Marijn Detiger
Naomi Schalken
Wouter van Andel
Marion van den Brakel
Sita Tan

Juli 2021

Colofon

Uitgever

Centraal Bureau voor de Statistiek
Henri Faasdreef 312, 2492 JP Den Haag
www.cbs.nl

Prepress

Centraal Bureau voor de Statistiek

Ontwerp

Edenspiekermann

Inlichtingen

Tel. 088 570 70 70
Via contactformulier: www.cbs.nl/infoservice

© Centraal Bureau voor de Statistiek, Den Haag/Heerlen/Bonaire, 2021.
Verveelvoudigen is toegestaan, mits het CBS als bron wordt vermeld.

Inhoudsopgave

Samenvatting 5

1 Inleiding 6

2 Data 7

- 2.1 Definitie armoede 7
- 2.2 Doelvariabele 7
- 2.3 Doelpopulaties 7
- 2.4 Kenmerken 8

3 Methoden 10

- 3.1 XGBoost 10
- 3.2 Schatten van het machine-learningmodel 11
- 3.3 Interpretatie van het model 13
- 3.4 Software 19

4 Resultaten 20

- 4.1 Inleiding 20
- 4.2 Arme populatie - geheel Nederland 22
- 4.3 Arme populatie - deelpopulatie met grootste kans om uit armoede te komen 32
- 4.4 Niet-arme populatie - geheel Nederland 40
- 4.5 Niet-arme populatie - deelpopulatie met grootste kans om in de armoede te komen 48

5 Conclusies en discussie 56

- 5.1 Conclusies 56
- 5.2 Discussie 57
- 5.3 Aanknopingspunten voor beleid 59

Afkortingen 60

Bijlagen 61

A Operationalisering variabelen 61

B Hyperparameter tuning 69

- B.1 Model arme populatie 69
- B.2 Model niet-arme populatie 69

C Modelkwaliteit 72

D Decompositie SHAP 75

E Deelpopulaties 77

- E.1 Arme populatie - Amsterdam 77
- E.2 Arme populatie - Heerlen 83
- E.3 Arme populatie - laagopgeleide jongeren 90
- E.4 Arme populatie - werkenden 98
- E.5 Arme populatie - alleenstaande ouders 105
- E.6 Niet-arme populatie - Amsterdam 112
- E.7 Niet-arme populatie - Heerlen 119
- E.8 Niet-arme populatie - laagopgeleide jongeren 125
- E.9 Niet-arme populatie - werkenden 132
- E.10 Niet-arme populatie - alleenstaande ouders 139

Referenties 147

Samenvatting

In Nederland leefde in 2019 ruim 6 procent van de inwoners onder de lage-inkomensgrens. Met het oog op effectief armoedebelied heeft het Ministerie van Binnenlandse Zaken en Koninkrijksrelaties (BZK) het Centraal Bureau voor de Statistiek (CBS) verzocht onderzoek te doen naar risicofactoren voor armoede. De twee onderzoeksvragen waren: hoe goed kunnen we op basis van (combinaties van) kenmerken de kans schatten dat iemand arm blijft dan wel arm wordt, en voor welke aspecten kan de combinatie van populatiedekkende registerdata en machine learning tot nieuwe inzichten over risicofactoren voor transitie in en uit armoede leiden?

Om deze vragen te beantwoorden hebben we met behulp van twee gradient boosting algoritmes de kans geschat dat iemand wel of niet arm blijft of wordt. Armoede is hier gedefinieerd op basis van zowel besteedbaar huishoudinkomen als vrij opneembaar vermogen. We hebben meer dan 500 potentiële risicofactoren voor armoede onderzocht, die zijn afgeleid uit registers over demografie, economie, gezondheid en criminaliteit. Daarbij hebben we gekeken naar zowel de meest recente status als gebeurtenissen in de afgelopen drie jaar. De kenmerken zijn gebruikt in twee algoritmes om de kans te schatten dat iemand die in het referentiejaar wel/niet arm is een jaar later arm is gebleven/geworden. Om de uitkomsten van deze algoritmes uitlegbaar te maken hebben we de gemiddelde bijdrage van een kenmerk aan de geschatte armoedekans van een persoon (SHAP) gebruikt. We doen dit niet alleen voor de gehele Nederlandse populatie maar zoomen ook in op de (datagedreven) deelpopulaties met de grootste kans om uit de armoede te komen en met de grootste kans om in de armoede te raken, en op de (expertgedreven) deelpopulaties Amsterdam, Heerlen, laagopgeleide jongeren, werkenden en alleenstaande ouders. Daarbij hebben we het verschil in belang van een kenmerk tussen deelpopulatie en gehele populatie ontleed in twee componenten: de verdeling van het kenmerk en het effect van het kenmerk. De verhouding tussen deze componenten geeft extra inzicht die de effectiviteit van beleid zou kunnen vergroten.

Beide modellen leveren redelijke tot goede schattingen op en presteren beduidend beter dan het gebruik van de waargenomen armoedefractie (landelijk: 71 procent blijft arm, 1 procent wordt arm). Vergeleken met het model voor de arme populatie scoort het model voor de niet-arme populatie wat hoger op onderscheidend vermogen (Somers' D) maar wat lager op de balans tussen zo veel mogelijk mensen die arm blijven/worden goed classificeren en zo min mogelijk fouten maken in het classificeren als 'blijft/wordt arm' (F1). De 500+ potentiële risicofactoren zijn gerangschikt op relatief belang. Een aantal zijn voor beide deelpopulaties belangrijk zoals alleenstaande-indicator, inkomensarmoede en leeftijd. Andere zijn vooral belangrijk voor de transitie uit armoede, zoals: recente veranderingen in bijstand en baan als werknemer; andere voor de transitie naar armoede, zoals: economische zelfstandigheid en eigen woningbezit. Verder laten de resultaten een gedetailleerd beeld zien van de relatie tussen bijvoorbeeld leeftijd of tijd sinds geboorte van een kind en de armoedekans, en de spreiding tussen mensen die hetzelfde scoren op een kenmerk. De deelpopulatie met de grootste kans om uit de armoede te komen zijn hoofdzakelijk mensen die niet het hele jaar in de bijstand zaten en recent een partner hebben gekregen. De deelpopulatie met de grootste kans om in de armoede te raken zijn mensen die economisch onzelfstandig zijn en in een huurwoning met huurtoeslag wonen. Mensen in deze deelpopulatie hebben niet alleen een hogere kans omdat de samenstelling ongunstiger is maar onder meer ook omdat het ongunstige effect van alleenstaand zijn sterker is en het gunstige effect van middelbare leeftijd zwakker dan in de gehele niet-arme populatie.

We bespreken de voor- en nadelen van de gekozen aanpak, de risico's op verkeerd gebruik en ideeën voor verbetering. We concluderen dat met registerdata en machine learning potentiële risicofactoren voor transitie in en uit armoede kunnen worden geïdentificeerd en dat dit tot nieuwe inzichten leidt.

Keywords: classificatie, gradient boosting, life event, machine learning, SHAP-decompositie, uitlegbaarheid

1 Inleiding

Iets meer dan 1 miljoen mensen leefden in 2019 in een huishouden onder de lage-inkomensgrens (6,2 procent van de Nederlandse bevolking), van wie 398 duizend (2,5 procent van de bevolking) langdurig - dat wil zeggen al ten minste vier jaar achtereen (CBS, 2020). Mogelijk kan vroegtijdige signalering van risicofactoren voor armoede via armoedepreventie en -bestrijding bijdragen aan effectief armoedebeleid.

Daarbij maakt vroegtijdige signalering van armoede deel uit van het Strategisch Actieplan voor Artificial Intelligence (AI) (EZK, 2019). Binnen *spoor 1 - Maatschappelijke en economische kansen benutten*, wil het kabinet de potentie van Artificial Intelligence (AI) voor de publieke taakuitvoering optimaal benutten, onder meer door *“het vroegtijdig signaleren van mogelijke armoede, waardoor leed en hoge kosten kunnen worden voorkomen”*. Ook worden armoede en schulden opgepakt in een door de Nederlandse AI coalitie erkend ELSA (Ethical, Legal and Societal Aspects) lab, dat ontwikkeld wordt door een consortium op initiatief van Brightlands Smart Services Campus (Brightlands Smart Services Campus, 2021). Dit ELSA lab heeft als doel om data science en AI in te zetten voor het terugdringen en voorkomen van armoede en problematische schulden in Nederland.

Naar aanleiding van de toenemende aandacht voor dit onderwerp heeft het Centraal Bureau voor de Statistiek (CBS) op verzoek van het Ministerie van Binnenlandse Zaken en Koninkrijksrelaties (BZK) een eerste versie van twee AI modellen ontwikkeld, waarbij we op basis van (combinaties van) kenmerken de kans schatten dat iemand arm blijft dan wel arm wordt. Deze modellen zijn getraind op historische data en geven inzicht in welke kenmerken tot een hogere kans op armoede leiden. De modellen hebben dus nadrukkelijk niet tot doel om op individueel niveau te schatten of iemand arm wordt (*profiling*). Als de modellen de situatie in het verleden goed schatten, geeft dit zicht op risicofactoren voor armoede. In een vervolgtraject zal onderzocht worden wat ervoor nodig is om dit model een plek te geven in het armoedebeleid van gemeentes en hoe dit verder ingebed kan worden in lopende armoede-initiatieven.

Onderzoeksvraag

Met behulp van AI heeft het CBS onderstaande onderzoeksvragen onderzocht:

1. Hoe goed kunnen we op basis van (combinaties van) kenmerken de kans schatten dat iemand arm blijft dan wel arm wordt?
2. Voor welke aspecten kan de combinatie van populatiedekkende registerdata en machine learning tot nieuwe inzichten over risicofactoren voor transitie in en uit armoede leiden?

Leeswijzer

Er zijn verschillende perspectieven op armoede, zie (Van den Brakel and Otten, 2019). Hoofdstuk 2 gaat in op de in dit onderzoek gehanteerde definitie. Daarnaast gaat dit hoofdstuk in op de doelvariabele, doelpopulaties en potentiële risicofactoren (kenmerken) die meegenomen zijn in het model. Hoofdstuk 3 beschrijft het gebruikte AI-model—eXtreme Gradient Boosting (XGBoost)—het schatten van de modellen en het interpreteren van de uitkomsten. Vervolgens geeft hoofdstuk 4 een beschrijving van de belangrijkste bevindingen. Hierbij gaan we eerst in op de resultaten voor de totale arme populatie in Nederland en zoomen we in op de deelpopulatie met de grootste kans om uit de armoede te komen. Vervolgens gaan we in op de totale niet-arme populatie in Nederland en zoomen we in op de deelpopulatie met de grootste kans om arm te worden. Tenslotte beschrijven we in hoofdstuk 5 de conclusies en (on)mogelijkheden van het onderzoek, evenals aanknopingspunten voor beleid. In dit rapport is veel aandacht voor de methodologische kant van het onderzoek. Een samenvatting van de inhoudelijke resultaten is te vinden in (CBS, 2021a).

2 Data

2.1 Definitie armoede

Het fenomeen armoede kan op verschillende manieren worden geoperationaliseerd. In dit rapport is - in lijn met de CBS definitie van inkomens- en vermogensarmoede - een financiële maat gekozen. Iemand is volgens deze definitie arm wanneer aan onderstaande twee voorwaarden is voldaan (Van den Brakel and Otten, 2019):

1. Het besteedbaar jaarinkomen exclusief huurtoeslag ligt onder de lage-inkomensgrens. Hierbij wordt het besteedbaar inkomen gestandaardiseerd, zodat het vergelijkbaar is met de koopkracht van een alleenstaande zonder kinderen. De lage-inkomensgrens is een vast koopkrachtbedrag, namelijk de bijstandsuitkering van een alleenstaande in 1979 gecorrigeerd voor inflatie. Daarmee is de lage-inkomensgrens vergelijkbaar over de tijd, want onafhankelijk van beleidskeuzes.
2. Het liquide (vrij opneembare) vermogen (stand op op 1 januari) ligt onder het vermogensplafond. Hierbij is het liquide vermogen het vermogen exclusief de waarde van het eigen huis, de hypotheekschuld, ondernemingsvermogen en aanmerkelijk belang. Analooq aan het besteedbaar inkomen wordt ook het liquide vermogen gestandaardiseerd. Verder is het vermogensplafond de helft van de lage-inkomensgrens, wat ongeveer gelijk is aan het maximale vrijstellingsbedrag van eigen vermogen volgens de Bijstandswetgeving.

Daarmee is deze definitie van welvaartsarmoede dus nauwer dan de definitie van inkomensarmoede. Daarbij zijn de inkomens- en vermogensgegevens hoofdzakelijk gebaseerd op belastinggegevens.

2.2 Doelvariabele

De doelvariabele is een indicator voor armoede in het jaar na het referentiejaar ($t + 1$), waarbij het referentiejaar het meest recente jaar is waarvoor data beschikbaar is:

$$y_i = \begin{cases} 1 & \text{als persoon } i \text{ in jaar } t + 1 \text{ arm is} \\ 0 & \text{anders} \end{cases}$$

De modellen schatten de kans dat persoon i in het jaar na het referentiejaar ($t + 1$) arm blijft (arme doelpopulatie) of wordt (niet-arme doelpopulatie): $P(y_i = 1)$.

2.3 Doelpopulaties

De doelpopulatie bestaat uit alle personen die op 31 december van referentiejaar (t) 18 jaar of ouder waren, in Nederland woonden in een particulier huishouden waarvan het inkomen en vermogen over het gehele referentiejaar (t) en het daaropvolgende jaar ($t + 1$) bekend waren. Omdat bij start van het onderzoek inkomens- en vermogensgegevens tot en met 2018 ($t + 1$) bekend waren, is 2017 als meest recente referentiejaar (t) gekozen. Daarnaast is een tweede referentiejaar (t) 2013 toegevoegd. Omdat dit het eind van de economische crisis betreft, zit er meer variatie in de data, waardoor de modellen robuuster worden.

Merk op dat onder andere minderjarigen zijn uitgesloten uit de doelpopulatie, evenals personen in institutionele of studentenhuishoudens, personen die in referentiejaar (t) zijn geïmmigreerd of verhuisd van

een institutioneel of studentenhuishouden, en personen die in het jaar na het referentiejaar ($t + 1$) zijn overleden, geëmigreerd of verhuisd naar een institutioneel of studentenhuishouden.

De doelpopulatie is geplitst in twee subpopulaties: de arme en de niet-arme doelpopulatie. De arme doelpopulatie bestaat uit alle personen die in referentiejaar (t) arm waren (ongeveer 0,6 miljoen voor elk referentiejaar). Voor deze personen schatten we de kans dat ze in het jaar na het referentiejaar ($t + 1$) arm blijven, en daarmee ook de kans dat ze uit de armoede komen. De niet-arme doelpopulatie bestaat uit alle personen die in referentiejaar (t) niet arm waren (ongeveer 12 miljoen voor elk referentiejaar). Voor deze personen schatten we de kans dat ze in het jaar na het referentiejaar ($t + 1$) arm worden, en daarmee ook de kans dat ze niet-arm blijven.

2.4 Kenmerken

Wat in de statistiek vaak een onafhankelijke, verklarende of voorspellende variabele wordt genoemd, wordt in machine learning meestal 'kenmerk' genoemd. Als kenmerken hebben we op basis van een beknopte literatuurscan en input van domeinexperts meer dan 500 potentiële risicofactoren afgeleid uit het Stelsel van Sociaal-statistische Bestanden (SSB) (bijlage A). In deze eerste versie van de modellen zijn alleen kenmerken afgeleid op basis van CBS-data uit registraties. Het voordeel daarvan is dat de gegevens voor (vrijwel) de gehele bevolking beschikbaar zijn. De kenmerken kunnen op verschillende manieren worden geclassificeerd:

- Naar niveau:
 - persoonskenmerken zoals plaats in het huishouden
 - huishoudkenmerken zoals woningbezit
- Naar periodiciteit:
 - tijdonafhankelijke kenmerken zoals geboorteland
 - kenmerken die binnen het SSB op jaarbasis beschikbaar zijn, zoals opleidingsniveau
 - kenmerken die binnen het SSB op maandbasis beschikbaar zijn, zoals belangrijkste inkomensbron
- Naar thema:
 - demografische kenmerken zoals leeftijd
 - economische kenmerken zoals sociaaleconomische categorie
 - gezondheidskenmerken zoals zorgkosten
 - criminaliteitskenmerken zoals detentie
- Naar status of gebeurtenis:
 - status van persoonskenmerken zoals zelfstandig ondernemer
 - gebeurtenis in partnerschap zoals scheidingen
- Naar type:
 - numerieke/ fractionele kenmerken zoals de fractie maanden in referentiejaar dat persoon werkzaam was
 - integer/ geheeltallige kenmerken zoals aantal maanden sinds start van het huishouden—ook ordinale kenmerken zoals opleidingsniveau en generatie worden als integer meegenomen
 - dummies/ indicatoren zoals indicator of een persoon wel of niet in een schuldsaneringstraject zit
 - nominale kenmerken zoals sociaaleconomische categorie—deze worden met one-hot encoding omgezet naar dummies. Voor iedere categorie wordt een indicatorvariabele afgeleid.

We hebben gekeken in hoeverre het mogelijk is om ook gemeente-/ en wijkenkenmerken mee te nemen in deze eerste versie van de modellen. Dit zou namelijk toegevoegde waarde kunnen hebben, omdat dit iets zegt over de context waarin iemand leeft. Zo zou iemand met een hoog inkomen in een arme wijk misschien een andere kans op armoede kunnen hebben dan een persoon in een rijke wijk. Het meenemen van wijk- en gemeentekenmerken gaf echter implausibele resultaten: ze kregen een groter belang toegedicht dan persoons- en huishoudkenmerken. Een mogelijke verklaring is dat de hiërarchische structuur

in de data werd genegeerd. Iedereen in dezelfde wijk scoort hetzelfde op wijk- en gemeentekenmerken. Een niet-hiërarchisch model onderschat dan de standaardfouten van de effecten. Daarom zijn wijk- en gemeentekenmerken niet meegenomen in de modellen. In een vervolgonderzoek kan verder uitgezocht worden hoe dit soort kenmerken meegenomen kan worden.

Voor de meeste kenmerken bevat de brondata drie jaar aan historie. Per kenmerk is bepaald hoe deze informatie kan worden gebruikt om tot een zo informatief mogelijk kenmerk te komen. Is het van meerwaarde om niet alleen naar de meest actuele situatie te kijken (bijvoorbeeld de sector waar een persoon werkt), maar ook naar die in het verleden (bijvoorbeeld de economische zelfstandigheid van een persoon)? En is het van meerwaarde om mee te nemen hoe lang een bepaalde situatie heeft voortgeduurd (bijvoorbeeld het aantal maanden bijstand in een jaar)? Bijlage A bevat een totaaloverzicht van de kenmerken.

3 Methoden

3.1 XGBoost

Voor ieder van de twee populaties (arme en niet-arme populatie; zie hoofdstuk 2) wordt een model geschat dat op basis van een set kenmerken de kans schat dat iemand arm blijft c.q. wordt in jaar $t + 1$. De methode die gebruikt wordt is *eXtreme Gradient Boosting* (XGBoost) (Chen and Guestrin, 2016). XGBoost combineert eenvoudige beslisbomen om tot gedetailleerde schattingen te komen. Een standaard beslisboom is een eenvoudige machine-learningmethode (Hastie et al., 2009, hoofdstuk 9). In een beslisboom wordt de data op basis van één kenmerk herhaaldelijk in twee groepen verdeeld. In iedere groep is de geschatte waarde het groepsgemiddelde. De groepen worden zo bepaald dat iedere keer als er subgroepen gemaakt worden de toename in de modelkwaliteit het grootst is. De methode zal proberen om groepen zo groot mogelijk te maken en te zorgen dat de groepen zoveel mogelijk gelijke records bevat wat betreft de doelvariabele. In een beslisboom kan bijvoorbeeld gekozen worden om eerst twee groepen te maken aan de hand van leeftijd, bijvoorbeeld in jongeren (leeftijd < 25) en ouderen (leeftijd ≥ 25), omdat dat de grootste toename van de modelkwaliteit oplevert. Vervolgens kan een groep opnieuw in twee groepen worden verdeeld op basis van hetzelfde of een ander kenmerk. Bijvoorbeeld de jongeren in mannen en vrouwen en de ouderen (leeftijd > 65) en 27 tot en met 65-jarigen. Dit proces wordt herhaald tot een stopcriterium is bereikt. Het stopcriterium is een combinatie van voldoende grote groepen en voldoende toename van de modelkwaliteit.

Uiteindelijk schat de beslisboom voor iedere persoon i een kans om arm te worden/blijven. Deze schrijven we als $\hat{p}_i$ en is gelijk aan de gemiddelde kans van de personen in de groep waarin persoon i zit. Om de kwaliteit van het model te meten vergelijken we de geschatte kans voor iedere persoon met de daadwerkelijk waargenomen waarde. De daadwerkelijk waargenomen waarde voor persoon i wordt geschreven als y_i (y_i is gelijk aan nul als iemand niet arm is in $t + 1$ en gelijk aan één als i wel arm is). Idealiter willen we dat het model een kans gelijk aan één schat als iemand arm blijft of geworden is en een kans gelijk aan nul wanneer dat niet zo is. Op dat moment weet het model perfect onderscheid te maken tussen personen die arm worden/blijven en personen die dat niet doen. De modelkwaliteit van de beslisboom meet dus in welke mate de geschatte kans verschilt van de waargenomen waarde. Omdat we hier te maken hebben met binaire variabelen en kansen die altijd tussen nul en één moeten liggen wordt de modelkwaliteit gemeten aan de hand van de negatieve log-likelihood (NLL), welke in het geval van binaire classificatie gelijk is aan de totale binaire cross entropie:

$$NLL = - \sum_i [y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i)].$$

Dit is in dezelfde maat die gebruikt wordt om logistische regressiemodellen te schatten.

Varianten op de beslisboom, zoals *random forest* en *gradient tree boosting*, zijn ensemble-methoden die worden ontwikkeld om het algoritme stabiel te maken (kleinere variantie), robuuster tegen het modelleren van ruis (overfitting; zie ook paragraaf 3.2) en tegelijkertijd schaalbaar te houden voor grote datasets. Bij ensemble-methoden wordt niet één beslisboom geschat maar een set beslisbomen. Bij een random forest worden complete beslisbomen onafhankelijk (parallel) van elkaar getraind met subsets van de trainingsdata en kenmerken en worden de schattingen gemiddeld over bomen (Hastie et al., 2009, hoofdstuk 15). Bij gradient tree boosting—gebruikt in dit onderzoek—worden eenvoudige beslisbomen na elkaar (sequentieel) getraind waarbij elke opeenvolgende boom de schattingen van de voorgaande bomen verfijnd (Hastie et al., 2009, hoofdstuk 10, in het bijzonder algoritme 10.3). Chen en Guestrin (Chen and Guestrin, 2016) hebben een variant van gradient tree boosting gemaakt die zeer schaalbaar is voor grote datasets. Hun systeem heet XGBoost en is de methode die in dit onderzoek is gebruikt.

3.2 Schatten van het machine-learningmodel

Bij het schatten van een machine-learningmodel moet in het algemeen een afweging gemaakt worden tussen overfit en onderfit. Bij overfit bevat het model teveel detail. Het gaat allerlei eigenschappen beschrijven die toevallig in de dataset zitten waarmee je het model geschat hebt. Het model generaliseert daardoor niet goed naar andere datasets. Bij een nieuwe dataset zal het model slechtere schattingen opleveren. Bij onderfit bevat het model te weinig detail. Het model is te grof en zal in het algemeen minder goed schatten. Machine-learningmethodes hebben bijna altijd een aantal zogenaamde hyperparameters waarmee deze twee effecten tegen elkaar afgewogen kunnen worden. Bij het schatten van het model moeten verschillende waardes van deze hyperparameters uitgeprobeerd worden om een goed model te vinden. Dit proces van zoeken naar de juiste hyperparameters wordt het tunen van het model genoemd.

XGBoost heeft een aantal parameters waarmee de complexiteit van het model bepaald kan worden. De belangrijkste zijn (in dit onderzoek onderzocht):

- *ntree*: het aantal beslisbomen dat geschat wordt. Dit is de belangrijkste parameter. Hoe meer bomen, hoe meer detail de methode kan schatten.
- *boomdiepte*: maximale diepte van de bomen, dus het aantal keer dat de dataset gesplitst wordt. Bij een boomdiepte van 4 kan de dataset in maximaal $2^4 = 16$ groepen verdeeld worden. Hier ook weer: hoe dieper de boom hoe meer detail.
- *leersnelheid*: hoe snel leert het model. Als deze hoog is dan hebben de eerst geschatte beslisbomen heel veel invloed op de schattingen. Omdat er een zekere mate van toeval zit in deze beslisbomen kan het model hiermee minder goed zijn. Een lagere leersnelheid is in het algemeen beter maar zorgt er wel voor dat het aantal bomen (*ntree*) hoger moet zijn en dat het trainen langer duurt.

Om de optimale waarde te vinden van deze hyperparameters is geneste kruisvalidatie toegepast. Aangezien we dit maar één keer gedaan hebben komt dit erop neer dat we de dataset in drie delen hebben verdeeld:

- *Testset*: bevat 10% van personen.
- *Tuningset*: bevat de overige 90% van de personen. Deze wordt gebruikt om je hyperparameters te tunen en om het uiteindelijke model te schatten. Deze is verder onderverdeeld in:
 - *Trainingset*: bevat 90% van de tuningset dus 81% van de gehele dataset. Deze wordt gebruikt om het model te schatten. Hierbij worden verschillende hyperparameters uitgeprobeerd.
 - *Validatieset*: bevat 10% van de tuningset dus 9% van de gehele dataset. De wordt gebruikt om voor de met verschillende hyperparameters geschatte modellen de modelkwaliteit uit te rekenen om zo de optimale set hyperparameters te vinden.

Als met behulp van de trainingset en validatieset de optimale set parameters is gevonden, wordt het model geschat op de gehele tuningset. De kwaliteit van dat uiteindelijke model wordt geëvalueerd met behulp van de testset.

Vaak wordt het verdelen in test, validatie en trainingsset meerdere keren herhaald om zo meerdere schattingen te krijgen voor de modelkwaliteit van het model. Bij testen bleek er dankzij het formaat van de dataset waarmee we hier te maken hebben, nauwelijks variatie in de schattingen te zitten. Daarom zijn in verband met de rekentijden de hyperparameters geoptimaliseerd met slechts 1 validatieset en 1 testset.

Metrieken modelkwaliteit

Voor het tunen van de hyperparameters en het evalueren van de modelkwaliteit van het uiteindelijke model zijn kwaliteitsmaten nodig. Het model schat een kans $\hat{p}_i$ dat persoon i in jaar $t + 1$ arm blijft of wordt. Er bestaan veel maten die de kwaliteit van deze schattingen kwantificeren (Powers, 2011). Reden dat er

meerdere maten hiervoor zijn is onder andere dat het vaak van de toepassing van het model afhangt welke maten het meest geschikt zijn.

Hoewel het model dat geschat wordt een classificatiemodel is omdat het model bepaald in welke klasse (blijft/wordt wel/niet arm) iemand valt, wordt het model hier niet expliciet gebruikt als classificatiemodel. Het doel is niet om personen in klassen in te delen. We willen zo goed mogelijk de kans op armoede te schatten en zijn vooral geïnteresseerd hoe deze kansen beïnvloed worden door de kenmerken van de personen. Daarom zijn maten die meten hoeveel fouten het model maakt als personen in klassen ingedeeld worden iets minder interessant en zijn maten die meer kijken naar de geschatte kansen interessanter. De maten die te maken hebben met classificatie zullen echter toch wel gerapporteerd worden omdat deze iets gebruikelijker zijn in de machine-learningliteratuur en omdat deze ook wel iets zeggen over de kwaliteit van de geschatte kansen. Het gaat hierbij om de maten *MCC*, *F1*, *Recall* en *Precision* die verderop geïntroduceerd worden.

Er zijn twee maten die de geschatte kansen direct vergelijken met de waargenomen armoedeindicator y_i . De eerste is de negatieve log-likelihood (zie boven). Deze ligt tussen 0 en oneindig en is vooral bruikbaar om modellen te vergelijken. Een hogere waarde is daarbij beter. De tweede is de *area under the ROC curve* (AUC). De ROC (Receiver Operating Characteristic) curve is de grafiek van de fractie personen die arm blijven/worden die correct geïdentificeerd wordt tegen de fractie personen die niet arm blijven/worden die incorrect geïdentificeerd wordt. Een eenvoudige interpretatie is deze: de AUC meet de fractie van paren van records—één record met $y_i = 1$ en één record met $y_i = 0$ —waarbij voor het record met $y_i = 1$ een hogere kans wordt geschat dan voor het record met $y_i = 0$. De AUC ligt tussen 0,5 en 1. Somers' D is een maat die de AUC schaalst zodat deze tussen 0 en 1 ligt:

$$\text{Somers' D} = 2 \cdot \text{AUC} - 1.$$

Tabel 3.1 Confusiematrix

Werkelijk \ Geschat	Arm	Niet-arm
Arm	True Positives (TP)	False Negatives (FN)
Niet-arm	False Positives (FP)	True Negatives (TN)

De andere maten zijn gebaseerd op de confusiematrix (zie tabel 3.1) met de volgende cellen

- TP: het aantal *true positives* of correct als arm geïdentificeerde personen
- FN: het aantal *false negatives* of incorrect als niet-arm geïdentificeerde personen
- FP: het aantal *false positives* of incorrect als arm geïdentificeerde personen
- TN: het aantal *true negatives* of correct als niet-arm geïdentificeerde personen

Voor het maken van een confusiematrix is het nodig dat de geschatte kansen vertaald worden naar klassen (arm/niet-arm). Dit wordt gedaan door de waarnemingen met een geschatte kans boven een drempel c te classificeren als 'arm' en waarnemingen met een kans kleiner van c te classificeren als 'niet-arm' (in $t + 1$). AUC hangt niet van de drempelwaarde af omdat elke drempelwaarde een punt op de ROC-curve geeft. Voor de andere maten wordt vaak een drempel c van 0,5 gekozen, maar het kan zijn dat het optimum van een bepaalde maat bij een hogere of lagere drempelwaarde ligt (zie bijlage C). De optimale waarde van de drempel voor een bepaalde maat wordt aangeduidt met c^* .

Tabel 3.2 Maten modelkwaliteit

Metriek	Definitie
MCC	$\frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$
F1	$\frac{2 \cdot TP}{TP + 0.5(FP + FN)}$
- Recall	$\frac{TP}{TP + FN}$
- Precision	$\frac{TP}{TP + FP}$

Matthews' correlatiecoëfficiënt (MCC) is een soort correlatiecoëfficiënt tussen de waargenomen en de geschatte klassen (tabel 3.2). Een waarde van 0 betekent dat het model even slecht is als gokken, een waarde van 1 betekent een foutloos, deterministisch model. Deze maat neemt alle cellen van de confusiematrix mee en is ongevoelig voor de werkelijke fractie personen die arm blijft/wordt.

De F1-score is het harmonisch gemiddelde van de *recall* en *precision* (tabel 3.2). De recall meet de fractie personen die arm blijven/worden die ook als zodanig worden geïdentificeerd (een classificatiekans). De precision meet de fractie correct geschatte klassen in die records die geïdentificeerd worden als 'arm' (een calibratiekans). Hoe hoger drempelwaarde c wordt gekozen, hoe groter de precision maar hoe kleiner de recall (zie bijlage C). Door het harmonisch gemiddelde te nemen (en niet het rekenkundig of geometrisch gemiddelde) ligt de F1 dicht bij de laagste van de *precision* en *recall*. In de praktijk hangen de precision, recall en F1 af van de fractie personen die in werkelijkheid arm blijft/wordt. Hoe lager de fractie personen die arm blijft/wordt hoe lager de F1 bijvoorbeeld zal zijn. Hierdoor zijn deze maten niet goed te gebruiken om de modelkwaliteit te vergelijken tussen populaties die sterk verschillen in de fractie personen die arm blijft/wordt. Wanneer de F1-score gebruikt wordt voor vergelijkingen tussen deelpopulaties zal deze daarom genormaliseerd zoals beschreven in Burger and Meertens (2020). Voor de precision, recall, F1 geldt verder dat een score van 1 het best is en dat de laagste waarde 0 is. Verder geldt dat al deze drie maten focussen op correct bepalen of iemand arm blijft/wordt. De maten zijn relatief ongevoelig voor incorrect voorspellen van het niet-arm worden/blijven.

3.3 Interpretatie van het model

Deelpopulaties

XGBoost is gemaakt om de modelkwaliteit zo groot mogelijk te maken. Nadeel is dat het model meerdere beslisbomen combineert en daardoor moeilijk uitlegbaar is. Een techniek om inzicht te krijgen in wat er in een complex machine learning model gebeurt, is om een eenvoudiger model te schatten dat de schattingen van het complexe model probeert te beschrijven. Dit wordt onder andere ook wel een *global surrogate model* (Molnar, 2019) genoemd. Dit surrogaatmodel beschrijft dus niet de waarnemingen maar het complexe model en zal ook vanwege zijn eenvoud de schattingen van het complexe model maar beperkt kunnen beschrijven. De mate waarin dit lukt kan wel gemeten worden (voor de twee modellen voor armoede wordt circa 40-60% verklaard door het vereenvoudigde model; zie hoofdstuk 4). Het vereenvoudigde model zal wel een beeld geven van de belangrijkste effecten in het complexe model. Daarnaast kan het vereenvoudigde gebruikt worden om datagedreven deelpopulaties te definiëren die verder bestudeerd kunnen worden.

Voor ieder van de twee populaties (mensen in armoede in jaar t en mensen niet in armoede in jaar t) wordt een vereenvoudigd model geschat. Hiervoor wordt een eenvoudige beslisboom gebruikt die de 20 belangrijkste kenmerken van elk XGBoost model gebruikt (voor hoe het belang bepaald is zie onder). Aan de hand van deze beslisbomen worden de populaties verdeeld in circa tien deelpopulaties. Deze deelpopulaties zijn dus bepaald aan de hand van de belangrijkste kenmerken en zo homogeen mogelijk wat betreft hun

kans op armoede in jaar $t + 1$. Deze deelpopulaties zijn aangevuld met vijf in overleg met BZK en experts gedefinieerde deelpopulaties:

- de gemeente Amsterdam,
- de gemeente Heerlen,
- laagopgeleide jongeren (≤ 26 jaar conform het hoofdstuk Werk in de Landelijke Jeugdmonitor) die niet thuis wonen en geen opleiding volgen,
- werkenden (werknemer, directeur-grotaandeelhouder, zelfstandig ondernemer, overige zelfstandige of meewerkend gezinslid, en ≤ 74 jaar conform de CBS-definitie van beroepsbevolking),
- ouders in een eenouderhuishouden.

De eerste twee populaties (de twee gemeenten) zijn gekozen om na te gaan of het mogelijk is om met het machine-learningmodel gemeentspecifieke verbanden te identificeren. Van de overige aangedragen deelpopulaties is bekend dat deze een hoog risico lopen en/of minder goed in beeld zijn bij instanties (bijv. werkende armen).

Vervolgens kan binnen deze deelpopulaties gekeken worden welke kenmerken voor die deelpopulaties belangrijk zijn, hoe de kans op armoede beïnvloed wordt door de kenmerken van de personen in deze deelpopulaties, en in hoeverre het belang van een kenmerk in een deelpopulatie verschilt van de gehele populatie.

SHAP-waardes

Bij een machine-learningmodel is het wat lastiger dan bij een eenvoudig regressiemodel om te zien hoe belangrijk een kenmerk is en welk effect kenmerken hebben op de schattingen. Dit heeft een aantal oorzaken:

- Vaak is het aantal kenmerken in een machine-learningmodel veel hoger dan gebruikt wordt in een traditioneel model (zoals logistische regressie). Deze kenmerken hangen vaak ook nog samen (inkomen hangt bijvoorbeeld samen met leeftijd).
- Machine-learningmodellen nemen meestal automatisch interacties tussen kenmerken op waardoor het lastiger wordt kenmerken op zichzelf te bestuderen.
- Traditionele modellen hebben vaak voor iedere kenmerk een aparte parameter die aangeeft wat het effect is van dat kenmerk. Machine-learningmodellen hebben vaak een heel groot aantal parameters per kenmerk.

Een manier om toch te zien hoe een kenmerk de schatting beïnvloed, is om om te kijken hoe de schattingen veranderen als een kenmerk wordt toegevoegd aan het machine-learningmodel. Eén zo'n methode is SHAP (SHapley Additive exPlanation) (Lundberg and Lee, 2017). Voor een dataset worden allerlei schattingen uitgerekend door verschillende combinaties van kenmerken weg te laten uit het model. Er wordt dan gekeken hoe de schattingen veranderen. Stel iemands kans op armoede neemt toe als leeftijd wordt toegevoegd aan het model. Dan heeft voor die persoon zijn leeftijd een armoedekansverhogend effect. Door ook combinaties van kenmerken weg te laten, dus niet alleen leeftijd maar ook leeftijd en inkomen of alleen inkomen, wordt het beter mogelijk om te zien welk effect een kenmerk heeft, los van de andere kenmerken in het model (ook al blijft dat lastig). Dus om het effect van een kenmerk A uit te rekenen wordt A toegevoegd aan alle modellen met alle mogelijke combinaties van kenmerken exclusief A. Vervolgens wordt bepaald hoe de schattingen veranderen. Het effect van A is dan een gewogen gemiddelde over al deze combinaties.

Met behulp van SHAP kan er dus voor iedere persoon en voor ieder kenmerk in het model bepaald worden wat voor effect dit kenmerk heeft voor die persoon. Dus voor de ene persoon kan leeftijd een

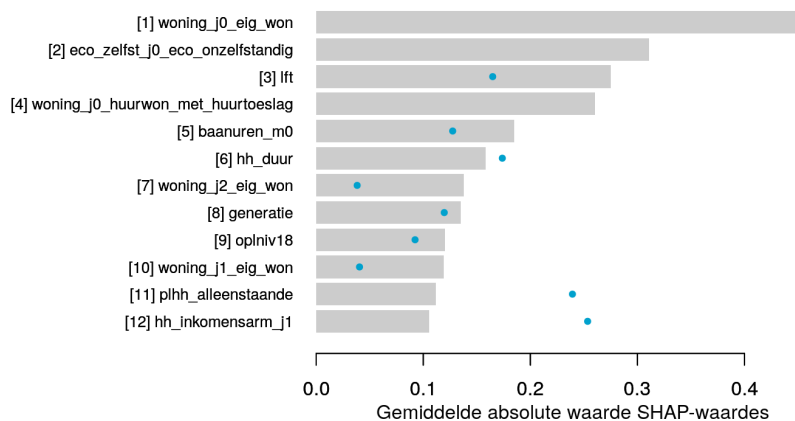
armoedekansverhogend effect hebben en voor de andere een armoedekansverlagend effect. Doordat in machine-learningmodellen kenmerken elkaar ook kunnen beïnvloeden (interacties), kan het zelfs zo zijn dat dit zo is voor twee personen met dezelfde leeftijd. Die twee personen zullen dan op andere kenmerken verschillen waardoor leeftijd voor hen een ander effect heeft.

Het gemiddelde van de SHAP-waardes over de hele populatie is ongeveer nul (SHAP-waardes worden geschat waardoor deze niet exact zijn en dus ook niet exact nul). Positieve waardes geven aan dat dit kenmerk voor die persoon een kansverhogend effect heeft en negatieve waardes geven een kansverlagend effect aan.

Belang van kenmerken

Als een kenmerk voor veel mensen een kansverhogend of een kansverlagend effect heeft, dan is dit een belangrijk kenmerk. Daarom kan het belang van een kenmerk in het model bepaald worden door voor een kenmerk de gemiddelde absolute SHAP-waarde te bepalen. Hiermee wordt gemeten in welke mate de SHAP-waardes van nul afwijken.

Dit kunnen we doen voor zowel de gehele populatie als voor deelpopulaties. We hebben de SHAP-waardes op persoonsniveau. We kunnen dus de personen binnen zo'n deelpopulatie selecteren en daarvan de gemiddelde absolute waarde van de SHAP-waardes bepalen. In onderstaande figuur is hiervan een voorbeeld gegeven zoals deze ook is opgenomen in de resultaten voor de deelpopulaties (zie paragraaf 4.3).



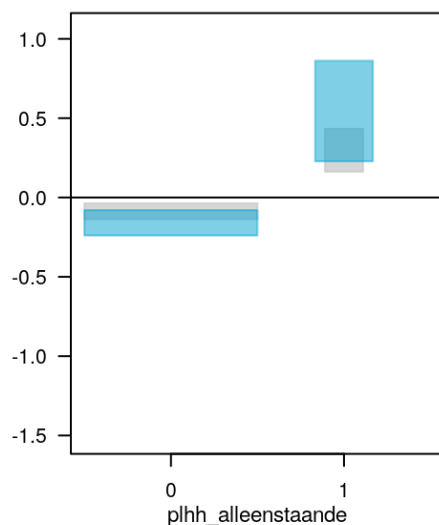
De grijze balken geven het belang van de kenmerken aan voor de hele populatie (van niet-armen in dit geval). De blauwe punten geven het belang van kenmerken aan voor de voorbeelddeelpopulatie. Bij een paar kenmerken ontbreken de punten omdat hierop geselecteerd is en dus iedereen in de voorbeelddeelpopulatie dezelfde waarde heeft voor dit kenmerk. Er is bijvoorbeeld te zien dat leeftijd (lft) in deze voorbeelddeelpopulatie minder belangrijk is dan in de gehele niet-arme populatie. Het kenmerk dat aangeeft of iemand alleenstaand is (plhh_alleenstaande) is daarentegen in deze voorbeelddeelpopulatie belangrijker dan in de gehele niet-arme populatie.

Effect van kenmerken

Met behulp van de SHAP-waardes kunnen we bepalen welke kenmerken belangrijk zijn en welke minder belangrijk. We kunnen nu ook kijken hoe deze kenmerken de kans op armoede beïnvloeden. Het wel of niet alleenstaand zijn is blijkbaar belangrijk, maar lopen mensen die alleenstaand zijn een hogere kans op armoede of juist een kleinere kans op armoede? Dit kunnen we zien als we de SHAP-waardes bekijken voor verschillende waardes van een kenmerk. Zoals eerder aangegeven variëren SHAP-waardes per persoon zelfs voor dezelfde waarde van een kenmerk. We kunnen een gemiddelde SHAP-waarde uitrekenen voor

alleenstaanden en voor niet-alleenstaanden. Het is echter ook nuttig om te kijken naar de spreiding in de SHAP-waardes. Het wel of niet alleenstaand zijn kan bijvoorbeeld ook belangrijk zijn als de SHAP-waardes gemiddeld nul zijn maar er veel mensen zijn met hoge of lage waardes. Daarom is ervoor gekozen om het gebied aan te geven waar 80% van de personen in de populatie hun SHAP-waardes hebben (lijn van 10-de tot 90-ste percentiel).

In onderstaande figuur is voor het kenmerk dat aangeeft of iemand wel of niet alleenstaand is (`plhh_alleenstaand`) aangegeven wat de SHAP-waardes zijn voor de verschillende waarden van dit kenmerk. In grijs zijn hier de waardes voor de gehele niet-arme populatie weergegeven en in blauw die van een voorbeelddeelpopulatie. Er is te zien dat in beide populaties alleenstaanden (`plhh_alleenstaand` = 1) een hogere kans op armoede hebben: de SHAP-waardes liggen boven de nul (het populatiegemiddelde). Niet-alleenstaanden (`plhh_alleenstaand` = 0) hebben een lagere kans op armoede: de SHAP-waardes liggen onder de nul.

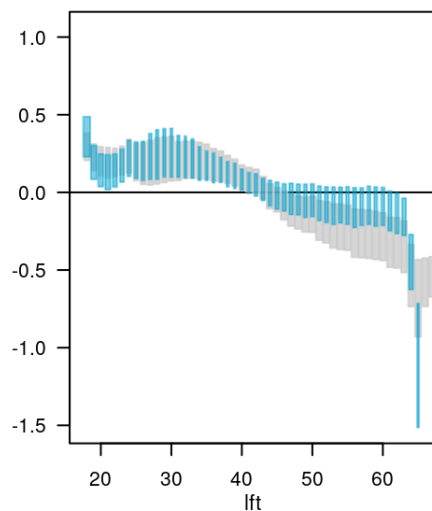


Naast de SHAP-waardes geeft dit plaatje ook weer hoe vaak de waardes voorkomen in de twee populaties met behulp van de breedte van de balken. Voor de voorbeelddeelpopulatie is deze breder voor alleenstaanden dan voor de gehele populatie. Dit geeft aan dat in deze voorbeelddeelpopulatie relatief veel alleenstaanden voorkomen.

We zagen eerder dat dit kenmerk in deze voorbeelddeelpopulatie belangrijker is dan voor de gehele populatie. In deze voorbeelddeelpopulatie liggen de SHAP-waardes dus gemiddeld verder van de nul af. Dit is hier ook te zien. In deze voorbeelddeelpopulatie hebben de niet-alleenstaanden relatief lagere armoedekansen en de alleenstaanden relatief hogere armoedekansen dan in de gehele populatie (de blauwe balken liggen verder van de nul af dan de grijze balken). Verder is er meer spreiding in armoedekansen tussen alleenstaanden in de voorbeelddeelpopulatie dan in de gehele populatie (de rechter blauwe balk is hoger dan de rechter grijze balk).

De figuur hieronder geeft de SHAP-waardes voor leeftijd weer. Dit is een kenmerk dat in de voorbeelddeelpopulatie minder effect heeft dan in de gehele populatie. Dat laatste is ook te zien in het effect dat dit kenmerk heeft in deze voorbeelddeelpopulatie. Vooral voor de wat oudere mensen (40+) liggen de SHAP-waardes dicht bij de nul. In het algemeen (in de gehele populatie) hebben ouderen dus een lager risico op armoede dan gemiddeld maar in deze voorbeelddeelpopulatie is dat effect een stuk minder. De uitschieter rond 65 heeft te maken met het met pensioen gaan van mensen wat de kans op armoede verlaagt. Merk op dat bij kenmerken met veel verschillende waardes de dikte van de balkjes niet goed afleesbaar is o.a. doordat de balkjes altijd een minimale breedte hebben van 1 pixel. Daarnaast geven

de SHAP-waardes dus niet de kans weer, maar alleen de (gemiddelde) verandering in de geschatte kans als het kenmerk wordt meegenomen. Hier neemt de geschatte kans op armoede dus toe voor mensen rond de 30 als leeftijd wordt meegenomen in het model.



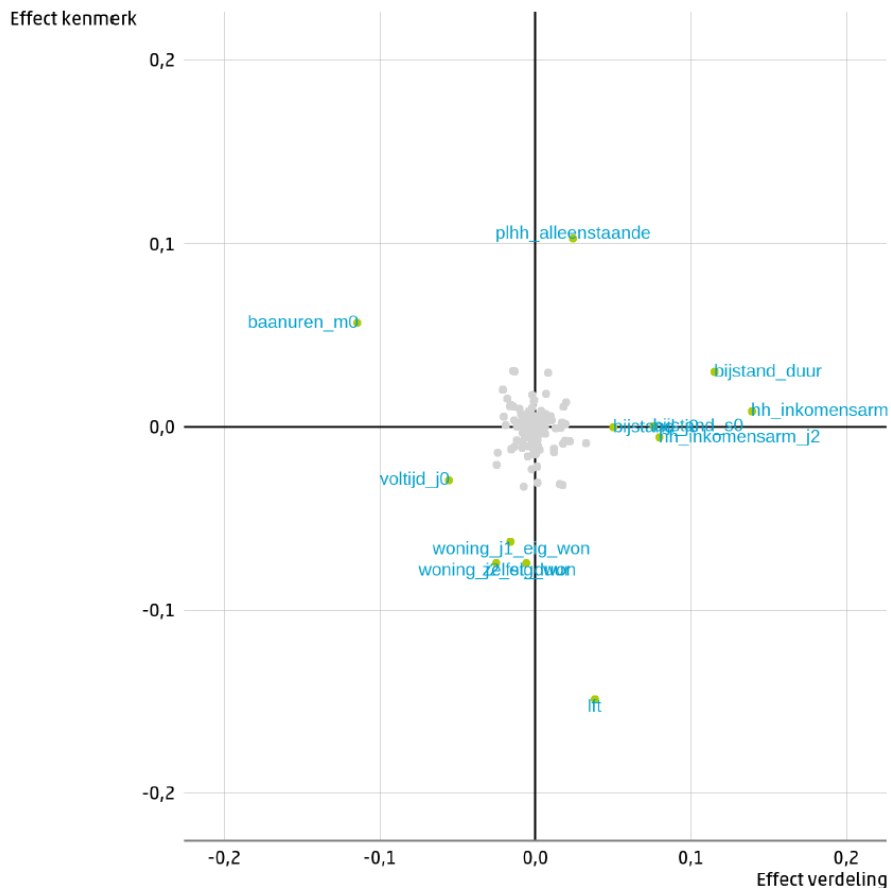
Decompositie

In de voorgaande discussie zagen we dat het kenmerk dat aangeeft of iemand alleenstaand is voor deze voorbeelddeelpopulatie een groter effect heeft dan in de gehele populatie. Dit lijkt enerzijds te komen doordat er in deze voorbeelddeelpopulatie meer alleenstaanden zijn (die een groter risico op armoede hebben dan niet-alleenstaanden). Anderzijds hebben alleenstaanden in deze voorbeelddeelpopulatie een hoger risico op armoede dan alleenstaanden in de gehele populatie.

Het effect voor leeftijd is in deze populatie kleiner. Dit lijkt vooral te komen doordat het verlaagde risico voor mensen boven de 40—that te zien is in de gehele populatie—voor een flink gedeelte verdwijnt in deze voorbeelddeelpopulatie. Het effect van leeftijd op de kans op armoede is in deze populatie dus anders.

Dit geldt voor ieder kenmerk: het verschil in belang dat een kenmerk heeft in een deelpopulatie ten opzichte van de gehele populatie kan komen doordat de verdeling van dit kenmerk anders is in de deelpopulatie, en dit kan komen doordat het effect dat het kenmerk heeft op de kans op armoede anders is in de deelpopulatie. Het kunnen zien waardoor de verschillen komen geeft extra inzicht.

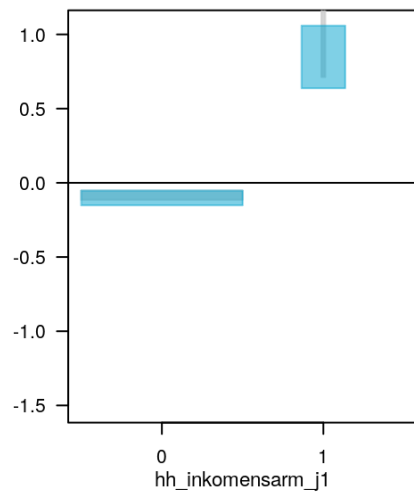
Het is mogelijk om het verschil in belang dat zichtbaar was in de eerste figuur op te splitsen in een component die het belang van de verdeling van het kenmerk aangeeft en een component die het belang van het effect dat het kenmerk heeft op de geschatte kans aangeeft (zie bijlage D). Onderstaande figuur geeft deze twee componenten weer voor de kenmerken in de eerste figuur.



De verticale as geeft de component weer die het effect van het kenmerk op de geschatte kans meet. De horizontale as geeft het effect van de verdeling aan. In de figuur zien we bovenin het kenmerk dat aangeeft of iemand alleenstaand is of niet (plhh_alleenstaande). Het kenmerk plhh_alleenstaande ligt boven de horizontale lijn, wat betekent dat het effect van dit kenmerk op de armoedekans groter is in deze voorbeelddeelpopulatie dan in de gehele niet-arme populatie. Daarnaast ligt dit kenmerk iets rechts van de verticale lijn, wat betekent dat deze voorbeelddeelpopulatie wat meer mensen bevat waarbij het kenmerk effect heeft vergeleken met de gehele niet-arme populatie. Leeftijd (lft) ligt helemaal onder. Ondanks een klein verhogend effect van de verdeling, is leeftijd is dus minder belangrijk in deze populatie omdat het effect van leeftijd op de geschatte kans kleiner is dan in de gehele populatie.

Bij beide kenmerken komen de verschillen met de hele populatie dus voornamelijk door het effect dat de kenmerken hebben op de geschatte kans. Voor een kenmerk als hh_inkomensarm_j1, dat aangeeft of het huishouden van een persoon inkomensarm is in jaar $t - 1$, komt het verschil voornamelijk doordat dit kenmerk in deze voorbeelddeelpopulatie een andere verdeling heeft. Dit is duidelijk te zien in onderstaande figuur. Het aantal personen uit een huishouden dat inkomensarm is in $t - 1$ is in deze voorbeelddeelpopulatie veel groter dan in de gehele populatie (balkbreedte). Het effect dat dat heeft op armoede is echter wel min of meer gelijk aan dat in de populatie (afstand tot nullijn).

Een aanvullend detail: het ('horizontale') verschil in frequentie tussen deelpopulatie en gehele populatie wordt gewogen met het effect in de gehele populatie, en het ('verticale') verschil in effect tussen deelpopulatie en gehele populatie wordt gewogen met de frequentie in de deelpopulatie (zie de formule van de decompositie in bijlage D). Een ('horizontaal') verschil in frequentie is dus relevanter naarmate het kenmerk belangrijker is in de gehele populatie, c.q. de grijze balk verder van 0 af ligt. Een ('verticaal') verschil in effect is dus relevanter naarmate de frequentie van het kenmerk groter is in de deelpopulatie, c.q. de blauwe balk breder is.



3.4 Software

De belangrijkste R packages die zijn gebruikt in dit onderzoek zijn xgboost (Chen et al., 2021), SHAPforxgboost (Liu and Just, 2021), rpart (Therneau and Atkinson, 2019) en ggplot2 (Wickham, 2016).

4 Resultaten

4.1 Inleiding

In dit hoofdstuk bespreken we van beide modellen de modelkwaliteit, de belangrijkste kenmerken en hun relatie met de geschatte kansen voor de arme populatie (paragraaf 4.2) en de niet-arme populatie (paragraaf 4.4). Naast het landelijk beeld wordt per model één datagedreven deelpopulatie uitgelicht: de deelpopulatie met de grootste kans om uit de armoede te komen (paragraaf 4.3) en de deelpopulatie met de grootste kans om in de armoede te komen (paragraaf 4.5). De expertgedreven deelpopulaties worden in bijlage E besproken.

Omdat veel kenmerken lastig te omschrijven zijn in één of enkele woorden, bevatten de figuren afkortingen van de kenmerken. Onderstaande tabel geeft een omschrijving van de belangrijkste kenmerken die voorkomen in de figuren voor de arme populatie. Zie bijlage A voor een operationalisering van alle kenmerken.

Merk op dat de kenmerken over zorgkosten (zorgkosten_j0, _j1, _j2) met voorzichtigheid dienen te worden geïnterpreteerd. Zo hangen de zorgkosten af van de kosten die in de basisverzekering zijn opgenomen, wat van jaar op jaar verschilt. Ook zijn (veranderingen in) zorgkosten niet direct te relateren aan (veranderingen in) gezondheid.

Tabel 4.1 Belangrijkste kenmerken voor arme populatie

Variabele	Omschrijving
plhh_alleenstaande	Plaats in huishouden: alleenstaande
hh_inkomensarm_j1	Persoon behoort tot een huishouden met een inkomen tot de lage inkomensgrens (referentiejaar - 1)
secmsbi2008v21_j0_Onbekend	Indicator voor niet werkzaam (mensen met een uitkering, studenten of mensen zonder inkomen) (referentiejaar)
lft	Leeftijd
bijstand_j0	Fractie van maanden dat persoon bijstand had (referentiejaar)
bijstand_s0	Fractie van maanden dat persoon bijstand had (laatste half jaar van referentiejaar)
werkn_duur	Aantal maanden geleden dat persoon werknemer was
dpartner_duur	Aantal maanden sinds wijziging in partnerschap
hh_inkomensarm_j2	Persoon behoort tot een huishouden met een inkomen tot de lage inkomensgrens (referentiejaar - 2)
hh_duur	Aantal maanden sinds start van huishouden
kindgeboren_duur	Aantal maanden sinds geboorte kind
fpartner_j0	Fractie van maanden dat een persoon een partner heeft (referentiejaar)
woning_j0_eig_won	Eigen woning (referentiejaar)
oplniv18	Opleidingsniveau
generatie	Generatie migratieachtergrond
woning_j0_huurwon_met_huurtoeslag	Huurwoning met huurtoeslag (referentiejaar)
secm_ontv_uitk_bijstand	Bijstand als belangrijkste inkomstenbron
aantalkindhh	Aantal kinderen in huishouden
inkbronhh_j0_loon	Loon als voornaamste inkomensbron van huishouden (referentiejaar)

Loopt door op volgende pagina

Variabele	Omschrijving
baanuren_m0	Aantal gewerkte uren (december van referentiejaar)
inkbronn_hh_j0_bijst	Bijstand als voornaamste inkomensbron van huishouden (referentiejaar)
nkinderbij_j0	Fractie van maanden in betreffende periode dat aantal kinderen in huishouden toenam (referentiejaar)
vrouw	Indicator voor vrouw
zorgkosten_j2	Zorgkosten (referentiejaar - 2)
zorgkosten_j0	Zorgkosten (referentiejaar)
plhh_thuiswkind	Plaats in huishouden: thuiswonend kind
zorgkosten_j1	Zorgkosten (referentiejaar - 1)
bijstand_j2	Fractie van maanden dat persoon bijstand had (referentiejaar - 2)
plhh_ouder_eeenouder	Plaats in huishouden: ouder in eenouderhuishouden
fpartner_j1	Fractie van maanden dat persoon een partner had (referentiejaar - 1)
scholstud_duur	Aantal maanden geleden dat persoon scholier/student was
baanuren_j0	Aantal gewerkte uren (referentiejaar)
ETNGROEP3PBL_3	Indicator voor niet-westerse migratieachtergrond
plhh_partnr_huw_zkind	Plaats in huishouden: partner in gehuwd paar zonder kinderen
flex_j2	Fractie van maanden in betreffende periode dat persoon werkzaam was met flex contract (referentiejaar - 2)
fpartner_j2	Fractie van maanden dat persoon een partner had (referentiejaar - 2)
bijstand_duur	Aantal maanden geleden dat persoon bijstand had
bijstand_m0	Bijstand (december van referentiejaar)
eco_zelfst_j0_eco_onzelfstandig	Niet economisch zelfstandig (referentiejaar)
typhh_eeenouder	Type huishouden: eenouderhuishouden
plhh_overig	Plaats in huishouden: overig

Onderstaande tabel geeft een omschrijving van de belangrijkste kenmerken die voorkomen in de figuren voor de niet-arme populatie.

Tabel 4.2 Belangrijkste kenmerken voor niet-arme populatie

Variabele	Omschrijving
woning_j0_eig_won	Eigen woning (referentiejaar)
eco_zelfst_j0_eco_onzelfstandig	Economisch onzelfstandig (referentiejaar)
lft	Leeftijd
woning_j0_huurwon_met_huurtoeslag	Huurwoning met huurtoeslag (referentiejaar)
baanuren_m0	Aantal gewerkte uren (december referentiejaar)
hh_duur	Aantal maanden sinds start van huishouden
woning_j2_eig_won	Eigen woning (referentiejaar - 2)
generatie	Generatie migratieachtergrond
oplniv18	Opleidingsniveau
woning_j1_eig_won	Eigen woning (referentiejaar - 1)
plhh_alleenstaande	Plaats in huishouden: alleenstaande
hh_inkomensarm_j1	Persoon behoort tot een huishouden met een inkomen tot de lage inkomensgrens (referentiejaar - 1)
zelfst_duur	Aantal maanden geleden dat persoon zelfstandige was
plhh_thuiswkind	Plaats in huishouden: thuiswonend kind
inkbronn_hh_j0_loon	Loon als voornaamste inkomensbron van huishouden (referentiejaar)

Loopt door op volgende pagina

Variabele	Omschrijving
voltijd_j0	Fractie van maanden dat persoon voltijd werkzaam was (referentiejaar)
baanuren_m1	Aantal gewerkte uren (november referentiejaar)
hh_inkomensarm_j2	Persoon behoort tot een huishouden met een inkomen tot de lage inkomensgrens (referentiejaar - 2)
fpartner_j0	Fractie van maanden dat persoon een partner had (referentiejaar)
pensioen_s0	Fractie van maanden dat persoon een pensioen had (laatste half jaar van referentiejaar)
inkbronn_hh_j2_loon	Loon als voornaamste inkomensbron van huishouden (referentiejaar - 2)
woning_j0_huurwon_z_htoeslag	Huurwoning zonder huurtoeslag (referentiejaar)
typhh_huw_paar_mkind	Type huishouden: gehuwd paar met kinderen
vrouw	Indicator voor vrouw
typhh_huw_paar_zkind	Type huishouden: gehuwd paar zonder kinderen
zorgkosten_j0	Zorgkosten (referentiejaar)
baanuren_j0	Aantal gewerkte uren (referentiejaar)
flex_j2	Fractie van maanden in betreffende periode dat persoon werkzaam was met flex contract (referentiejaar - 2)
bijstand_duur	Aantal maanden geleden dat persoon bijstand had
inkbronn_hh_j1_loon	Loon als voornaamste inkomensbron van huishouden (referentiejaar - 1)
nkinderaf_j0	Fractie van maanden in betreffende periode dat aantal kinderen in huishouden afnam (referentiejaar)
eco_zelfst_j1_niet_doelpopulatie	Behoort niet tot de doelpopulatie economische zelfstandigheid
woning_j1_huurwon_z_htoeslag	Huurwoning zonder huurtoeslag (referentiejaar - 1)
rest_j2	Fractie van maanden in betreffende periode dat persoon werkzaam was als overig werknemer (referentiejaar - 2)
inkbronn_hh_j1_pensioen	Pensioen als voornaamste inkomensbron van huishouden (referentiejaar - 1)
ETNGROEP3PBL_3	Indicator voor niet-westerse migratieachtergrond
aantalkindhh	Aantal kinderen in huishouden
woning_j2_huurwon_z_htoeslag	Huurwoning zonder huurtoeslag (referentiejaar - 2)
secm_ov_zonder_ink	Belangrijkste inkomensbron van persoon: overig zonder inkomen (december van referentiejaar)
dpartner_duur	Aantal maanden sinds wijziging in partnerschap
plhh_ouder_eenouder	Plaats in huishouden: ouder in eenouderhuishouden
bijstand_s0	Fractie van maanden dat persoon bijstand had (laatste half jaar van referentiejaar)
bijstand_j0	Fractie van maanden dat persoon bijstand had (referentiejaar)
npartnereraf	Totaal aantal afnames in partnerschap in afgelopen drie jaar

4.2 Arme populatie - geheel Nederland

Deze populatie bestaat uit personen die in het referentiejaar arm zijn. Van deze populatie is 71% het daaropvolgende jaar arm.

Modelkwaliteit

De onderstaande tabellen tonen de kwaliteit van het model voor de arme populatie (testset $k = 1$; 121 duizend waarnemingen), en de bijbehorende confusiematrices. Zie paragraaf 3.2 voor methodes gebruikt voor de evaluatie van het model.

Volgens de AUC zijn de modelschattingen goed. Voor twee willekeurige personen, één die arm is gebleven en een die uit de armoede is gekomen, schat het model in 83 procent van de gevallen een hogere

armoedekans voor degene die arm is gebleven. In het geval van gokken zou dat in 50 procent van de gevallen zijn. 83 procent op een schaal van 0,5 tot 1 (AUC) komt overeen met een score van 67 procent op een schaal van 0 tot 1 (Somers' D).

De andere metrieken hangen af van de drempelwaarde waarboven een kans wordt geclassificeerd als 'blijft arm' (zie ook bijlage C). Bij de meest voor de hand liggende drempelwaarde van $c = 0,5$ is er een correlatie (MCC) van 47 procent tussen de geschatte en werkelijke klassen. De correlatie is maximaal 48 procent bij een iets hogere drempelwaarde van $c = 0,61$. De confusiematrix laat dan een goede balans zien tussen het aantal foutnegatieven en het aantal foutpositieven.

De F1 richt zich meer op de arme categorie en zoekt een balans tussen een zo hoog mogelijke recall (classificatiekans van mensen die arm blijven) en een zo hoog mogelijke precision (calibratiekans van mensen die als 'blijft arm' worden geclassificeerd). De F1 is maximaal 86 procent bij een iets lagere drempelwaarde van $c = 0,42$, omdat de recall harder stijgt dan de precision daalt. De confusiematrix laat zien dat bij deze metriek de goedpositieven en foutnegatieven zwaarder wegen dan de goednegatieven en foutpositieven (*better safe than sorry*).

Voor de vergelijking met deelpopulaties en de niet-arme populatie zijn de scores min-max genormaliseerd (kolom mmn in onderstaande tabel), met als minimum de score die je zou halen als je de categorie 'blijft arm' zou gokken met een kans gelijk aan de waargenomen fractie die arm blijft (hier $\frac{86.044}{121.473} = 0,71$). De F1 zakt dan richting MCC: 53 procent op de weg van beredeneerd gokken naar een foutloos, deterministisch model. De modellen leren dus bruikbare informatie uit de kenmerken. Het model kan zeker gebruikt worden om de relaties tussen kenmerken en armoede in $t + 1$ beter te begrijpen (zie volgende paragrafen) maar de kwaliteit van de schattingen is niet voldoende om op persoonsniveau te gebruiken.

Tabel 4.3 Kwaliteitsscores bij verschillende drempelwaardes volgens een aantal metrieken, voor en na min-max normalisatie (mmn).

Drempelwaarde	Metriek	Score	mmn
N.v.t.	AUC	0,83	0,67
$c = 0,5$	MCC	0,47	0,47
	F1	0,86	0,53
	- recall	0,91	0,69
	- precision	0,82	0,38
$c^*(MCC) = 0,61$	MCC(c^*)	0,48	0,48
	F1	0,85	0,48
	- recall	0,85	0,48
	- precision	0,85	0,48
$c^*(F_1) = 0,42$	MCC	0,45	0,45
	F1(c^*)	0,86	0,53
	- recall	0,94	0,80
	- precision	0,80	0,31

Tabel 4.4 Confusion matrices bij een aantal drempelwaardes.

Drempelwaarde	Werkelijk\Voorspeld	Blijft arm	Wordt niet-arm
$c = 0,5$	Blijft arm	78.300	7.744
	Wordt niet-arm	17.312	18.117
$c^*(MCC) = 0,61$	Blijft arm	72.869	13.175
	Wordt niet-arm	13.071	22.358
$c^*(F_1) = 0,42$	Blijft arm	80.937	5.107
	Wordt niet-arm	20.323	15.106

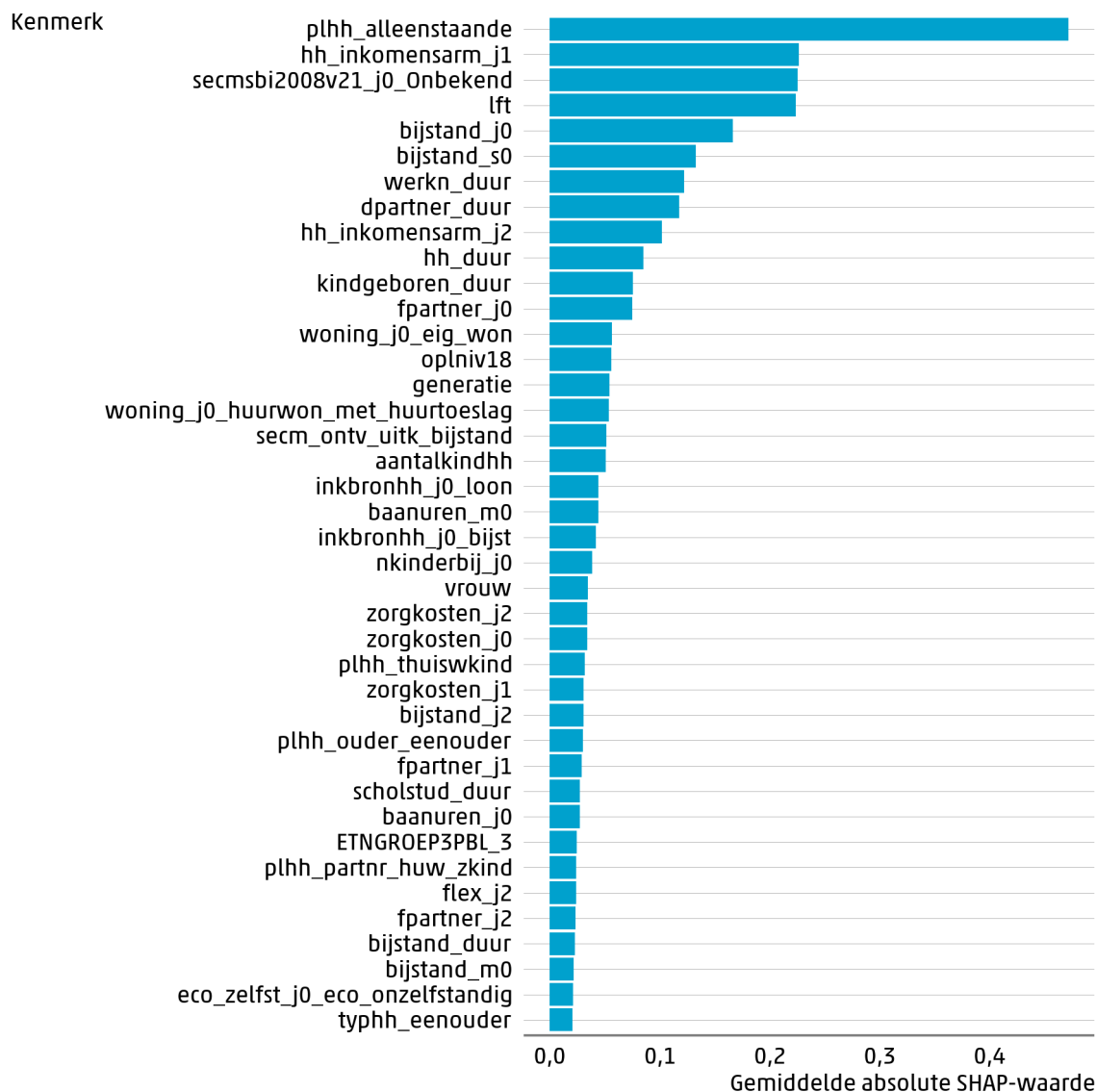
Bijlage C bevat nog twee aanvullende analyses die we in deze en de volgende alinea kort bespreken, één over het effect van de steekproefomvang en één over het effect van het aantal potentiële kenmerken op de kwaliteit van de geschatte kansen. In dit onderzoek hebben we de beschikking over kenmerken en doelvariabele over de gehele populatie. Dat heeft als voordeel dat het model complexere relaties kan vinden maar als nadeel dat het trainen rekenintensief is. In andere onderzoeken zijn de kenmerken en doelvariabele vaak alleen bekend voor een steekproef. De derde figuur in bijlage C laat het effect zien van de steekproefomvang op de kwaliteit. Een grotere steekproef leidt tot hogere gemiddelde kwaliteitsscores en kleinere variatie tussen testsets maar de positieve effecten vlakken snel af. Als rekencapaciteit beperkt is zou overwogen kunnen worden om een steekproef te trekken van bijvoorbeeld $n = 0,2N$. De analyse laat ook zien dat de machine-learningaanpak ook met steekproefdata zou kunnen werken.

In dit onderzoek is veel aandacht besteed aan het verzamelen en afleiden van kenmerken. Hoe meer kenmerken, hoe nauwkeuriger het model de relaties en interacties in kaart kan brengen, wanneer er voldoende personen in de dataset zijn. Nadelen zijn dat het veel werk is en dat het risico op overfitting toeneemt. De vierde figuur in bijlage C laat het effect op de kwaliteit zien van het aantal kenmerken dat wordt aangeboden. Ook hier nemen de gemiddelde kwaliteitsscores toe met het aantal kenmerken en volgt de wet van de verminderde meeropbrengst. Na de top-100 belangrijkste kenmerken is er vrijwel geen meeropbrengst meer. We weten echter van tevoren meestal niet hoe de kenmerken te sorteren op belang. Om het model hiervoor te gebruiken is het belangrijk dat het model niet gevoelig is voor irrelevante kenmerken. Somers' D blijft stijgen met het aantal kenmerken, MCC en F1 laten een minieme daling zien in de derde decimaal. Het model negeert dus irrelevante kenmerken en kan gebruikt worden om kenmerken te sorteren op belang.

Belangrijkste kenmerken

Onderstaande figuur geeft van de veertig belangrijkste kenmerken het belang weer volgens de SHAP-methode (zie paragraaf 3.3). Zie paragraaf 4.1 voor een omschrijving van deze kenmerken. De alleenstaande-indicator (`plhh_alleenstaande`) springt eruit, gevolgd door de indicator voor inkomensarmoede in $t - 1$ (`hh_inkomensarm_j1`), werkend-indicator (`secmsbi2008v21_j0_Onbekend`) en leeftijd (`lft`). `secmsbi2008v21_j0_Onbekend` geeft aan of de economische activiteit (SBI) bekend is; dit kenmerk scheidt de werkenden (werknemers, directeuren-grootaandeelhouders, zelfstandigen en meewerkende gezinsleden) van de niet-werkenden (mensen met een uitkering, studenten of mensen zonder inkomen). In de top-40 belangrijkste kenmerken staan verder enkele life events: het aantal maanden geleden dat iemand werknemer was (`werknemer_duur`) of ouder werd (`kindgeboren_duur`), veranderde van partnerschap (`dpartner_duur`) of van huishoud-ID (`hh_duur`).

De gemiddelde absolute SHAP-waarde voor de alleenstaande-indicator is 0,47. Dit betekent dat toevoegen van dit kenmerk aan het model de odds voor armoede gemiddeld met een factor $e^{0,47} = 1,60$ verandert. In de volgende paragraaf wordt ingezoomd op de verdeling van de individuele SHAP-waardes als functie van iemands score op het kenmerk. Dan wordt ook de richting (toe- of afname) duidelijk.



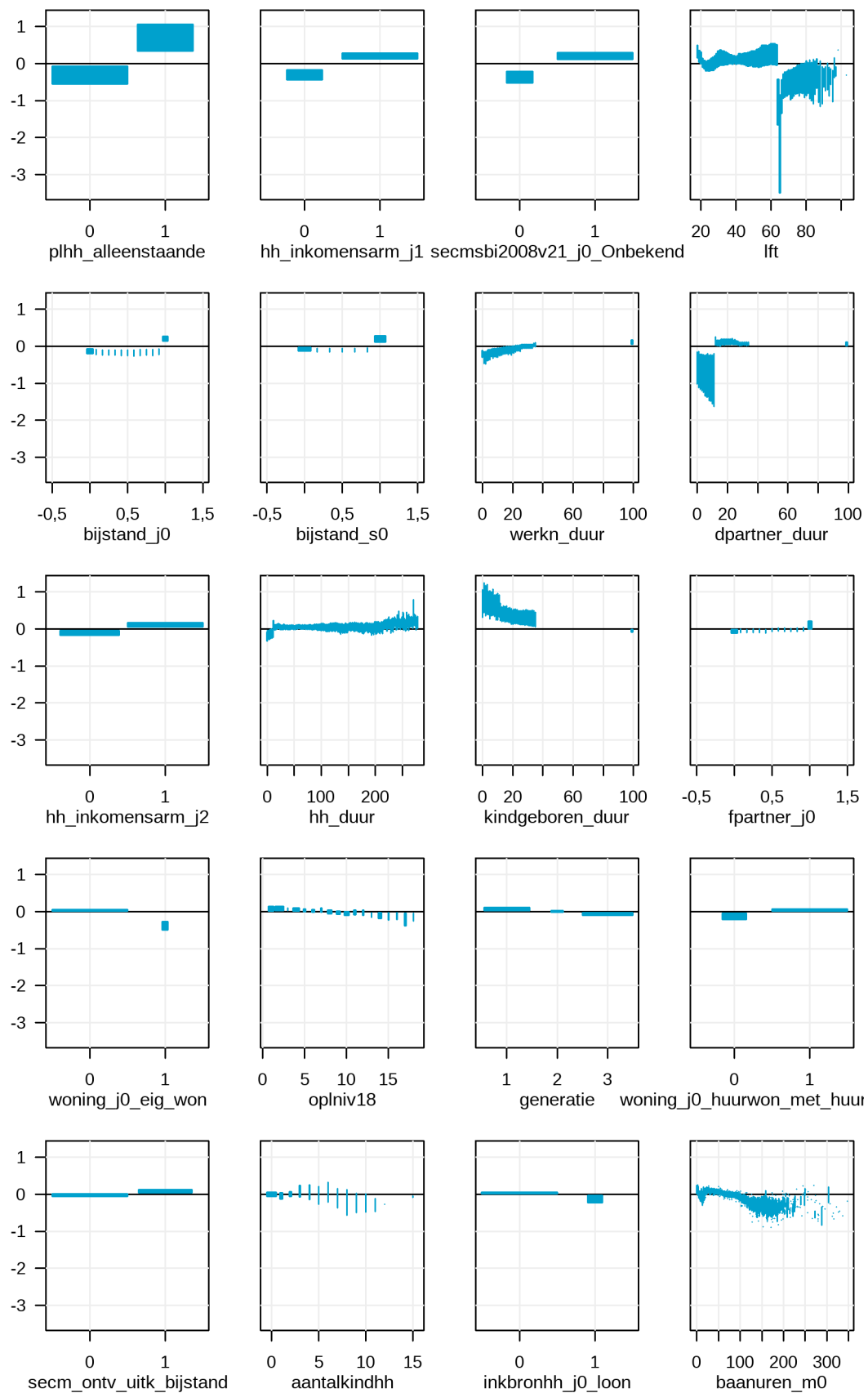
Effect van de 20 belangrijkste kenmerken op de geschatte kans

Zoals beschreven in paragraaf 3.3, geeft de hoogte van de balkjes het gebied aan waarin 80% van de betreffende populatie zich bevindt. De breedte komt overeen met de relatieve grootte van deze waarde van het kenmerk in de populatie. Een negatieve waarde op de y-as betekent dat personen met dit kenmerk volgens het model een lagere kans op armoede lopen dan gemiddeld. Een positieve waarde betekent dat personen een hogere kans lopen dan gemiddeld. Zoals beschreven in paragraaf 2.4 is bij de meeste duurkenmerken drie jaar terug gekeken: 0 t/m 35 maanden geleden; als er niets gebeurd is, is de duur van 36+ maanden gecodeerd met 99. Kenmerk x_{j0} betekent de fractie van het aantal maanden in het jaar t dat iemand tot categorie x behoorde (x_{s0} voor afgelopen semester, x_{m0} voor afgelopen maand, x_{j1} voor jaar $t - 1$, etc).

Het model schat een hogere kans om arm te blijven voor personen die alleenstaand zijn ($plhh_alleenstaande = 1$), vorig jaar ook al inkomensarm waren ($hh_inkomensarm_j1 = 1$), niet werken ($secmsbi2008v21_j0_Onbekend = 1$), net volwassen zijn ($lft = 18$) of tegen de pensioenleeftijd aan zitten ($lft = 63$ of 64), het hele jaar in de bijstand zaten ($bijstand_j0 = 1$), het afgelopen halfjaar in de bijstand zaten ($bijstand_s0 = 1$), tot een huishouden behoren waarin de afgelopen drie jaar een kind is geboren ($kindgeboren_duur < 36$), etc. Mensen met de pensioenleeftijd hebben volgens het model de kleinste kans om arm te blijven. Ouderen die na de pensioenleeftijd nog steeds arm zijn, een relatief kleine groep,

hebben ook een lagere kans om arm te blijven. Een andere potentieel beschermende factor is een recente verandering in partnerschap ($d_{\text{partner_duur}} < 12$). Mensen die relatief kort geleden een partner hebben gekregen of verloren ($d_{\text{partner_duur}}$) hebben een grotere kans om uit armoede te komen, terwijl mensen die langer geleden een partner hebben gekregen of verloren juist een grotere kans hebben om arm te blijven. Een mogelijke verklaring hiervoor is dat als het langer geleden is dat je een partner gekregen hebt en je bent nog steeds arm, dan heeft je partner je blijkbaar niet uit de armoede geholpen.

In Armoede en Sociale Uitsluiting (Van den Brakel and Otten, 2019) staat dat 1 op de 5 alleenstaanden tot de AOW-leeftijd een risico op inkomensarmoede lopen. Ons model bevestigt dat alleenstaande armen en arme mensen tot de AOW-leeftijd een hoger risico lopen. Het laat daarnaast zien hoe groot de variatie in armoedekansen is binnen alleenstaanden en mensen met dezelfde leeftijd, het kwantificeert de relatie tussen kenmerk en armoedekans, en vergelijkt het belang van deze kenmerken met andere mogelijke risicofactoren. Een wijziging in partnerschap kan van partner naar niet-partner zijn of andersom. Uit de vervolganalyse blijkt dat de beschermende factor meestal een wijziging is van alleenstaand zijn naar een partner hebbend. Bij verweduwning kan het inkomen en vermogen stijgen (Van den Brakel et al., 2021) en dus de kans op armoede verlagen. Dit zou bevestigd kunnen worden door de SHAP-waardes voor $d_{\text{partner_duur}}$ te bekijken voor verweduwden.

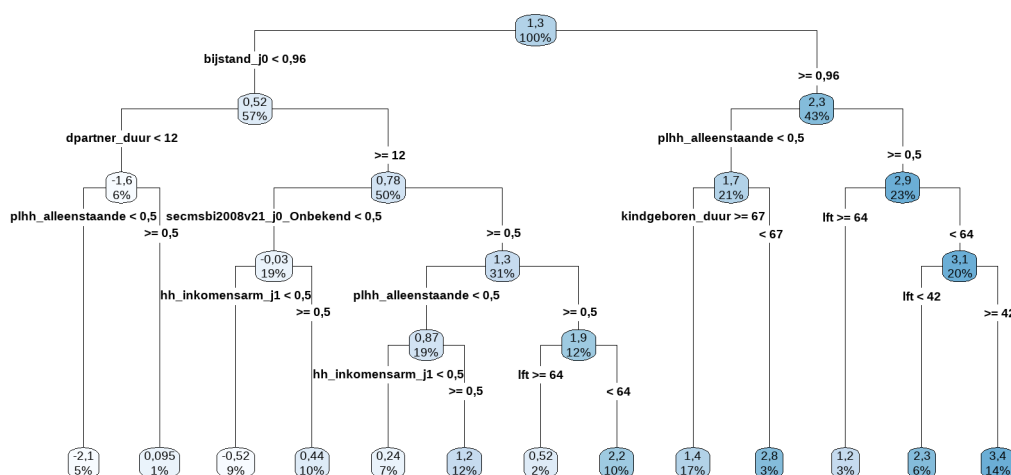


Deelpopulaties

Zoals uitgelegd in paragraaf 3.3 is met behulp van de 20 belangrijkste kenmerken een enkelvoudige beslisboom gefit op de (logit van de) door het machine-learningalgoritme geschatte kansen (zie onderstaande figuur). In elke knoop van de beslisboom wordt de gemiddelde armoedekans gegeven (op een logit-schaal, dus een logit van 1.3 betekent een gemiddelde armoedekans van $\frac{e^{1,3}}{1+e^{1,3}} = 0,79$), en het percentage van de populatie die in deze knoop zit.

Aan de hand van het vereenvoudigde model kunnen we de arme populatie indelen in groepen met relatief hoge en lage armoedekansen, om vervolgens binnen die deelpopulaties te kijken welke kenmerken daar een rol spelen. Het vereenvoudigde model verklaart ongeveer 43 procent van de totale variatie van het XGBoost-model. De bladeren (deelpopulaties) zijn genummerd van links naar rechts.

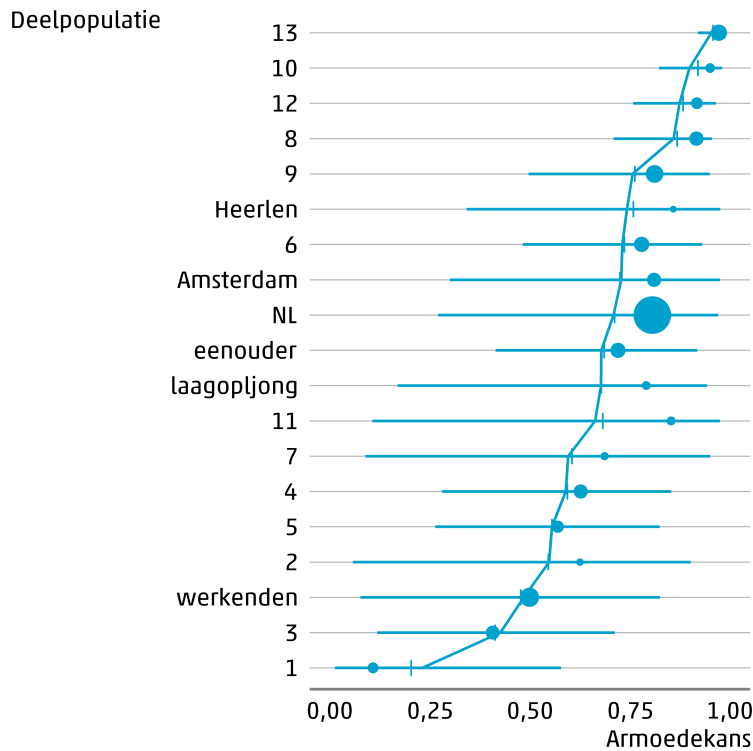
Naast deze datagedreven deelpopulaties worden nog een aantal expertgedreven deelpopulaties onderscheiden: Heerlen, Amsterdam, ouders in een eenouderhuishouden, laagopgeleide jongeren die niet thuiswonen en geen opleiding volgen, en werkenden. In de tabel 4.5 zijn de deelpopulaties gesorteerd van hoge naar lage waargenomen armoedefractie.



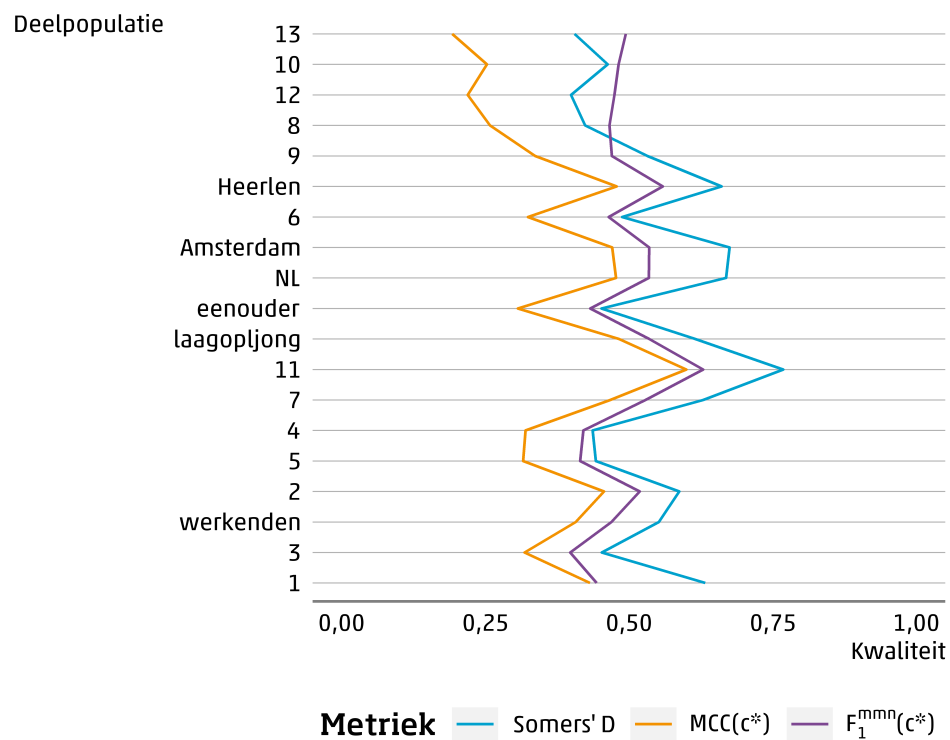
Tabel 4.5 Overzicht van de deelpopulaties voor de arme populatie.

Deelpopulatie	Beschrijving
13	hele jaar bijstand, alleenstaand, 42-63
10	hele jaar bijstand, niet alleenstaand, kind geboren minder dan 3 jaar geleden
12	hele jaar bijstand, alleenstaand, 18-41
8	niet hele jaar bijstand, partnerschap niet gewijzigd, niet werkend, alleenstaand, 18-63
9	hele jaar bijstand, niet alleenstaand, kind geboren ten minste 3 jaar geleden
Heerlen	inwoner gemeente Heerlen
6	niet hele jaar bijstand, partnerschap niet gewijzigd, niet werkend, niet alleenstaand, vorig jaar inkomensarm
Amsterdam	inwoner gemeente Amsterdam
eenouder	ouder in eenouderhuishouden
laagopljong	laag opleidingsniveau, niet thuiswonend, 18-26, volgt geen opleiding
11	hele jaar bijstand, alleenstaand, 64+
7	niet hele jaar bijstand, partnerschap niet gewijzigd, niet werkend, alleenstaand, 64+
4	niet hele jaar bijstand, partnerschap niet gewijzigd, werkend, vorig jaar inkomensarm
5	niet hele jaar bijstand, partnerschap niet gewijzigd, niet werkend, niet alleenstaand, vorig jaar niet inkomensarm
2	niet hele jaar bijstand, partnerschap gewijzigd, alleenstaand
werkenden	werknemer, groot aandeelhouder, zelfstandig ondernemer, overig zelfstandige of meewerkend gezinslid, 18-74
3	niet hele jaar bijstand, partnerschap niet gewijzigd, werkend, vorig jaar niet inkomensarm
1	niet hele jaar bijstand, partnerschap gewijzigd, niet alleenstaand

In onderstaande figuur staat per deelpopulatie de gemiddelde waargenomen armoedeindicator (lijn) en de verdeling van de geschatte armoedekans (gemiddelde (verticaal streepje), mediaan (punt) en 10e en 90e percentiel (horizontale streep)). Het gemiddelde van de geschatte armoedekans komt meestal goed overeen met het gemiddelde van de waargenomen armoedeindicator. De deelpopulatie met het laagste risico (1) bevat arme mensen die niet het hele jaar in de bijstand zitten, waarvan het partnerschap is gewijzigd en die niet alleenstaand zijn, c.q. partner zijn geworden (of eenouder maar in deelrapport blijkt dat dit zeldzaam is). Het model schat ook relatief lage kansen voor arme personen die niet het hele jaar in de bijstand zitten, waarvan het partnerschap niet gewijzigd is maar die wel werken en vorig jaar niet inkomensarm waren (deelpopulatie 3). Het model schat relatief hoge kansen voor arme personen die het hele jaar in de bijstand zitten, alleenstaand zijn en van middelbare leeftijd (deelpopulatie 13) of arme personen die het hele jaar in de bijstand zitten, niet alleenstaand zijn maar in een huishouden wonen waarin in de afgelopen drie jaar een kind is geboren (deelpopulatie 10). De spreiding in de geschatte kansen is opvallend klein in de deelpopulaties waar bijna iedereen arm blijft (13, 10, 12, 8, zie tabel 4.5).



Onderstaande figuur laat een aantal kwaliteitsmaten zien voor de verschillende deelpopulaties. Allen hebben een bereik van 0 tot 1 (eventueel door min-max normalisatie) en zijn daardoor vergelijkbaar tussen deelpopulaties, ondanks de verschillen in de waargenomen fractie die arm blijft. Alle drie de maten laten een vergelijkbaar beeld zien. Er is geen duidelijk verband met de grootte van de deelpopulatie of de waargenomen fractie. De kwaliteit voor de deelpopulatie met de grootste kans om uit de armoede te komen (1) is vergelijkbaar met de kwaliteit voor de gehele populatie. De kwaliteit is wel het laagst voor de deelpopulaties met de grootste kans om arm te blijven (13, 10, 12, 8). In deze deelpopulaties is de armoedekans hoog en is het moeilijk om gokken te verslaan.



4.3 Arme populatie - deelpopulatie met grootste kans om uit armoede te komen

Deze paragraaf zoomt in op de deelpopulatie met de grootste kans om uit armoede te komen:

- hoe is deze deelpopulatie afgebakend,
- in hoeverre verschilt deze deelpopulatie qua kenmerken van de hele arme populatie,
- hoe groot is de kans om arm te blijven,
- in hoeverre verschillen de kenmerken die het belangrijkste zijn voor het schatten van de armoedekans van het landelijke beeld,
- welk effect hebben deze kenmerken op de kans om arm te blijven.

Afbakening

- In referentiejaar in de armoede;
- In referentiejaar niet hele jaar bijstand;
- In referentiejaar partnerschap gewijzigd;
- In referentiejaar niet alleenstaand.

Deze deelpopulatie betreft personen die in referentiejaar t arm zijn en die de grootste kans hebben om uit de armoede te komen in het volgende jaar ($t + 1$). Deze subpopulatie bestaat uit niet-alleenstaande personen die niet het hele jaar bijstand hebben ontvangen en waarbij het partnerschap is gewijzigd in het afgelopen jaar. Dit betreft zowel personen die een partner gekregen hebben, als personen die een eenouderhuishouden hebben gevormd. Maar in de volgende figuren is te zien dat het voornamelijk personen betreft die een partner hebben gekregen.

Om beter zicht te krijgen op deze deelpopulatie hebben we gekeken naar een aantal achtergrondkenmerken:

- leeftijd (lft)
- opleidingsniveau (oplniv3)
- generatie migratieachtergrond
 - eerste generatie migratieachtergrond (generatie = 1)
 - tweede generatie (generatie = 2)
 - Nederlandse achtergrond (generatie = 3)
- geslacht (vrouw: vrouw = 1)
- type huishouden
 - alleenstaande (plhh_alleenstaande = 1)
 - thuiswonend kind (plhh_thuiswkind = 1)
 - ongehuwd partner zonder kinderen (plhh_partn_zkind = 1)
 - ongehuwd partner met kinderen (plhh_partn_mkind = 1)
 - gehuwd partner zonder kinderen (plhh_huw_zkind = 1)
 - gehuwd partner met kinderen (plhh_huw_mkind = 1)
 - ouder in eenouderhuishouden (plhh_ouder_eenouder = 1)
 - bijstand (bijstand_j0)
 - duur sinds wijziging in partnerschap (dpartner_duur).

Onderstaande figuren tonen de verdeling van deze kenmerken in deze deelpopulatie (blauw) in vergelijking met de gehele arme populatie (grijs). Merk op dat het opleidingsniveau deels op registraties en deels op enquêtes gebaseerd is, waardoor dit niet voor de gehele populatie bekend is. Hierdoor is de groep met een onbekend opleidingsniveau relatief groot. Ook is dit een selectieve groep. Zo is het opleidingsniveau relatief vaak onbekend voor ouderen en personen met een migratieachtergrond. Dit kan leiden tot vertekening in

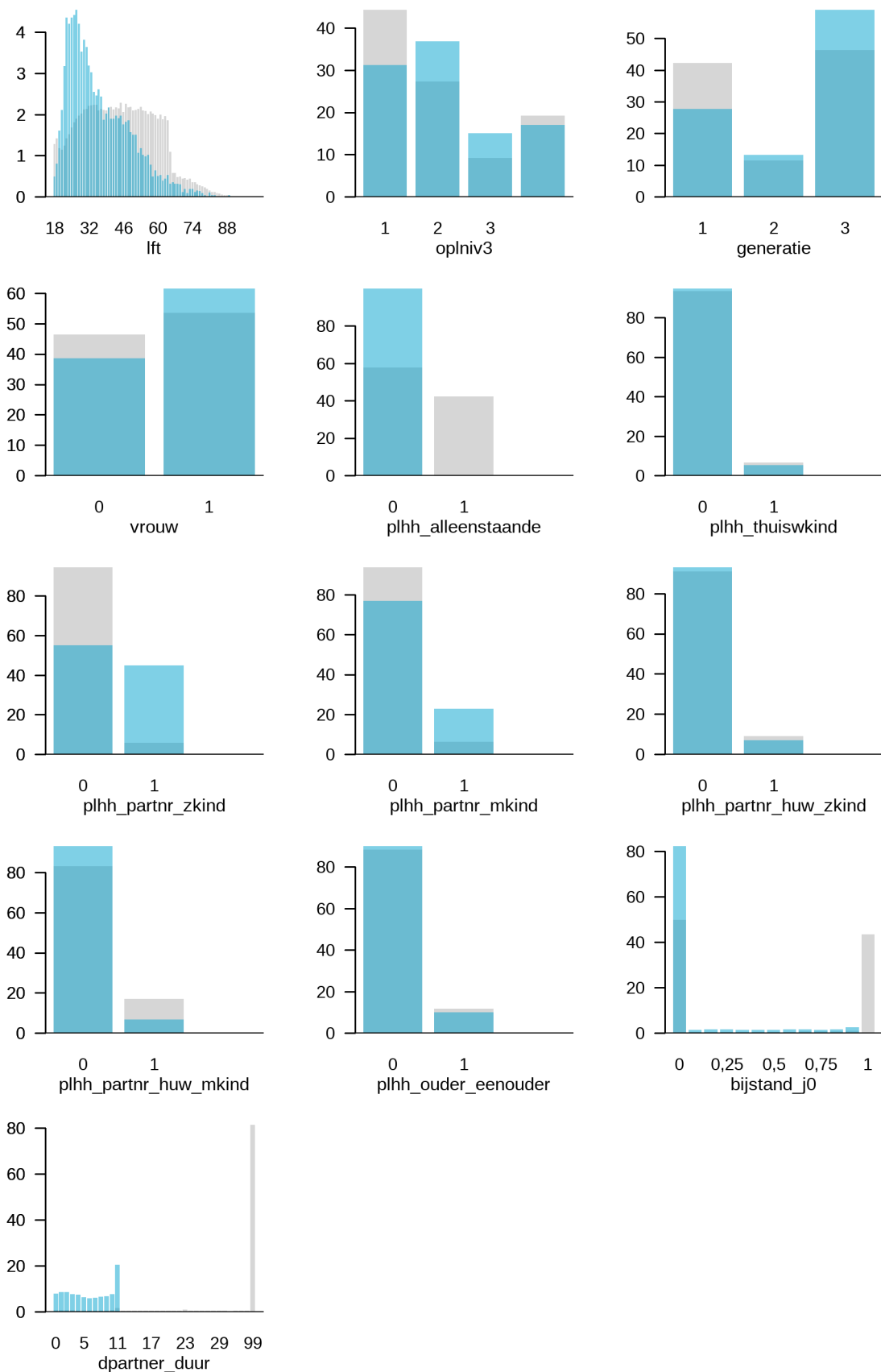
de opleidingsniveauverdeling. Daarom laten we dit buiten beschouwing bij het beschrijven van de populatiekenmerken.

In onderstaande figuren is te zien dat er in deze deelpopulatie met name relatief meer jongeren zijn ($20 < 1ft < 40$), meer personen met een Nederlandse achtergrond zijn ($generatie = 3$) en meer vrouwen ($vrouw = 1$). Dat er relatief meer vrouwen in deze deelpopulatie zitten is in lijn met ander CBS-onderzoek dat naar verwachting eind 2021 verschijnt (van den Brakel et al., in voorbereiding). Daaruit blijkt onder meer dat gescheiden vrouwen er in koopkracht meer op vooruit gaan dan gescheiden mannen die een nieuwe partner vinden.

Ook is te zien dat relatief veel ongehuwde partners voorkomen (zowel zonder kinderen: $plhh_partner_zkind = 1$ als met kinderen: $plhh_partner_mkind = 1$) en relatief weinig gehuwde partners met kinderen: $plhh_partner_huw_mkind = 1$. Daarbij blijkt dat deze deelpopulatie voornamelijk bestaat uit mensen die partner in het huishouden zijn; slechts een kleine groep betreft eenouders ($plhh_ouder_eenouder = 1$) en thuiswonende kinderen ($plhh_thuiswkind$). Dus het kenmerk gewijzigd partnerschap betreft met name personen die een partner gekregen hebben. Wanneer een persoon een partner krijgt is te verwachten dat de financiële situatie daardoor verbetert door lagere lasten en/of meer inkomen. Dit maakt het plausibel dat deze personen een hogere kans hebben om uit de armoede te komen.

Verder is te zien dat de ruime meerderheid van de personen geen enkele maand in het jaar een bijstandsuitkering heeft ontvangen ($bijstand_j0 = 0$). Dit ligt in lijn der verwachting, aangezien de personen die geen bijstand ontvangen vaker financieel sterker staan en daardoor een grotere kans hebben om uit de armoede te komen.

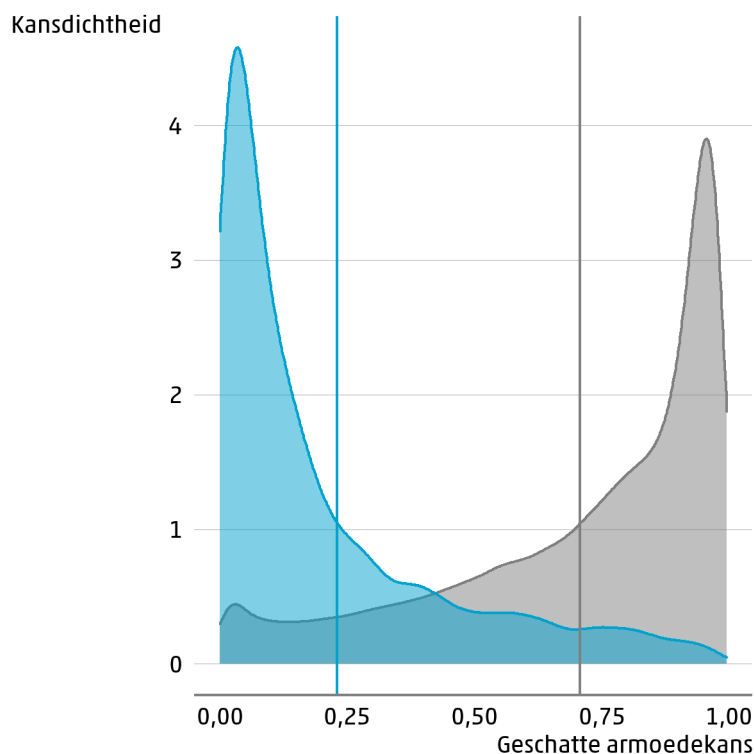
In de figuur over wijziging in partnerschap in het afgelopen jaar ($dpartner_duur$) is een kleine piek te zien bij 11 maanden. Dit wordt veroorzaakt door de samenstelling van de data om de plaats in het huishouden van een persoon te bepalen. Voor een deel van Nederland is het niet bekend wat voor plaats in het huishouden iemand heeft. Deze gegevens worden vervolgens geschat op basis van achtergrondkenmerken om toch een beeld te krijgen van de plaats in het huishouden van een persoon. Een limitatie van deze methode is echter dat rond de jaarovergang, de schatting weer kan veranderen en er daarom relatief meer partnerwijzigingen zijn rondom de jaarovergang. Deze cijfers moeten dus met voorzichtigheid geïnterpreteerd worden.



Kans om arm te blijven

Onderstaande figuur toont de verdeling van de geschatte kans om arm te blijven voor deze deelpopulatie (blauw) en voor de gehele arme populatie in Nederland (grijs). De verticale lijnen tonen de gemiddelde kans

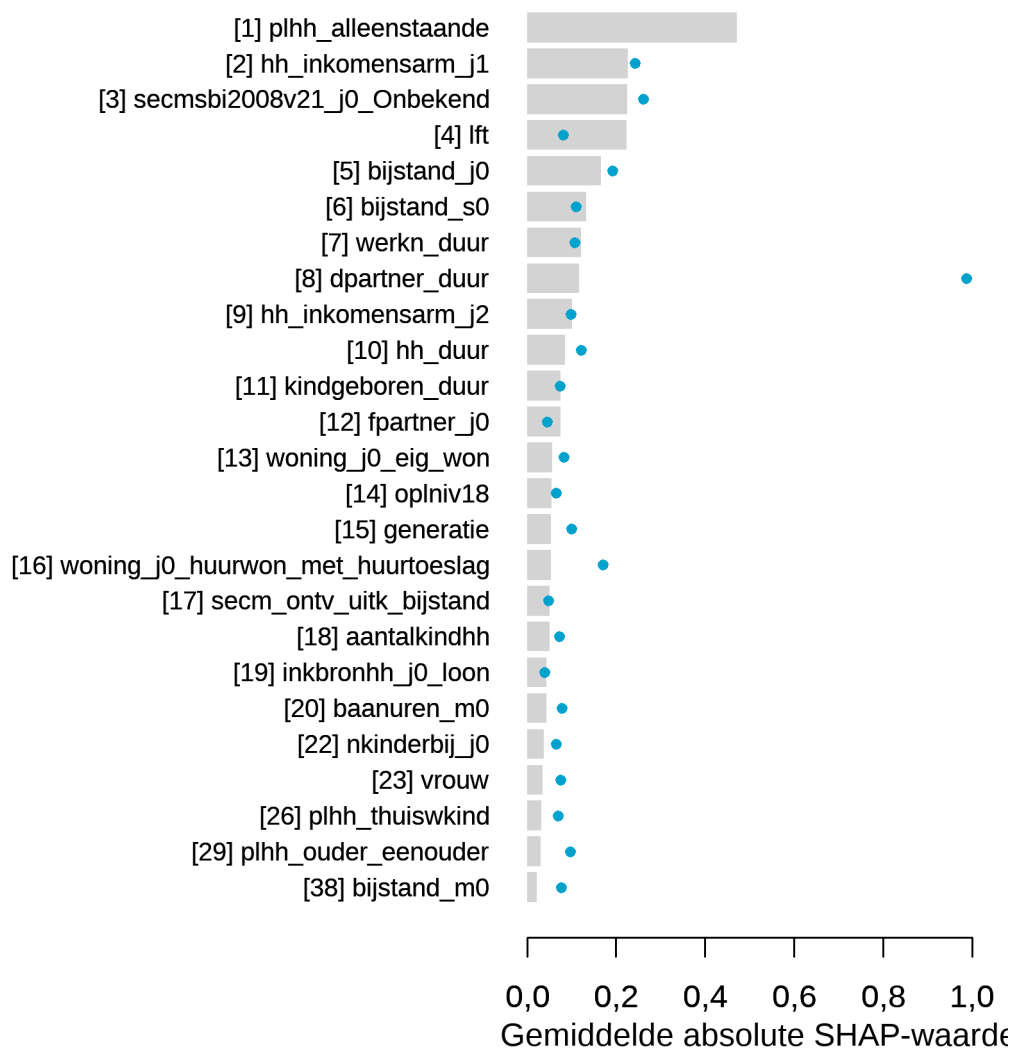
om arm te blijven. De gemiddelde kans in deze deelpopulatie is met 23% een stuk kleiner dan in de gehele arme populatie in Nederland (71%).



Belangrijkste kenmerken

Onderstaande figuur geeft de belangrijkste kenmerken weer op basis van SHAP-waarden. De 20 belangrijkste kenmerken in de gehele arme populatie (grijze balken) zijn aangevuld met de 20 belangrijkste kenmerken voor de deelpopulatie (blauwe punten). Bij kenmerken waarbij iedereen in de deelpopulatie dezelfde waarde heeft, ontbreekt de blauwe punt, aangezien dit kenmerk geen effect heeft binnen de deelpopulatie. (Zie paragraaf 3.3 voor verdere toelichting.)

In de figuur is te zien dat de belangrijkste kenmerken voor deze deelpopulatie redelijk overeenkomen met de gehele arme populatie. (Zie paragraaf 4.1 voor een omschrijving van deze kenmerken.) Wel is te zien dat de blauwe punten voor een aantal kenmerken in meer of mindere mate afwijkt van de grijze balken. Hieronder gaan we in op een aantal kenmerken die *belangrijker* zijn dan voor de Nederlandse arme populatie, dus waarvoor de blauwe punt rechts van de grijze balk ligt. Dit is met name het geval voor het aantal maanden sinds wijziging in partnerschap (`dpartner_duur`) en wonen in een huurwoning met huurtoeslag (`woning_j0_huurwon_met_huurtoeslag`). In de volgende figuren bekijken we hoe deze kenmerken de kans om arm te blijven beïnvloeden.



Effect op kans om arm te blijven

Onderstaande figuren tonen het *effect* van belangrijke kenmerken op de kans om arm te blijven en de *decompositie* van het verschil in belang tussen deelpopulatie en gehele populatie. De staafdiagrammen tonen het effect van de 20 belangrijkste kenmerken op de geschatte kans om arm te blijven voor de deelpopulatie (blauwe balken) en de gehele niet-arme populatie (grijze balken). Een negatieve waarde op de y-as betekent dat personen met dit kenmerk een lagere kans dan gemiddeld hebben om arm te worden, een positieve waarde betekent dat personen een hogere kans dan gemiddeld hebben. De hoogte van de balk geeft de range van de kans weer voor 80% van de betreffende populatie. De breedte van de balk komt overeen met de relatieve grootte van de populatie met deze waarde. Een smalle balk betekent dus relatief weinig mensen en een brede balk relatief veel mensen. Zie paragraaf 3.3 voor verdere toelichting.

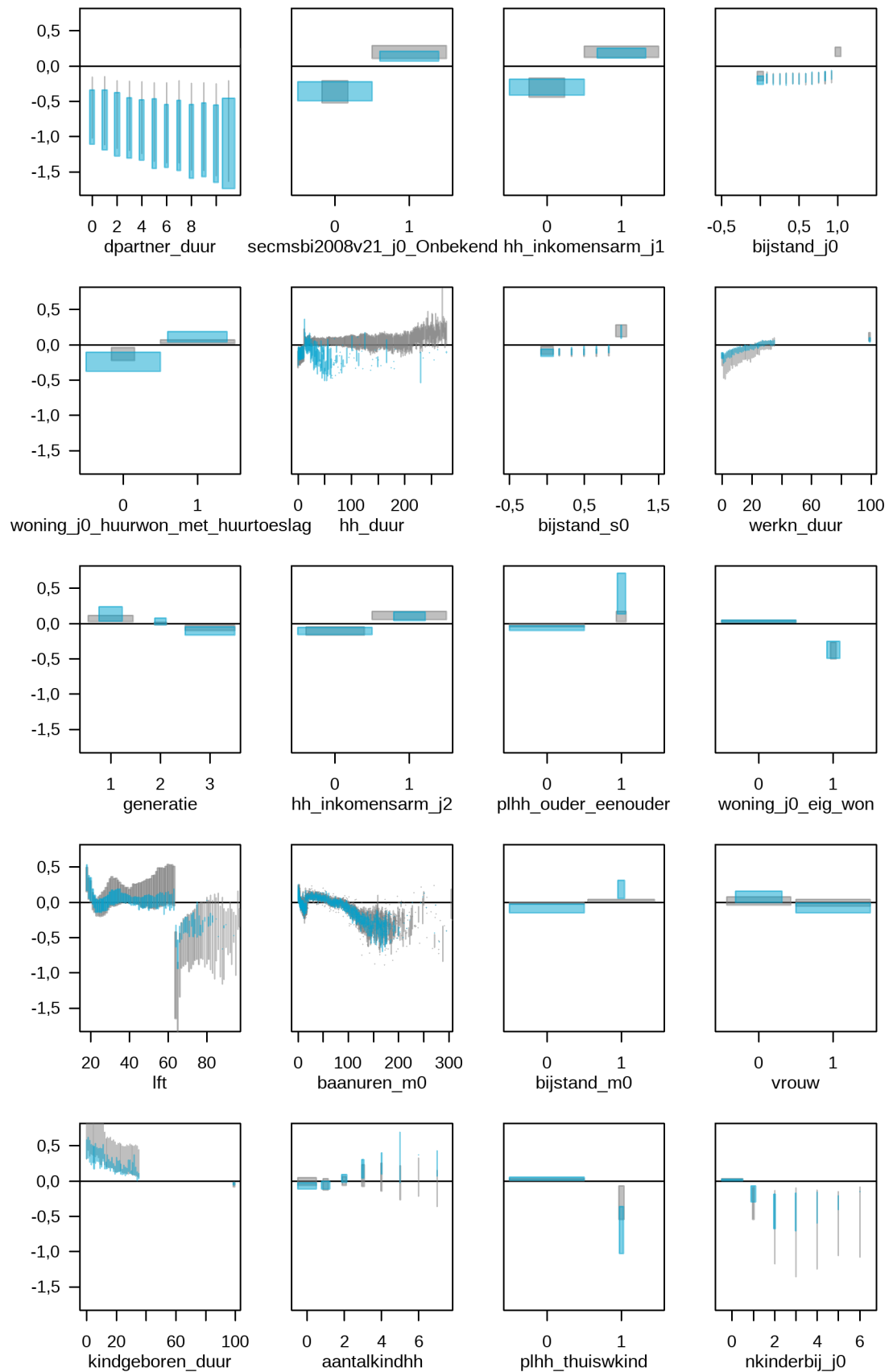
De tweede figuur toont de decompositie van het verschil in belang van de belangrijkste kenmerken tussen de deelpopulatie en de gehele populatie. Het verschil in belang wordt opgesplitst in een component die het belang van de *verdeling* van het kenmerk weergeeft (horizontale as) en een component die het belang weergeeft van het *effect* dat het kenmerk heeft op de geschatte kans (verticale as). Zie paragraaf 3.3 voor verdere toelichting.

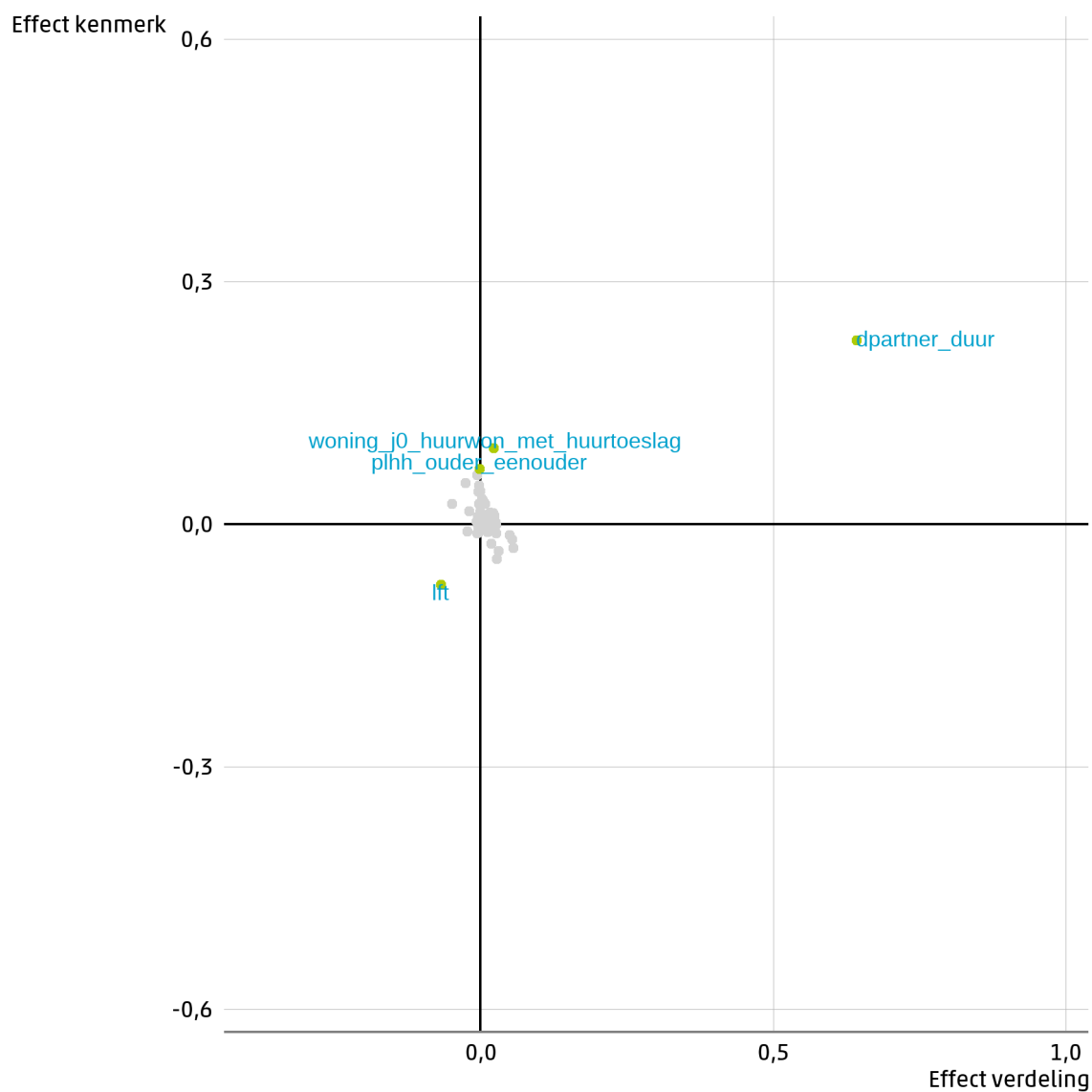
Merk verder op dat bij de meeste duurkenmerken drie jaar terug is gekeken: 0 t/m 35 maanden geleden. Als er niets gebeurd is, is de duur van 36+ maanden gecodeerd met 99. Omdat in deze deelpopulatie geselecteerd is op wijziging in partnerschap in het afgelopen jaar (`dpartner_duur < 12`), is de duur hier afgekapt op 12 maanden.

In deze deelpopulatie is iedereen per definitie afgelopen jaar van partnerschap gewijzigd. Dit is terug te zien in de brede blauwe balken bij `dpartner_duur` en het grote positieve verdelingseffect in de decompositiefiguur. Maar personen die het afgelopen jaar van partnerschap zijn gewijzigd hebben ook vaker een grotere kans om uit de armoede te komen dan in de gehele arme populatie. De blauwe balken voor `dpartner_duur` liggen namelijk lager dan de grijze. Ook dit verschil is terug te zien in het positieve kenmerkeffect in de decompositiefiguur.

Verder laat de decompositie zien dat het grotere belang van huurtoeslag (`woning_j0_huurwon_met_huurtoeslag`) voornamelijk is toe te schrijven aan een groter effect van het kenmerk (positief kenmerkeffect). De blauwe balken liggen inderdaad verder van nul af dan de grijze. In deze deelpopulatie hebben mensen zonder huurwoning over het algemeen een nog kleinere armoedekans dan in de gehele populatie, en mensen met huurwoning een nog grotere. Er is ook een klein positief verdelingseffect: deze deelpopulatie bevat meer mensen zonder huurwoning (sterk effect op armoedekans) en minder mensen met huurwoning (zwak effect op armoedekans) dan de gehele populatie.

De indicator voor ouder in eenouderhuishouden (`p1hh_ouder_eenouder`) is in deze deelpopulatie belangrijker dan in de gehele arme populatie omdat ouders in een eenouderhuishouden in deze deelpopulatie een nog grotere armoedekans hebben en de andere categorie een nog kleinere armoedekans (positief kenmerkeffect).





4.4 Niet-arme populatie - geheel Nederland

Deze populatie bestaat uit personen die in het referentiejaar niet arm zijn. Van deze populatie is 1,2% het daaropvolgende jaar arm.

Modelkwaliteit

De onderstaande tabellen tonen de kwaliteit van het model voor de niet-arme populatie. Deze is geëvalueerd op een testset van waarnemingen die niet gebruikt zijn voor het schatten van het model. Deze dataset bevat 10 procent van de waarnemingen (testset bevat 2.387.499 waarnemingen). Zie paragraaf 3.2 voor methodes gebruikt voor de evaluatie van het model.

Het model heeft een AUC van 0,95, wat een erg hoge waarde is. Deze score ligt tussen 0,5 en 1 waarbij een waarde van 1 aangeeft dat het model perfect in staat is onderscheid te maken tussen personen. Een waarde van 0,95 geeft aan dat het model voor twee willekeurige personen waarvan één arm is geworden en één niet, in 95 procent van de gevallen een hogere kans schat voor degene die arm is geworden.

Tabel 4.6 Kwaliteitsscores bij verschillende drempelwaardes volgens een aantal metriecken, voor en na min-max normalisatie (mmn).

Drempelwaarde	Metriek	Score	mmn
N.v.t	AUC	0,95	0,90
$c = 0,5$	MCC	0,38	0,38
	F1	0,32	0,32
	- recall	0,21	0,20
	- precision	0,70	0,70
	MCC(c^*)	0,42	0,42
$c^*(MCC) = 0,23$	F1	0,43	0,42
	- recall	0,39	0,38
	- precision	0,46	0,46
	MCC	0,42	0,42
	F1(c^*)	0,43	0,42
$c^*(F_1) = 0,19$	- recall	0,44	0,43
	- precision	0,41	0,40

Tabel 4.7 Confusion matrices bij een aantal drempelwaardes.

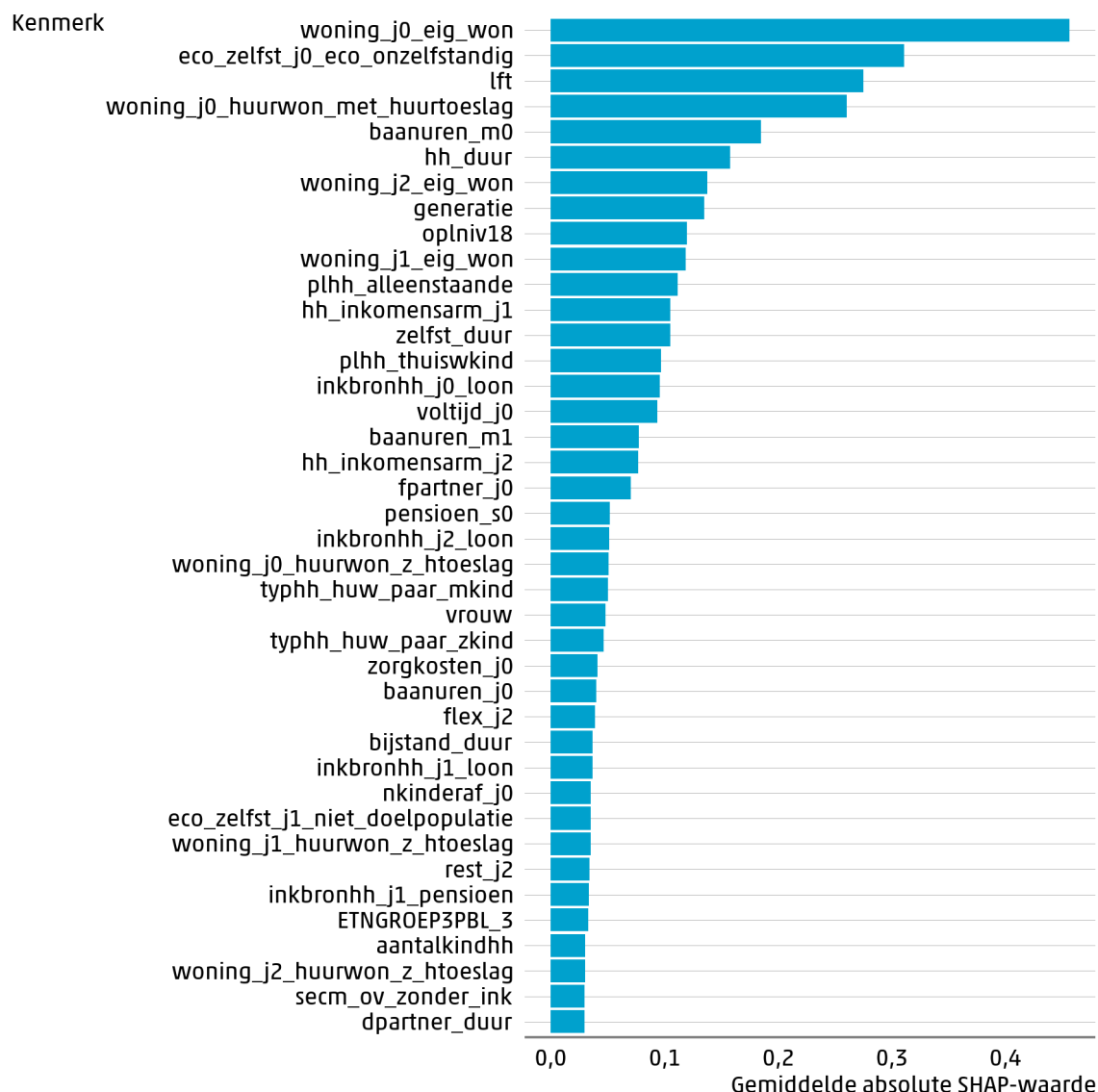
Drempelwaarde	Werkelijk\Voorspeld	Wordt arm	Blijft niet-arm
$c = 0,5$	Wordt arm	6.507	24.679
	Blijft niet-arm	2.782	2.353.531
$c^*(MCC) = 0,23$	Wordt arm	12.229	18.957
	Blijft niet-arm	14.133	2.342.180
$c^*(F_1) = 0,19$	Wordt arm	13.791	17.395
	Blijft niet-arm	19.758	2.336.555

De maten gebaseerd op de confusiematrix zijn lager. De MCC is 0,38 voor een drempel van 0,5 en 0,42 bij de optimale drempel van 0,23. De F1-score is 0,32 bij een drempel van 0,5 en 0,43 bij de optimale drempel van 0,19. Deze twee maten geven aan in welke mate het model gebruikt kan worden om personen te classificeren. Dus om op individueel niveau te voorspellen of iemand wel of niet arm wordt. Waarschijnlijk komt het verschil tussen de hoge AUC en wat lagere F1-score doordat er een redelijk grote groep is waarvan we redelijk zeker weten dat deze niet arm worden. Voor de groep die overblijft is het lastiger om op individueel niveau onderscheid te maken tussen personen die wel en die niet arm worden. Op zich is dat geen probleem. Stel we hebben een groep die 50 procent kans heeft om in de armoede te komen. Op individueel niveau zullen we met het model maar in 50 procent van de gevallen correct voorspellen dat

iemand arm wordt. Dit is echter nog steeds een groep met een erg hoog risico waarvoor het interessant kan zijn om beleid voor te ontwikkelen.

Belangrijkste kenmerken

Onderstaande figuur geeft van de veertig belangrijkste kenmerken het belang weer. Zie paragraaf 4.1 voor een omschrijving van deze kenmerken. Er staan meerdere kenmerken gerelateerd aan woningbezit in de top veertig. Samenstelling en plaats in huishouden (*plhh* en *typhh*) en veranderingen (life-events) in het huishouden (*hh_duur*, *fpartner*, *nkinderaf*) komen ook regelmatig voor in de lijst.



Effect van de 20 belangrijkste kenmerken op de geschatte kans

De hoogte van de balkjes geeft het gebied aan waarin 80% van de betreffende populatie zich bevindt. De breedte komt overeen met de relatieve grootte van deze waarde van de dit kenmerk in de populatie. Een negatieve waarde op de y-as betekent dat personen met dit kenmerk een lagere kans lopen dan gemiddeld. Een positieve waarde betekent dat personen een hogere kans dan gemiddeld lopen. Voor meer informatie zie paragraaf 3.3.

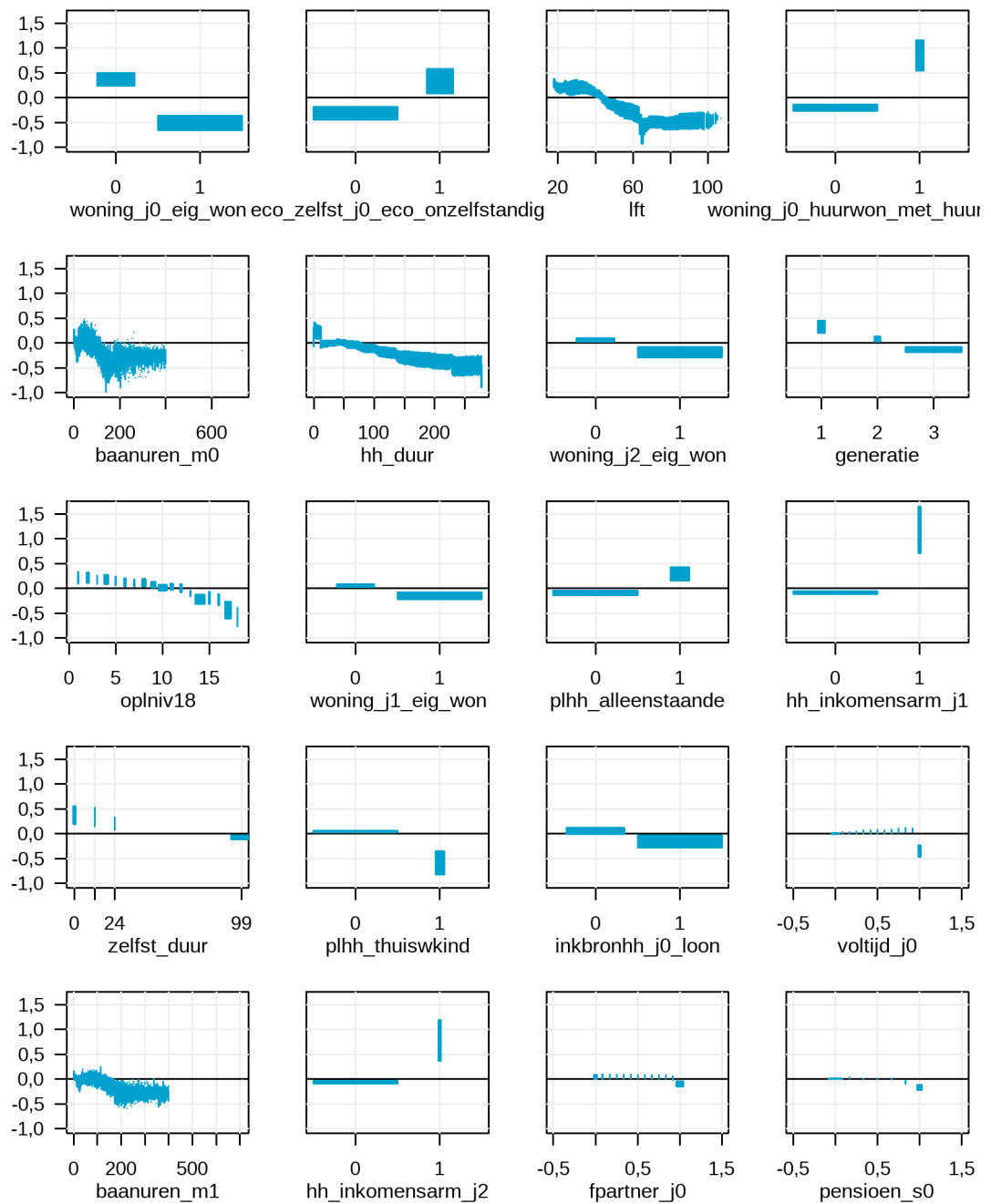
Bij de meeste duurkenmerken is drie jaar terug gekeken: 0 t/m 35 maanden geleden; als er niets gebeurd is, is de duur van 36+ maanden gecodeerd met 99. Kenmerk *x_j0* betekent de fractie van het aantal maanden

in het jaar t dat iemand tot categorie x behoorde (x_{s0} voor afgelopen semester, x_{m0} voor afgelopen maand, x_{j1} voor jaar $t - 1$, etc. Zie paragraaf 4.1 en 2.4 voor meer informatie over de kenmerken.

Het hebben van een eigen woning ($woning_{j0_eig_won} = 1$) is een belangrijke voorspeller voor armoede: mensen met een eigen woning lopen een lager risico arm te worden. Dit zou kunnen komen doordat eigenwoningbezit samenhangt met met (langdurige) arbeidshistorie en baan zekerheid: als iemand al langere tijd een lager inkomen heeft of onzekerheid heeft in zijn baan, is de kans groter dat deze geen eigen woning heeft. Analooch hebben mensen die in een woning met huurtoeslag wonen een grotere armoedekans. Dit wil echter niet zeggen, dat dit een causaal verband is. Een huurwoning zal eerder een gevolg dan een oorzaak van armoede zijn.

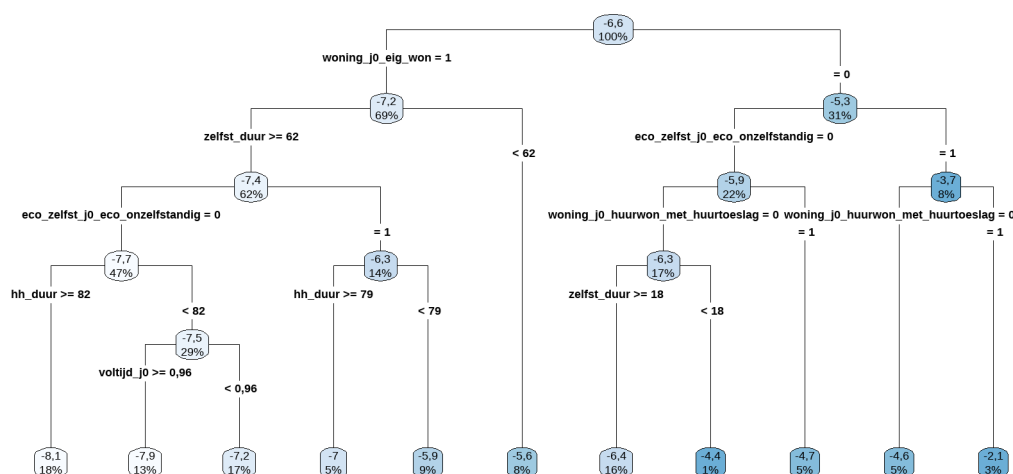
Met de variabelen over arbeidshistorie in het model wordt 'slechts' drie jaar teruggekeken. Mensen die economisch onzelfstandig zijn ($eco_zelfst_{j0_eco_onzelfstandig} = 1$) lopen ook een groter risico. Deze zijn (ook) afhankelijk van het inkomen van een partner of andere huisgenoot, of van een uitkering. Als dit wegvalt lopen zij een groter risico op armoede. De kans om arm te worden neemt af met leeftijd en is het grootst tussen de 20 en de 40. Bij deze groep is de kans om arm te blijven wel kleiner (zie paragraaf 4.2) dan bij anderen tussen de 20 en 65. Net als bij de arme populatie neemt de kans om arm te worden sterk af vlak voor de pensioengerechtigde leeftijd in verband met het gaan ontvangen van AOW en aanvullend pensioen. De kale AOW ligt boven de lage-inkomensgrens. Mogelijk speelt hier ook mee dat bij verweduwing het inkomen en vermogen kan stijgen (Van den Brakel et al., 2021), waardoor de kans op armoede afneemt.

Daarnaast zijn er ook een aantal life-events die de kans om arm te worden beïnvloeden: recente veranderingen in het huishouden en recent zelfstandige geweest of geworden zijn verhogen de kans op armoede.



Deelpopulaties

Zoals uitgelegd in paragraaf 3.3 is met behulp van de 20 belangrijkste kenmerken een enkelvoudige beslisboom gefit op de (logit van de) door het machine-learningalgoritme geschatte kansen (zie onderstaande figuur). Het vereenvoudigde model verklaard ongeveer 56,03% van de totale variatie van het XGBoost-model. In elke knoop van de beslisboom wordt de gemiddelde armoedekans gegeven (op een logit-schaal, dus een logit van 1.3 betekent een gemiddelde armoedekans van $\frac{e^{1.3}}{1+e^{1.3}} = 0.79$), en het percentage van de populatie die in deze knoop zit.



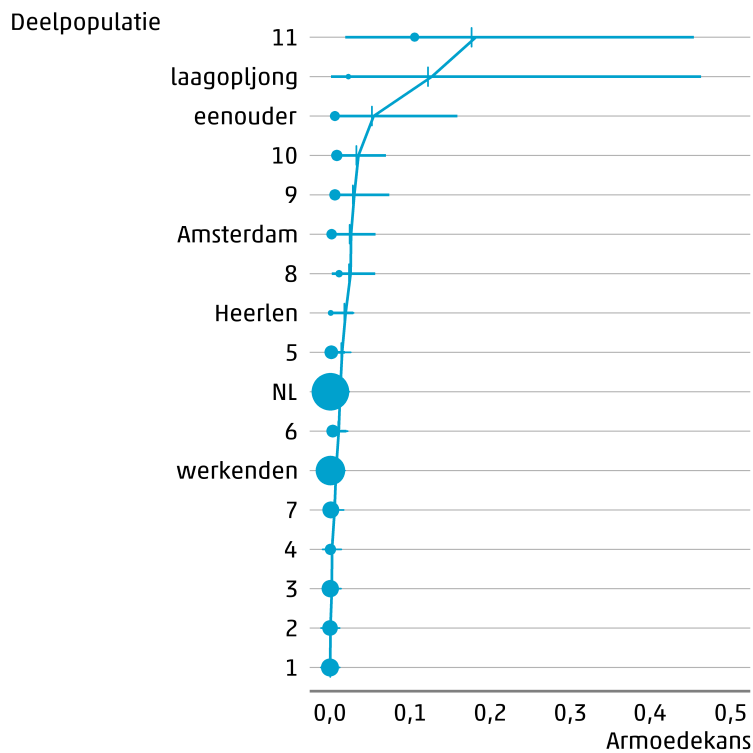
Aan de hand van het vereenvoudigde model kunnen we de populatie indelen in deelpopulaties om vervolgens binnen die deelpopulaties te kijken welke kenmerken daar een rol spelen. De deelpopulatie met het hoogste risico bevat bijvoorbeeld mensen die in een huurwoning wonen, huurtoeslag ontvangen en niet economisch zelfstandig zijn. Als we inzoomen op deze populatie kunnen we zien welke kenmerken nog meer een rol spelen in deze populatie. Naast de deelpopulaties zoals uit het simpele model volgen worden ook nog de volgende (deels overlappende) deelpopulaties onderscheiden: Amsterdam, Heerlen, laagopgeleide jongeren, werkenden en ouders in een eenouderhuishouden. Tabel 4.8 geeft een overzicht van de deelpopulaties.

Tabel 4.8 Overzicht van de deelpopulaties voor de niet-arme populatie.

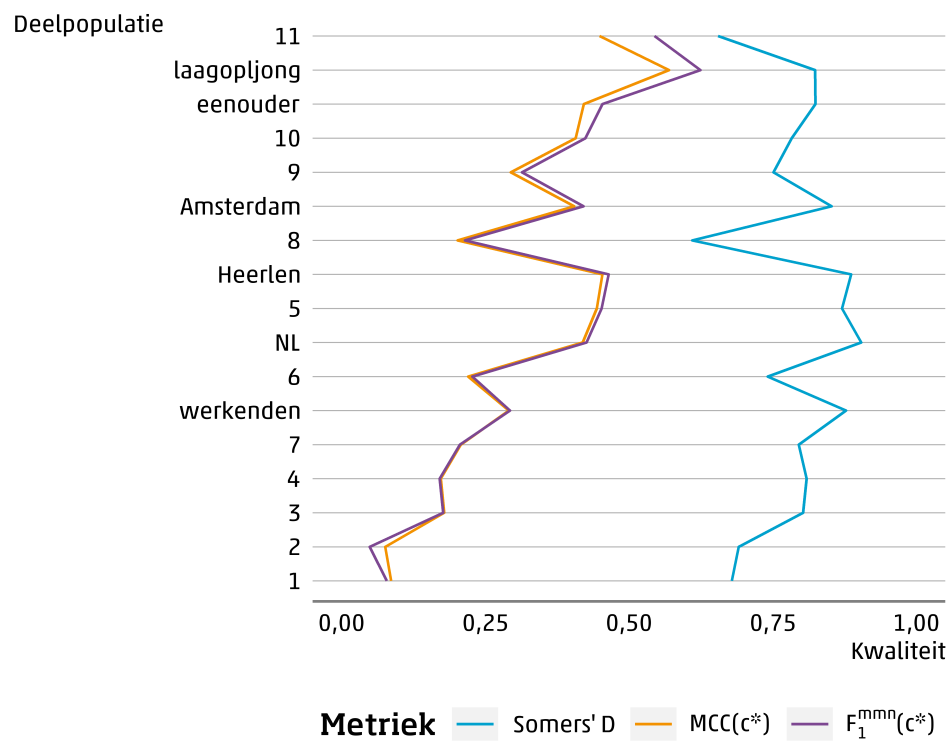
Deelpopulatie	Beschrijving
11 laagopljong eenouder	Geen eigen woning, economisch onzelfstandig, huurwoning met huurtoeslag Laag opleidingsniveau, niet thuiswonend, 18-26, volgt geen opleiding Ouder in eenouderhuishouden
10 9 Amsterdam	Geen eigen woning, economisch onzelfstandig, geen huurwoning met huurtoeslag Geen eigen woning, economisch zelfstandig, huurwoning met huurtoeslag Inwoner gemeente Amsterdam
8	Geen eigen woning, economisch zelfstandig, geen huurwoning met huurtoeslag, zelfstandig ondernemer of zelfstandig ondernemer geweest in de afgelopen 1,5 jaar
Heerlen	Inwoner gemeente Heerlen
5	Eigen woning, geen zelfstandig ondernemer of langer dan 3 jaar geleden zelfstandig ondernemer, economisch onzelfstandig, woont in een huishouden dat korter dan 79 maanden bestaat
6	Eigen woning, zelfstandig ondernemer of zelfstandig ondernemer geweest in de afgelopen 3 jaar
werkenden	Werknemer, groot aandeelhouder, zelfstandig ondernemer, overig zelfstandige of meewerkend gezinslid, 18-74
7	Geen eigen woning, economisch zelfstandig, geen huurwoning met huurtoeslag, geen zelfstandig ondernemer of ergens tussen de 1,5 en 3 jaar geleden zelfstandig ondernemer geweest
4	Eigen woning, geen zelfstandig ondernemer of langer dan 3 jaar geleden zelfstandig ondernemer, economisch onzelfstandig, woont in een huishouden dat langer dan 79 maanden bestaat
3	Eigen woning, geen zelfstandig ondernemer of langer dan 3 jaar geleden zelfstandig ondernemer, economisch zelfstandig, woont in een huishouden dat minder dan 82 maanden bestaat, niet hele jaar voltijd baan
2	Eigen woning, geen zelfstandig ondernemer of langer dan 3 jaar geleden zelfstandig ondernemer, economisch zelfstandig, woont in een huishouden dat minder dan 82 maanden bestaat, voltijd baan in het afgelopen jaar
1	Eigen woning, geen zelfstandig ondernemer of langer dan 3 jaar geleden zelfstandig ondernemer, economisch zelfstandig, woont in huishouden dat langer dan 82 maanden bestaat

In onderstaande figuur staan de geschatte kansen van de deelpopulaties weergegeven. De horizontale lijnen lopen vanaf het 10-de percentiel van de geschatte kansen naar het 90-ste. De punt is de mediane kans op armoede en de verticale lijntjes geven het gemiddelde aan. De doorgetrokken lijn van boven naar beneden geeft de waargenomen fractie personen aan die arm geworden zijn in ieder van de populaties. De grootte van de punten geeft de relatieve grootte van de deelpopulatie aan.

De geschatte kansen (de doorgetrokken lijn) volgt de gemiddelde waargenomen kans binnen de deelpopulaties goed wat betekent dat het model gemiddeld zuiver is in deze deelpopulaties. Er is verder te zien dat voor de meeste deelpopulaties bijna iedereen een zeer kleine kans armoede loopt. Verder zijn de verdelingen binnen de groepen ook scheef: de mediaan ligt bij relatief lage kansen; de verdeling heeft een staart richting hogere kansen.



Omdat de deelpopulaties kunnen verschillen in hun samenstelling kan het ook zo zijn dat de modellen beter of minder goed voorspellen in de deelpopulaties. Onderstaande figuur geeft voor iedere deelpopulatie een aantal modelkwaliteitsmetrieken weer. Zie 3.2 voor meer informatie over de kwaliteitsmetrieken. Somers' D is hoog binnen alle populaties. Dit betekent dat het model goed in staat is onderscheid te maken tussen mensen met een hoog risico en mensen met een laag risico op armoede. De overige, op classificatie gebaseerde, maten variëren meer. Zij zijn vooral lager in de populaties met een zeer kleine kans op armoede: armoede wordt zeldzaam en daardoor moeilijker te voorspellen. Omdat Somers' D in die populaties nog steeds hoog is, is het model binnen die populaties nog steeds wel bruikbaar.



4.5 Niet-arme populatie - deelpopulatie met grootste kans om in de armoede te komen

Deze paragraaf zoomt in op de deelpopulatie met de grootste kans om in de armoede te komen:

- hoe is deze deelpopulatie afgebakend,
- in hoeverre verschilt deze deelpopulatie qua kenmerken van de hele niet-arme populatie,
- hoe groot is de kans om arm te worden,
- in hoeverre verschillen de kenmerken die het belangrijkste zijn voor het schatten van de armoedekans van het landelijke beeld,
- welk effect hebben deze kenmerken op de kans om arm te worden.

Afbakening

- In referentiejaar niet arm;
- In referentiejaar in een huurwoning met huurtoeslag;
- In referentiejaar economisch onzelfstandig.

Deze deelpopulatie betreft personen die niet arm zijn in het referentiejaar (t) en die de grootste kans hebben om in de armoede te komen het volgende jaar ($t + 1$). In het referentiejaar (t) woont deze deelpopulatie in een huurwoning met huurtoeslag en is economisch onzelfstandig.

Kenmerken

Om beter zicht te krijgen op deze deelpopulatie hebben we gekeken naar een aantal achtergrondkenmerken:

- leeftijd (lft)
- opleidingsniveau (opl_niv3)
- generatie migratieachtergrond
 - eerste generatie migratieachtergrond (generatie = 1)
 - tweede generatie (generatie = 2)
 - Nederlandse achtergrond (generatie = 3)
- geslacht (vrouw: vrouw = 1)
- type huishouden
 - alleenstaande (plhh_alleenstaande = 1)
 - thuiswonend kind (plhh_thuiswkind = 1)
 - ongehuwd partner zonder kinderen (plhh_partn_zkind = 1)
 - ongehuwd partner met kinderen (plhh_partn_mkind = 1)
 - gehuwd partner met kinderen (plhh_huw_zkind = 1)
 - gehuwd partner zonder kinderen (plhh_huw_mkind = 1)
 - ouder in eenouderhuishouden (plhh_ouder_eenouder = 1)
- Persoon woont in een huishouden met armoede één jaar voor het referentiejaar (hh_inkomensarm_j1 = 1)
- Persoon woont in een huishouden met armoede twee jaar voor het referentiejaar (hh_inkomensarm_j2 = 1)

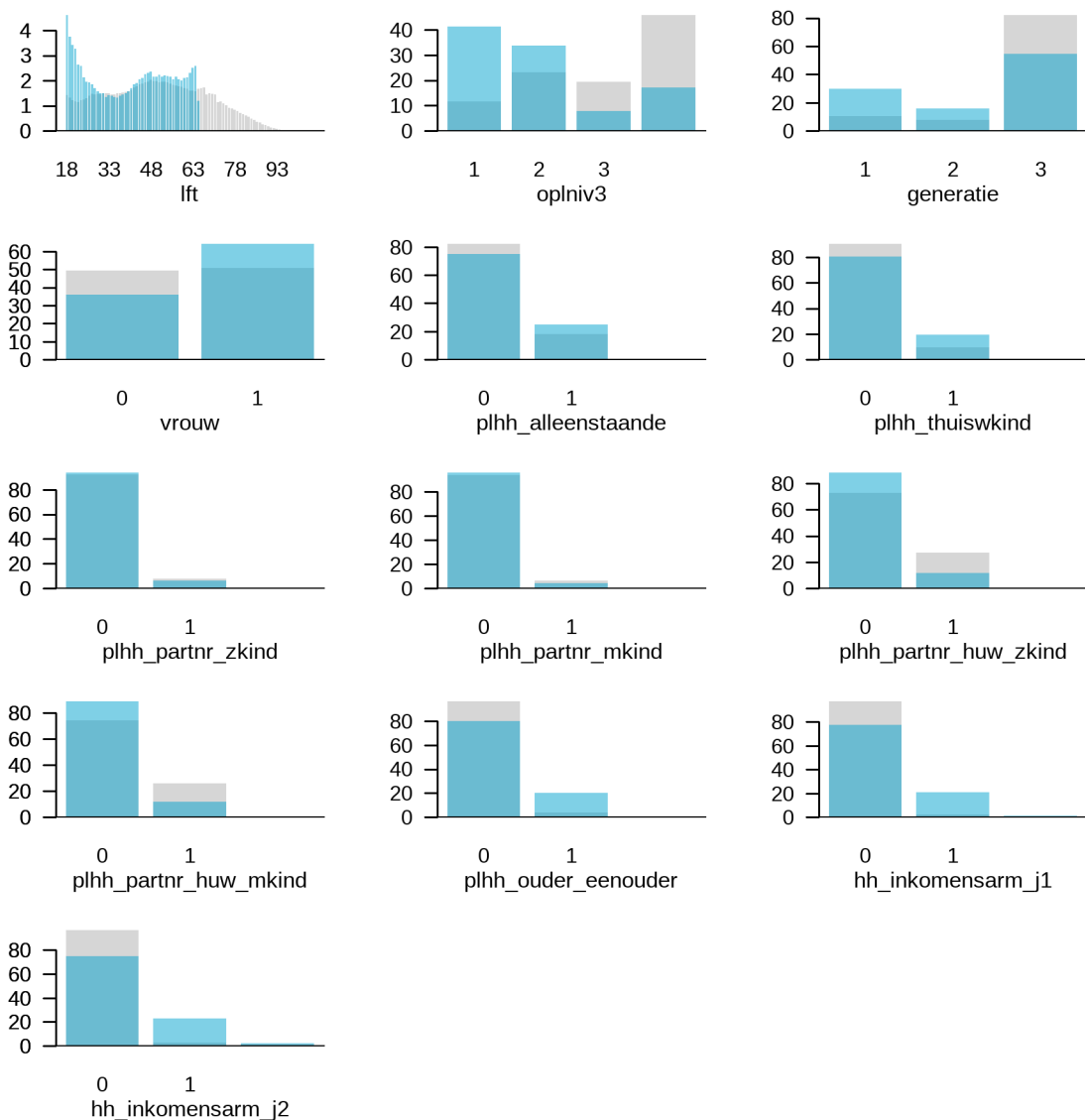
Onderstaande figuren tonen de verdeling van deze kenmerken in deze deelpopulatie (blauw) in vergelijking met de gehele niet-arme populatie (grijs). Merk op dat het opleidingsniveau deels op registraties en deels op enquêtes gebaseerd is, waardoor dit niet voor de gehele populatie bekend is. Hierdoor is de groep met een onbekend opleidingsniveau relatief groot. Ook is dit een selectieve groep. Zo is het opleidingsniveau relatief vaak onbekend voor ouderen en personen met een migratieachtergrond. Dit kan leiden tot

vertekening in de opleidingsniveauverdeling. Daarom laten we dit buiten beschouwing bij het beschrijven van de populatiekenmerken.

Uit de figuren blijkt dat in vergelijking met de gehele populatie, de deelpopulatie relatief veel jongeren bevat tussen ongeveer 18 en de 30 jaar ($18 < 1ft < 30$). Dat juist jongeren een groter risico lopen om in de armoede terecht te komen in deze deelpopulatie, is volgens verwachting, omdat zij vaker nog economisch onzelfstandig zijn. Dit komt bijvoorbeeld omdat ze studeren en een bijbaan of flexibel contract hebben (Van den Brakel and Otten, 2019). Daarnaast hebben ze relatief een lager vermogen (CBS, 2021b).

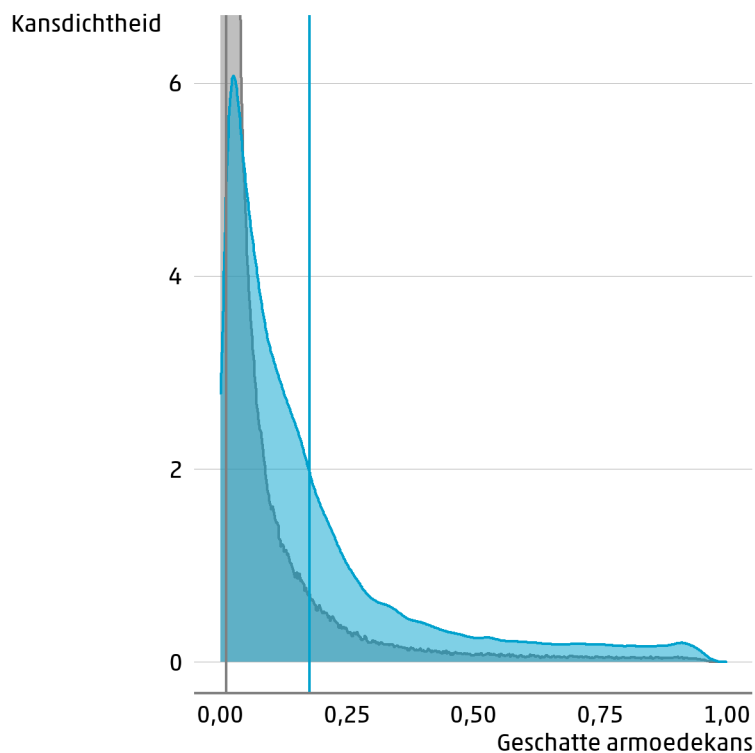
Ook bevat de deelpopulatie relatief meer vrouwen ($vrouw = 1$) en eerste en tweede generatie migranten ($generatie < 3$). Verder zien we relatief meer alleenstaanden ($plhh_alleenstaande = 1$), eenoudergezinnen ($plhh_ouder_eenouder = 1$) en thuiswonende kinderen ($plhh_thuiswkind = 1$) in de deelpopulatie. Voor deze huishoudsituaties is het naar verwachting dat zij een groter risico lopen om in de armoede terecht te komen, omdat zij alleen het inkomen moeten verdienen en geen partner hebben die het huishoudinkomen aan kan vullen (Van den Brakel and Otten, 2019). Gehuwde partners komen in deze deelpopulatie juist relatief weinig voor (zowel met kinderen: $plhh_partn_huw_mkind = 1$ als zonder kinderen $plhh_partn_huw_zkind = 1$).

Verder heeft deze deelpopulatie in jaar $t - 1$ en $t - 2$ relatief vaker te maken gehad met armoede ($hh_inkomensarm_j1 = 1$ en $hh_inkomensarm_j2 = 1$) dan de gehele niet-arme populatie. Dit betekent dat deze personen niet alleen een risico lopen om in de armoede terecht te komen in het referentiejaar, ze zijn in het verleden dus ook relatief vaker arm geweest.



Kans om arm te worden

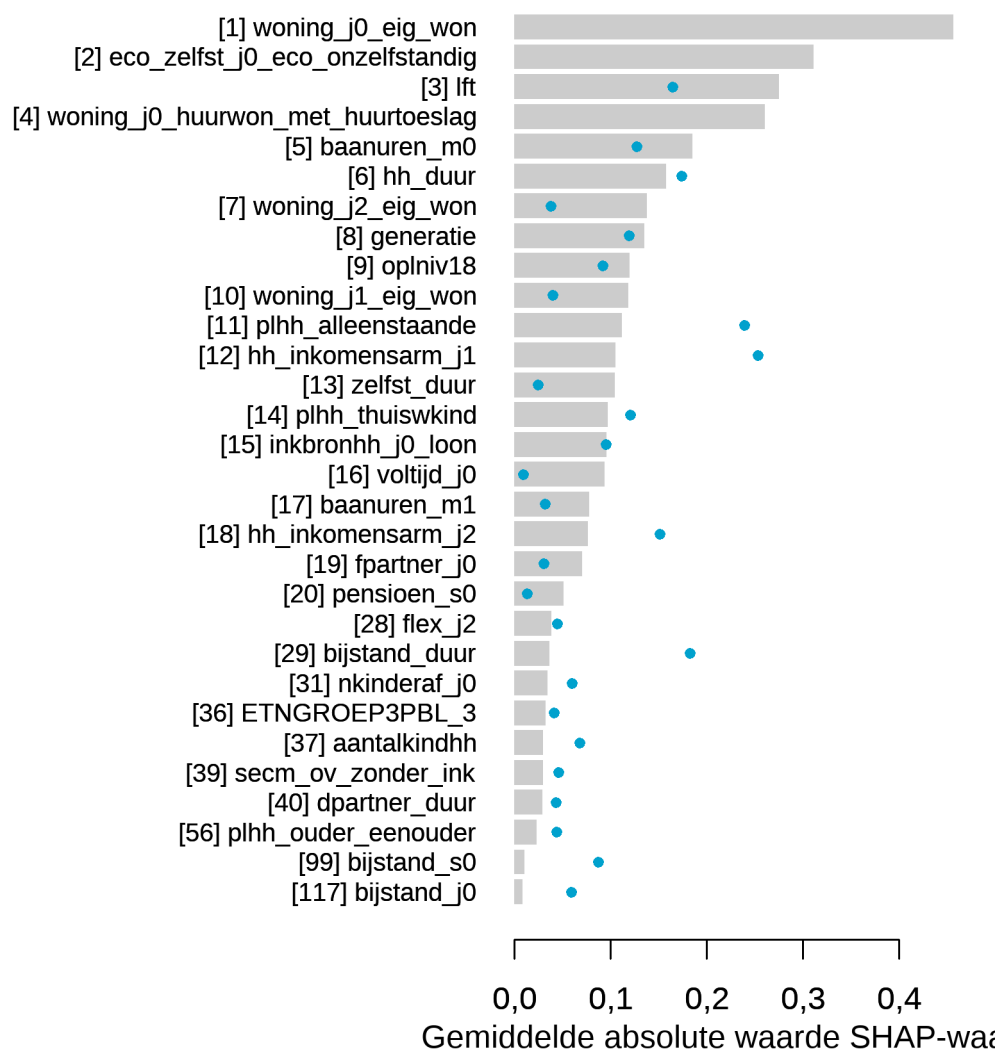
Onderstaande figuur toont de verdeling van de geschatte kans om arm te worden voor de niet-arme deelpopulatie met de grootste kans om in de armoede te komen (blauw) en voor de gehele niet-arme populatie in Nederland (grijs). De verticale lijnen tonen de gemiddelde kans om arm te worden. De gemiddelde kans voor deze deelpopulatie (18%) is hoger dan in de gehele Nederlandse niet-arme populatie (1,2%). Waar de gemiddelde kans op armoede dus relatief laag is voor de gehele niet-arme populatie, is deze relatief hoog voor deze deelpopulatie.



Belangrijkste kenmerken

Onderstaande figuur geeft de belangrijkste kenmerken weer op basis van SHAP-waarden. (Zie paragraaf 3.3 voor verdere toelichting.) De 20 belangrijkste kenmerken in de gehele niet-arme populatie (grijze balken) zijn aangevuld met de 20 belangrijkste kenmerken voor de deelpopulatie (blauwe punten). Bij kenmerken waarbij iedereen in de deelpopulatie dezelfde waarde heeft, ontbreekt de blauwe punt, aangezien dit kenmerk geen effect heeft binnen de deelpopulatie.

In onderstaande figuur is te zien dat de belangrijkste kenmerken voor deze deelpopulatie redelijk overeenkomen met de gehele arme populatie. (Zie paragraaf 4.1 voor een omschrijving van deze kenmerken.) Wel is te zien dat de blauwe punten voor een aantal kenmerken in meer of mindere mate afwijkt van de grijze balken. Hieronder gaan we in op een aantal kenmerken die belangrijker zijn dan voor de Nederlandse niet-arme populatie, dus waarvoor de blauwe punt rechts van de grijze balk ligt. Dit is met name het geval voor het kenmerk `plhh_alleenstaande`, wat een veel groter belang heeft in deze deelpopulatie. Daarnaast zijn de kenmerken die aangeven of een huishouden de twee jaren voorafgaand aan het referentiejaar in de armoede zat (`hh_inkomensarm_j1` en `_j2`), van groter belang voor de huidige deelpopulatie. Daarnaast is ook de duur in maanden waarin een persoon een bijstandsuitkering had in de afgelopen 3 jaar van belang (`bijstand_duur`). In de volgende figuren bekijken we hoe deze kenmerken de kans om arm te worden beïnvloeden.



Effect op kans om arm te worden

Onderstaande figuren tonen het *effect* op de kans om arm te worden en de *decompositie* van het effect. De staafdiagrammen tonen het effect van de 20 belangrijkste kenmerken op de geschatte kans om arm te worden voor de deelpopulatie (blauwe balken) en de gehele niet-arme populatie in Nederland (grijze balken). Een negatieve waarde op de y-as betekent dat personen met dit kenmerk een lagere kans dan gemiddeld hebben om arm te worden, een positieve waarde betekent dat personen een hogere kans dan gemiddeld hebben. De hoogte van de balk geeft de range van de kans weer voor 80% van de betreffende populatie. De breedte van de balk komt overeen met de relatieve grootte van de populatie met deze waarde. Een smalle balk betekent dus relatief weinig mensen en een brede balk relatief veel mensen. Zie paragraaf 3.3 voor verdere toelichting.

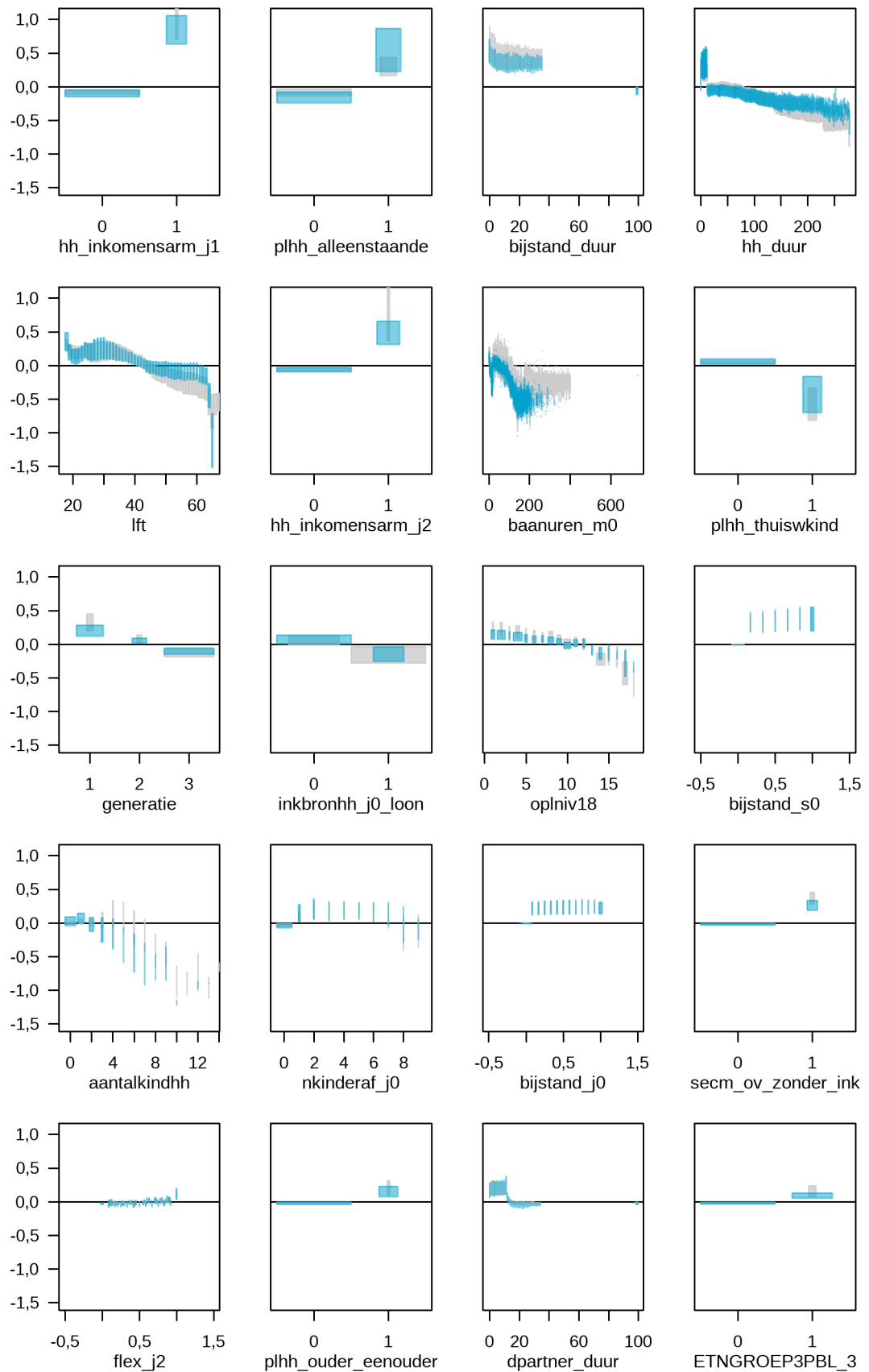
De tweede figuur toont de decompositie van het verschil in belang van de belangrijkste kenmerken. Het verschil in belang wordt opgesplitst in een component die het belang van de *verdeling* van het kenmerk weergeeft (horizontale as) en een component die het belang weergeeft van het *effect* dat het kenmerk heeft op de geschatte kans (verticale as). Zie paragraaf 3.3 voor verdere toelichting.

In onderstaande figuren is te zien dat armoede in het verleden ($hh_inkomensarm_j1 = 1$ en $hh_inkomensarm_j2 = 1$) vaker voorkomt bij deze deelpopulatie, omdat de blauwe balken breder zijn dan

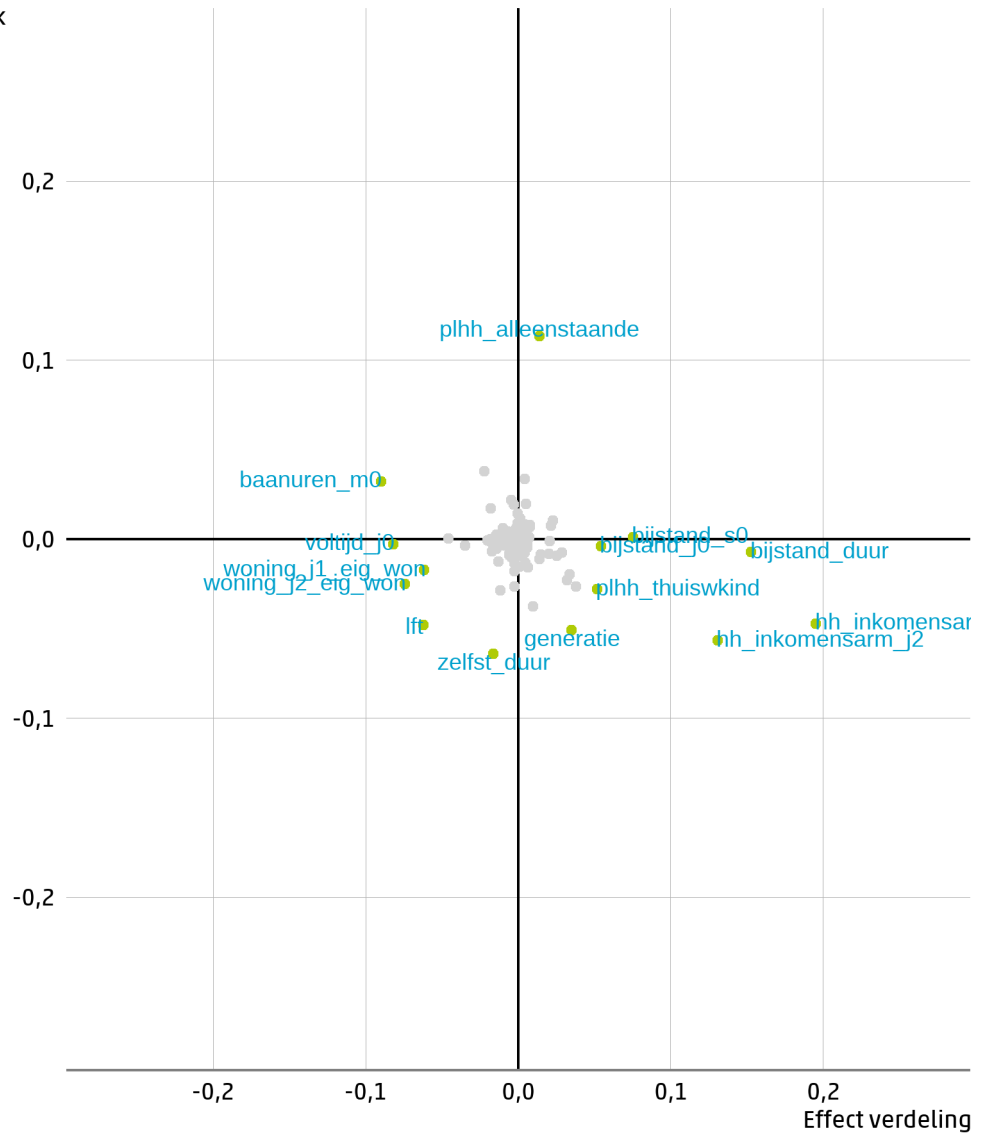
de grijze balken. Daarnaast is te zien dat de blauwe balken iets lager liggen dan de grijze. Dit betekent dat de kans op armoede voor personen in deze deelpopulatie die eerder arm waren, iets lager is dan voor personen in de gehele niet-arme populatie die eerder arm waren. Dit zien we ook terug in de decompositiefiguur: er is een groot positief verdelingseffect en een wat kleiner negatief kenmerkeffect. Netto is het kenmerk belangrijker in deze deelpopulatie.

Daarnaast laten de figuren zien dat personen die minder dan drie jaar geleden een bijstandsuitkering gehad hebben (`bijstand_duur < 36`) net als in de gehele arme populatie een hogere kans hebben om arm te worden. De blauwe en grijze balken liggen namelijk boven nul. De kans om arm te worden voor bijstandsontvangers lijkt in de deelpopulatie kleiner dan voor bijstandsontvangers in de gehele niet-arme populatie. De blauwe balken liggen namelijk dichterbij nul dan de grijze. De decompositie figuur laat echter vooral een positief effect zien van de verdeling. Omdat dit kenmerk uit veel mogelijke waarden bestaat, verschillen de blauwe en grijze balken in de voorgaande figuur niet zichtbaar in breedte. Hierdoor is dit verdelingseffect niet goed te zien in de eerste figuur.

Verder is te zien dat de kans op armoede hoger is voor de alleenstaanden (`plhh_alleenstaande = 1`) in de deelpopulatie, in vergelijking met alleenstaanden in de gehele populatie. We zien dat de blauwe balk een stuk hoger ligt dan de grijze, wat betekent dat alleenstaanden in deze deelpopulatie een grotere kans hebben om arm te worden dan in de gehele niet-arme populatie. We zien ook dat de blauwe balk iets breder dan de grijze is, wat aangeeft dat deze deelpopulatie relatief meer alleenstaanden heeft. Dit is ook terug te zien in de decompositie figuur, met een groot positief effect van het kenmerk en een klein positief effect van de verdeling. Niet-alleenstaanden (`plhh_alleenstaande = 0`) hebben in deze deelpopulatie juist een lagere armoedekans dan in de gehele populatie, wat bijdraagt aan het positieve kenmerkeffect in de decompositie.



Effect kenmerk



5 Conclusies en discussie

5.1 Conclusies

Met het oog op effectief armoedebelief heeft het CBS op verzoek van BZK een eerste versie van twee machine-learningmodellen ontwikkeld, waarmee risicofactoren voor transities in en uit armoede in kaart worden gebracht. Via deze modellen heeft het CBS onderstaande twee onderzoeksvragen onderzocht:

1. Hoe goed kunnen we op basis van (combinaties van) kenmerken de kans schatten dat iemand arm blijft dan wel arm wordt?

Uit de resultaten blijkt aan de ene kant dat de modellen minder goed zijn in het *classificeren* van individuen. Daarbij schat het model beter of iemand arm blijft, dan of iemand arm wordt. Aan de andere kant kunnen de modellen arme en niet-arme personen wel goed van elkaar *onderscheiden*; de modellen schatten in een zeer ruime meerderheid van de gevallen voor arme personen een hogere armoedekans dan voor niet-arme personen. Dit betekent dat de modellen minder geschikt zijn om op individueel niveau personen te ‘profilieren’ als wel of niet arm—los van ethische en juridische aspecten. Wel zijn de modellen geschikt om in kaart te brengen hoe potentiële risicofactoren de armoedekans beïnvloeden.

2. Voor welke aspecten kan de combinatie van populatiedekkende registerdata en machine learning tot nieuwe inzichten over risicofactoren voor voor transities in en uit armoede leiden?

De modellen blijken nieuwe inzichten te geven voor de volgende aspecten:

- *life events*: doorgaans worden risicogroepen beschreven aan de hand van ‘statische’ kenmerken van de arme populatie, zoals demografische of werkgerelateerde kenmerken. Naast inzicht in dergelijke statische risicofactoren bieden de modellen ook inzicht in verschillende life events zoals duur sinds het krijgen van een partner, geboorte van een kind, ontslag.
- *deelpopulaties*: de modellen bieden inzicht in risicofactoren voor specifieke deelpopulaties. Daarbij is het mogelijk om in te zoomen op zowel data- als expertgedreven deelpopulaties. Ook is het mogelijk om een verschil tussen deelpopulatie en gehele populatie te ontleden in een verschil in samenstelling van de (deel)populatie en een verschil in effect van een risicofactor.
- *ranking*: op basis van SHAP-waardes kunnen de ruim 500 potentiële risicofactoren in de modellen gerangschikt worden.
- *effecten*: op basis van SHAP-waardes kunnen gedetailleerde niet-lineaire verbanden tussen kenmerk en armoedekans in kaart worden gebracht.
- *overgangen*: de modellen bieden inzicht in risicofactoren die samenhangen met overgangen van en naar armoede. Dit in tegenstelling tot eerder onderzoek dat zich vooral richt op kenmerken van arme populaties.

5.2 Discussie

Doel van dit onderzoek was om potentiële risicofactoren voor transitie in en uit armoede te vinden. Daartoe zijn twee machine-learningmodellen ontwikkeld: een die de kans schat dat een arm persoon arm blijft en een die de kans schat dat een niet-arm persoon arm wordt. Voor de complete doelpopulaties in twee jaren zijn meer dan 500 persoons- en huishoudenkenmerken afgeleid, die informatie bevatten over de huidige status en gebeurtenissen in de afgelopen drie jaar binnen de thema's demografie, economie, gezondheid en criminaliteit. Het onderzoek richt zich op de modelkwaliteit, het rangschikken van potentiële risicofactoren en het transparant maken van de geleerde verbanden. In deze paragraaf bespreken we een aantal voor- en nadelen van de gekozen aanpak, de risico's op verkeerd gebruik en ideeën voor verbetering of uitbreiding.

Er is gekozen voor een machine-learningmodel in plaats van een (logistisch) regressiemodel, omdat dit beter schaalbaar is naar een groot aantal kenmerken. Bij een regressiemodel moet vooraf worden gespecificeerd welke parameters geschat moeten worden. Is er bijvoorbeeld een lineair verband tussen leeftijd en de logit van de armoedekans of is er een hogere orde polynoom nodig? Moet leeftijd getransformeerd worden om het verband lineair te maken? Of maken we leeftijd categoriaal? Welke categorieën kiezen we dan en schatten we elke categorie onafhankelijk, of behandelen we een ordinale variabele als numerieke? Welke interacties nemen we mee en hoe? Hoe meer variabelen, hoe meer werk en hoe ingewikkelder de modelselectie wordt. Bij het op beslisbomen gebaseerde machine-learningmodel (XGBoost) hoeven niet-lineaire verbanden en interacties niet vooraf gespecificeerd te worden, maar ontstaan ze als het ware vanzelf als bijproduct van het algoritme. Nadeel is dat er geen regressiecoëfficiënten geschat worden, die de richting en sterkte van elke term aangeven. Machine learning is gericht op het maximaliseren van de modelkwaliteit, ten koste van de transparantie. Een ander nadeel is dat er niet expliciet rekening wordt gehouden met de hiërarchie in de data, wat multilevel/hiërarchische regressiemodellen wel doen. Regionale kenmerken zijn daarom niet meegenomen.

De modellen zijn gefit op observationele data. Dat betekent dat alle verbanden slechts correlaties zijn en niet noodzakelijkerwijs causale verbanden. Een goed voorbeeld is de huurtoeslag-indicator (`woning_j0_huurwon_met_huurtoeslag`). Huurtoeslag is een goede voorspeller voor de kans dat iemand arm wordt. Maar het is niet waarschijnlijk dat huurtoeslag de oorzaak is. Het stoppen of verlagen van huurtoeslag zal geen effectief armoedepreventiebeleid zijn. Het model kan gebruikt worden om potentiële risicofactoren te identificeren, maar wijst geen oorzaken aan. Een beleidsinterventie is alleen effectief als het verband oorzakelijk is. Om oorzakelijkheid aan te tonen is aanvullend onderzoek nodig, bijvoorbeeld met structurele vergelijkmingsmodellen. Dit is een methode om hypothesen over causale verbanden te beschrijven en te toetsen (met weer andere beperkingen en nadelen).

De modellen schatten individuele kansen. De kwaliteitsmaten laten zien dat er voorspellende informatie in de kenmerken zit, maar ook dat de geschatte kansen niet perfect zijn (zie de foutpositieven en foutnegatieven). Wanneer is de modelkwaliteit goed genoeg? Voor het ordenen en begrijpen van potentiële risicofactoren denken we dat de modellen zeer bruikbaar zijn, voor individueel profileren niet. Daar zijn ze ook nooit voor bedoeld. De CBS-wet verbiedt het openbaar maken van herkenbare gegevens over een afzonderlijk persoon, en het fraudedetectiesysteem SyRI en de toeslagenaffaire hebben laten zien wat de gevaren en onethische aspecten van individueel profileren zijn. Verder moeten risicofactoren ook als zodanig geïnterpreteerd worden. Alleenstaanden hebben een hoger risico om in armoede te blijven dan niet-alleenstaanden. Los van de vraag of het een causaal verband is, is het een (stochastische) kans en geen (deterministische) regel. Dit onderkennen kan stigmatisering voorkomen.

De lijst met potentiële kenmerken is samengesteld op basis van beschikbare registerinformatie en domeinkennis. Er is hierbij geen rekening gehouden met de mate waarin deze kenmerken beïnvloedbaar zijn voor beleid. De modellen zouden opnieuw getraind kunnen worden met een selectie van *actionable* kenmerken, maar dat zal ten koste gaan van de modelkwaliteit. De lijst kenmerken zou beperkt kunnen

worden tot bijvoorbeeld de top-100, aangevuld met nieuwe informatie uit bijvoorbeeld politieregistraties, over het type zorg, over inkomensgerelateerde gemeentelijke regelingen of over intergenerationele armoede (inkomen van huishouden in iemands jeugd). Uitsplitsing naar kenmerken die niet in het model zitten, zoals gemeente, kan leiden tot verkeerde conclusies (zie De Waal, 2016). Dit probleem speelt hier minder omdat we geen vertekening zien in de gemiddelde geschatte armoedekans.

De modellen zijn getraind met kenmerken over 2011–2013 en 2015–2017 om de kans op armoede in 2014 en 2018 te schatten. Het combineren van twee referentie jaren maakt de modellen robuuster tegen conjuncturele schommelingen. Toepassen van de modellen om armoede in 2022 te voorspellen op basis van kenmerken over 2019–2021 kan niet om een aantal redenen. Ten eerste is het CBS gericht op het beschrijven en schatten van verleden en heden in plaats van op het voorspellen van de toekomst, terwijl de modellen sterk leunen op privacygevoelige CBS-data. Ten tweede levert het geen nieuwe inzichten in belangen en verbanden op zonder hertraining met meer recente data. Ten derde zijn niet alle kenmerken over 2019–2021 begin 2022 beschikbaar. Begin 2022 kunnen personen die verhuizen naar een institutioneel huishouden, studentenhuishouden of buitenland ook nog niet verwijderd worden, terwijl het model wel op die selectie getraind is. De analyse zou wel herhaald kunnen worden na 2022, zodra alle benodigde data beschikbaar is. Meer data over een andere periode kunnen wel tot nieuwe inzichten leiden, bijvoorbeeld over corona-gerelateerde effecten. Wat nu al zou kunnen is trainen met 2011–2014 data en testen met 2015–2018 data om inzicht te krijgen over hoe goed de modellen generaliseren over de tijd.

In dit onderzoek hebben we vooral SHAP-waardes gebruikt om te begrijpen hoe de modellen tot armoedekansen komen. Dit zou uitgebreid kunnen worden met aanvullende analyses. Friedman's H^2 (Friedman and Popescu, 2008) is een maat voor de sterkte van de interactie tussen twee of meer kenmerken, gebaseerd op *partial dependence* functies. *Counterfactuals* van iemand met een hoge c.q. lage armoedekans zijn fictieve personen met zo klein mogelijke wijzigingen in zo min mogelijk kenmerken, die de geschatte armoedekans onder c.q. boven een bepaalde drempel brengen (Molnar, 2019). Deze kunnen weer tot nieuwe inzichten leiden en zijn mogelijk interessant voor beleidsmakers als de kenmerken beïnvloedbaar zijn.

De modellen schatten de kans dat iemand het komende jaar arm blijft of wordt. Sommige personen zoals zelfstandigen met een onregelmatig inkomen zullen vaak wisselen tussen boven en onder de lage-inkomensgrens. Om die oscillaties te dempen hebben we in dit onderzoek armoede gedefinieerd op basis van zowel inkomen als vermogen. Daarnaast zou de definitie van armoede verder vernauwd kunnen worden tot langdurige armoede: er zouden modellen ontwikkeld kunnen worden voor de kans dat iemand die langdurig arm is ($t - 3$ tot en met t) uit de armoede komt, en voor de kans dat iemand die niet arm is langdurig arm wordt ($t + 1$ tot en met $t + 4$). Daarmee wordt het fenomeen relevanter, maar ook zeldzamer. Het is dus moeilijker te schatten; het tweede model kan alleen geschat worden op oudere gegevens, waardoor het minder goed recente effecten kan vangen. De definitie van armoede zou ook verbreed kunnen worden naar brede welvaart, waarbij ook rekening wordt gehouden met welzijn, gezondheid en dergelijke.

5.3 Aanknopingspunten voor beleid

Via de modellen is veel informatie beschikbaar. In samenwerking met gemeentes kan verkend worden wat er nodig is om ervoor te zorgen dat dit kan bijdragen aan beleid. Zo kunnen gemeentes de modellen gebruiken om specifieke doelgroepen voor beleid te bepalen, met onderscheid tussen de arme en niet-arme populatie. Daarbij kunnen ze op basis van de ranking van risicofactoren focus in hun beleid aanbrengen. Zo mogelijk kunnen gemeentes via maatregelen proberen deze risicofactoren te beïnvloeden. Voor risicofactoren die niet direct beïnvloedbaar zijn, kunnen mogelijk randvoorwaarden gecreëerd worden. Zo heeft de overheid uiteraard geen invloed op huishoudenssamenstelling, maar bijvoorbeeld wel op kosten/beschikbaarheid van kinderopvang en op betaalbare huur voor huishoudens met één inkomen. Daarbij zou het voor gemeentes nuttig zijn om verder in te zoomen op risicofactoren. Dit kan bijvoorbeeld door kenmerken verder te specificeren. Nu bevatten de modellen bijvoorbeeld kenmerken over de mate van bijstand in verschillende periodes. Alternatief is om in- en uitstroom in de bijstand expliciet in de modellen op te nemen. Een andere mogelijkheid is om (met behulp van Friedman's H^2) in te zoomen op interacties tussen kenmerken. Zo kunnen bijvoorbeeld (niet-lineaire) verbanden tussen leeftijd en huishoudenssamenstelling of tussen bijstand en geboorte in kaart worden gebracht. Verder heeft het CBS in het onderzoek ingezoomd op een aantal data- en expertgedreven deelpopulaties. Uit het onderzoek komen nog 22 andere datagedreven deelpopulaties naar voren, die het CBS wegens de scope van het onderzoek niet verder onderzocht heeft. Met name de deelpopulaties met relatief hoge en lage armoedekansen zijn interessant om verder te onderzoeken, evenals aanvullende expertgedreven deelpopulaties. Daarbij lag de focus in dit onderzoek met name op kenmerken die belangrijker zijn voor een deelpopulatie. Daarnaast kan ook ingezoomd worden op kenmerken die juist minder belangrijk zijn voor een deelpopulatie, of op kenmerken die weliswaar even belangrijk zijn, maar waarbij het verdelingseffect (doordat het kenmerk in de deelpopulatie anders verdeeld is dan in de gehele populatie) en het kenmerkeffect (doordat het kenmerk in deze deelpopulatie een ander effect heeft op de geschatte armoedekans dan in de gehele populatie) in tegengestelde richting werken.

Verder kunnen gemeentes de modeluitkomsten confronteren met gemeentelijke regelingen. Zo leeft bij gemeentes de vraag in hoeverre het bereik van inkomensgerelateerde gemeentelijke regelingen overeenkomt met de hoogrisicogroepen uit de modellen. Dit kan in kaart gebracht worden door data over gemeentelijke regelingen te koppelen aan hoogrisicokenmerken uit de modellen.

Afkortingen

- AOW - Algemene Ouderdomswet
- AI - Artificial Intelligence
- AUC - Area Under the ROC Curve
- BZK - Binnenlandse Zaken en Koninkrijksrelaties
- CBS - Centraal Bureau voor de Statistiek
- ELSA - Ethical Legal and Societal Aspects
- EZK - Economische Zaken en Klimaat
- XGBoost - eXtreme Gradient Boosting
- MCC - Matthews' Correlatiecoëfficiënt
- NLL - Negatieve Log-Likelihood
- ROC - Receiver Operating Characteristic
- SHAP - SHapley Additive exPlanation
- SSB - Stelsel van Sociaal-statistische Bestanden

Bijlagen

A Operationalisering variabelen

Onderstaande tabel bevat een totaaloverzicht van de kenmerken, waarbij de gemaakte duuraflleidingen als achtervoegsels zijn opgenomen. Voor nominale maandkenmerken is maximaal vastgesteld:

- x_{j0} : fractie van maanden in referentiejaar t dat persoon tot categorie x behoorde $[0, \frac{1}{12}, \dots, 1]$
- x_{j1} : fractie van maanden in jaar voor referentiejaar ($t - 1$) dat persoon tot categorie x behoorde $[0, \frac{1}{12}, \dots, 1]$
- x_{j2} : fractie van maanden in twee jaar voor referentiejaar ($t - 2$) dat persoon tot categorie x behoorde $[0, \frac{1}{12}, \dots, 1]$
- x_{s0} : fractie van maanden in laatste halfjaar van referentiejaar t (jul–dec) dat persoon tot categorie x behoorde $[0, \frac{1}{6}, \dots, 1]$
- x_{m0} : indicator voor behoorde tot categorie x in laatste maand van referentiejaar t (dec) $[0, 1]$
- x_{m1} : indicator voor behoorde tot categorie x in een-na-laatste maand van referentiejaar t (nov) $[0, 1]$
- x_{m2} : indicator voor behoorde tot categorie x in twee-na-laatste maand van referentiejaar t (okt) $[0, 1]$
- x_{m3} : indicator voor behoorde tot categorie x in drie-na-laatste maand van referentiejaar t (sep) $[0, 1]$
- x_{m4} : indicator voor behoorde tot categorie x in vier-na-laatste maand van referentiejaar t (aug) $[0, 1]$
- x_{m5} : indicator voor behoorde tot categorie x in vijf-na-laatste maand van referentiejaar t (jul) $[0, 1]$
- x_{duur} : aantal maanden geleden sinds december van referentiejaar t dat persoon tot categorie x behoorde $[0, \dots, 35, 99]$

Hierbij is de categorie onbekend gecodeerd als waarde 99.

Tabel A.1 Beschrijving en operationalisering van modelkenmerken, gesorteerd op niveau, periodiciteit, thema en status

Variabele	Type	Omschrijving	Indeling
lft_m0	i	Leeftijd	
vrouw	d	Indicator voor vrouw	
generatie	i	Generatie	1 - Ten minste één ouder en zelf in buitenland geboren 2 - Ten minste één ouder in buitenland maar zelf in Nederland geboren 3 - Beide ouders in Nederland geboren
ETNGRP_<cat>	[7]	Migratieachtergrond (1e generatie: eigen geboorteland; 2e generatie: geboorteland moeder als buitenland, anders geboorteland vader; overig: Nederland)	0 Nederland 1 Marokko 2 Turkije 3 Suriname 4 Voormalige Nederlandse Antillen en Aruba 5 Overige niet-westerse landen 6 Overige westerse landen
ETNGROEP3PBL_3	d	Indicator voor niet-westerse migratieachtergrond	
plhh_<cat>_m0	[10]	Plaats in huishouden	1 Thuiswonend kind 2 Alleenstaande 3 Partner in niet-gehuwd paar z. kinderen 4 Partner in gehuwd paar zonder kinderen 5 Partner in niet-gehuwd paar met kinderen 6 Partner in gehuwd paar met kinderen 7 Ouder in eenouderhuishouden 8 Referentiepersoon in overig huishouden 9 Overig lid van een huishouden 10 Lid van institutioneel huishouden
plhh_partner_[j0,...,j2,s0,m0,...,m5]	n	Fractie van maanden in betreffende periode dat persoon 'partner' was	
vrlverblijfsdoel_<cat>_m0	[13]	Migratiemotief	10 Gezinsmigratie 11 Gezinsvorming 12 Gezinshereniging 13 Adoptie- en pleegkinderen 20 Asiel 30 Arbeid 31 Kennismigrant 32 Onderzoeker 33 Au pair 34 Stagiair 40 Studie 50 Overig 90 Onbekend; binnen zelfde jr geëmigreerd

Loopt door op volgende pagina

Voortgezet van vorig pagina

Variabele	Type	Omschrijving	Indeling
npartnererbij	i	Totaal aantal toenames in partnerschap in afgelopen drie jaar	
npartnereraf	i	Totaal aantal afnames in partnerschap in afgelopen drie jaar	
fpartner_[j0,...,j2]	i	Fractie van maanden in betreffende periode dat persoon partner heeft	
dpartner_duur	i	Aantal maanden sinds wijziging in partnerschap	
gescheiden_[j0,...,j2,s0,m0,...,m5,duur]	n	Fractie van maanden in betreffende periode en aantal maanden geleden dat samenwoonperiode is geëindigd wegens verhuizing van persoon, partner of beide	
verweduwing_[j0,...,j2,s0,m0,...,m5,duur]	n	Fractie van maanden in betreffende periode en aantal maanden geleden dat samenwoonperiode is geëindigd wegens overlijden partner	
secm_<cat>	[13]	Belangrijkste inkomensbron van persoon in december van jaar t	11 Werknemer 12 Directeur-grootaandeelhouder 13 Zelfstandig ondernemer 14 Overige zelfstandige 15 Meewerkend gezinslid 21 Ontvanger werkloosheidsuitkering 22 Ontvanger bijstandsuitkering 23 Ontvanger uitkering sociale voorz. overig 24 Ontvanger uitkering ziekte/AO 25 Ontvanger pensioenuitkering 26 Nog niet schoolg./schol./stud. met ink. 31 Nog niet schoolg./schol./stud. geen ink. 32 Overig zonder inkomen
secm_<cat>_[j0,...,j2,s0,m0,...,m5,duur]	[11]	Fractie van maanden in betreffende periode en aantal maanden geleden dat persoon sociaaleconomische categorie <cat> was	11 Werknemer 12 Directeur-grootaandeelhouder 13 Zelfstandig ondernemer 14 Overige zelfstandige 15 Meewerkend gezinslid 21 Ontvanger werkloosheidsuitkering 22 Ontvanger bijstandsuitkering 23 Ontvanger uitkering sociale voorz. overig 24 Ontvanger uitkering ziekte/AO 25 Ontvanger pensioenuitkering 26,31 Nog niet schoolg./schol./stud.

Loopt door op volgende pagina

Voortgezet van vorig pagina

Variabele	Type	Omschrijving	Indeling
vast_[j0,..j2,s0,m0,...m5]	n	Fractie van maanden in betreffende periode dat persoon werkzaam was met vast contract (dga, wsw, of rest met onbepaalde tijd)	
flex_[j0,..j2,s0,m0,...m5]	n	Fractie van maanden in betreffende periode dat persoon werkzaam was met flex contract (stagiair, uitzendkracht, oproep, wsw of rest met bepaalde tijd)	
voltijd_[j0,..j2,s0,m0,...m5]	n	Fractie van maanden in betreffende periode dat persoon werkzaam was in voltijd	
deeltijd_[j0,..j2,s0,m0,...m5]	n	Fractie van maanden in betreffende periode dat persoon werkzaam was in deeltijd	
stagiaire_[j0,..j2,s0,m0,...m5]	n	Fractie van maanden in betreffende periode dat persoon werkzaam was als stagiaire	
wsw_[j0,..j2,s0,m0,...m5]	n	Fractie van maanden in betreffende periode dat persoon werkzaam was als WSW	
uitzend_[j0,..j2,s0,m0,...m5]	n	Fractie van maanden in betreffende periode dat persoon werkzaam was als uitzendkracht	
oproep_[j0,..j2,s0,m0,...m5]	n	Fractie van maanden in betreffende periode dat persoon werkzaam was als oproepkracht	
rest_[j0,..j2,s0,m0,...m5]	n	Fractie van maanden in betreffende periode dat persoon werkzaam was als overig werknemer	
onbepaaldetijd_[j0,..j2,s0,m0,...m5]	n	Fractie van maanden in betreffende periode dat persoon contract voor onbepaalde tijd had	
bepaaldetijd_[j0,..j2,s0,m0,...m5]	n	Fractie van maanden in betreffende periode dat persoon contract voor bepaalde tijd had	
baanuren_[j0,..j2,s0,m0,...m5]	n	Gemiddeld aantal uren in betreffende periode dat persoon werkzaam was	
werkuitk_[j0,...,j2,s0,m0,...,m5,duur]	n	Fractie van maanden in betreffende periode en aantal maanden geleden dat persoon een overgang maakt van werkzaam (SECM = 11, 12, 13 of 14) naar een uitkeringssituatie vanuit de Werkloosheidswet, Bijstand of Arbeidsongeschiktheidswet (SECM = 21, 22 of 24)	

Loopt door op volgende pagina

Voortgezet van vorig pagina

Variabele	Type	Omschrijving	Indeling
werkuitk_partner_[j0,...,j2,s0,m0,...,m5,duur]	n	Fractie van maanden in betreffende periode en aantal maanden geleden dat partner van persoon een overgang maakt van werkzaam (SECM = 11, 12, 13 of 14) naar een uitkeringssituatie vanuit de Werkloosheidswet, Bijstand of Arbeidsongeschiktheidswet (SECM = 21, 22 of 24)	
ontslag_[j0,...,j2,s0,m0,...,m5,duur]	n	Fractie van maanden in betreffende periode en aantal maanden geleden dat voor persoon ontslag is aangevraagd bij het Uitvoeringsinstituut Werknemersverzekeringen (UWV) en het ontslag is ingewilligd waardoor de arbeidsverhouding is verbroken	
oplniv18_j0	i	Opleidingsniveau fijn	1 - 1111 Basisonderwijs gr1-2 2 - 1112 Basisonderwijs gr3-8 3 - 1211 Praktijkonderwijs 4 - 1212 Vmbo-b/k 5 - 1213 Mbo1 6 - 1221 Vmbo-g/t 7 - 1222 Havo-, vwo-onderbouw 8 - 2111 Mbo2 9 - 2112 Mbo3 10 - 2121 Mbo4 11 - 2131 Havo-bovenbouw 12 - 2132 Vwo-bovenbouw 13 - 3111 Hbo-associate degree 14 - 3112 Hbo-bachelor 15 - 3113 Wo-bachelor 16 - 3211 Hbo-master 17 - 3212 Wo-master 18 - 3213 Doctor NA - 9999 Weet niet of onbekend
oplniv8_j0	i	Opleidingsniveau middel	1 - 1110 Basisonderwijs z.n.d. 2 - 1210 Vmbo-b/k, mbo1 z.n.d. 3 - 1220 Vmbo-g/t, havo-, vwo-onderbouw 4 - 2110 Mbo2 en mbo3 z.n.d. 5 - 2120 Mbo4 z.n.d. 6 - 2130 Havo, vwo z.n.d. 7 - 3110 Hbo, wo-bachelor z.n.d. 8 - 3210 Hbo-, wo-master, doctor z.n.d. NA - 9999 Weet niet of onbekend

Loopt door op volgende pagina

Voortgezet van vorig pagina

Variabele	Type	Omschrijving	Indeling
oplniv3_j0	i	Opleidingsniveau grof	1 - 1110 Basisonderwijs z.n.d., 1210 Vmbo-b/k, mbo1 z.n.d., 1220 Vmbo-g/t, havo-, vwo-onderbouw 2 - 2110 Mbo2 en mbo3 z.n.d., 2120 Mbo4 z.n.d., 2130 Havo, vwo z.n.d. 3 - 3110 Hbo, wo-bachelor z.n.d., 3210 Hbo-, wo-master, doctor z.n.d. NA - 9999 Weet niet of onbekend
schuldanering_[j0,...,j2]	d	Indicator voor toegelaten tot wettelijke schuldsanering natuurlijke personen (Wsnp)	
wanb_zv_[j0,...,j2]	d	Indicator voor wanbetaler zorgpremie	
vrtg_schlvrltr_[j0,...,j2]	d	Indicator voor voortijdig schoolverlater	
eco_zelfst_<cat>_[j0,...,j2]	[4]	Economische zelfstandigheid	0 Niet economisch zelfstandig 1 Economisch zelfstandig 8 Behoort niet tot de doelpopulatie economische zelfstandigheid 9 Geen inkomen
zelfstandig_<cat>_[j0,...,j2]	[6]	Arbeidspositie van zelfstandigen	0 Geen zelfstandig ondernemer 1 Zelfstandig ondernemer met personeel 2 Zelfstandig ondernemer zonder personeel, eigen arbeid 3 Zelfstandig ondernemer zonder personeel, producten 4 Zelfstandig ondernemer, type onbekend 9 Behoort tot een huishouden zonder waargenomen inkomen
secmsbi2008v21_<cat>_j0	[22]	Economische activiteit van werkgever van persoon	A Landbouw, bosbouw en visserij B Delfstoffenwinning C Industrie D Energievoorziening E Waterbedrijven en afvalbeheer F Bouwnijverheid G Handel H Vervoer en opslag I Horeca J Informatie en communicatie K Financiële dienstverlening L Verhuur en handel van onroerend goed M Specialistische zakelijke diensten N Verhuur en overige zakelijke diensten O Openbaar bestuur en overheidsdiensten P Onderwijs Q Gezondheids- en welzijnszorg R Cultuur, sport en recreatie S Overige dienstverlening T Huishoudens U Extraterritoriale organisaties Onbekend

Loopt door op volgende pagina

Voortgezet van vorig pagina

Variabele	Type	Omschrijving	Indeling
zorgkosten_[j0,...,j2]	n	Jaarlijkse zorgkosten (euro)	
wmo_[j0,...,j2]	d	Indicator voor ondersteuning via Wet Maatschappelijke Ondersteuning (WMO)	
wet_langd_zorg_[j0,...,j2]	d	Indicator voor Wet langdurige zorg (Wlz)-indicatie	
verdacht_crimi_[j0,...,j2]	d	Indicator voor verdacht van criminaliteit	
gedetineerd_[j0,...,j2]	d	Indicator voor gedetineerd	
typhh_<cat>_m0	[8]	Type huishouden	1 Eenpersoonshuishouden 2 Niet-gehuwd paar zonder kinderen 3 Gehuwd paar zonder kinderen 4 Niet-gehuwd paar met kinderen 5 Gehuwd paar met kinderen 6 Eenouderhuishouden 7 Overig huishouden 8 Institutioneel huishouden
hh_duur	i	Aantal maanden sinds start van huishouden	
nkinderbij_[j0,...,j2,s0,m0,...,m5]	n	Fractie van maanden in betreffende periode dat aantal kinderen in huishouden toenam	
nkinderaf_[j0,...,j2,s0,m0,...,m5]	n	Fractie van maanden in betreffende periode dat aantal kinderen in huishouden afnam	
kindgeboren_[j0,...,j2,s0,m0,...,m5,duur]	n	Fractie van maanden in betreffende periode en aantal maanden geleden dat in het huishouden van persoon een kind is geboren	
inkbronhh_<cat>_[j0,...,j2]	[12]	Voornaamste inkomensbron van huishouden	11 Loon 12 Loon directeur-grootaandeelhouder 13 Winst zelfstandig ondernemer 14 Inkomen overige zelfstandige 21 Werkloosheidsuitkering 22 Bijstandsuitkering 23 Uitkering sociale voorziening overig 24 Uitkering ziekte/arbeidsongeschiktheid 25 Pensioenuitkering 26 Studiefinanciering 30 Inkomen uit vermogen 99 Huishoudensinkomen onbekend

Loopt door op volgende pagina

Voortgezet van vorig pagina

Variabele	Type	Omschrijving	Indeling
woning_<cat>_[j0,...,j2]	[4]	Woningbezit	1 Eigen woning 2 Huurwoning, met huurtoeslag 3 Huurwoning, zonder huurtoeslag 9 Onbekend
hh_inkomensarm_[j0,...,j2]	d	Persoon behoort tot een huishouden met een inkomen tot de lage inkomensgrens	
lang_arm_[j0,...,j2]	d	Langdurig inkomen tot de lage inkomensgrens	
jeugdhulp_[j0,...,j2]	d	Indicator voor gebruik jeugdhulp, -bescherming of re-classering	

B Hyperparameter tuning

Zoals in paragraaf 3.2 uitgelegd, hebben machine-learningmodellen in het algemeen zogenaamde hyperparameters die bepaald moeten worden. Dit wordt gedaan met behulp van geneste kruisvalidatie. In deze paragraaf worden de resultaten van het vaststellen/tunen van de hyperparameters van de twee modellen besproken.

B.1 Model arme populatie

Onderstaande figuur toont het effect van de verschillende hyperparameters op de negatieve log-likelihood (NLL; zie paragraaf 3.1) voor het model voor de arme populatie. Op de x-as staat het aantal bomen. In de kolommen staat de maximale boomdiepte en in de rijen de leersnelheid. De minimale waarde van de negatieve loglikelihood wordt aangegeven met de horizontale lijn en de punt geeft de combinatie van hyperparameters aan waarop dit minimum ligt. De figuur daaronder toont hetzelfde maar dan voor de Somers' D.

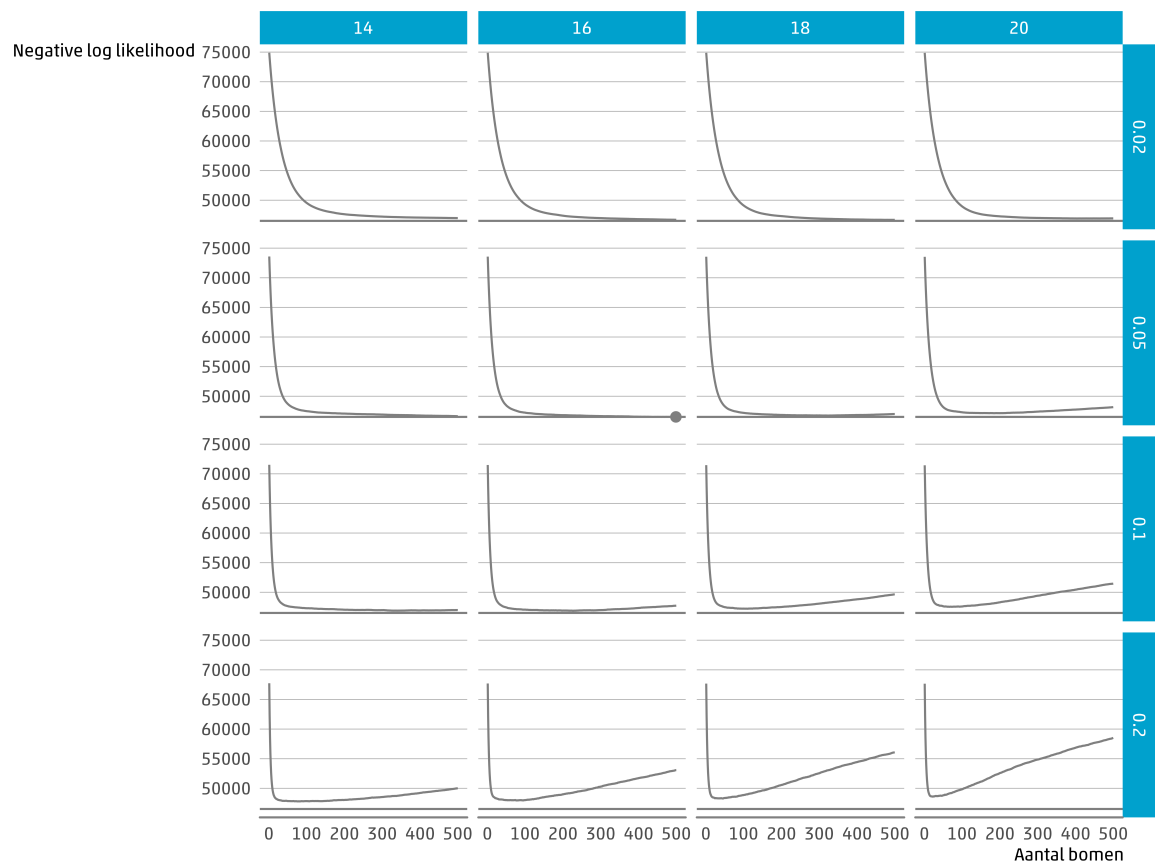
Hoe hoger de leersnelheid (rijpanelen) en hoe dieper de bomen (kolompanelen), hoe gevoeliger het model voor het aantal beslisbomen (x-as). Een lage leersnelheid in combinatie met veel bomen geeft over het algemeen een beter resultaat maar verhoogt ook de rekentijd. Veel andere kwaliteitsmaten blijven toenemen met modelcomplexiteit (bijvoorbeeld Somers' D), terwijl de negatieve loglikelihood aangeeft dat er op een gegeven moment overfit optreedt en de modelkwaliteit slechter wordt. Omdat de negatieve loglikelihood de overall fit van het model weergeeft is besloten dit criterium te gebruiken bij selectie van de optimale set hyperparameters. Hiermee komen we dus uit op een maximale boomdiepte van 16, een leersnelheid van 0,05 en 499 bomen.

Een van de dingen die mogelijk opvalt is dat de onderzochte parameters afwijken van de parameters die standaard in xgboost gebruikt worden (voornamelijk de maximale boom diepte die standaard op 6 staat). In verkennende analyses met subsets van de data en variabelen bleek dat het optimum niet bij deze waarden te vinden zou zijn. Dit houdt waarschijnlijk verband met het grote aantal variabelen in het model en het grote aantal waarnemingen.

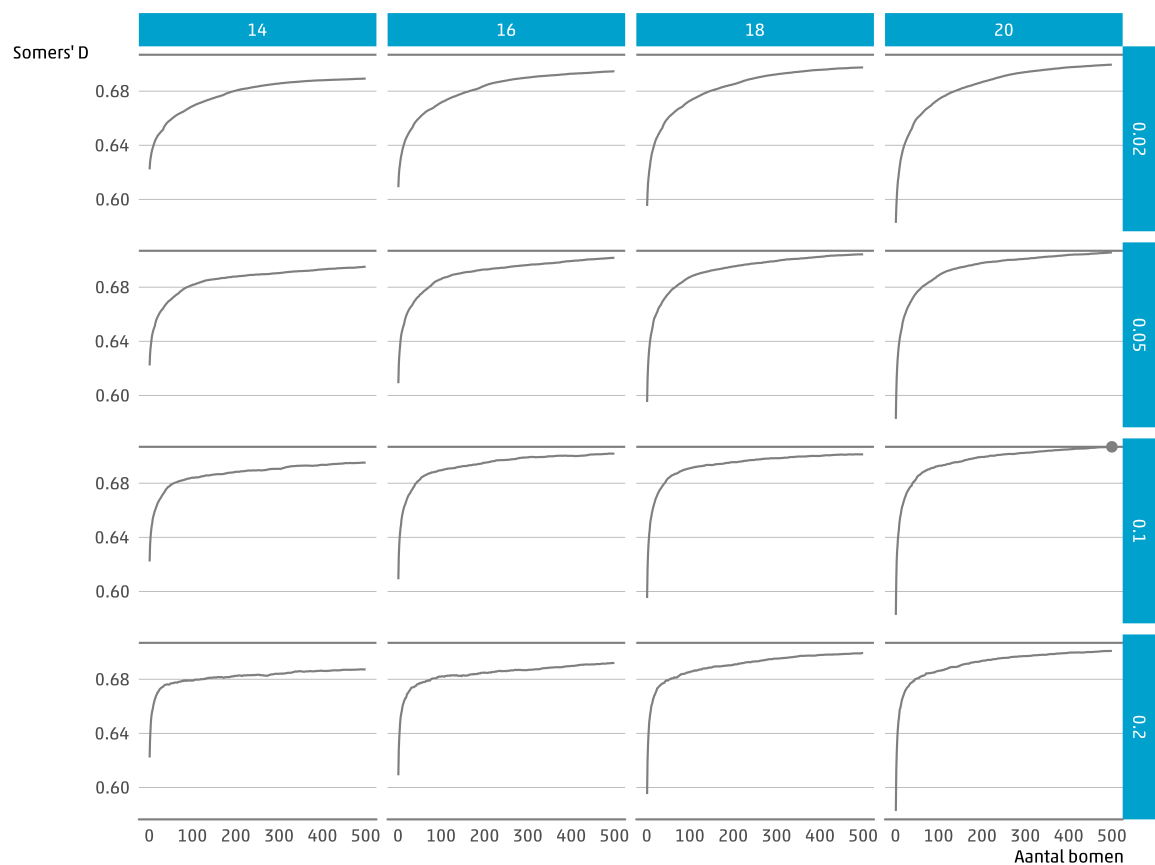
B.2 Model niet-arme populatie

Doordat de niet-arme populatie veel groter is dan de arme populatie, zijn de rekentijden voor het schatten van modellen voor de niet-arme populatie veel groter (> 3 dagen per model). Daarom is er een minder ruime set hyperparameters uitgeprobeerd bij het tunen van dit model. Daarnaast zijn in eerste instantie bij het tunen van het model buurt- en wijkenmerken meegenomen. Later is besloten deze niet mee te nemen in het model. Onderstaande figuren tonen de resultaten van het tunen van hyperparameters voor het model voor de niet-arme populatie (bij de arme populatie is het hele tunen over gedaan hetgeen geen probleem was in verband met de kortere rekentijden bij die populatie). Het optimum ligt bij een maximale boomdiepte van 14, een leersnelheid van 0,05 en 662 bomen.

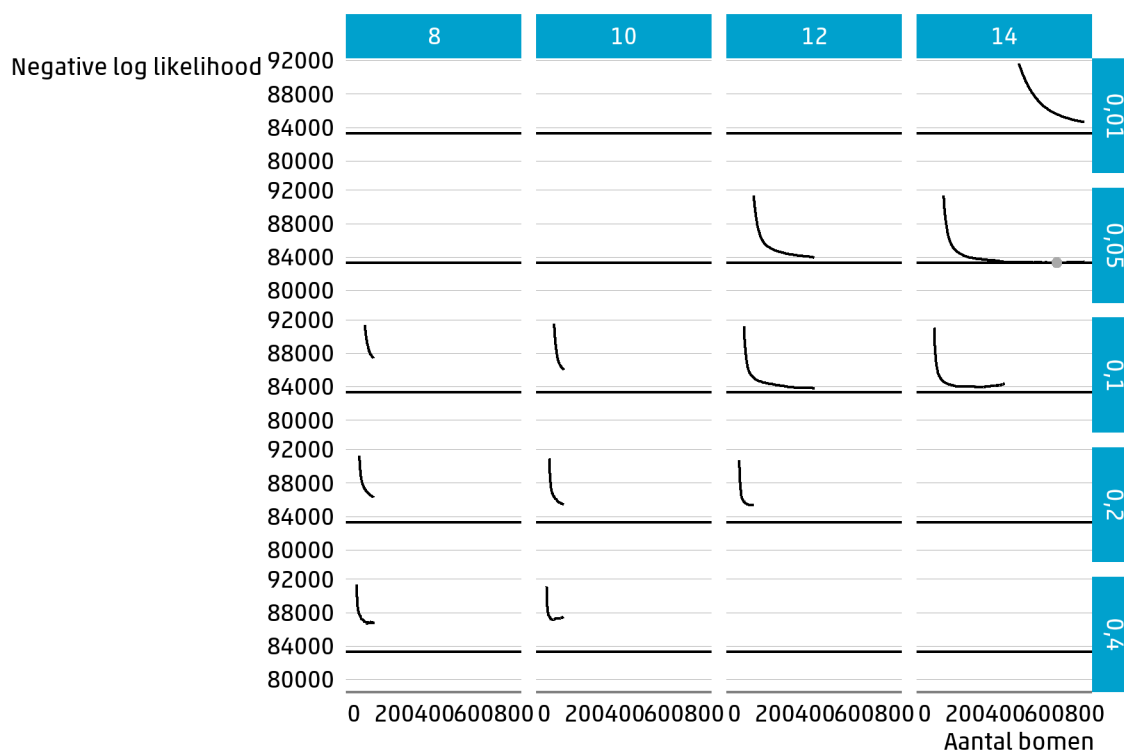
Aangezien met het weglaten van de buurt- en wijkenmerken er slechts enkele tientallen variabelen uit het model verdwijnen die ook nog sterk gecorreleerd zijn, is het niet waarschijnlijk dat voor het model zonder deze kenmerken op erg afwijkende hyperparameters uitgekomen wordt. Daarom is voor het model zonder wijk- en buurtkenmerken alleen het aantal bomen getuned en zijn de maximale boomdiepte en de leersnelheid vastgehouden op respectievelijk 14 en 0,05. De resultaten hiervan staan in figuur B.5. Het optimum ligt nu bij 450 bomen. Dit is ook het aantal dat uiteindelijk gebruikt is.



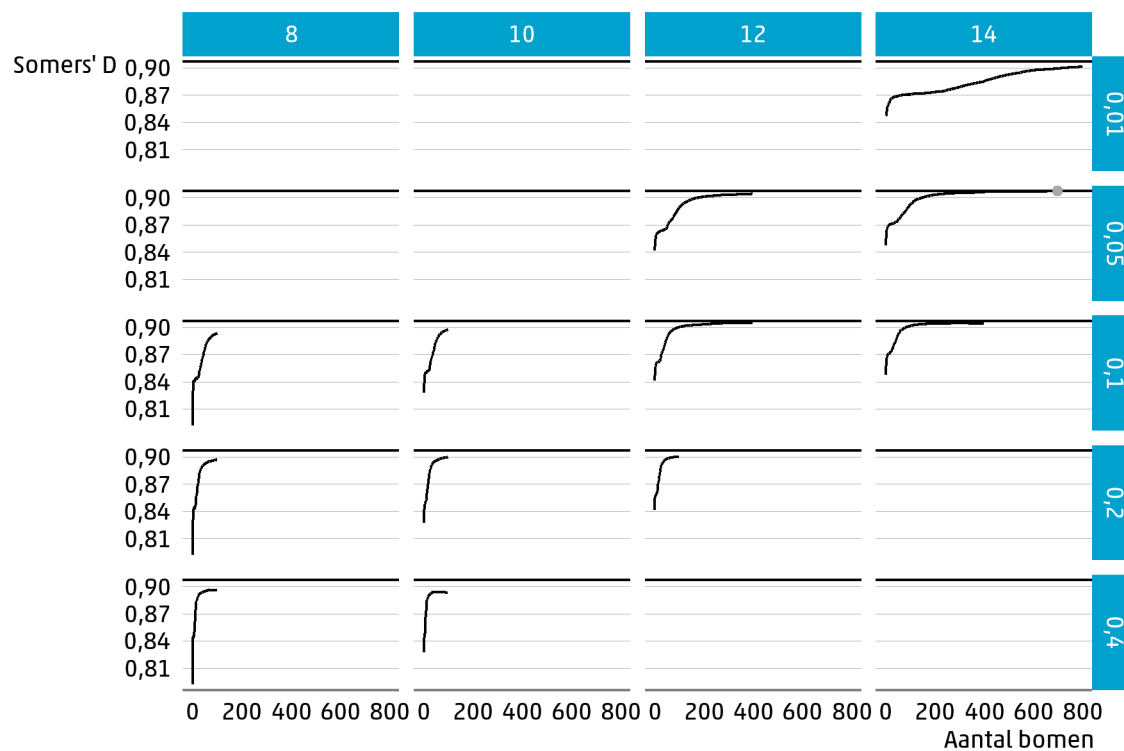
Figuur B.1 Hyperparameter tuning voor arme populatie: negatieve log likelihood



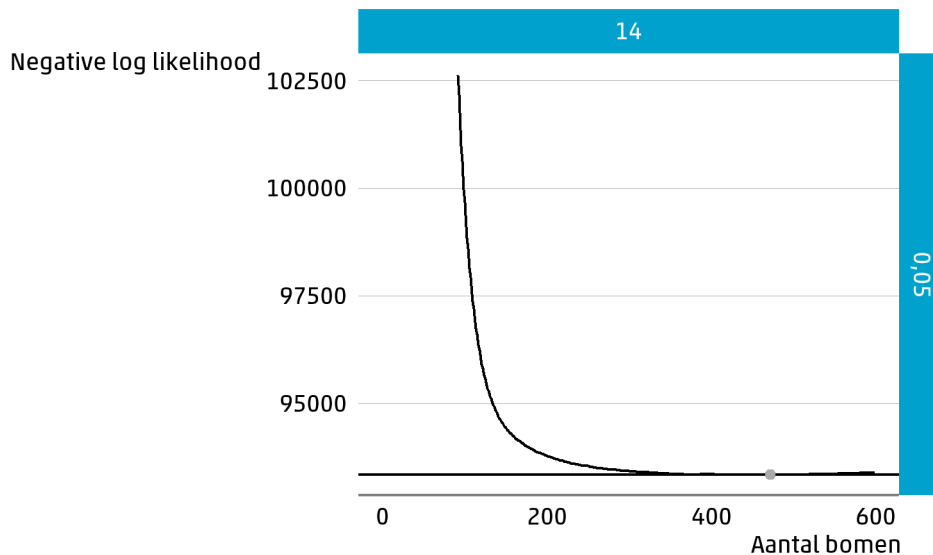
Figuur B.2 Hyperparameter tuning voor arme populatie: Somers' D .



Figuur B.3 Hyperparameter tuning voor niet-arme populatie: negative log likelihood. Model met buurtkenmerken.



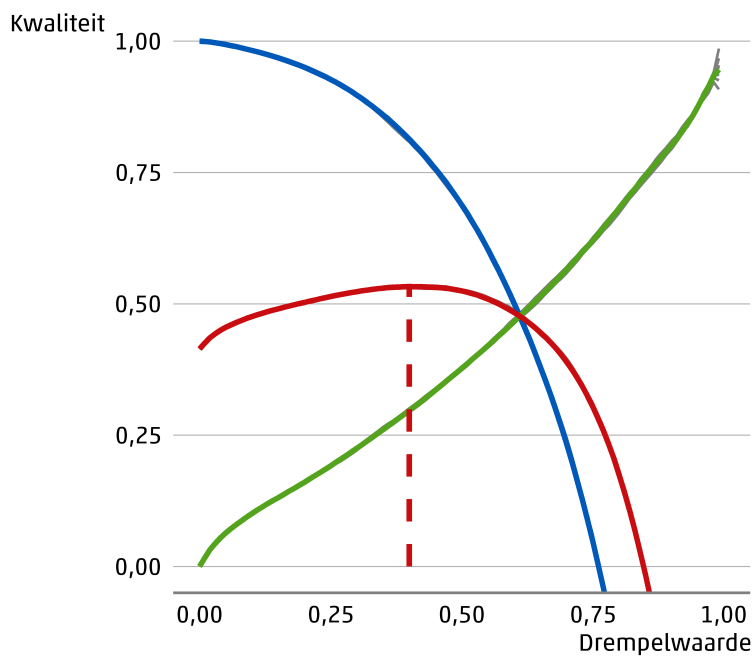
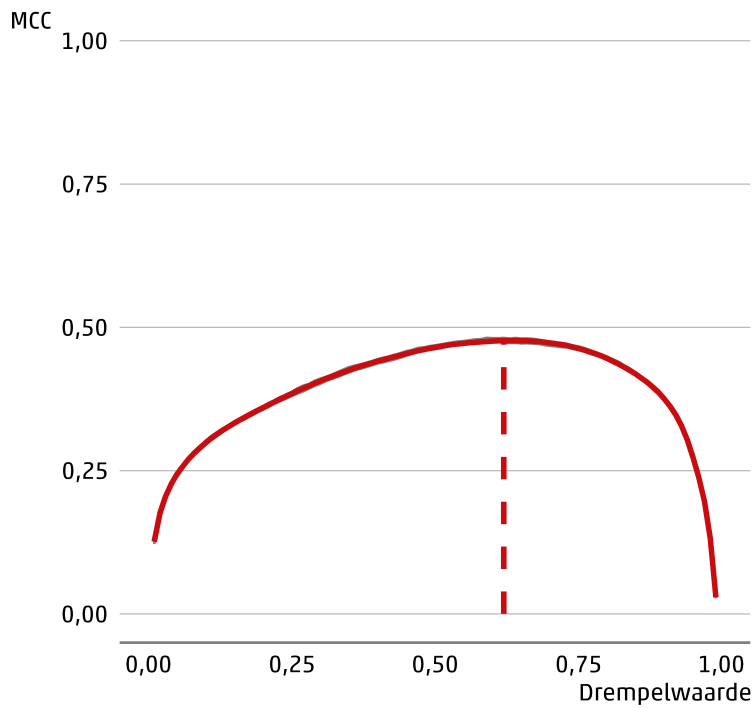
Figuur B.4 Hyperparameter tuning voor niet-arme populatie: Somers'D. Model met buurtkenmerken.



Figuur B.5 Hyperparameter tuning voor niet-arme populatie: negative log likelihood. Model zonder buurtkenmerken.

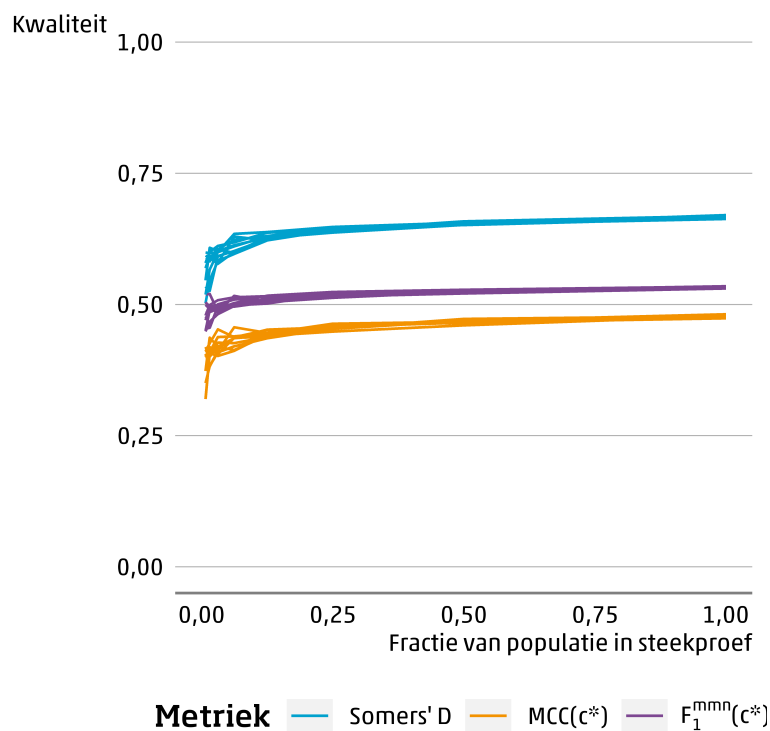
C Modelkwaliteit

Onderstaande twee figuren illustreren de samenhang tussen een aantal kwaliteitmaten en de drempelwaarde waarboven de geschatte armoedekans wordt vertaald naar het label 'blijft arm'. Hoe hoger de drempelwaarde, hoe minder goed- en foutpositieven en meer goed- en foutnegatieven. Daardoor neemt de recall (classificatiekans voor 'blijft arm') af maar de precision (calibratiekans voor 'blijft arm') toe. MCC en F1 zoeken allebei naar een balans, waarbij MCC de nadruk legt op de balans tussen foutpositieven en -negatieven, en F1 tussen recall en precision. De optimale drempelwaarde (verticale stippellijn) hangt dus af van de kwaliteitmaat. De gekleurde lijnen zijn gemiddeldes over $K = 10$ testsets (grijze lijnen). Er is vrijwel geen variatie tussen de testsets.

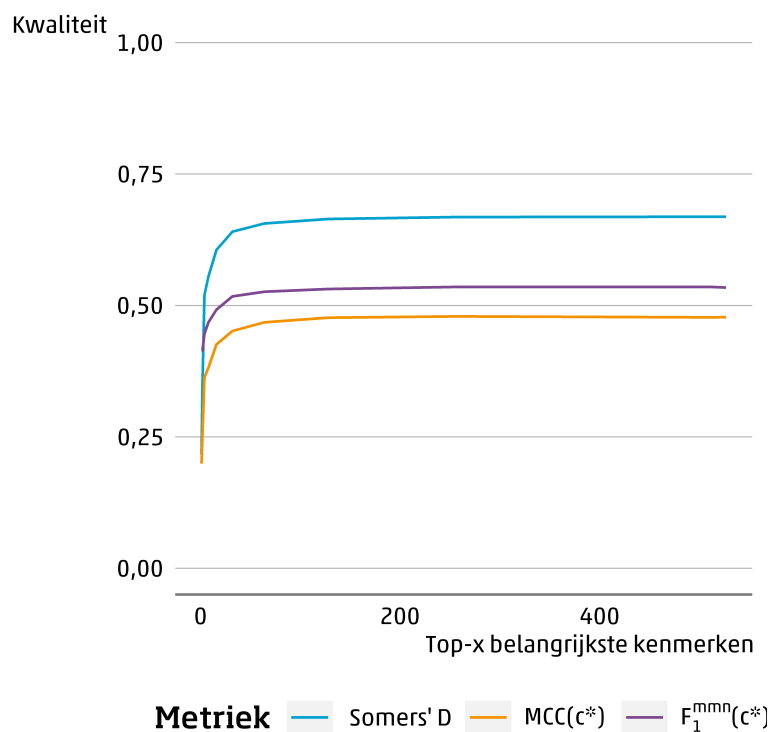


Metriek — Recall^{mmn} — Precision^{mmn} — F₁^{mm}

In onderstaande figuur is het effect van steekproefomvang op de modelkwaliteit getest voor de arme populatie ($K = 10$ testsets). Dit (empirische) verband is relevant voor studies waarbij niet de gehele populatie wordt waargenomen of wanneer geheugen of rekentijd beperkt is. Hoe kleiner de (enkelvoudig aselechte) steekproef, hoe lager de gemiddelde kwaliteit en hoe groter de spreiding tussen testsets, maar de effecten worden pas zichtbaar wanneer minder dan een vijfde van de populatie wordt waargenomen.



In onderstaande figuur is het effect van het aantal potentiële kenmerken op de modelkwaliteit getest voor de arme populatie ($k = 1$ testset). Kenmerken zijn gesorteerd op belang volgens de gemiddelde absolute SHAP-waarden. Hoewel slechts een vijfde van de kenmerken iets bijdraagt aan de modelkwaliteit, heeft het meenemen van redundante kenmerken er geen negatief effect op.



D Decompositie SHAP

In paragraaf 3.3 wordt het verschil in belang van een variabele in een deelpopulatie ten opzicht van het belang van die variabele in de gehele populatie opgesplitst in twee componenten. Eén component beschrijft het effect van verschillen in de verdeling van de variabele in een deelpopulatie: een variabele kan bijvoorbeeld minder belangrijk worden als er meer jongeren in een deelpopulatie zitten. De andere component beschrijft het verschil in het effect van een variabele in de deelpopulatie.

De SHAP-waarde s_{ij} geeft het belang van kenmerk x_j voor persoon i . Voor de leesbaarheid laten we de indicator j hier weg. De absolute SHAP-waardes gesommeerd over alle personen is een maat voor het algemene belang (*importance* I) van een kenmerk. In deelpopulatie h kan het belang van een kenmerk, I_h , groter of kleiner zijn dan in de gehele populatie. Dat kan twee oorzaken hebben: deelpopulatie h heeft een andere frequentieverdeling van het kenmerk, bijvoorbeeld meer jongeren en minder ouderen, of deelpopulatie h bevat een groep met dezelfde waarde maar een ander effect van het kenmerk, bijvoorbeeld meer jongeren met hoge SHAP-waardes en minder jongeren met lage SHAP-waardes. We zijn dus geïnteresseerd in het verschil in het belang van een kenmerk tussen deelpopulatie h en de gehele populatie, $I_h - I$, en de decompositie van dit verschil in een component die het belang van de verdeling van het kenmerk weergeeft en een component die het belang van het effect van het kenmerk weergeeft. Het belang van een variabele kan geschreven worden als

$$I = \sum_i |s_i|$$

De variabele x kan een eindig aantal waardes g aannemen. We kunnen bovenstaande dus ook schrijven als

$$I = \sum_g f_g S_g$$

waarin f_g de relatieve frequentie van waarde $x = g$:

$$f_g = \frac{n_g}{n},$$

n_g is het aantal personen met $x = g$ en n het aantal personen. Verder gebruiken we

$$S_g = \frac{1}{n_g} \sum_{i:x=g} |s_i|.$$

waarbij we dus de absolute SHAP-waarde middelen over de personen die waarde g hebben op kenmerk x .

Het verschil in belang van de variabele wordt gegeven door

$$\Delta I = I^{(h)} - I$$

waarin het superscript (h) aangeeft dat het deelpopulatie h betreft. Dit is uit te werken als:

$$\begin{aligned}
\Delta I &= I^{(h)} - I \\
&= \sum_g f_g^{(h)} S_g^{(h)} - \sum_g f_g S_g \\
&= \sum_g f_g^{(h)} S_g^{(h)} - \sum_g f_g^{(h)} S_g + \sum_g f_g^{(h)} S_g - \sum_g f_g S_g \\
&= \sum_g f_g^{(h)} (S_g^{(h)} - S_g) + \sum_g S_g (f_g^{(h)} - f_g) \\
&= \Delta I^{\text{kenm}} + \Delta I^{\text{verd}}
\end{aligned}$$

waarin

$$\Delta I^{\text{kenm}} = \sum_g f_g^{(h)} (S_g^{(h)} - S_g)$$

het kenmerkeffect beschrijft, dus het deel van het verschil dat wordt veroorzaakt doordat het kenmerk een groter of kleiner effect heeft op de geschatte armoedekans, en

$$\Delta I^{\text{verd}} = \sum_g S_g (f_g^{(h)} - f_g).$$

het verdelingseffect, dus het deel van het verschil dat wordt veroorzaakt doordat het kenmerk in de deelpopulatie anders verdeeld is dan in de gehele populatie. Het verschil in effect $S_g^{(h)} - S_g$ wordt dus gewogen met de frequentie in de deelpopulatie $f_g^{(h)}$ en het verschil in frequentie $f_g^{(h)} - f_g$ met het effect in de gehele populatie S_g .

E Deelpopulaties

E.1 Arme populatie - Amsterdam

Deze paragraaf zoomt in op de arme deelpopulatie Amsterdam:

- hoe is deze deelpopulatie afgebakend,
- in hoeverre verschilt deze deelpopulatie qua kenmerken van de hele arme populatie,
- hoe groot is de kans om arm te blijven,
- in hoeverre verschillen de kenmerken die het belangrijkste zijn voor het schatten van de armoedekans van het landelijke beeld,
- welk effect hebben deze kenmerken op de kans om arm te blijven.

Afbakening

- In referentiejaar arm;
- In referentiejaar inwoner gemeente Amsterdam.

Deze deelpopulatie betreft personen die arm zijn in het referentiejaar (t) en die woonachtig zijn in de gemeente Amsterdam in het referentiejaar (t).

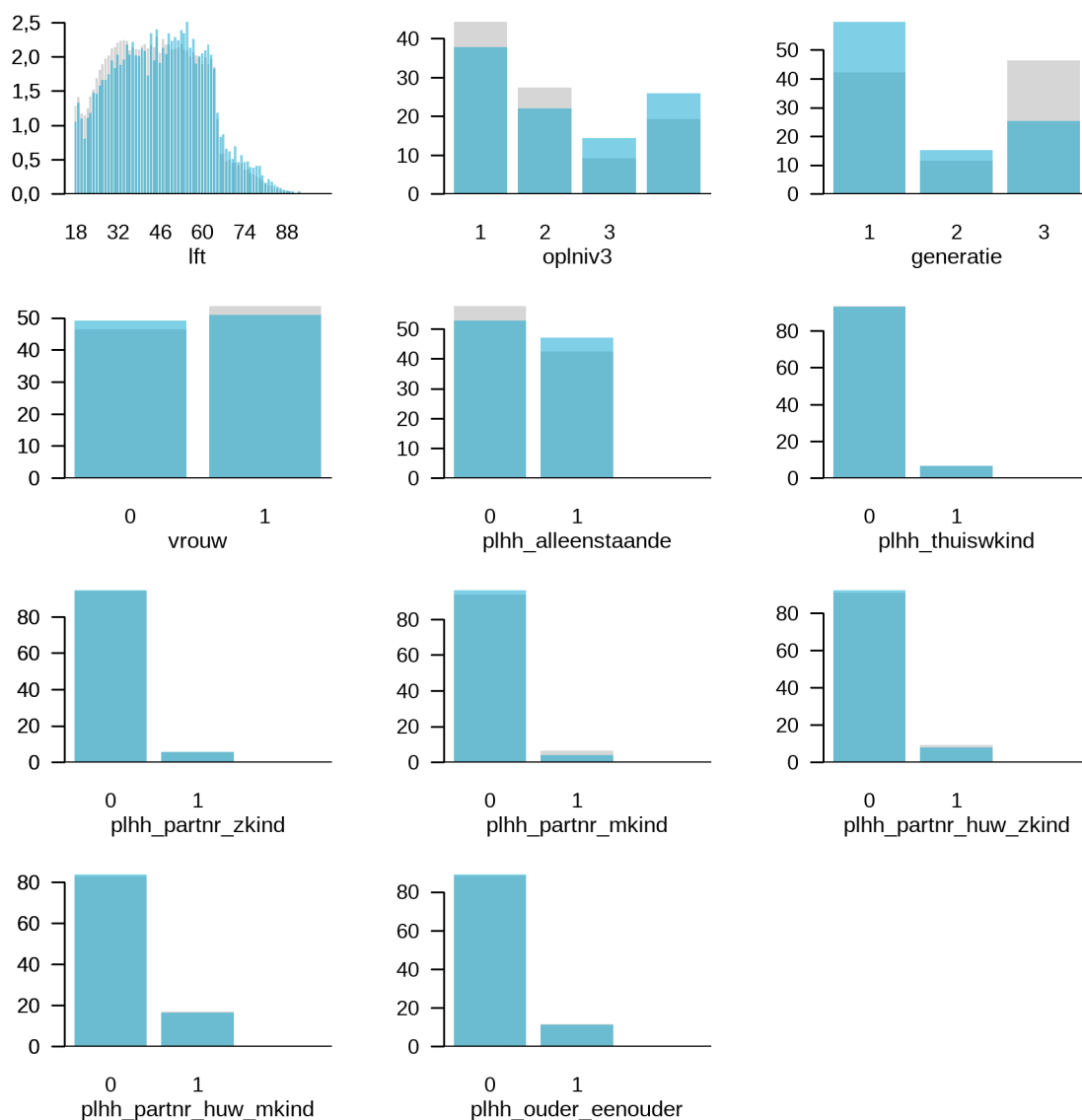
Kenmerken

Om beter zicht te krijgen op deze deelpopulatie hebben we gekeken naar een aantal achtergrondkenmerken:

- leeftijd ($1ft$)
- opleidingsniveau ($opl\text{niv}3$)
- generatie migratieachtergrond
 - eerste generatie migratieachtergrond ($generatie = 1$)
 - tweede generatie ($generatie = 2$)
 - Nederlandse achtergrond ($generatie = 3$)
- geslacht (vrouw: $vrouw = 1$)
- type huishouden
 - alleenstaande ($plhh_alleenstaande = 1$)
 - thuiswonend kind ($plhh_thuiswkind = 1$)
 - ongehuwd partner zonder kinderen ($plhh_partn_zkind = 1$)
 - ongehuwd partner met kinderen ($plhh_partn_mkind = 1$)
 - gehuwd partner met kinderen ($plhh_huw_zkind = 1$)
 - gehuwd partner zonder kinderen ($plhh_huw_mkind = 1$)
 - ouder in eenouderhuishouden ($plhh_ouder_eenouder = 1$)

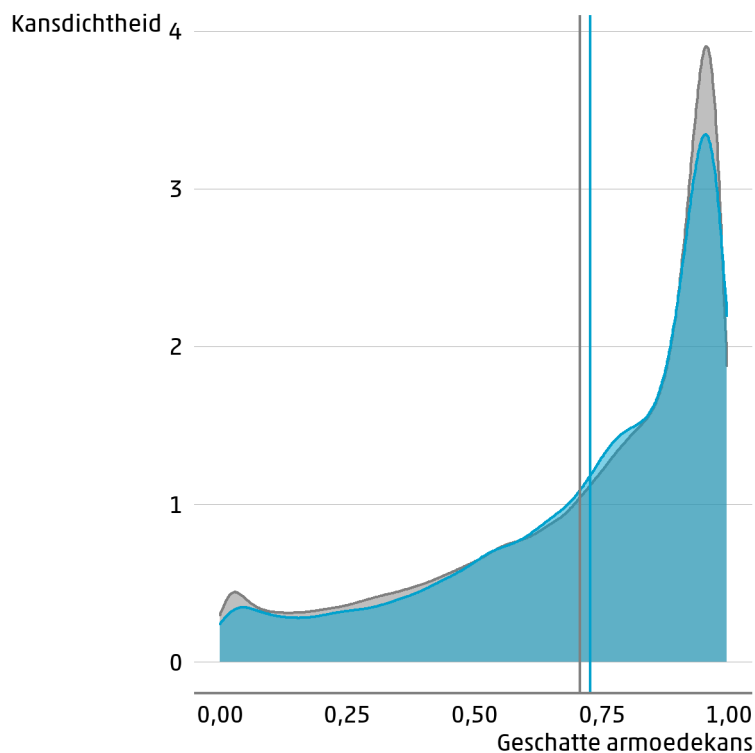
Onderstaande figuren tonen de verdeling van deze kenmerken in deze deelpopulatie (blauw) in vergelijking met de gehele arme populatie (grijs). Merk op dat het opleidingsniveau deels op registraties en deels op enquêtes gebaseerd is, waardoor dit niet voor de gehele populatie bekend is. Hierdoor is de groep met een onbekend opleidingsniveau relatief groot. Ook is dit een selectieve groep. Zo is het opleidingsniveau relatief vaak onbekend voor ouderen en personen met een migratieachtergrond. Dit kan leiden tot vertekening in de opleidingsniveauverdeling. Daarom laten we dit buiten beschouwing bij het beschrijven van de populatiekenmerken.

Uit de figuren blijkt dat er in Amsterdam met name relatief meer personen met een eerste generatie migratieachtergrond zijn (generatie = 1), meer alleenstaanden (plhh_alleenstaande = 1) en relatief minder personen met een Nederlandse achtergrond (generatie = 3)



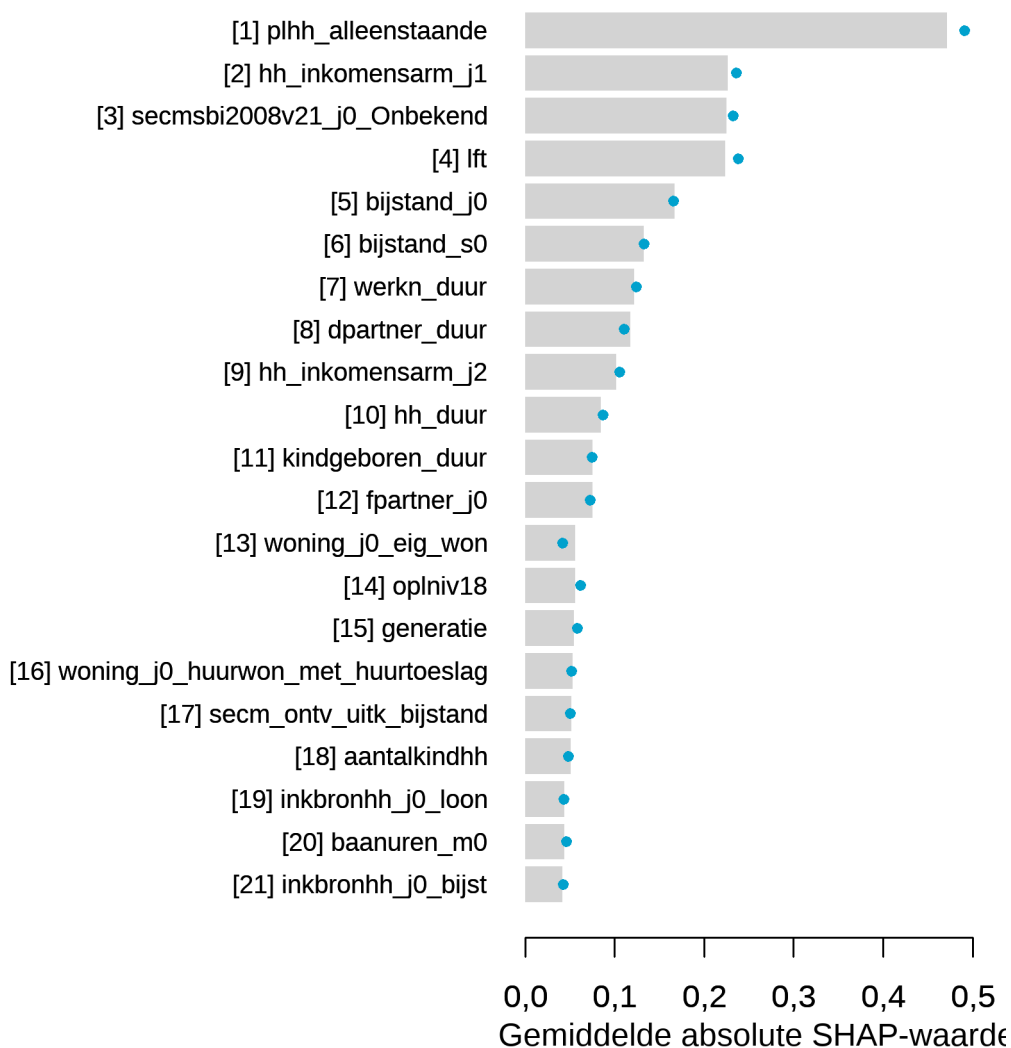
Kans om arm te blijven

Onderstaande figuur toont de verdeling van de geschatte kans om arm te blijven voor de arme deelpopulatie in Amsterdam (blauw) en voor de gehele arme populatie in Nederland (grijs). De verticale lijnen tonen de gemiddelde kans om arm te blijven. Hieruit blijkt dat de gemiddelde kans om arm te blijven in Amsterdam (73%) wat groter is dan in de gehele arme populatie in Nederland (71%).



Belangrijkste kenmerken

Onderstaande figuur geeft de belangrijkste kenmerken weer op basis van SHAP-waarden. (Zie paragraaf 3.3 voor verdere toelichting.) De 20 belangrijkste kenmerken in de gehele arme populatie (grijze balken) zijn aangevuld met de 20 belangrijkste kenmerken voor de deelpopulatie (blauwe punten). Bij kenmerken waarbij iedereen in de deelpopulatie dezelfde waarde heeft, ontbreekt de blauwe punt, aangezien dit kenmerk geen effect heeft binnen de deelpopulatie. In onderstaande figuur is te zien dat de belangrijkste kenmerken voor Amsterdam in grote mate overeenkomen met de gehele arme populatie. Zie paragraaf 4.1 voor een omschrijving van deze kenmerken. In de volgende figuren bekijken we hoe deze kenmerken de kans om arm te blijven beïnvloeden.



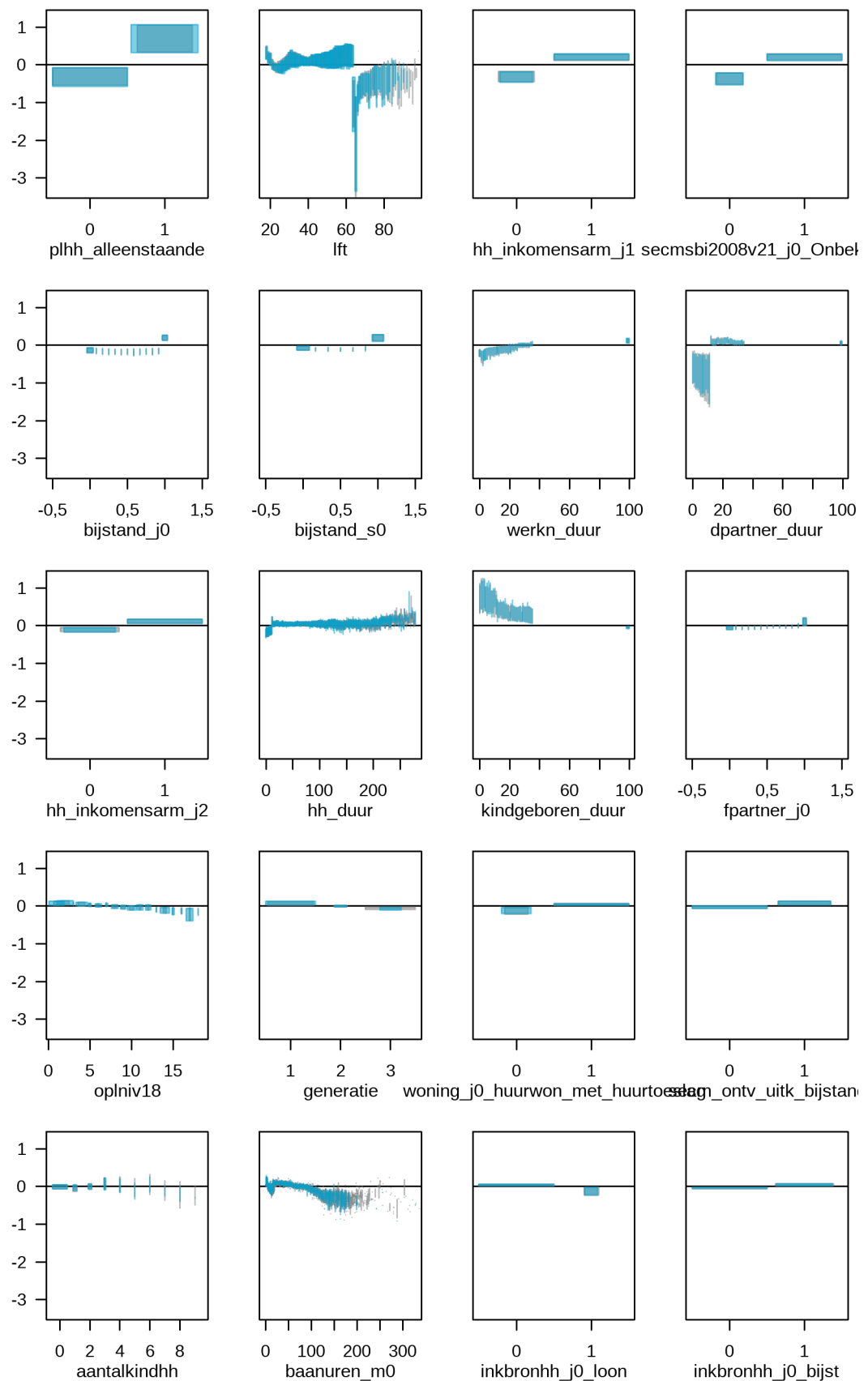
Effect op kans om arm te blijven

Onderstaande figuren tonen het *effect* op de kans om arm te blijven en de *decompositie* van het effect. De staafdiagrammen tonen het effect van de 20 belangrijkste kenmerken op de geschatte kans om arm te blijven voor de deelpopulatie (blauwe balken) en de gehele niet-arme populatie in Nederland (grijze balken). Een negatieve waarde op de y-as betekent dat personen met dit kenmerk een lagere kans dan gemiddeld hebben om arm te worden, een positieve waarde betekent dat personen een hogere kans dan gemiddeld hebben. De hoogte van de balk geeft de range van de kans weer voor 80% van de betreffende populatie. De breedte van de balk komt overeen met de relatieve grootte van de populatie met deze waarde. Een smalle balk betekent dus relatief weinig mensen en een brede balk relatief veel mensen. Zie paragraaf 3.3 voor verdere toelichting.

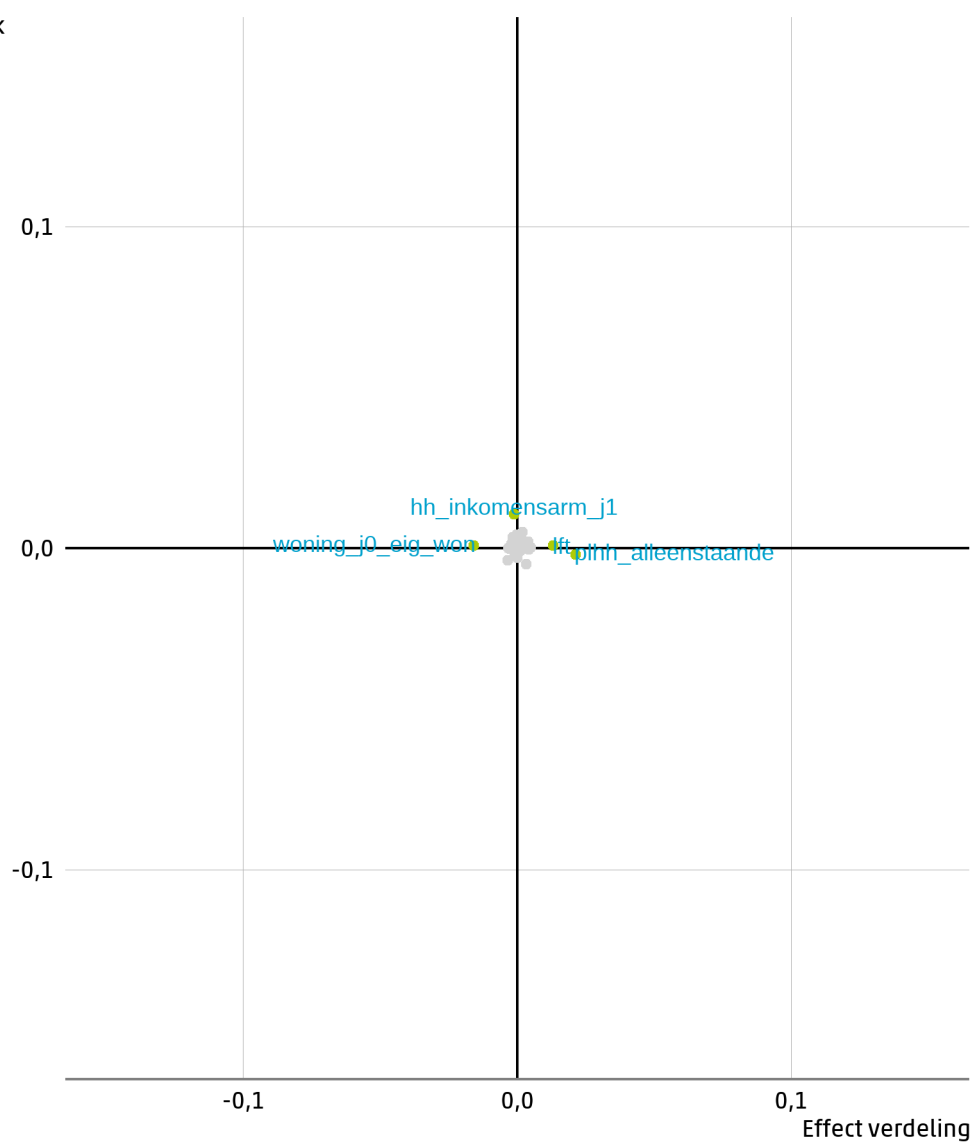
De tweede figuur toont de decompositie van het verschil in belang van de belangrijkste kenmerken. Het verschil in belang wordt opgesplitst in een component die het belang van de *verdeling* van het kenmerk weergeeft (horizontale as) en een component die het belang weergeeft van het *effect* dat het kenmerk heeft op de geschatte kans (verticale as). Zie paragraaf 3.3 voor verdere toelichting.

In onderstaande figuren is te zien dat de effecten van de belangrijkste kenmerken in Amsterdam in grote mate vergelijkbaar zijn met die van de gehele arme populatie, wat overeenkomt met de voorgaande figuur.

Ook de decompositie figuur laat zien dat alle kenmerken relatief dichtbij nul liggen. Dit betekent dat zowel het verschil in belang door de verdeling van de kenmerken als het verschil in belang door het effect van de kenmerken relatief klein is.



Effect kenmerk



E.2 Arme populatie - Heerlen

Deze paragraaf zoomt in op de arme deelpopulatie Heerlen:

- hoe is deze deelpopulatie afgebakend,
- in hoeverre verschilt deze deelpopulatie qua kenmerken van de hele arme populatie,
- hoe groot is de kans om arm te blijven,
- in hoeverre verschillen de kenmerken die het belangrijkste zijn voor het schatten van de armoedekans van het landelijke beeld,
- welk effect hebben deze kenmerken op de kans om arm te blijven.

Afbakening

- In referentiejaar arm;
- In referentiejaar inwoner gemeente Heerlen.

Deze deelpopulatie betreft personen die arm zijn in het referentiejaar (t) en die woonachtig zijn in de gemeente Heerlen in het referentiejaar (t).

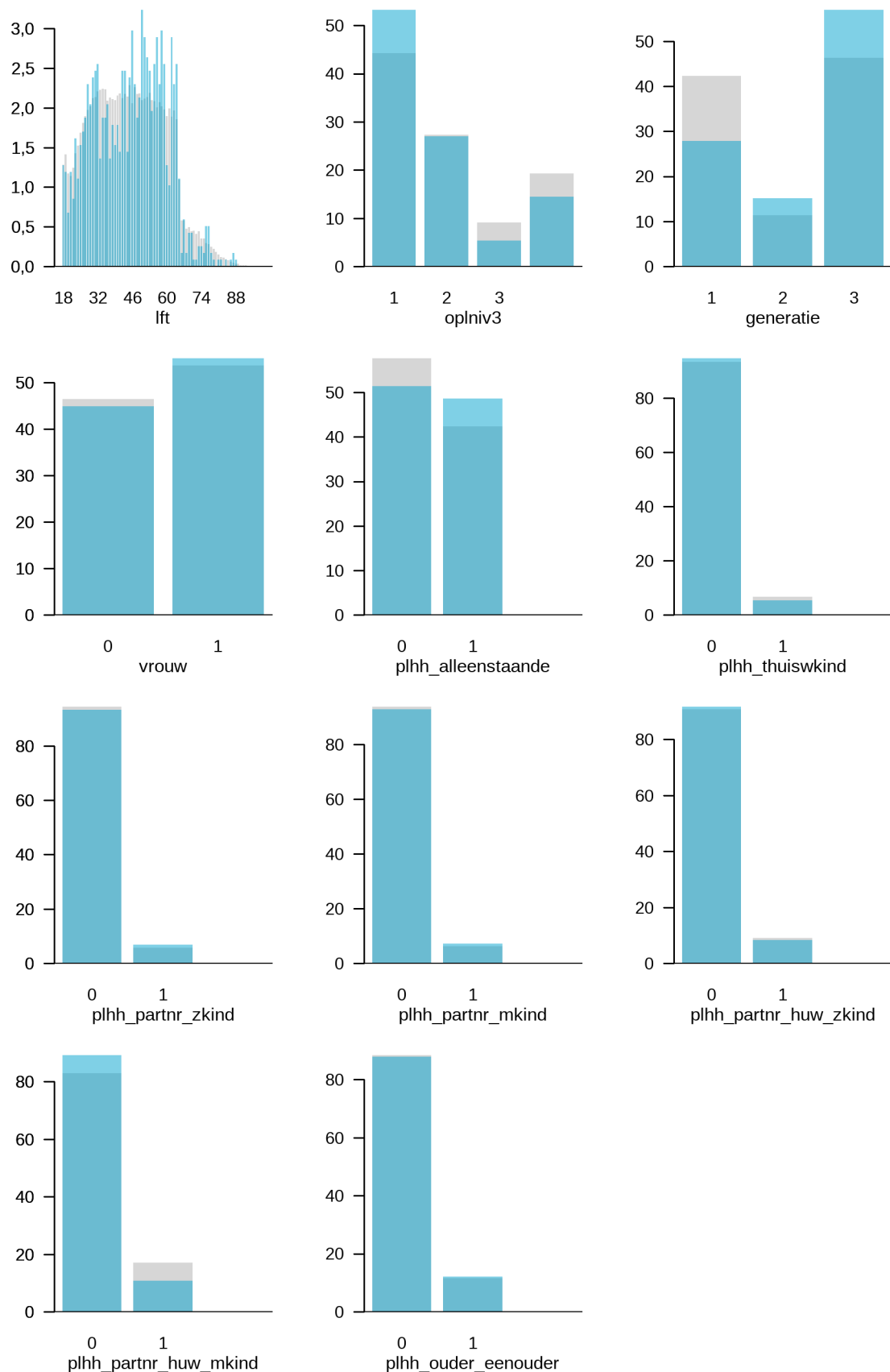
Kenmerken

Om beter zicht te krijgen op deze deelpopulatie hebben we gekeken naar een aantal achtergrondkenmerken:

- leeftijd (lft)
- opleidingsniveau ($oplniv3$)
- generatie migratieachtergrond
 - eerste generatie migratieachtergrond ($generatie = 1$)
 - tweede generatie ($generatie = 2$)
 - Nederlandse achtergrond ($generatie = 3$)
- geslacht (vrouw: $vrouw = 1$)
- type huishouden
 - alleenstaande ($plhh_alleenstaande = 1$)
 - thuiswonend kind ($plhh_thuiswkind = 1$)
 - ongehuwd partner zonder kinderen ($plhh_partn_zkind = 1$)
 - ongehuwd partner met kinderen ($plhh_partn_mkind = 1$)
 - gehuwd partner met kinderen ($plhh_huw_zkind = 1$)
 - gehuwd partner zonder kinderen ($plhh_huw_mkind = 1$)
 - ouder in eenouderhuishouden ($plhh_ouder_eenouder = 1$)

Onderstaande figuren tonen de verdeling van deze kenmerken in deze deelpopulatie (blauw) in vergelijking met de gehele arme populatie (grijs). Merk op dat het opleidingsniveau deels op registraties en deels op enquêtes gebaseerd is, waardoor dit niet voor de gehele populatie bekend is. Hierdoor is de groep met een onbekend opleidingsniveau relatief groot. Ook is dit een selectieve groep. Zo is het opleidingsniveau relatief vaak onbekend voor ouderen en personen met een migratieachtergrond. Dit kan leiden tot vertekening in de opleidingsniveauverdeling. Daarom laten we dit buiten beschouwing bij het beschrijven van de populatiekenmerken.

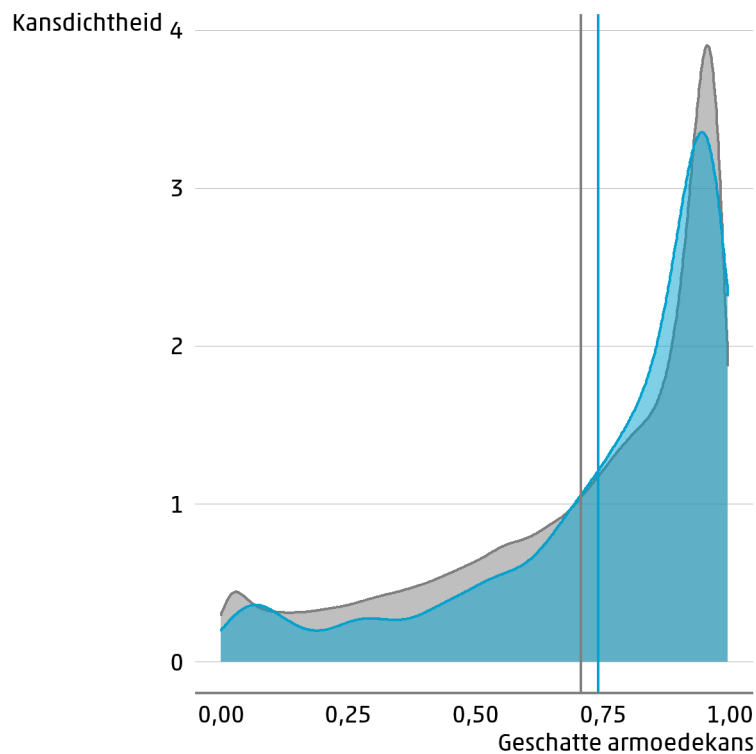
Uit de figuren blijkt dat er in Heerlen met name relatief meer personen met een tweede generatie migratieachtergrond ($generatie = 2$), met een Nederlandse achtergrond ($generatie = 3$) en meer alleenstaanden zijn ($plhh_alleenstaande = 1$) en relatief minder gehuwde partners met kinderen ($plhh_partner_huw_mkind = 1$).



Kans om arm te blijven

Onderstaande figuur toont de verdeling van de geschatte kans om arm te blijven voor de arme deelpopulatie in Heerlen (blauw) en voor de gehele arme populatie in Nederland (grijs). De verticale lijnen

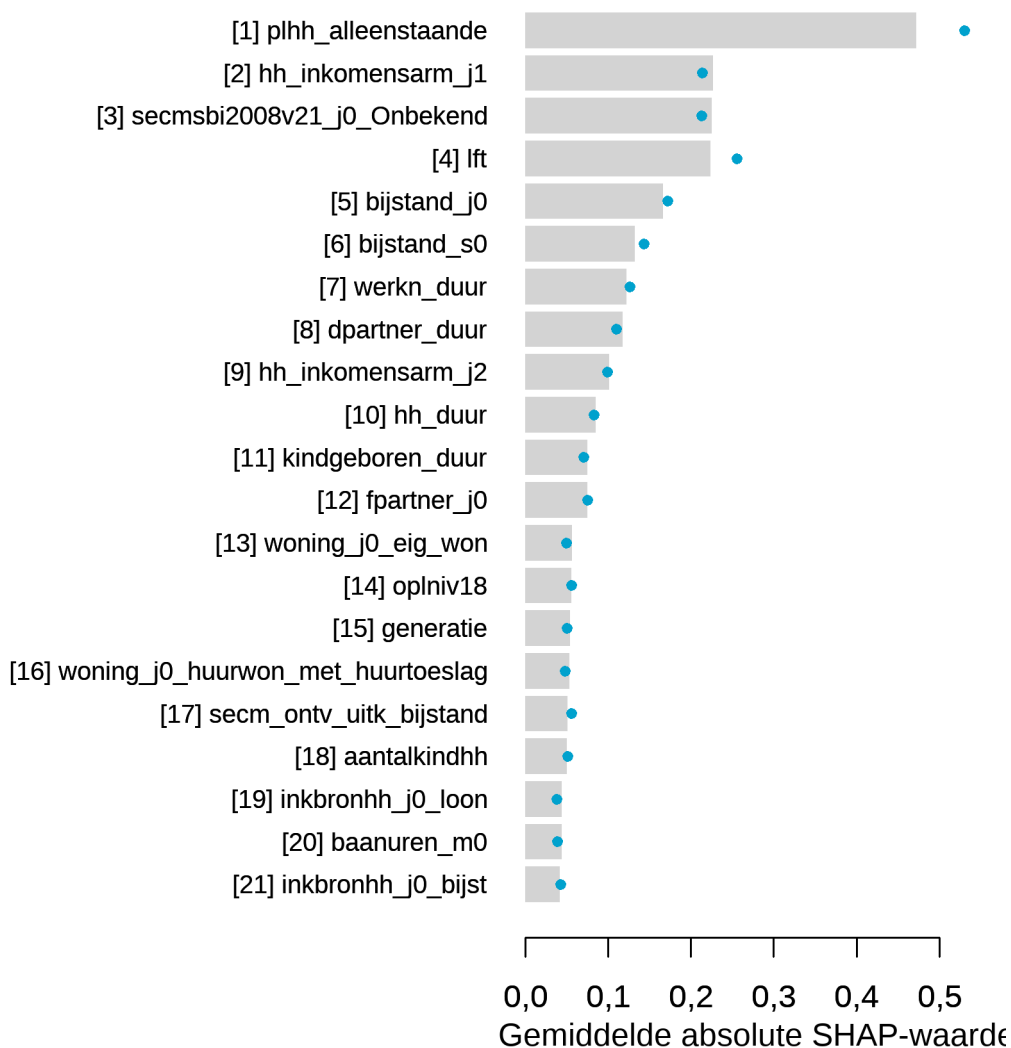
tonen de gemiddelde kans om arm te blijven. Hieruit blijkt dat de gemiddelde kans om arm te blijven in Heerlen (74%) wat groter is dan in de gehele arme populatie in Nederland (71%).



Belangrijkste kenmerken

Onderstaande figuur geeft de belangrijkste kenmerken weer op basis van SHAP-waarden. (Zie paragraaf 3.3 voor verdere toelichting.) De 20 belangrijkste kenmerken in de gehele arme populatie (grijze balken) zijn aangevuld met de 20 belangrijkste kenmerken voor de deelpopulatie (blauwe punten). Bij kenmerken waarbij iedereen in de deelpopulatie dezelfde waarde heeft, ontbreekt de blauwe punt, aangezien dit kenmerk geen effect heeft binnen de deelpopulatie.

In onderstaande figuur is te zien dat de belangrijkste kenmerken voor Heerlen in grote mate overeenkomen met de gehele arme populatie. (Zie paragraaf 4.1 voor een omschrijving van deze kenmerken.) Wel is er met name bij het eerder genoemde kenmerk alleenstaand (plhh_alleenstaande) een verschil met de gehele Nederlandse arme populatie te zien. In de volgende figuren bekijken we hoe deze kenmerken de kans om arm te blijven beïnvloeden.



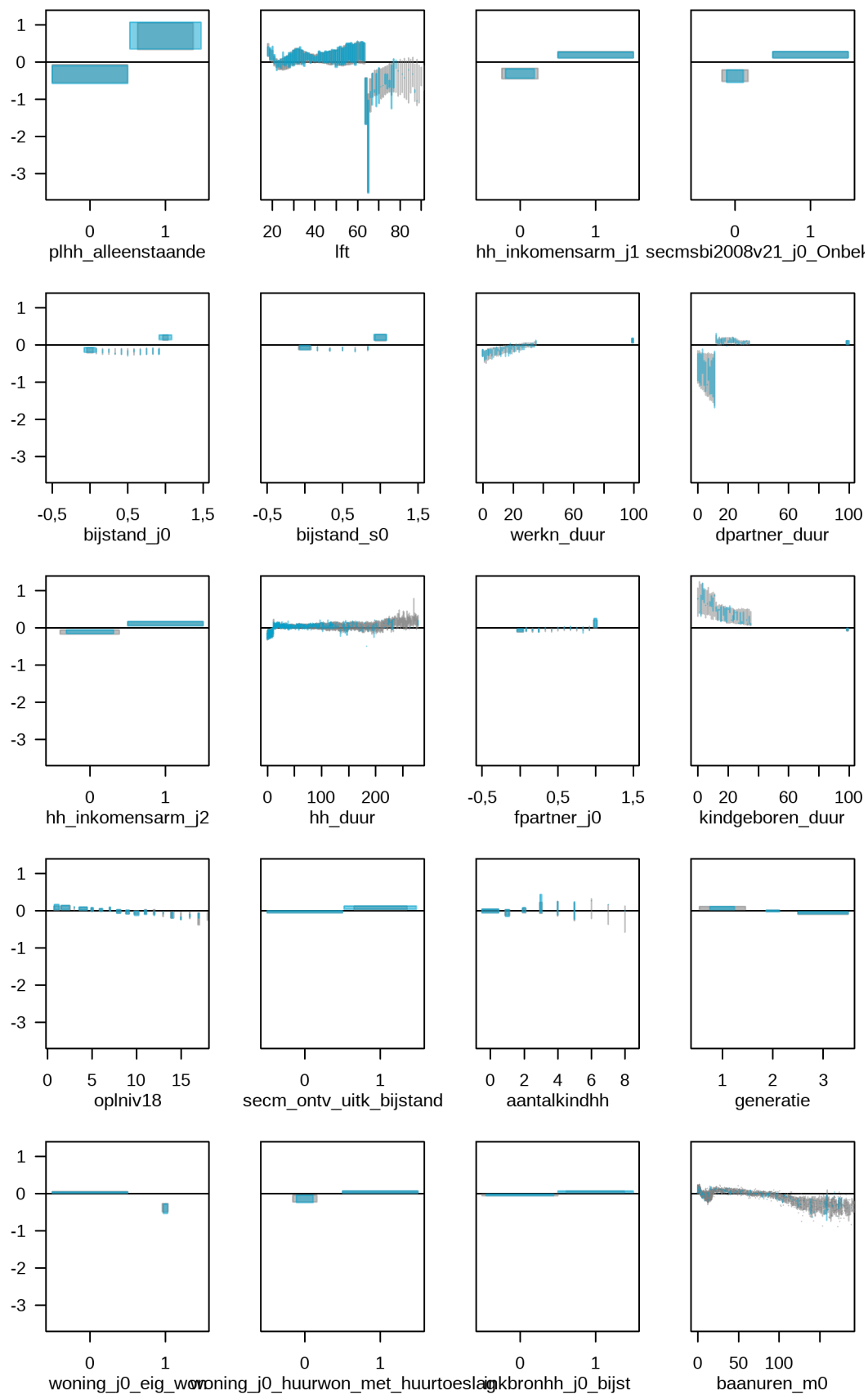
Effect op kans om arm te blijven

Onderstaande figuren tonen het *effect* op de kans om arm te blijven en de *decompositie* van het effect. De staafdiagrammen tonen het effect van de 20 belangrijkste kenmerken op de geschatte kans om arm te blijven voor de deelpopulatie (blauwe balken) en de gehele niet-arme populatie in Nederland (grijze balken). Een negatieve waarde op de y-as betekent dat personen met dit kenmerk een lagere kans dan gemiddeld hebben om arm te worden, een positieve waarde betekent dat personen een hogere kans dan gemiddeld hebben. De hoogte van de balk geeft de range van de kans weer voor 80% van de betreffende populatie. De breedte van de balk komt overeen met de relatieve grootte van de populatie met deze waarde. Een smalle balk betekent dus relatief weinig mensen en een brede balk relatief veel mensen. Zie paragraaf 3.3 voor verdere toelichting.

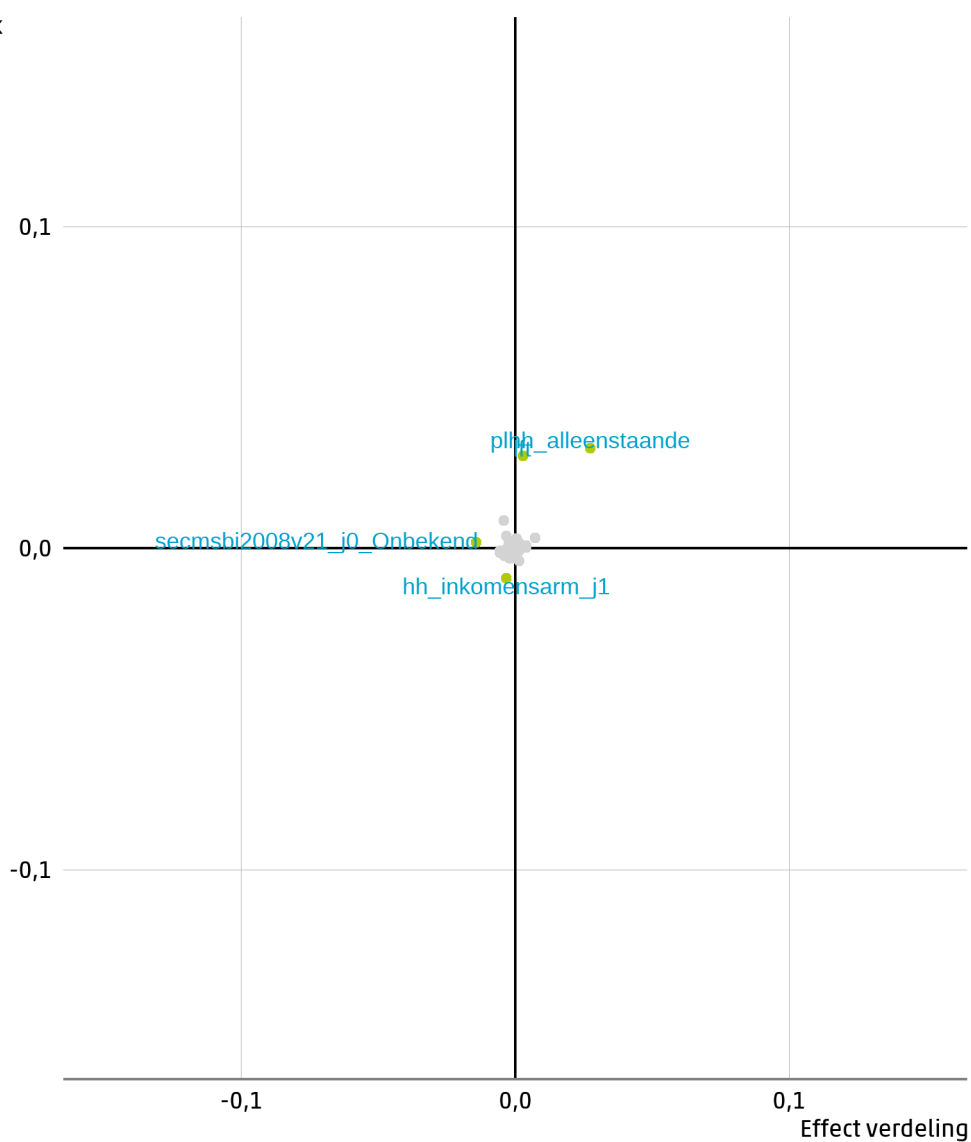
De tweede figuur toont de decompositie van het verschil in belang van de belangrijkste kenmerken. Het verschil in belang wordt opgesplitst in een component die het belang van de *verdeling* van het kenmerk weergeeft (horizontale as) en een component die het belang weergeeft van het *effect* dat het kenmerk heeft op de geschatte kans (verticale as). Zie paragraaf 3.3 voor verdere toelichting.

In staafdiagrammen is te zien dat de effecten van de belangrijkste kenmerken in Heerlen in grote mate vergelijkbaar zijn met die van de gehele niet-arme populatie. Wel is te zien dat het kenmerk alleenstaand in

deze deelpopulatie iets belangrijker is. Volgens de decompositie is er zowel een positief verdelingseffect als een positief kenmerkeffect. De deelpopulatie bevat meer alleenstaanden (met sterk effect op armoedekans) dan de gehele arme populatie. De blauwe balk is voor deze groep namelijk breder dan de grijze. Het positieve kenmerkeffect betekent dat alleenstaanden in deze deelpopulatie ook een grotere kans hebben om arm te blijven dan de gehele arme populatie, dan wel dat niet-alleenstaanden een kleinere kans hebben om arm te blijven. Dit effect is niet zichtbaar in het staafdiagram, waar alleen het 10e en 90e percentiel weergegeven worden.



Effect kenmerk



E.3 Arme populatie - laagopgeleide jongeren

Deze paragraaf zoomt in op de deelpopulatie arme laagopgeleide jongeren:

- hoe is deze deelpopulatie afgebakend,
- in hoeverre verschilt deze deelpopulatie qua kenmerken van de hele arme populatie,
- hoe groot is de kans om arm te blijven,
- in hoeverre verschillen de kenmerken die het belangrijkste zijn voor het schatten van de armoedekans van het landelijke beeld,
- welk effect hebben deze kenmerken op de kans om arm te blijven.

Afbakening

- In referentiejaar in de armoede;
- In referentiejaar laag opleidingsniveau;
- In referentiejaar jong (18+ en niet ouder dan 26 jaar);
- In referentiejaar niet schoolgaand of studerend;
- In referentiejaar geen thuiswonend kind.

Deze deelpopulatie betreft personen die in het referentiejaar t arm zijn, een laag opleidingsniveau hebben en jong zijn, maar niet schoolgaand/studerend zijn en niet meer thuis wonen.

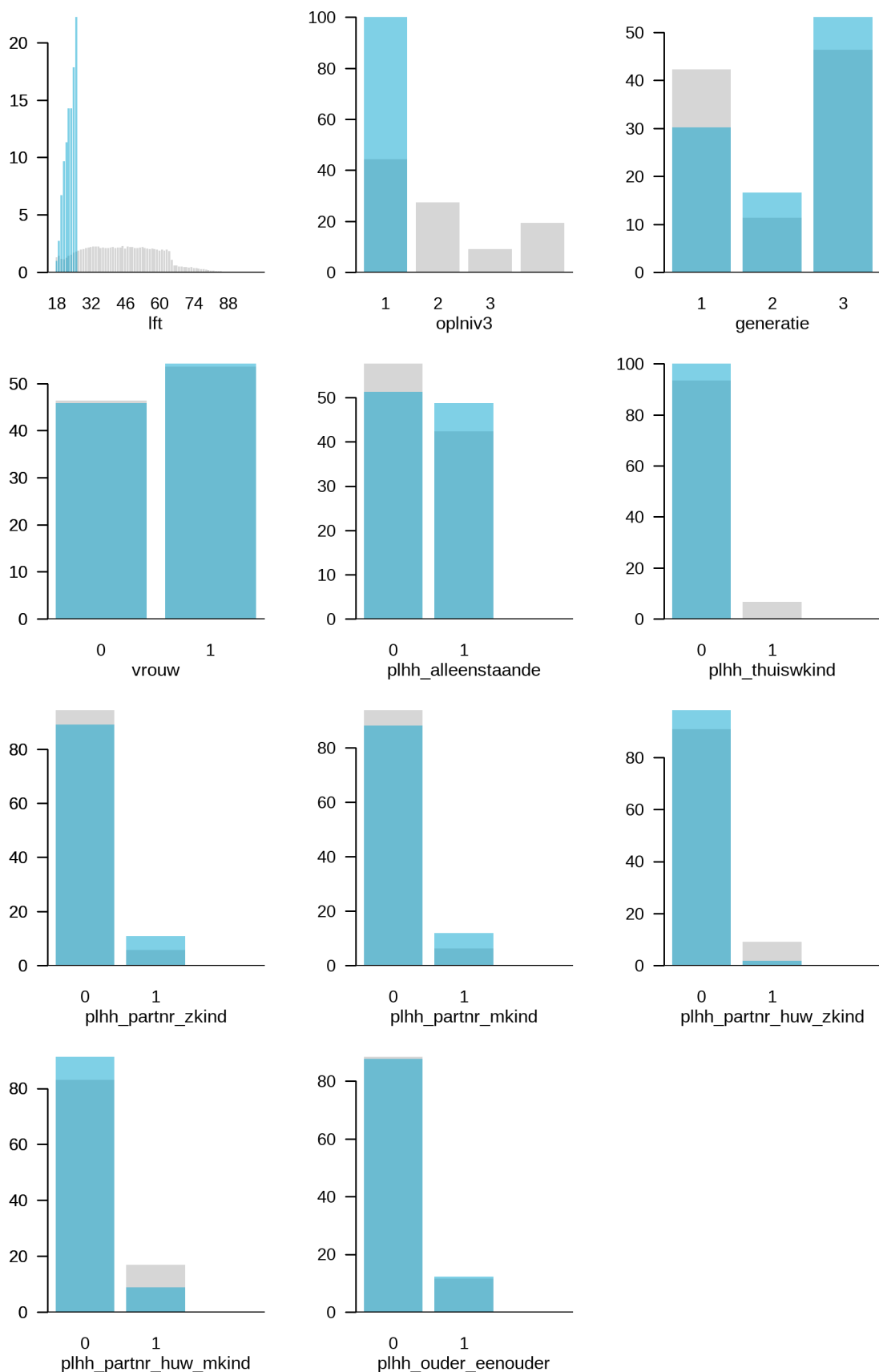
Kenmerken

Om beter zicht te krijgen op deze deelpopulatie hebben we gekeken naar een aantal achtergrondkenmerken:

- leeftijd (lft)
- opleidingsniveau (opl_niv3)
- generatie migratieachtergrond
 - eerste generatie migratieachtergrond (generatie = 1)
 - tweede generatie (generatie = 2)
 - Nederlandse achtergrond (generatie = 3)
- geslacht (vrouw: vrouw = 1)
- type huishouden
 - alleenstaande (plhh_alleenstaande = 1)
 - thuiswonend kind (plhh_thuiswkind = 1)
 - ongehuwd partner zonder kinderen (plhh_partn_zkind = 1)
 - ongehuwd partner met kinderen (plhh_partn_mkind = 1)
 - gehuwd partner met kinderen (plhh_huw_zkind = 1)
 - gehuwd partner zonder kinderen (plhh_huw_mkind = 1)
 - ouder in eenouderhuishouden (plhh_ouder_eenouder = 1)

Onderstaande figuren tonen de verdeling van deze kenmerken in deze deelpopulatie (blauw) in vergelijking met de gehele arme populatie (grijs). Merk op dat het opleidingsniveau deels op registraties en deels op enquêtes gebaseerd is, waardoor dit niet voor de gehele populatie bekend is. Hierdoor is de groep met een onbekend opleidingsniveau relatief groot. Ook is dit een selectieve groep. Zo is het opleidingsniveau relatief vaak onbekend voor ouderen en personen met een migratieachtergrond. Dit kan leiden tot vertekening in de opleidingsniveauverdeling. Daarom laten we dit buiten beschouwing bij het beschrijven van de populatiekenmerken. Mogelijk heeft dit ook invloed op de selectie van deze deelpopulatie, hoewel dit bij jongeren minder speelt. Wel kan dit er bijvoorbeeld toe leiden dat personen met een eerste generatie migratieachtergrond oververtegenwoordigd zijn.

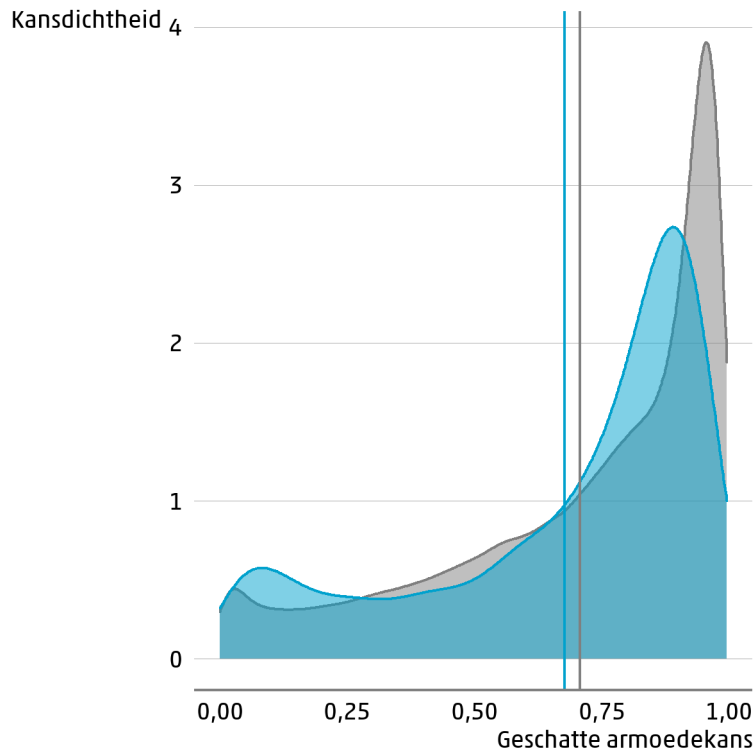
Uit de figuren blijkt dat er onder arme laagopgeleide jongeren met name relatief meer personen met een tweede generatie migratieachtergrond zijn (generatie = 2), personen met een Nederlandse achtergrond (generatie = 3), meer alleenstaanden (plhh_alleenstaande = 1), meer ongehuwde partners (zowel zonder kinderen: plhh_partner_zkind = 1 als met kinderen: plhh_partner_mkind = 1) en minder gehuwde partners (zonder kinderen: plhh_partner_huw_zkind = 1 en met kinderen: plhh_partner_huw_mkind = 1)



Kans om arm te blijven

Onderstaande figuur toont de verdeling van de geschatte kans om arm te blijven voor arme laagopgeleide jongeren (blauw) en voor de gehele arme populatie in Nederland (grijs). De verticale lijnen tonen de

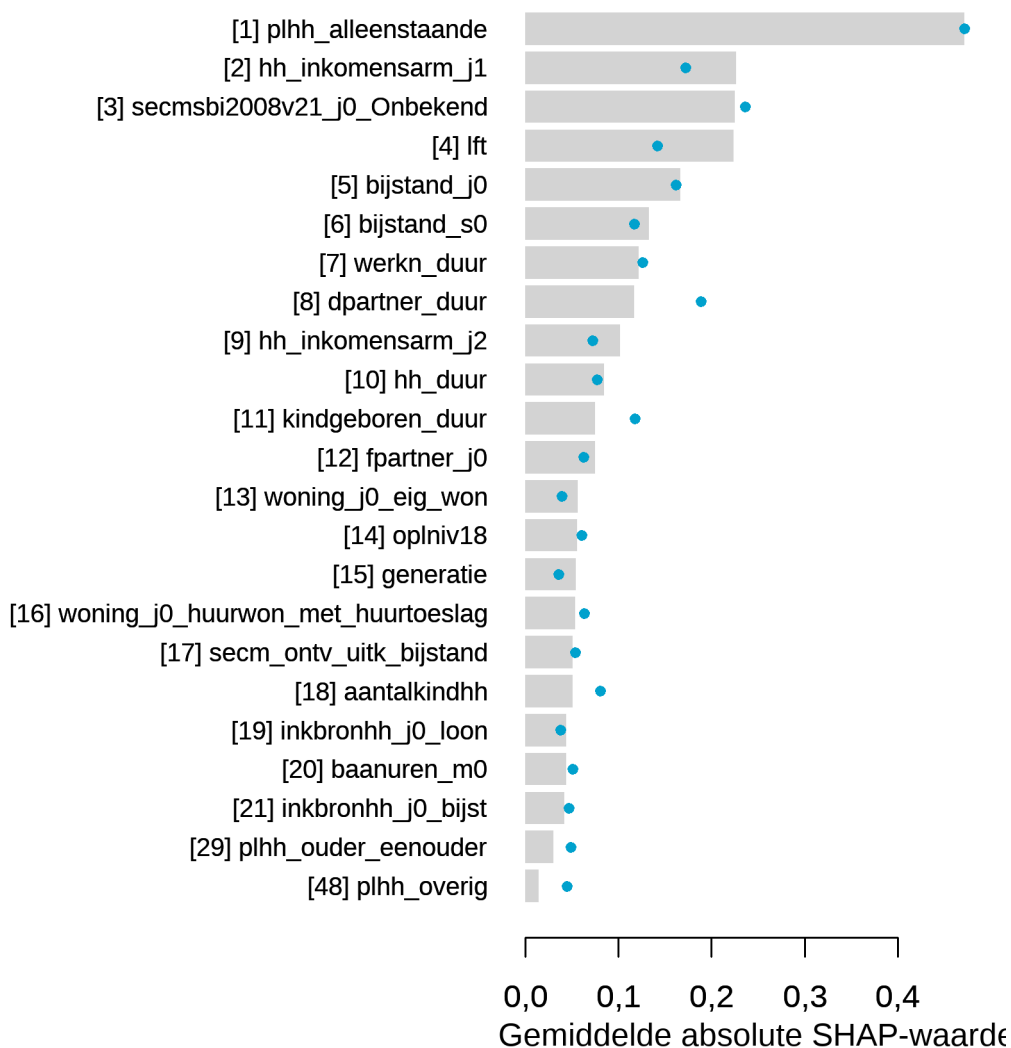
gemiddelde kans om arm te blijven. De gemiddelde kans voor arme laagopgeleide jongeren is met 68% wat kleiner dan in de gehele arme populatie in Nederland (71%). Niet-arme laagopgeleide jongeren hebben juist een hogere kans om arm te worden (zie bijlage E.8). Mogelijk speelt hierbij de vertegenwoordiging van risicogroepen onder arme en niet-arme laagopgeleide jongeren mee. Zo zijn personen met een eerste generatie migratieachtergrond ondervertegenwoordigd onder arme laagopgeleide jongeren en oververtegenwoordigd onder niet-arme laagopgeleide jongeren.



Belangrijkste kenmerken

Onderstaande figuur geeft de belangrijkste kenmerken weer op basis van SHAP-waarden. (Zie paragraaf 3.3 voor verdere toelichting.) De 20 belangrijkste kenmerken in de gehele arme populatie (grijze balken) zijn aangevuld met de 20 belangrijkste kenmerken voor de deelpopulatie (blauwe punten). Bij kenmerken waarbij iedereen in de deelpopulatie dezelfde waarde heeft, ontbreekt de blauwe punt, aangezien dit kenmerk geen effect heeft binnen de deelpopulatie.

In onderstaande figuur is te zien dat de belangrijkste kenmerken voor deze deelpopulatie redelijk overeenkomen met de gehele arme populatie. (Zie paragraaf 4.1 voor een omschrijving van deze kenmerken.) Wel is te zien dat de blauwe punten voor een aantal kenmerken in meer of mindere mate afwijkt van de grijze balken. Hieronder gaan we in op een aantal kenmerken die *belangrijker* zijn dan voor de Nederlandse arme populatie, dus waarvoor de blauwe punt rechts van de grijze balk ligt. Dit is met name het geval voor de kenmerken *duur sinds wijziging in partnerschap* (*dpartner_duur*), *duur sinds er een kind is geboren* (*kindgeboren_duur*) en het aantal kinderen in het huishouden (*aantal_kindh*). Mogelijk speelt hierbij mee dat jongeren meestal weinig inkomen en vermogen hebben, waardoor het krijgen van een partner of een kind meer invloed hebben dan bij oudere personen. In de volgende figuren bekijken we hoe deze kenmerken de kans om arm te blijven beïnvloeden.



Effect op kans om arm te blijven

Onderstaande figuren tonen het *effect* op de kans om arm te blijven en de *decompositie* van het effect. De staafdiagrammen tonen het effect van de 20 belangrijkste kenmerken op de geschatte kans om arm te blijven voor de deelpopulatie (blauwe balken) en de gehele niet-arme populatie in Nederland (grijze balken). Een negatieve waarde op de y-as betekent dat personen met dit kenmerk een lagere kans dan gemiddeld hebben om arm te worden, een positieve waarde betekent dat personen een hogere kans dan gemiddeld hebben. De hoogte van de balk geeft de range van de kans weer voor 80% van de betreffende populatie. De breedte van de balk komt overeen met de relatieve grootte van de populatie met deze waarde. Een smalle balk betekent dus relatief weinig mensen en een brede balk relatief veel mensen. Zie paragraaf 3.3 voor verdere toelichting.

De tweede figuur toont de decompositie van het verschil in belang van de belangrijkste kenmerken. Het verschil in belang wordt opgesplitst in een component die het belang van de *verdeling* van het kenmerk weergeeft (horizontale as) en een component die het belang weergeeft van het *effect* dat het kenmerk heeft op de geschatte kans (verticale as). Zie paragraaf 3.3 voor verdere toelichting.

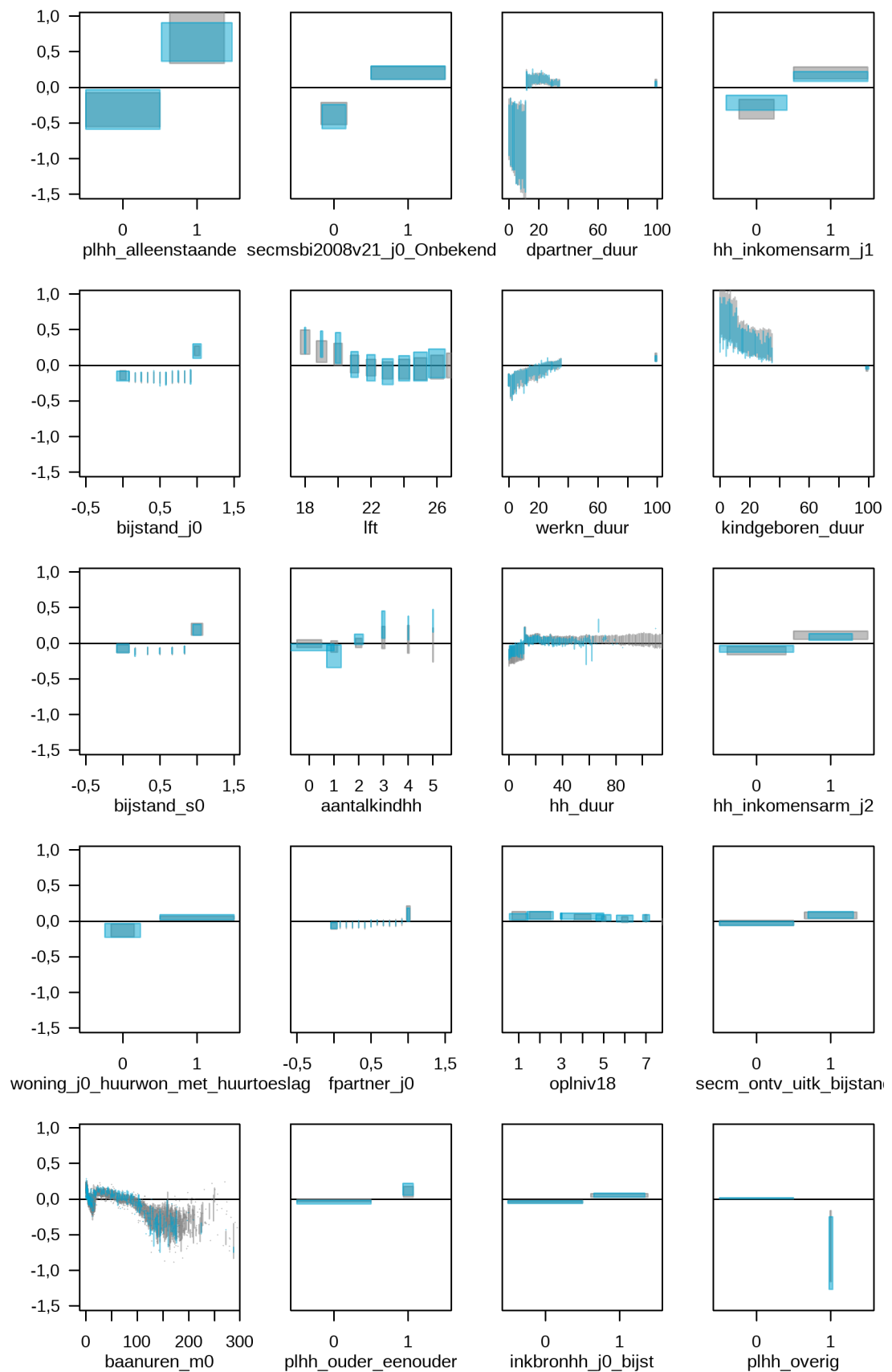
De decompositiefiguur laat zien dat wijziging in partnerschap (dpartner_duur) belangrijker is voor deze deelpopulatie vanwege de samenstelling van deze deelpopulatie (positief verdelingseffect). Dit is niet direct

terug te zien in de andere figuur, omdat deze uit veel balken bestaat waardoor de verschillen in de breedte moeilijker zichtbaar zijn. Deze deelpopulatie zal meer mensen bevatten met een recente verandering van partnerschap (met sterk effect op armoedekans) dan de gehele arme populatie.

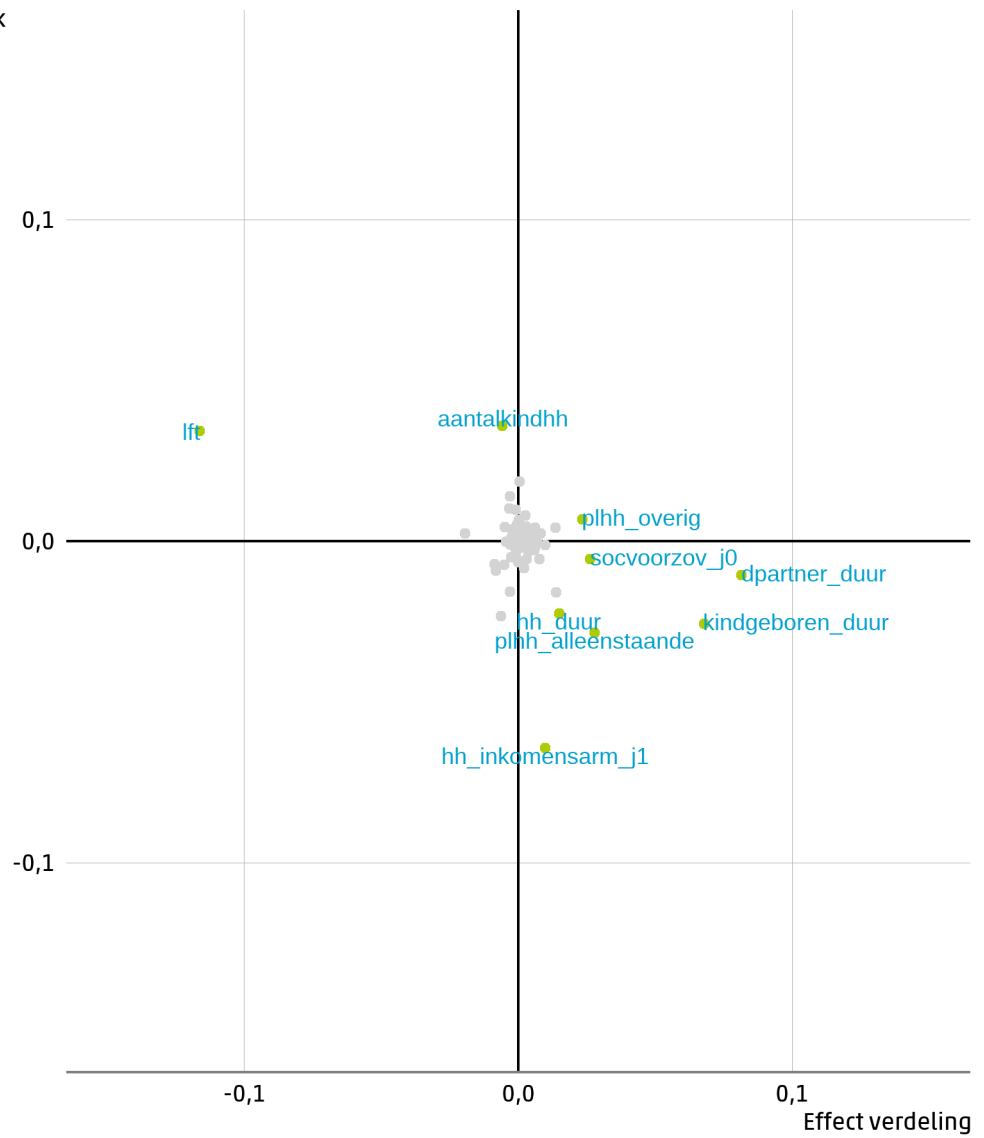
De kans om arm te blijven is groter naarmate het korter geleden is dat er een kind is geboren (kindgeboren_duur). Dit kenmerk is belangrijker voor deze deelpopulatie omdat de deelpopulatie meer (jonge) mensen bevat met een pasgeboren kind (positief verdelingseffect). Het effect van dit kenmerk is voor deze deelpopulatie wel wat minder sterk dan voor de gehele arme populatie (negatief kenmerkeffect). De blauwe balken liggen namelijk wat dichterbij nul dan de grijze. Mogelijk speelt hier mee dat het inkomen van jongeren normaal gesproken groeit, waardoor ze een kleinere kans hebben om arm te blijven. Netto is het kenmerk belangrijker.

In de decompositiefiguur is te zien dat het aantal kinderen (aantal_kindhh) voornamelijk belangrijker is door een positief kenmerkeffect. Dit is ook terug te zien in de staafdiagram: de blauwe balken liggen over het algemeen verder van nul dan de grijze. In deze deelpopulatie hebben mensen met minder dan twee kinderen een kleinere armoedekans dan in de gehele arme populatie. Mogelijk speelt hier mee dat het inkomen van jongeren normaal gesproken groeit, waardoor ze een grotere kans hebben om uit de armoede komen. Ook hebben arme laagopgeleide jongeren met één kind mogelijk recent een partner gekregen, waardoor het inkomen stijgt en daarmee de kans om uit armoede te komen. Arme laagopgeleide jongeren met twee of meer kinderen hebben juist een grotere kans om arm te blijven. Hier zal het meestal gaan om een paar dat al langer bij elkaar is en dan speelt ook mee dat het relatief lage inkomen al langere tijd met meer personen gedeeld moet worden. Merk hierbij wel op dat met name 3 of meer kinderen weinig voorkomt.

Ook zien we in het staafdiagram dat arme laagopgeleide jongeren relatief vaak minder dan twee kinderen hebben. Voor deze groepen zijn de blauwe balken namelijk breder dan de grijze balken. Maar dit negatieve verdelingseffect is klein, omdat het aantal kinderen minder belangrijk is in de gehele populatie (de grijze balken liggen relatief dicht bij nul).



Effect kenmerk



E.4 Arme populatie - werkenden

Deze paragraaf zoomt in op de deelpopulatie werkende armen:

- hoe is deze deelpopulatie afgebakend,
- in hoeverre verschilt deze deelpopulatie qua kenmerken van de hele arme populatie,
- hoe groot is de kans om arm te blijven,
- in hoeverre verschillen de kenmerken die het belangrijkste zijn voor het schatten van de armoedekans van het landelijke beeld,
- welk effect hebben deze kenmerken op de kans om arm te blijven.

Afbakening

- In referentiejaar in de armoede;
- In referentiejaar werkend: werknemer, directeur-grotoaandeelhouder, zelfstandig ondernemer, overig zelfstandige of meewerkend gezinslid.

Deze deelpopulatie bestaat uit personen die in het referentiejaar (t) arm zijn en werkend zijn.

Kenmerken

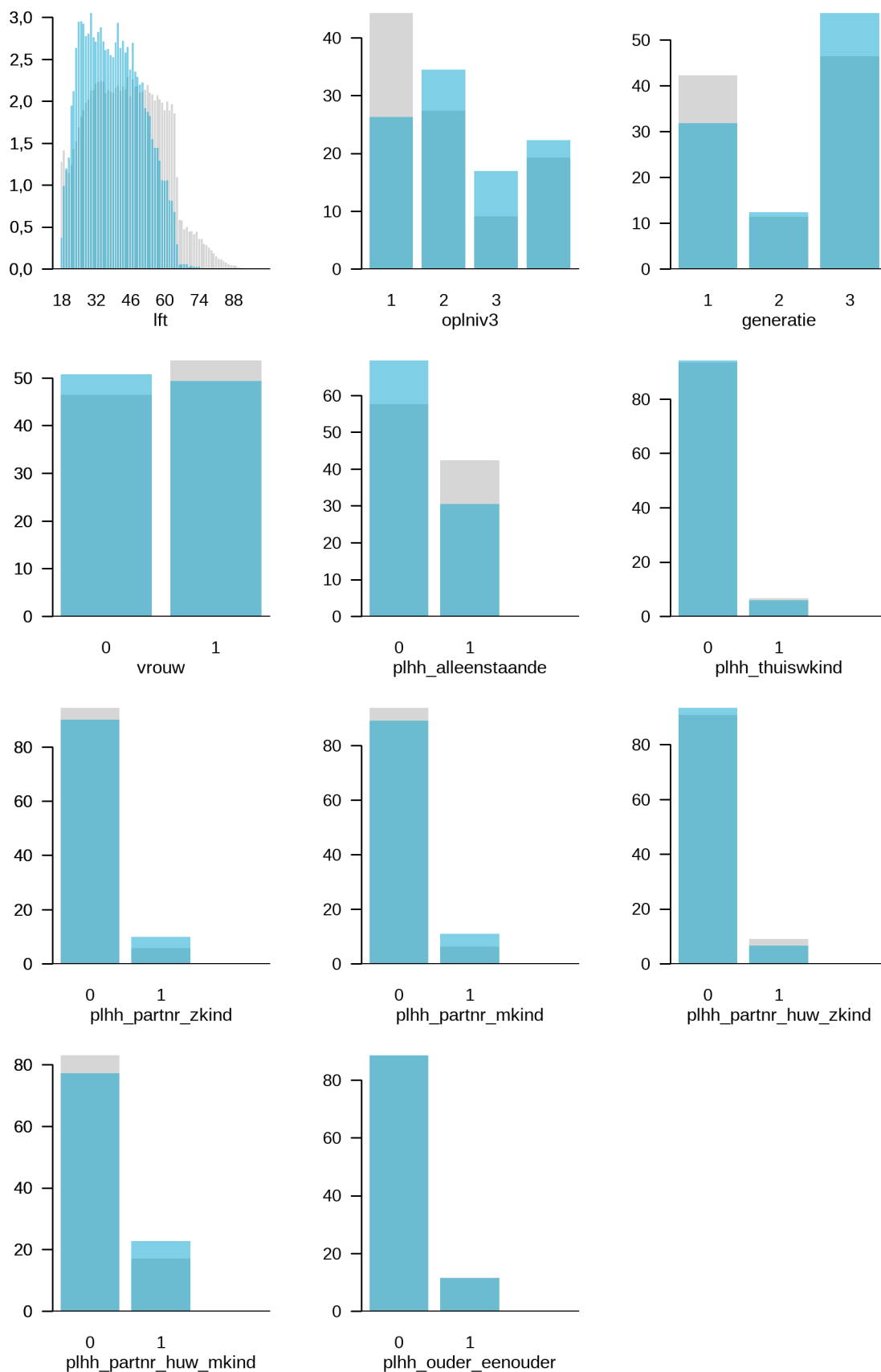
Om beter zicht te krijgen op deze deelpopulatie hebben we gekeken naar een aantal achtergrondkenmerken:

- leeftijd (lft)
- opleidingsniveau ($oplniv3$)
- generatie migratieachtergrond
 - eerste generatie migratieachtergrond ($generatie = 1$)
 - tweede generatie ($generatie = 2$)
 - Nederlandse achtergrond ($generatie = 3$)
- geslacht (vrouw: $vrouw = 1$)
- type huishouden
 - alleenstaande ($plhh_alleenstaande = 1$)
 - thuiswonend kind ($plhh_thuiswkind = 1$)
 - ongehuwd partner zonder kinderen ($plhh_partn_zkind = 1$)
 - ongehuwd partner met kinderen ($plhh_partn_mkind = 1$)
 - gehuwd partner met kinderen ($plhh_huw_zkind = 1$)
 - gehuwd partner zonder kinderen ($plhh_huw_mkind = 1$)
 - ouder in eenouderhuishouden ($plhh_ouder_eenouder = 1$)

Onderstaande figuren tonen de verdeling van deze kenmerken in deze deelpopulatie (blauw) in vergelijking met de gehele arme populatie (grijs). Merk op dat het opleidingsniveau deels op registraties en deels op enquêtes gebaseerd is, waardoor dit niet voor de gehele populatie bekend is. Hierdoor is de groep met een onbekend opleidingsniveau relatief groot. Ook is dit een selectieve groep. Zo is het opleidingsniveau relatief vaak onbekend voor ouderen en personen met een migratieachtergrond. Dit kan leiden tot vertekening in de opleidingsniveauverdeling. Daarom laten we dit buiten beschouwing bij het beschrijven van de populatiekenmerken.

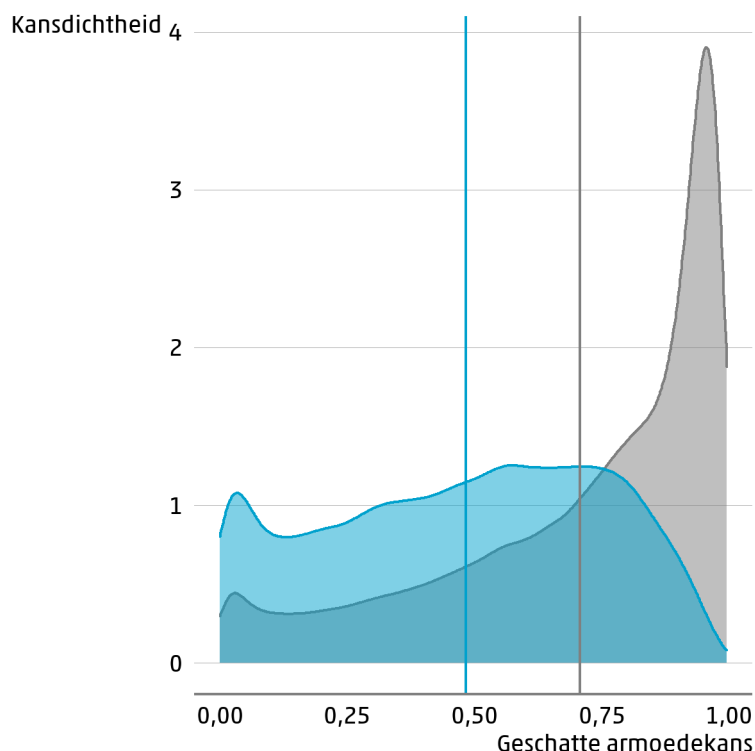
Uit de figuren blijkt dat er onder werkende armen met name relatief meer personen van ongeveer 20-50 zijn ($20 < lft < 50$), personen met een Nederlandse achtergrond ($generatie = 3$), meer mannen ($vrouw = 0$), meer partners met kinderen (zowel ongehuwd: $plhh_partner_zkind = 1$ als gehuwd: $plhh_partner_huw_mkind = 1$), meer ongehuwde partners zonder kinderen ($plhh_partner_mkind =$

1), en met name minder personen met een eerste generatie migratieachtergrond (generatie = 1) en minder alleenstaanden zijn (plhh_alleenstaande = 1).



Kans om arm te blijven

Onderstaande figuur toont de verdeling van de geschatte kans om arm te blijven voor werkende armen (blauw) en voor de gehele arme populatie in Nederland (grijs). De verticale lijnen tonen de gemiddelde kans om arm te blijven. De gemiddelde kans voor werkende armen is met 48% een stuk kleiner dan in de gehele arme populatie in Nederland (71%).

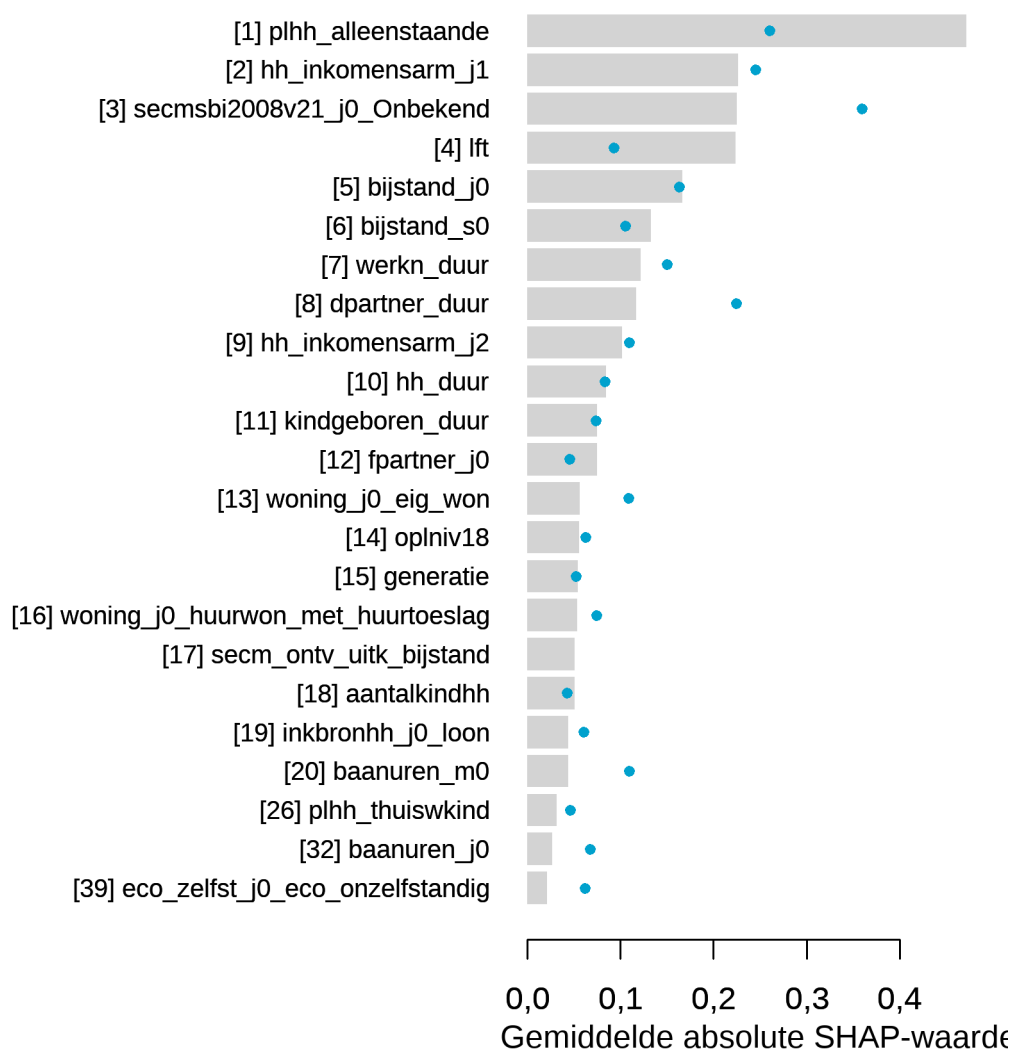


Belangrijkste kenmerken

Onderstaande figuur geeft de belangrijkste kenmerken weer op basis van SHAP-waarden. (Zie paragraaf 3.3 voor verdere toelichting.) De 20 belangrijkste kenmerken in de gehele arme populatie (grijze balken) zijn aangevuld met de 20 belangrijkste kenmerken voor de deelpopulatie (blauwe punten). Bij kenmerken waarbij iedereen in de deelpopulatie dezelfde waarde heeft, ontbreekt de blauwe punt, aangezien dit kenmerk geen effect heeft binnen de deelpopulatie.

In onderstaande figuur is te zien dat de belangrijkste kenmerken voor deze deelpopulatie redelijk overeenkomen met de gehele arme populatie. (Zie paragraaf 4.1 voor een omschrijving van deze kenmerken.) Wel is te zien dat de blauwe punten voor een aantal kenmerken in meer of mindere mate afwijken van de grijze balken. Hieronder gaan we in op een aantal kenmerken die *belangrijker* zijn dan voor de Nederlandse arme populatie, dus waarvoor de blauwe punt rechts van de grijze balk ligt. Dit is met name het geval voor `secmsbi2008v21_j0_Onbekend`. Dit kenmerk geeft aan of de economische activiteit volgens de Standaard Bedrijfsindeling (SBI) bekend is. Dit kenmerk scheidt de werkenden (werknemers, directeuren-grotoaandeelhouders, zelfstandigen en meewerkende gezinsleden) van de niet-werkenden (mensen met een uitkering, studenten of mensen zonder inkomen). Dit is te verwachten, aangezien deze deelpopulatie uit werkenden bestaat.

Daarnaast zijn de tijd sinds het partnerschap is gewijzigd (`dpartner_duur`), het hebben van een eigen woning (`woning_j0_eig_won`) en het aantal baanuren (`baanuren_m0`) van groter belang voor deze deelpopulatie in vergelijking met de hele arme populatie. In de volgende figuren bekijken we hoe deze kenmerken de kans om arm te blijven beïnvloeden.



Effect op kans om arm te blijven

Onderstaande figuren tonen het *effect* op de kans om arm te blijven en de *decompositie* van het effect. De staafdiagrammen tonen het effect van de 20 belangrijkste kenmerken op de geschatte kans om arm te blijven voor de deelpopulatie (blauwe balken) en de gehele niet-arme populatie in Nederland (grijze balken). Een negatieve waarde op de y-as betekent dat personen met dit kenmerk een lagere kans dan gemiddeld hebben om arm te worden, een positieve waarde betekent dat personen een hogere kans dan gemiddeld hebben. De hoogte van de balk geeft de range van de kans weer voor 80% van de betreffende populatie. De breedte van de balk komt overeen met de relatieve grootte van de populatie met deze waarde. Een smalle balk betekent dus relatief weinig mensen en een brede balk relatief veel mensen. Zie paragraaf 3.3 voor verdere toelichting.

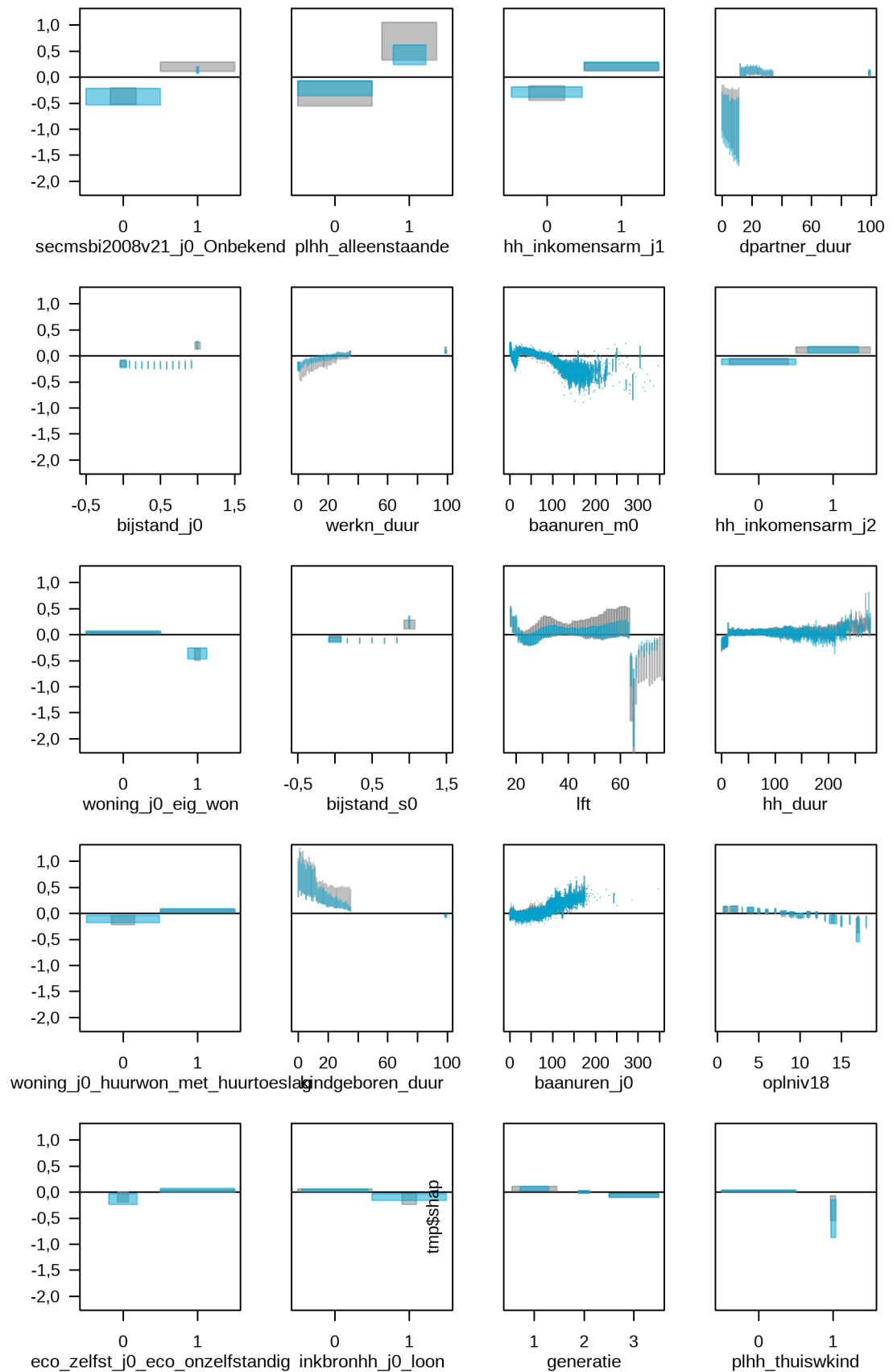
De tweede figuur toont de decompositie van het verschil in belang van de belangrijkste kenmerken. Het verschil in belang wordt opgesplitst in een component die het belang van de *verdeling* van het kenmerk weergeeft (horizontale as) en een component die het belang weergeeft van het *effect* dat het kenmerk heeft op de geschatte kans (verticale as). Zie paragraaf 3.3 voor verdere toelichting.

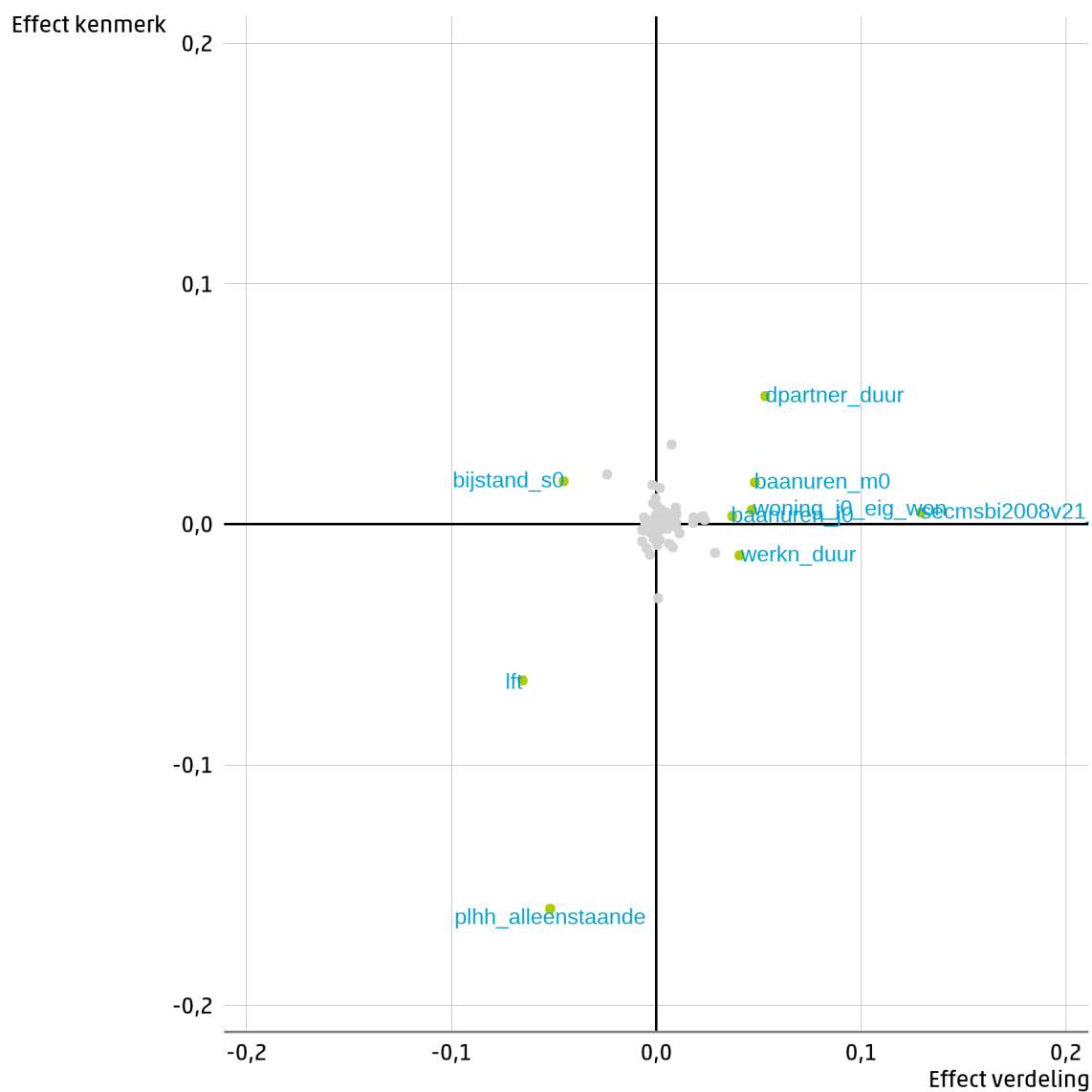
In onderstaande figuur is aan de breedte van de blauwe balken te zien dat vrijwel iedereen werkzaam is ($\text{secmsbi2008v21_j0_Onbekend} = 0$), wat aansluit bij de definitie van deze deelpopulatie op basis van

sociaaleconomische categorie en leeftijd. Dit kenmerk is dus belangrijker in deze deelpopulatie dan in de gehele populatie vanwege een positief verdelingseffect: minder niet-werkzame mensen (met zwakker effect op armoedekans) en meer werkzame mensen (met sterker effect op armoedekans).

In de decompositiefiguur is te zien dat het kenmerk wijziging in partnerschap (`dpartner_duur`) voor werkende armen belangrijker is zowel vanwege de samenstelling van deze deelpopulatie (positief verdelingseffect) als vanwege het grotere effect van het kenmerk (positief kenmerkeffect). Dat laatste zien we terug in het staafdiagram. Hier is te zien dat werkende armen die relatief kort geleden een wijziging in partnerschap hebben meegemaakt een grotere kans hebben om uit de armoede te komen en dat dit effect in de deelpopulatie sterker is dan voor de gehele arme populatie. De blauwe balken liggen namelijk wat lager dan de grijze. Wellicht speelt hier mee dat tweeverdieners belastingtechnisch gezien beter af zijn dan eenverdieners, of dat werkenden vaker een partner hebben die werkt. Het effect van de verdeling is in de staafdiagram niet goed zichtbaar, omdat er veel mogelijke waarden voorkomen, waardoor de blauwe en grijze balken niet zichtbaar in breedte verschillen.

Het aantal baanuren (`baanuren_m0`) laat voor deze deelpopulatie een vergelijkbaar beeld zien als voor de gehele Nederlandse populatie; de kans om arm te blijven neemt af naarmate het aantal baanuren toeneemt. De decompositiefiguur bevestigt dat dit kenmerk belangrijker is omdat werkenden meer baanuren hebben (positief verdelingseffect). Daarnaast is te zien dat werkende armen relatief vaak een eigen woning hebben (`woning_j0_eig_won = 1`). De blauwe balk is namelijk breder dan de grijze en de decompositiefiguur laat een positief verdelingseffect zien.





E.5 Arme populatie - alleenstaande ouders

Deze paragraaf zoomt in op de deelpopulatie arme alleenstaande ouders:

- hoe is deze deelpopulatie afgebakend,
- in hoeverre verschilt deze deelpopulatie qua kenmerken van de hele arme populatie,
- hoe groot is de kans om arm te blijven,
- in hoeverre verschillen de kenmerken die het belangrijkste zijn voor het schatten van de armoedekans van het landelijke beeld,
- welk effect hebben deze kenmerken op de kans om arm te blijven.

Afbakening

- In referentiejaar in de armoede;
- In referentiejaar ouder in eenouderhuishouden.

Deze deelpopulatie betreft personen die in het referentiejaar t arm zijn en in een eenouderhuishouden wonen.

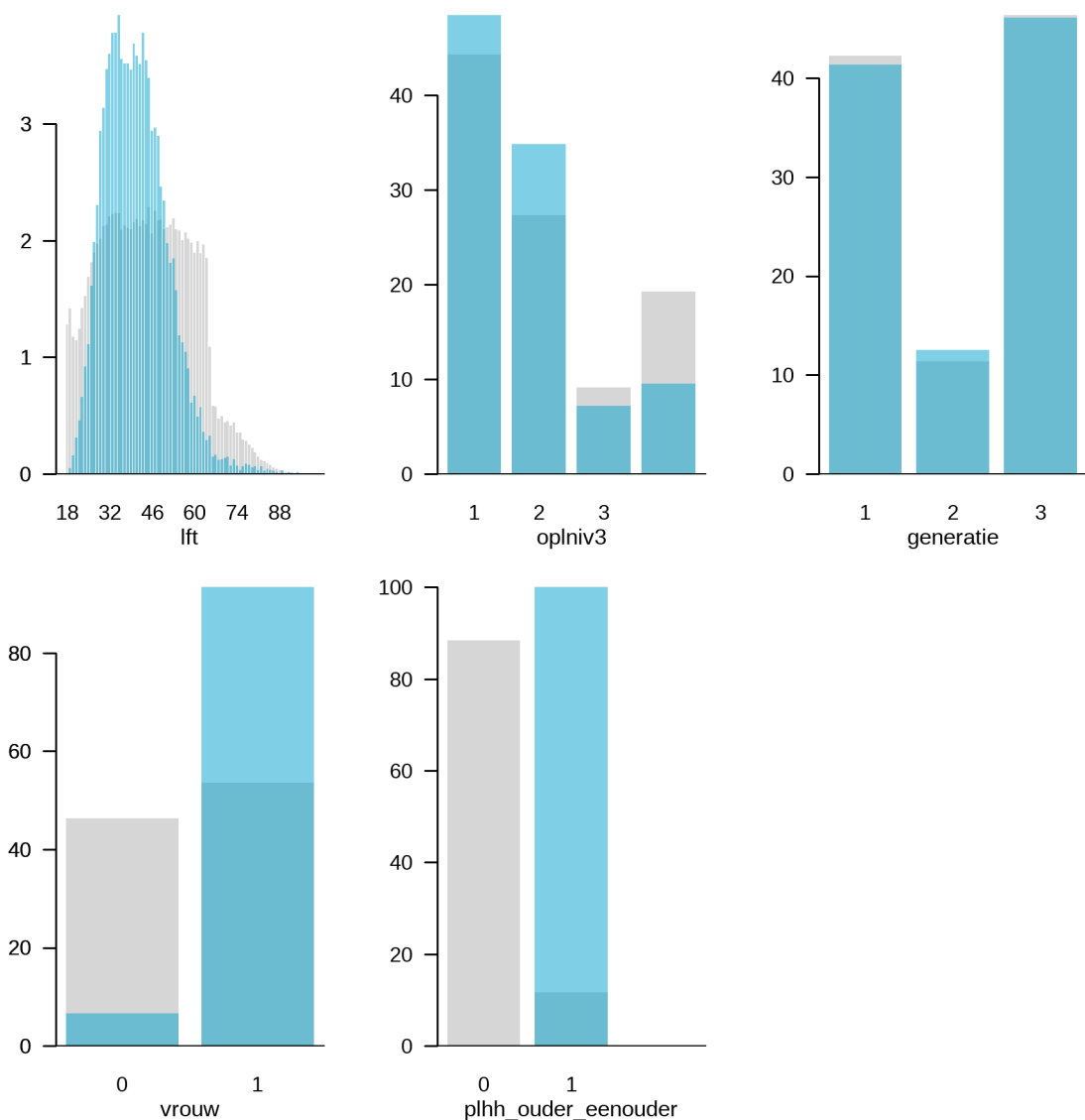
Kenmerken

Om beter zicht te krijgen op deze deelpopulatie hebben we gekeken naar een aantal achtergrondkenmerken:

- leeftijd ($1ft$)
- opleidingsniveau ($oplniv3$)
- generatie migratieachtergrond
 - eerste generatie migratieachtergrond ($generatie = 1$)
 - tweede generatie ($generatie = 2$)
 - Nederlandse achtergrond ($generatie = 3$)
- geslacht (vrouw: $vrouw = 1$)
- type huishouden
 - alleenstaande ($plhh_alleenstaande = 1$)
 - thuiswonend kind ($plhh_thuiswkind = 1$)
 - ongehuwd partner zonder kinderen ($plhh_partn_zkind = 1$)
 - ongehuwd partner met kinderen ($plhh_partn_mkind = 1$)
 - gehuwd partner met kinderen ($plhh_huw_zkind = 1$)
 - gehuwd partner zonder kinderen ($plhh_huw_mkind = 1$)
 - ouder in eenouderhuishouden ($plhh_ouder_eenouder = 1$)

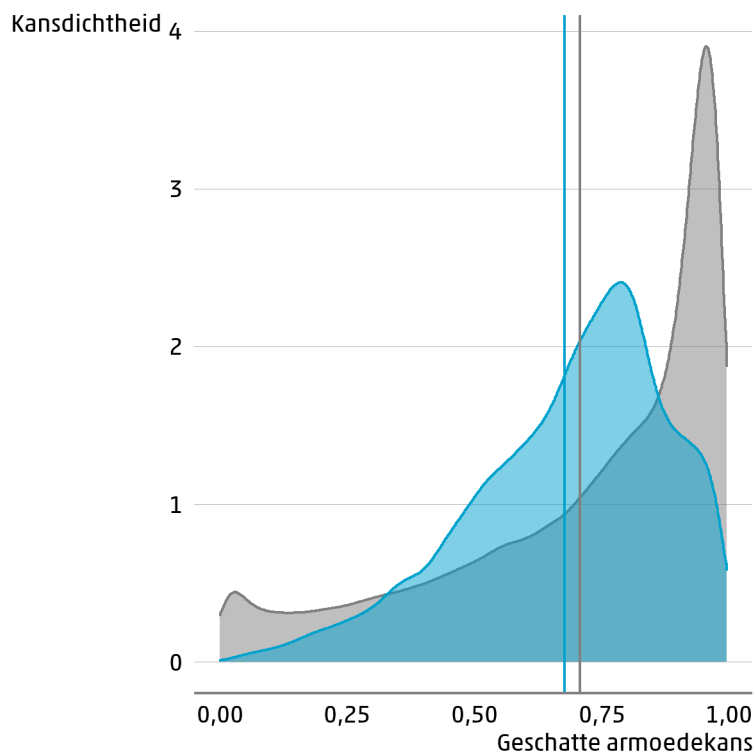
Onderstaande figuren tonen de verdeling van deze kenmerken in deze deelpopulatie (blauw) in vergelijking met de gehele arme populatie (grijs). Merk op dat het opleidingsniveau deels op registraties en deels op enquêtes gebaseerd is, waardoor dit niet voor de gehele populatie bekend is. Hierdoor is de groep met een onbekend opleidingsniveau relatief groot. Ook is dit een selectieve groep. Zo is het opleidingsniveau relatief vaak onbekend voor ouderen en personen met een migratieachtergrond. Dit kan leiden tot vertekening in de opleidingsniveauverdeling. Daarom laten we dit buiten beschouwing bij het beschrijven van de populatiekenmerken.

Uit de figuren blijkt dat arme alleenstaande ouders met name relatief meer personen van ongeveer 25-55 jaar bevat ($25 < 1ft < 55$), wat meer personen met een tweede generatie migratieachtergrond ($generatie = 2$) en meer vrouwen ($vrouw = 1$) en wat minder personen met een eerste generatie migratieachtergrond ($generatie = 1$).



Kans om arm te blijven

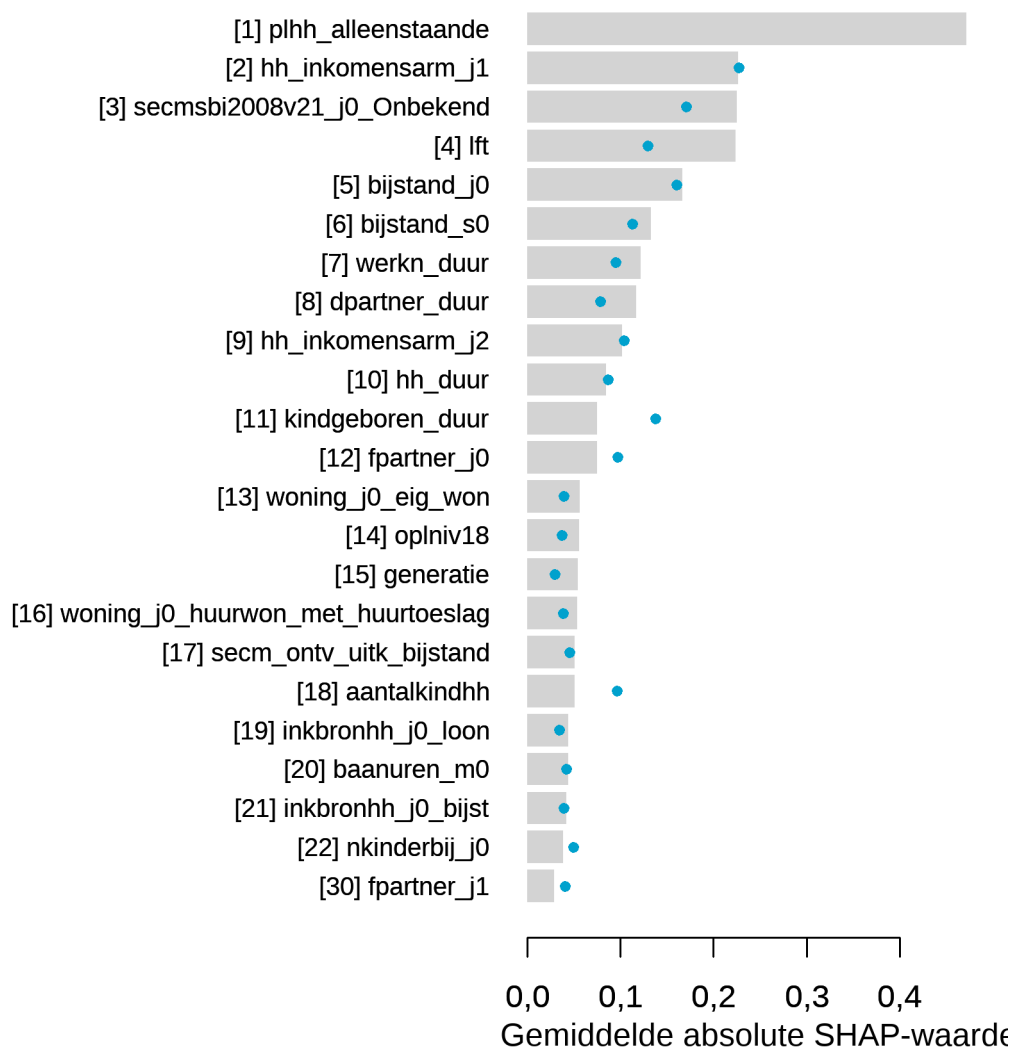
Onderstaande figuur toont de verdeling van de geschatte kans om arm te blijven voor arme alleenstaande ouders (blauw) en voor de gehele arme populatie in Nederland (grijs). De verticale lijnen tonen de gemiddelde kans om arm te blijven. De gemiddelde kans voor arme alleenstaande ouders is met 68% wat kleiner dan in de gehele arme populatie in Nederland (71%). Niet-arme alleenstaande ouders hebben juist een hogere kans om arm te worden (zie bijlage E.10). Mogelijk speelt hierbij de vertegenwoordiging van risicogroepen onder arme en niet-arme alleenstaande ouders mee. Zo zijn personen met een eerste generatie migratieachtergrond wat ondervertegenwoordigd onder arme alleenstaande ouders en oververtegenwoordigd onder niet-arme alleenstaande ouders.



Belangrijkste kenmerken

Onderstaande figuur geeft de belangrijkste kenmerken weer op basis van SHAP-waarden. (Zie paragraaf 3.3 voor verdere toelichting.) De 20 belangrijkste kenmerken in de gehele arme populatie (grijze balken) zijn aangevuld met de 20 belangrijkste kenmerken voor de deelpopulatie (blauwe punten). Bij kenmerken waarbij iedereen in de deelpopulatie dezelfde waarde heeft, ontbreekt de blauwe punt, aangezien dit kenmerk geen effect heeft binnen de deelpopulatie.

In onderstaande figuur is te zien dat de belangrijkste kenmerken voor deze deelpopulatie redelijk overeenkomen met de gehele arme populatie. (Zie paragraaf 4.1 voor een omschrijving van deze kenmerken.) Wel is te zien dat de blauwe punten voor een aantal kenmerken in meer of mindere mate afwijken van de grijze balken. Hieronder gaan we in op een aantal kenmerken die belangrijker zijn dan voor de Nederlandse arme populatie, dus waarvoor de blauwe punt rechts van de grijze balk ligt. Dit is met name het geval voor de kenmerken *duur* sinds een kind is geboren in het huishouden (*kindgeboren_duur*) en het aantal kinderen in het huishouden (*aantalkindhh*). Deze kenmerken hangen natuurlijk samen met de definitie van deze deelpopulatie *alleenstaande ouders*. In de volgende figuren bekijken we hoe deze kenmerken de kans om arm te blijven beïnvloeden.



Effect op kans om arm te blijven

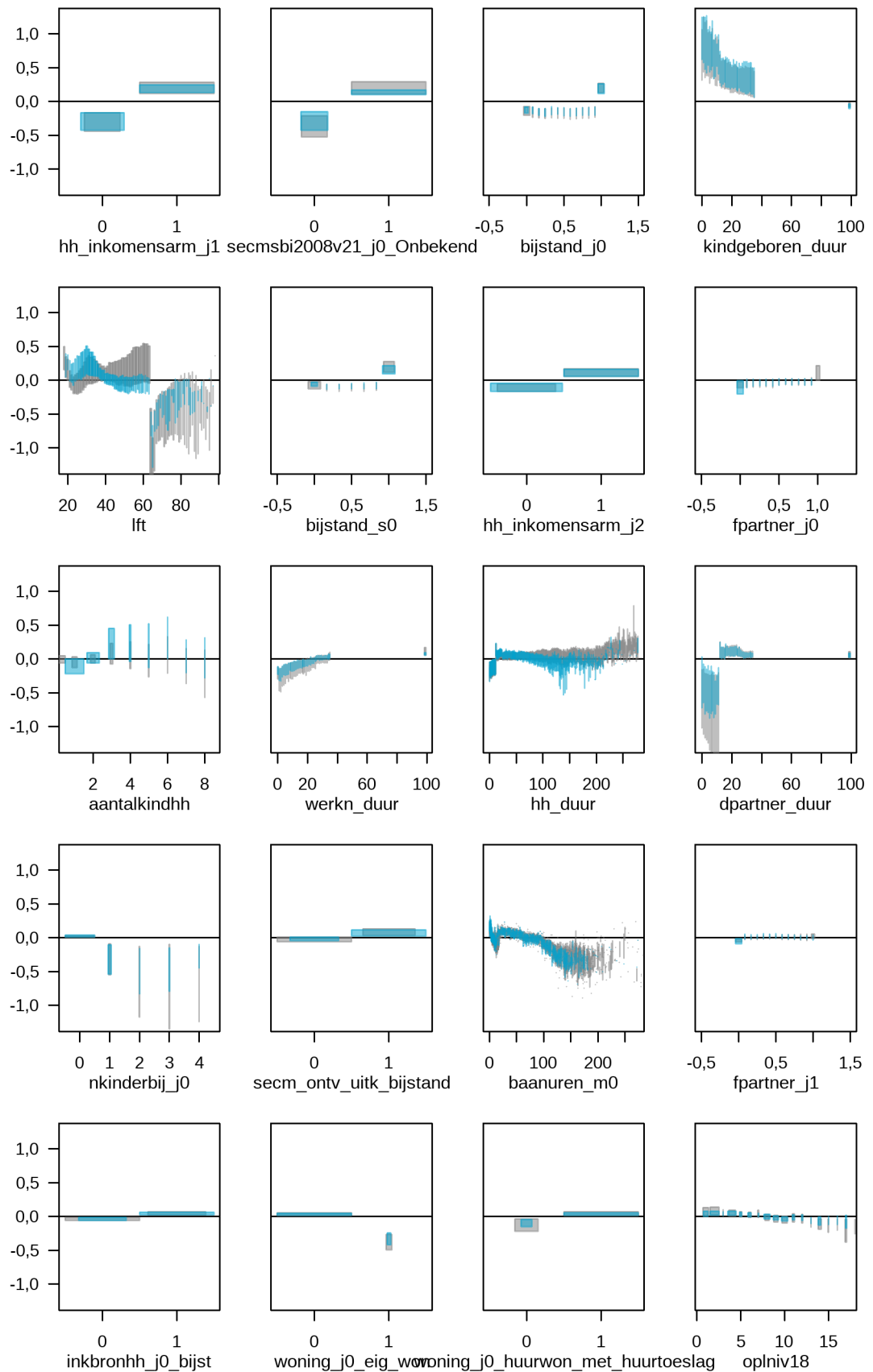
Onderstaande figuren tonen het *effect* op de kans om arm te blijven en de *decompositie* van het effect. De staafdiagrammen tonen het effect van de 20 belangrijkste kenmerken op de geschatte kans om arm te blijven voor de deelpopulatie (blauwe balken) en de gehele niet-arme populatie in Nederland (grijze balken). Een negatieve waarde op de y-as betekent dat personen met dit kenmerk een lagere kans dan gemiddeld hebben om arm te worden, een positieve waarde betekent dat personen een hogere kans dan gemiddeld hebben. De hoogte van de balk geeft de range van de kans weer voor 80% van de betreffende populatie. De breedte van de balk komt overeen met de relatieve grootte van de populatie met deze waarde. Een smalle balk betekent dus relatief weinig mensen en een brede balk relatief veel mensen. Zie paragraaf 3.3 voor verdere toelichting.

De tweede figuur toont de decompositie van het verschil in belang van de belangrijkste kenmerken. Het verschil in belang wordt opgesplitst in een component die het belang van de *verdeling* van het kenmerk weergeeft (horizontale as) en een component die het belang weergeeft van het *effect* dat het kenmerk heeft op de geschatte kans (verticale as). Zie paragraaf 3.3 voor verdere toelichting.

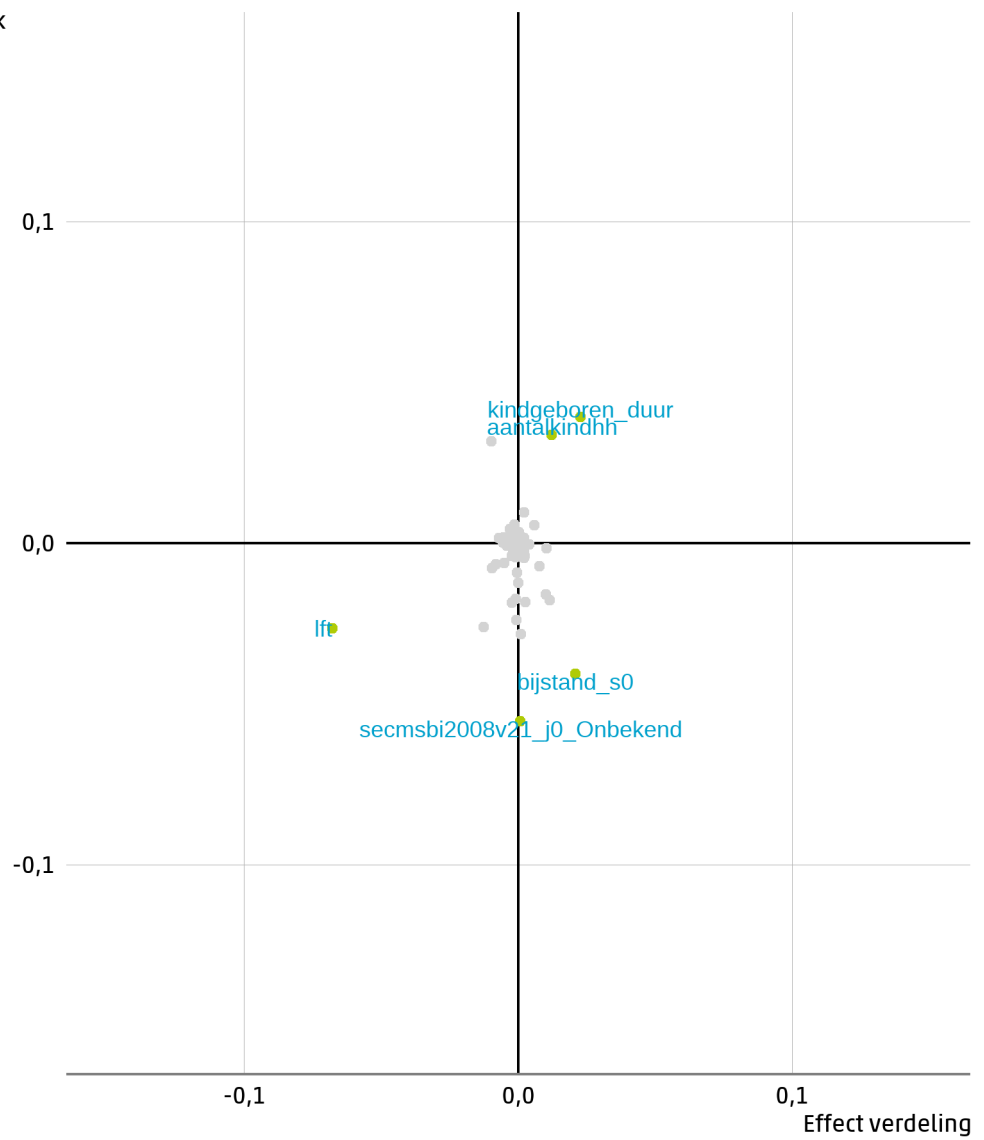
De tijd sinds geboorte van een kind (kindgeboren_duur) is belangrijker in deze deelpopulatie dan in de gehele arme populatie omdat het effect sterker is (positief kenmerkeffect) en omdat de deelpopulatie

relatief veel mensen bevat die recent een kind hebben gekregen (ongunstig effect op armoedekans) en relatief weinig mensen die meer dan drie jaar geleden of nooit een kind hebben gekregen (klein effect op armoedekans). Geboortes komen onder eenouders relatief vaker voor dan onder de gehele arme bevolking waar niet iedereen kinderen heeft. Dit positieve verdelingseffect komt niet naar voren in het staafdiagram, omdat dit kenmerk uit veel mogelijke waarden bestaat, waardoor de blauwe en grijze balk niet zichtbaar in breedte verschillen.

Ook het aantal kinderen (aantal_kindhh) is in deze deelpopulatie iets belangrijker dan in de gehele arme populatie. In deze deelpopulatie hebben mensen met één kind een kleinere armoedekans en mensen met drie of meer kinderen een grotere armoedekans (positief kenmerkeffect). Deze deelpopulatie bevat per definitie geen mensen zonder kinderen, voor wie het kenmerk onbelangrijk is, en alleen mensen met kinderen, voor wie het kenmerk wel belangrijk is (positief verdelingseffect).



Effect kenmerk



E.6 Niet-arme populatie - Amsterdam

Deze paragraaf zoomt in op de niet-arme deelpopulatie Amsterdam:

- hoe is deze deelpopulatie afgebakend,
- in hoeverre verschilt deze deelpopulatie qua kenmerken van de hele niet-arme populatie,
- hoe groot is de kans om arm te worden,
- in hoeverre verschillen de kenmerken die het belangrijkste zijn voor het schatten van de armoedekans van het landelijke beeld,
- welk effect hebben deze kenmerken op de kans om arm te worden.

Afbakening

- In referentiejaar niet arm;
- In referentiejaar inwoner gemeente Amsterdam.

Deze deelpopulatie betreft personen die in het referentiejaar t niet arm zijn en woonachtig in de gemeente Amsterdam.

Kenmerken

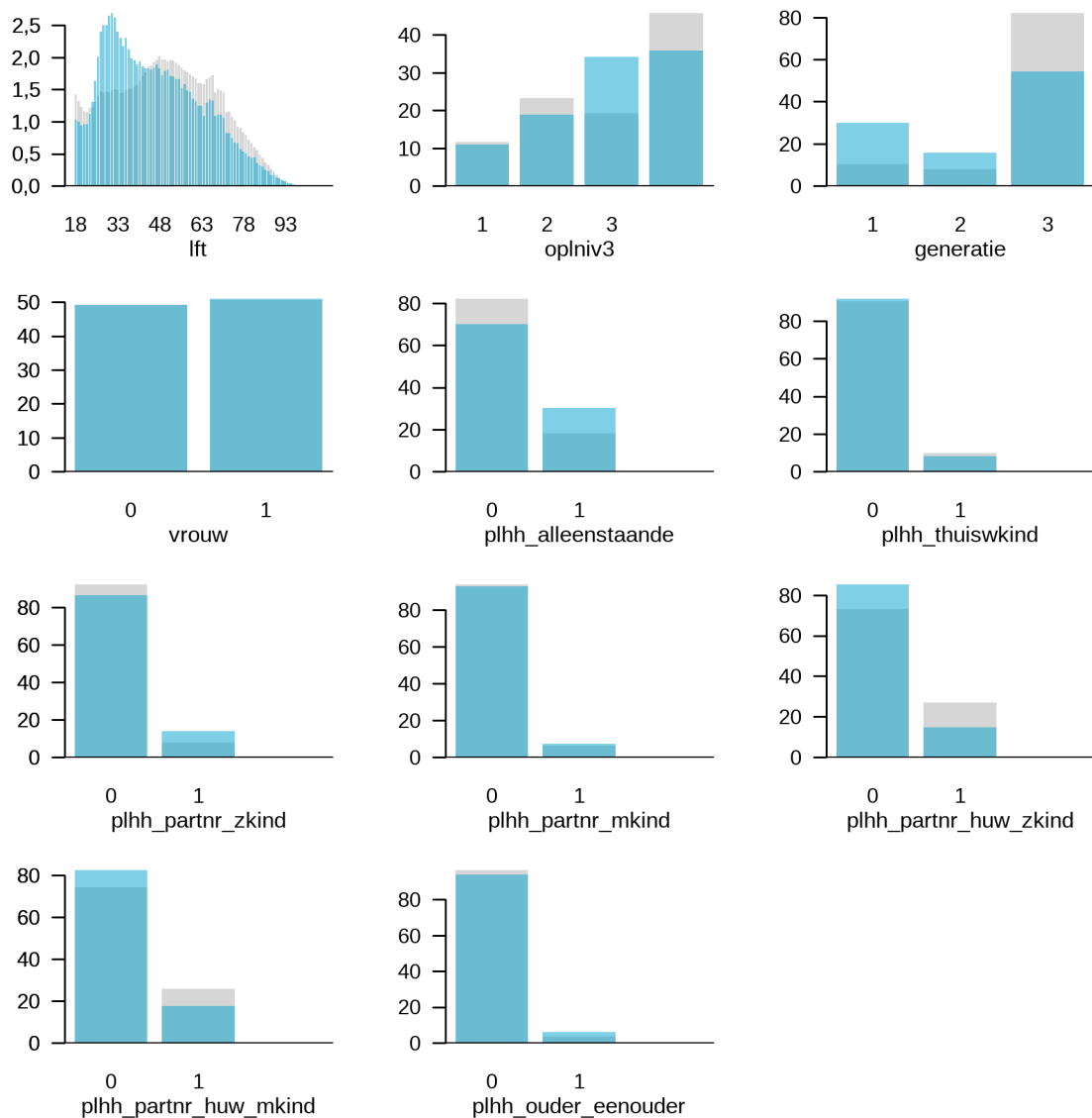
Om beter zicht te krijgen op deze deelpopulatie hebben we gekeken naar een aantal achtergrondkenmerken:

- leeftijd (lft)
- opleidingsniveau ($oplniv3$)
- generatie migratieachtergrond
 - eerste generatie migratieachtergrond ($generatie = 1$)
 - tweede generatie ($generatie = 2$)
 - Nederlandse achtergrond ($generatie = 3$)
- geslacht (vrouw: $vrouw = 1$)
- type huishouden
 - alleenstaande ($plhh_alleenstaande = 1$)
 - thuiswonend kind ($plhh_thuiswkind = 1$)
 - ongehuwd partner zonder kinderen ($plhh_partn_zkind = 1$)
 - ongehuwd partner met kinderen ($plhh_partn_mkind = 1$)
 - gehuwd partner met kinderen ($plhh_huw_zkind = 1$)
 - gehuwd partner zonder kinderen ($plhh_huw_mkind = 1$)
 - ouder in eenouderhuishouden ($plhh_ouder_eenouder = 1$)

Onderstaande figuren tonen de verdeling van deze kenmerken in deze deelpopulatie (blauw) in vergelijking met de gehele niet-arme populatie (grijs). Merk op dat het opleidingsniveau deels op registraties en deels op enquêtes gebaseerd is, waardoor dit niet voor de gehele populatie bekend is. Hierdoor is de groep met een onbekend opleidingsniveau relatief groot. Ook is dit een selectieve groep. Zo is het opleidingsniveau relatief vaak onbekend voor ouderen en personen met een migratieachtergrond. Dit kan leiden tot vertekening in de opleidingsniveauverdeling. Daarom laten we dit buiten beschouwing bij het beschrijven van de populatiekenmerken.

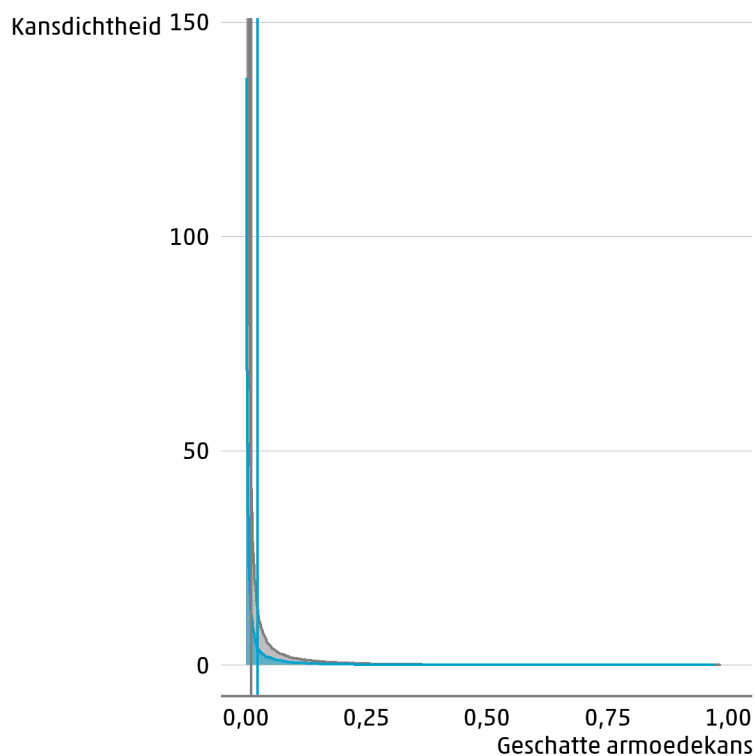
Uit de figuren blijkt dat er in Amsterdam met name relatief meer personen ongeveer tussen de 25 en de 45 jaar zijn ($25 < lft < 45$), personen met een eerste of tweede generatie migratieachtergrond ($generatie = 1$ of $generatie = 2$), alleenstaanden ($plhh_alleenstaande = 1$) en ongehuwde partners

zonder kinderen (plhh_partnr_zkind = 1), en relatief minder gehuwde partners (plhh_huw_zkind = 1 en plhh_huw_mkind = 1).



Kans om arm te worden

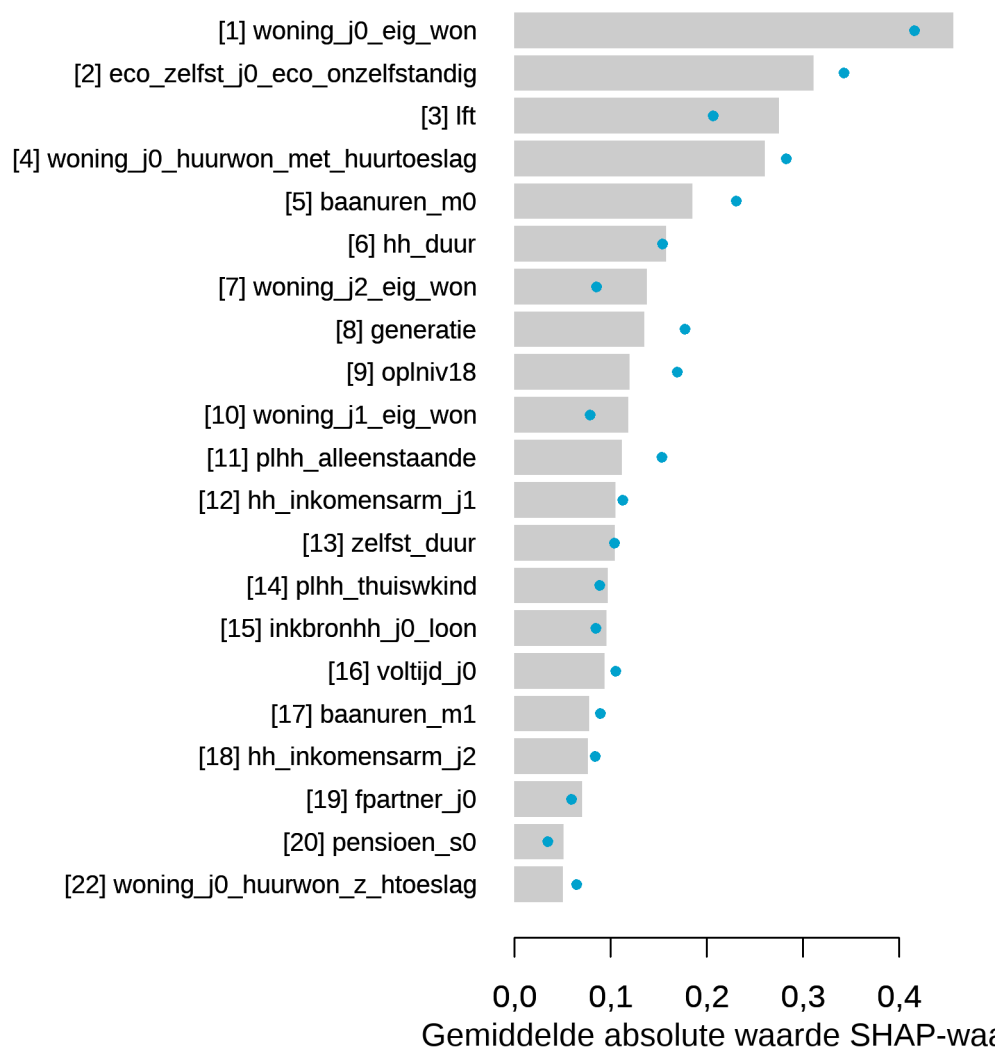
Onderstaande figuur toont de verdeling van de geschatte kans om arm te worden voor de niet-arme deelpopulatie in Amsterdam (blauw) en voor de gehele niet-arme populatie in Nederland (grijs). De verticale lijnen tonen de gemiddelde kans om arm te worden. Hieruit blijkt dat de gemiddelde kans om arm te worden in Amsterdam met 2,5% hoger is dan in de gehele niet-arme populatie (1,2%). De gemiddelde kans om arm te worden in Amsterdam is dus zo'n twee maal zo hoog als in de gehele niet-arme populatie.



Belangrijkste kenmerken

Onderstaande figuur geeft de belangrijkste kenmerken weer op basis van SHAP-waarden. (Zie paragraaf 3.3 voor verdere toelichting.) De 20 belangrijkste kenmerken in de gehele niet-arme populatie (grijze balken) zijn aangevuld met de 20 belangrijkste kenmerken voor de deelpopulatie (blauwe punten). Bij kenmerken waarbij iedereen in de deelpopulatie dezelfde waarde heeft, ontbreekt de blauwe punt, aangezien dit kenmerk geen effect heeft binnen de deelpopulatie.

In onderstaande figuur is te zien dat de belangrijkste kenmerken voor Amsterdam in grote mate overeenkomen met de gehele niet-arme populatie. (Zie paragraaf 4.1 voor een omschrijving van deze kenmerken.) Wel is te zien dat de blauwe punten voor een aantal kenmerken in meer of mindere mate afwijken van de grijze balken. Hieronder gaan we in op een aantal kenmerken die belangrijker zijn dan voor de Nederlandse niet-arme populatie, dus waarvoor de blauwe punt rechts van de grijze balk ligt. Zo zijn onder andere bij de eerder genoemde kenmerken leeftijd (lft), generatie migratieachtergrond (generatie) en alleenstaand (plhh_alleenstaande) wat verschillen met de gehele Nederlandse niet-arme populatie te zien. Maar bijvoorbeeld ook het niet economisch zelfstandig zijn (eco_zelfst_j0_eco_onzelfstandig) en het aantal gewerkte uren in de december van het referentiejaar (baanuren_m0). In de volgende figuren bekijken we hoe deze kenmerken de kans om arm te worden beïnvloeden.



Effect op kans om arm te worden

Onderstaande figuren tonen het *effect* op de kans om arm te worden en de *decompositie* van het effect. De staafdiagrammen tonen het effect van de 20 belangrijkste kenmerken op de geschatte kans om arm te worden voor de deelpopulatie (blauwe balken) en de gehele niet-arme populatie in Nederland (grijze balken). Een negatieve waarde op de y-as betekent dat personen met dit kenmerk een lagere kans dan gemiddeld hebben om arm te worden, een positieve waarde betekent dat personen een hogere kans dan gemiddeld hebben. De hoogte van de balk geeft de range van de kans weer voor 80% van de betreffende populatie. De breedte van de balk komt overeen met de relatieve grootte van de populatie met deze waarde. Een smalle balk betekent dus relatief weinig mensen en een brede balk relatief veel mensen. Zie paragraaf 3.3 voor verdere toelichting.

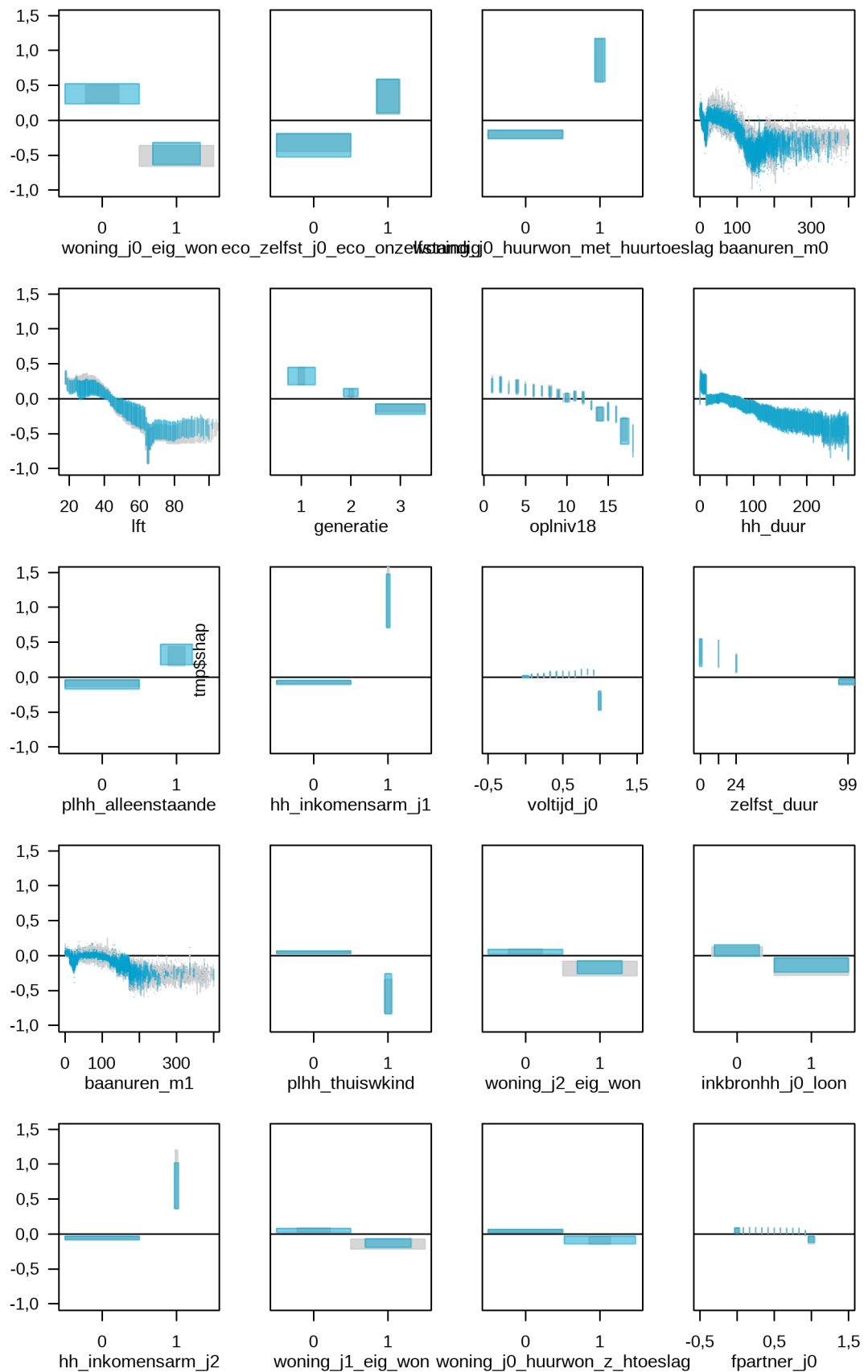
De tweede figuur toont de decompositie van het verschil in belang van de belangrijkste kenmerken. Het verschil in belang wordt opgesplitst in een component die het belang van de *verdeling* van het kenmerk weergeeft (horizontale as) en een component die het belang weergeeft van het *effect* dat het kenmerk heeft op de geschatte kans (verticale as). Zie paragraaf 3.3 voor verdere toelichting.

In onderstaande figuren is te zien dat de effecten van de belangrijkste kenmerken in Amsterdam in grote mate vergelijkbaar zijn met die van de gehele niet-arme populatie. We lichtten een paar kenmerk toe die

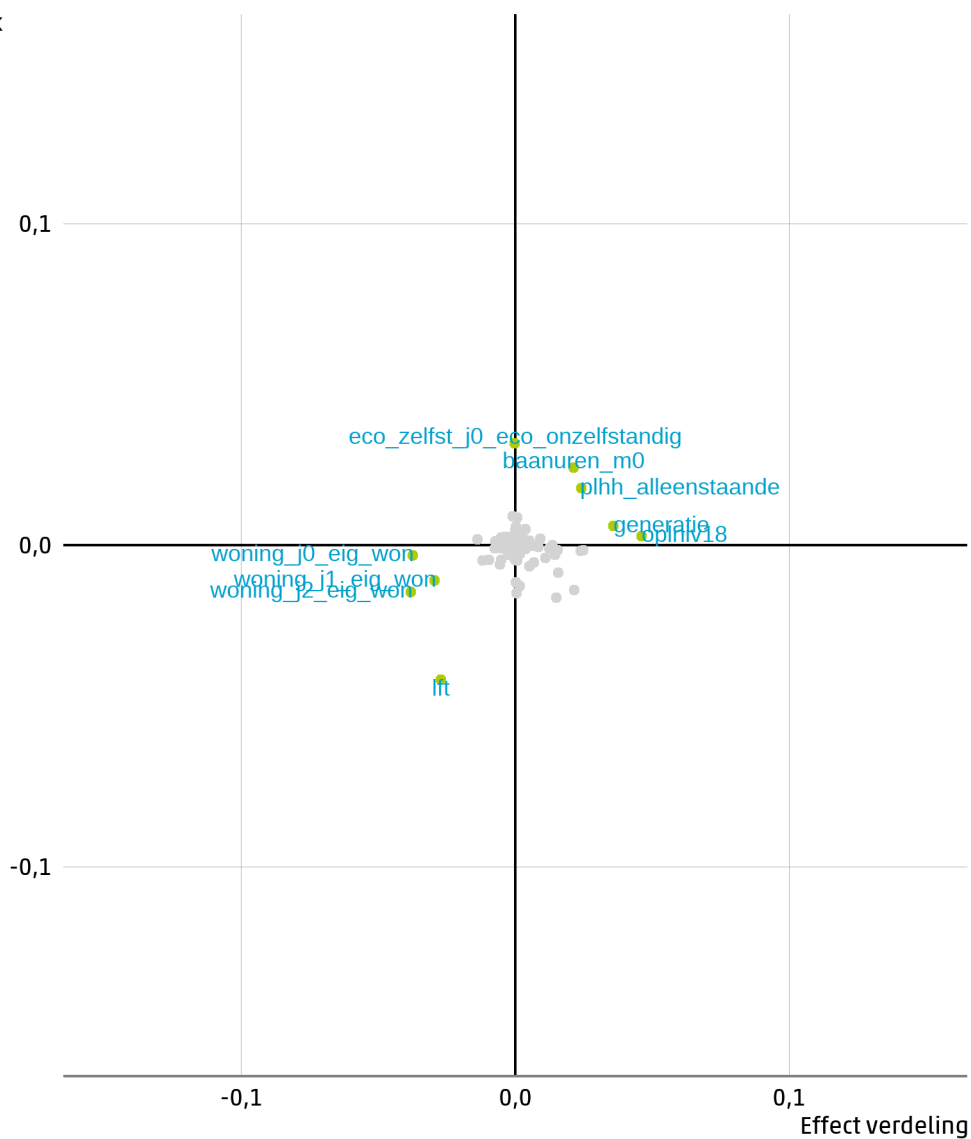
iets belangrijker zijn. Het aantal baanuren in december van het referentiejaar (`baanuren_m0`) is iets belangrijker omdat het effect sterker is (positief kenmerkeffect) en omdat de deelpopulatie meer mensen bevat voor wie het aantal baanuren effect heeft op de armoedekans (positief verdelingseffect).

Generatie migratieachtergrond (`generatie`) is in deze deelpopulatie belangrijker, met name doordat eerste en tweede generatie migratieachtergrond (met hoge armoedekans) relatief veel voorkomen (positief verdelingseffect). Het kenmerk alleenstaand is in deze deelpopulatie belangrijker, doordat alleenstaanden (`plhh_alleenstaande = 1`) relatief veel voorkomen. De blauwe balk is voor deze groep namelijk breder dan de grijze. De decompositie figuur laat ook dit positieve verdelingseffect zien. Daarnaast is het effect van dit kenmerk sterker voor de deelpopulatie dan voor de gehele niet-arme populatie, zoals vooral in de decompositie figuur is te zien.

Voor het al of niet economisch zelfstandig zijn (`eco_zelfst_j0_eco_onzelfstandig`) zien we in de decompositiefiguur dat het verschil in belang met name het belang van het effect betreft. In de verdeling zien we vooral iets gunstigere armoedekansen voor economische zelfstandigen.



Effect kenmerk



E.7 Niet-arme populatie - Heerlen

Deze paragraaf zoomt in op de niet-arme deelpopulatie Heerlen:

- hoe is deze deelpopulatie afgebakend,
- in hoeverre verschilt deze deelpopulatie qua kenmerken van de hele niet-arme populatie,
- hoe groot is de kans om arm te worden,
- in hoeverre verschillen de kenmerken die het belangrijkste zijn voor het schatten van de armoedekans van het landelijke beeld,
- welk effect hebben deze kenmerken op de kans om arm te worden.

Afbakening

- In referentiejaar niet arm;
- In referentiejaar inwoner gemeente Heerlen.

Deze deelpopulatie betreft personen die in het referentiejaar t niet arm zijn en woonachtig in de gemeente Heerlen.

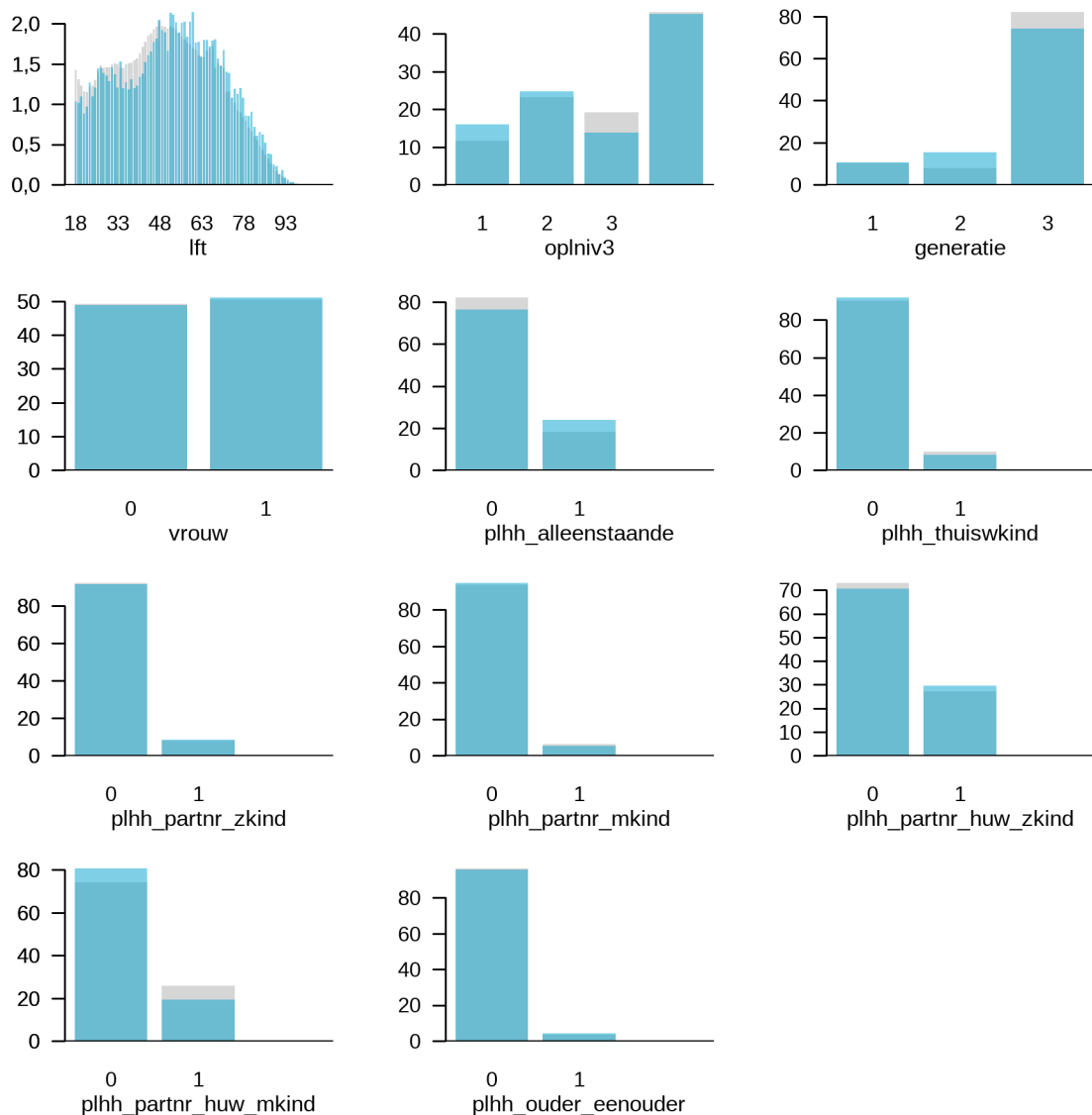
Kenmerken

Om beter zicht te krijgen op deze deelpopulatie hebben we gekeken naar een aantal achtergrondkenmerken:

- leeftijd (lft)
- opleidingsniveau ($oplniv3$)
- generatie migratieachtergrond
 - eerste generatie migratieachtergrond ($generatie = 1$)
 - tweede generatie ($generatie = 2$)
 - Nederlandse achtergrond ($generatie = 3$)
- geslacht (vrouw: $vrouw = 1$)
- type huishouden
 - alleenstaande ($plhh_alleenstaande = 1$)
 - thuiswonend kind ($plhh_thuiswkind = 1$)
 - ongehuwd partner zonder kinderen ($plhh_partn_zkind = 1$)
 - ongehuwd partner met kinderen ($plhh_partn_mkind = 1$)
 - gehuwd partner met kinderen ($plhh_huw_zkind = 1$)
 - gehuwd partner zonder kinderen ($plhh_huw_mkind = 1$)
 - ouder in eenouderhuishouden ($plhh_ouder_eenouder = 1$)

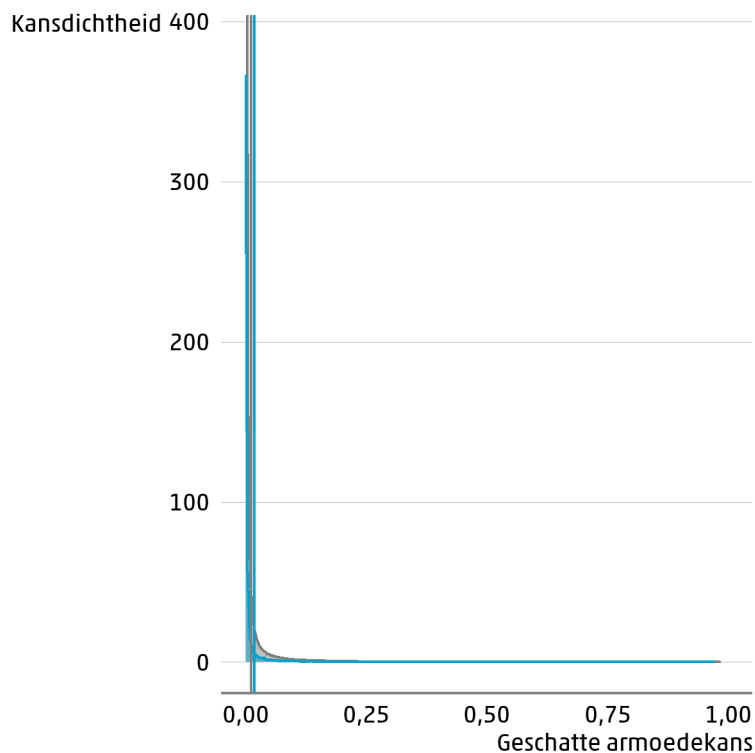
Onderstaande figuren tonen de verdeling van deze kenmerken in deze deelpopulatie (blauw) in vergelijking met de gehele niet-arme populatie (grijs). Merk op dat het opleidingsniveau deels op registraties en deels op enquêtes gebaseerd is, waardoor dit niet voor de gehele populatie bekend is. Hierdoor is de groep met een onbekend opleidingsniveau relatief groot. Ook is dit een selectieve groep. Zo is het opleidingsniveau relatief vaak onbekend voor ouderen en personen met een migratieachtergrond. Dit kan leiden tot vertekening in de opleidingsniveauverdeling. Daarom laten we dit buiten beschouwing bij het beschrijven van de populatiekenmerken.

Uit de figuren blijkt dat er in Heerlen met name relatief meer personen van ongeveer 55 of ouder zijn ($lft \geq 55$), personen met een tweede generatie migratieachtergrond ($generatie = 2$), alleenstaanden ($plhh_alleenstaande = 1$) en relatief minder gehuwde partners met kind ($plhh_partnr_huw_mkind = 1$).



Kans om arm te worden

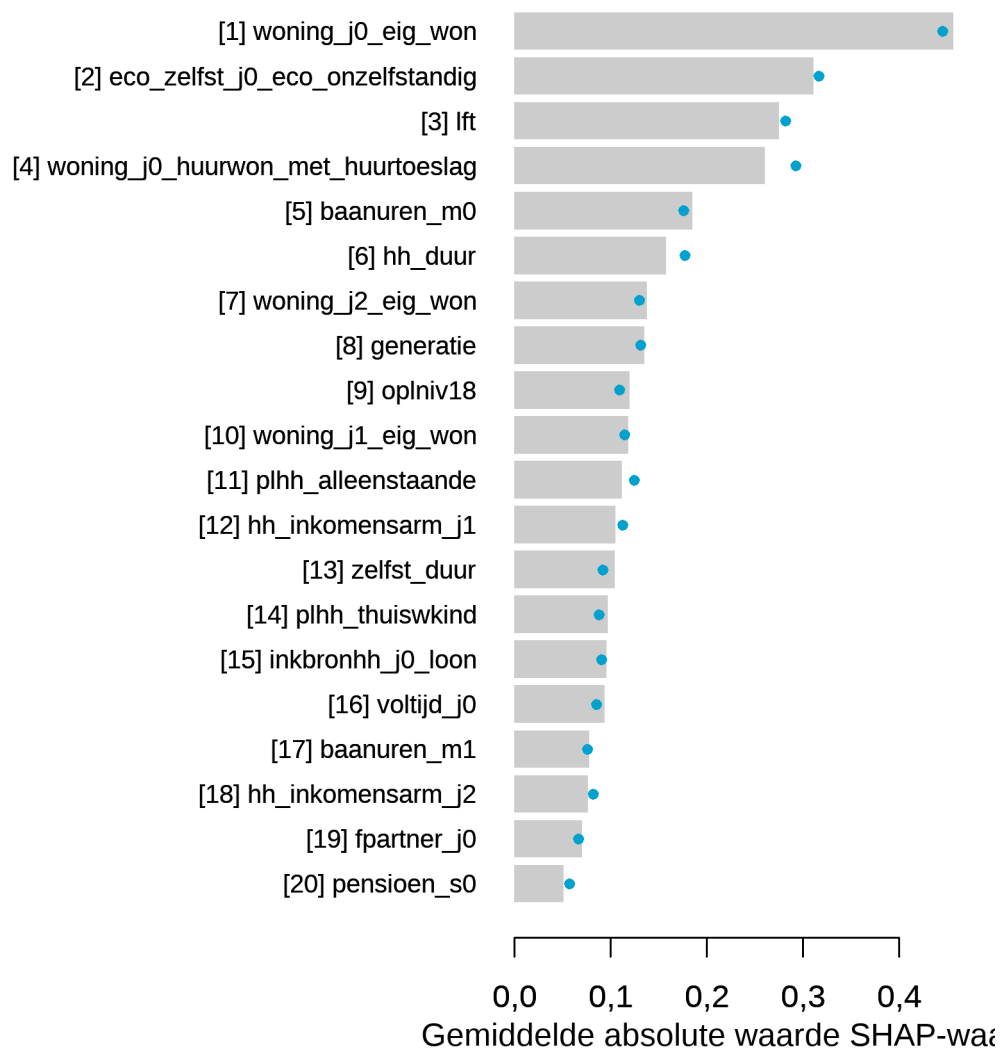
Onderstaande figuur toont de verdeling van de geschatte kans om arm te worden voor de niet-arme deelpopulatie in Heerlen (blauw) en voor de gehele niet-arme populatie in Nederland (grijs). De verticale lijnen tonen de gemiddelde kans om arm te worden. Hieruit blijkt dat de gemiddelde kans om arm te worden in Heerlen met 1,8% hoger is dan in de gehele niet-arme populatie (1,2%). De gemiddelde kans om arm te worden in Heerlen is dus zo'n 1,5 keer zo groot als in de gehele niet-arme populatie.



Belangrijkste kenmerken

Onderstaande figuur geeft de belangrijkste kenmerken weer op basis van SHAP-waarden. (Zie paragraaf 3.3 voor verdere toelichting.) De 20 belangrijkste kenmerken in de gehele niet-arme populatie (grijze balken) zijn aangevuld met de 20 belangrijkste kenmerken voor de deelpopulatie (blauwe punten). Bij kenmerken waarbij iedereen in de deelpopulatie dezelfde waarde heeft, ontbreekt de blauwe punt, aangezien dit kenmerk geen effect heeft binnen de deelpopulatie.

In onderstaande figuur is te zien dat de belangrijkste kenmerken voor Heerlen in grote mate overeenkomen met de gehele niet-arme populatie. (Zie paragraaf 4.1 voor een omschrijving van deze kenmerken.) Een kenmerk wat wel belangrijker is voor Heerlen, is het wonen in een huurwoning met huurtoeslag (`woning_j0_huurwon_met_huurtoeslag = 1`). In de volgende figuren bekijken we hoe dit kenmerk de kans om arm te worden beïnvloedt.



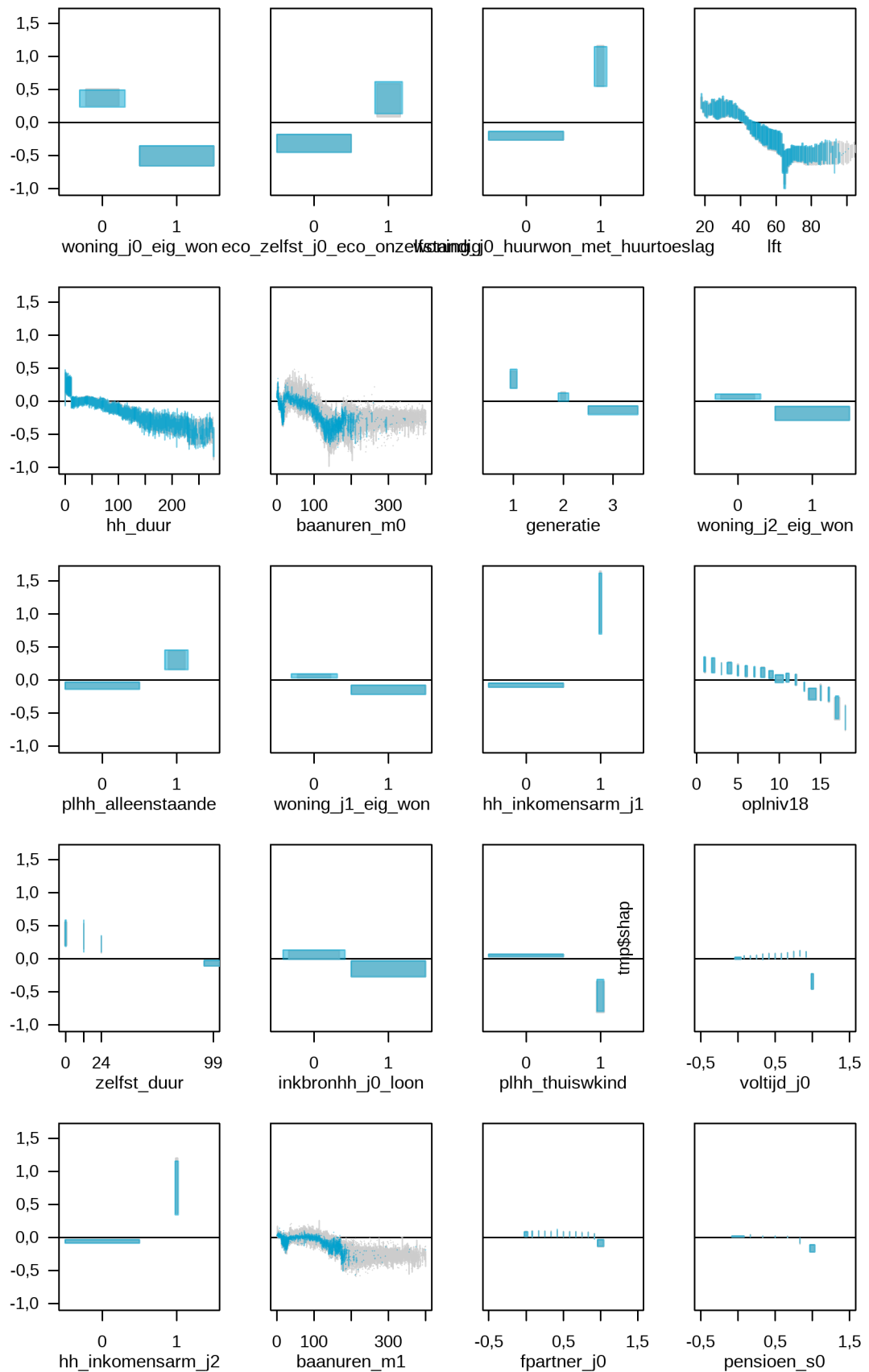
Effect op kans om arm te worden

Onderstaande figuren tonen het *effect* op de kans om arm te worden en de *decompositie* van het effect. De staafdiagrammen tonen het effect van de 20 belangrijkste kenmerken op de geschatte kans om arm te worden voor de deelpopulatie (blauwe balken) en de gehele niet-arme populatie in Nederland (grijze balken). Een negatieve waarde op de y-as betekent dat personen met dit kenmerk een lagere kans dan gemiddeld hebben om arm te worden, een positieve waarde betekent dat personen een hogere kans dan gemiddeld hebben. De hoogte van de balk geeft de range van de kans weer voor 80% van de betreffende populatie. De breedte van de balk komt overeen met de relatieve grootte van de populatie met deze waarde. Een smalle balk betekent dus relatief weinig mensen en een brede balk relatief veel mensen. Zie paragraaf 3.3 voor verdere toelichting.

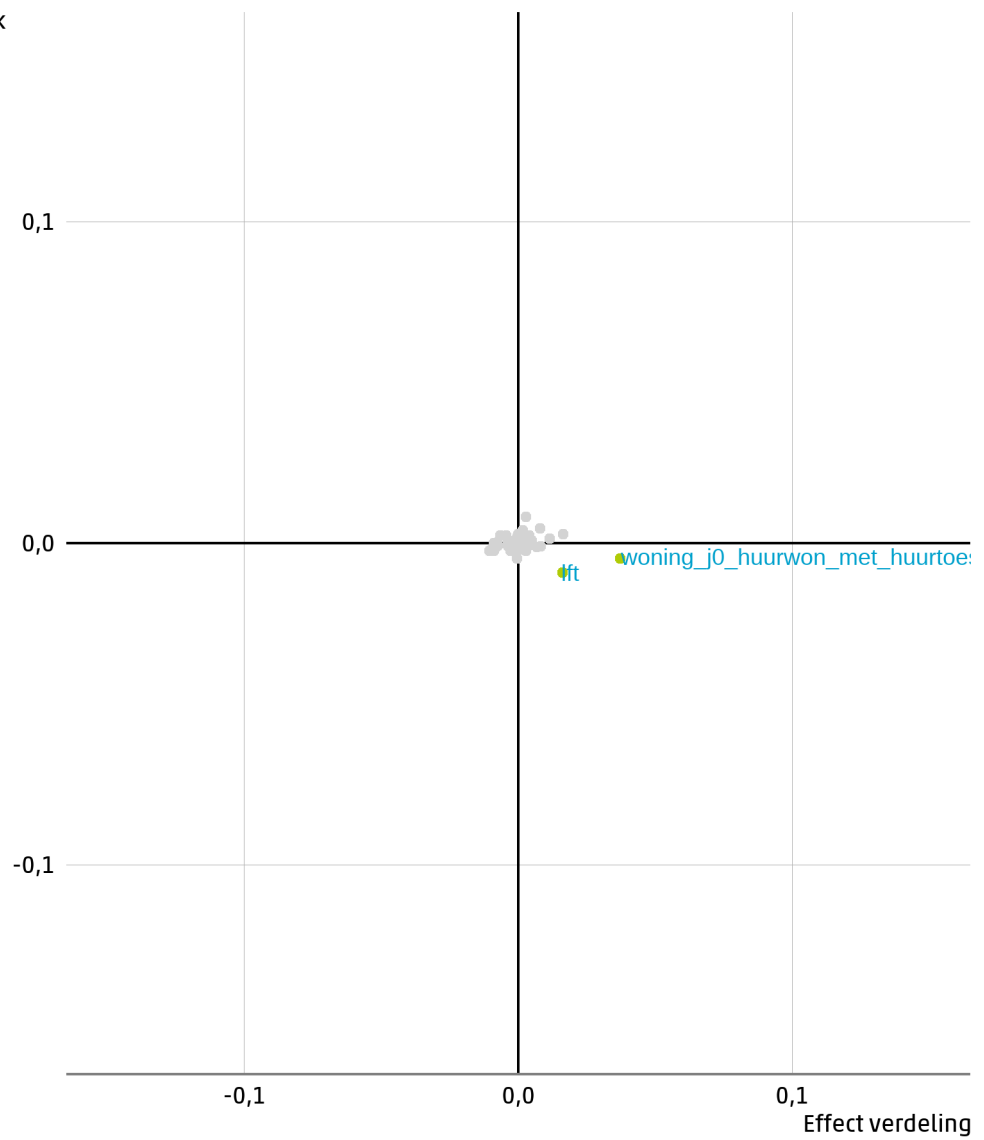
De tweede figuur toont de decompositie van het verschil in belang van de belangrijkste kenmerken. Het verschil in belang wordt opgesplitst in een component die het belang van de *verdeling* van het kenmerk weergeeft (horizontale as) en een component die het belang weergeeft van het *effect* dat het kenmerk heeft op de geschatte kans (verticale as). Zie paragraaf 3.3 voor verdere toelichting.

In onderstaande figuren is te zien dat in Heerlen de kenmerken over het algemeen even belangrijk zijn als in de gehele niet-arme populatie. Het wonen in een huurwoning met huurtoeslag

(woning_j0_huurwon_met_huurtoeslag = 1) komt relatief wat vaker voor in Heerlen, omdat de blauwe balk iets breder is dan de grijze balk. In de decompositiefiguur zien we ook dat dit kenmerk een klein positief verdelingseffect heeft.



Effect kenmerk



E.8 Niet-arme populatie - laagopgeleide jongeren

Deze paragraaf zoomt in op de deelpopulatie niet-arme laagopgeleide jongeren:

- hoe is deze deelpopulatie afgebakend,
- in hoeverre verschilt deze deelpopulatie qua kenmerken van de hele niet-arme populatie,
- hoe groot is de kans om arm te worden,
- in hoeverre verschillen de kenmerken die het belangrijkste zijn voor het schatten van de armoedekans van het landelijke beeld,
- welk effect hebben deze kenmerken op de kans om arm te worden.

Afbakening

- In referentiejaar niet in de armoede;
- In referentiejaar laag opleidingsniveau;
- In referentiejaar jong (18+ en niet ouder dan 26 jaar);
- In referentiejaar niet schoolgaand of studerend;
- In referentiejaar geen thuiswonend kind.

Deze deelpopulatie betreft personen die in het referentiejaar t niet arm zijn, een laag opleidingsniveau hebben, jong zijn, maar niet schoolgaand/studerend zijn en niet meer thuis wonen.

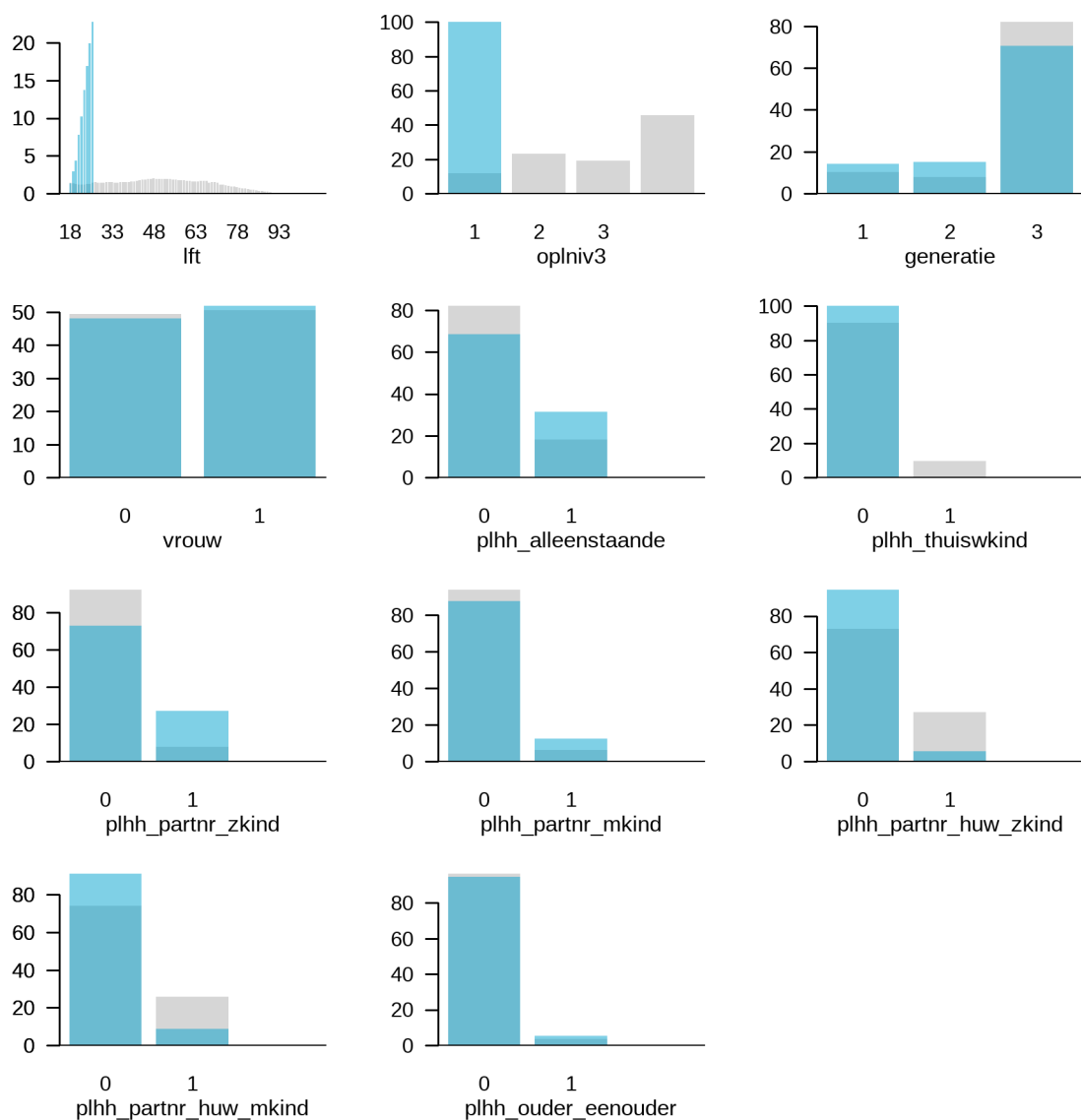
Kenmerken

Om beter zicht te krijgen op deze deelpopulatie hebben we gekeken naar een aantal achtergrondkenmerken:

- leeftijd (1ft)
- opleidingsniveau (opl_niv3)
- generatie migratieachtergrond
 - eerste generatie migratieachtergrond (generatie = 1)
 - tweede generatie (generatie = 2)
 - Nederlandse achtergrond (generatie = 3)
- geslacht (vrouw: vrouw = 1)
- type huishouden
 - alleenstaande (plhh_alleenstaande = 1)
 - thuiswonend kind (plhh_thuiswkind = 1)
 - ongehuwd partner zonder kinderen (plhh_partn_zkind = 1)
 - ongehuwd partner met kinderen (plhh_partn_mkind = 1)
 - gehuwd partner met kinderen (plhh_huw_zkind = 1)
 - gehuwd partner zonder kinderen (plhh_huw_mkind = 1)
 - ouder in eenouderhuishouden (plhh_ouder_eenouder = 1)

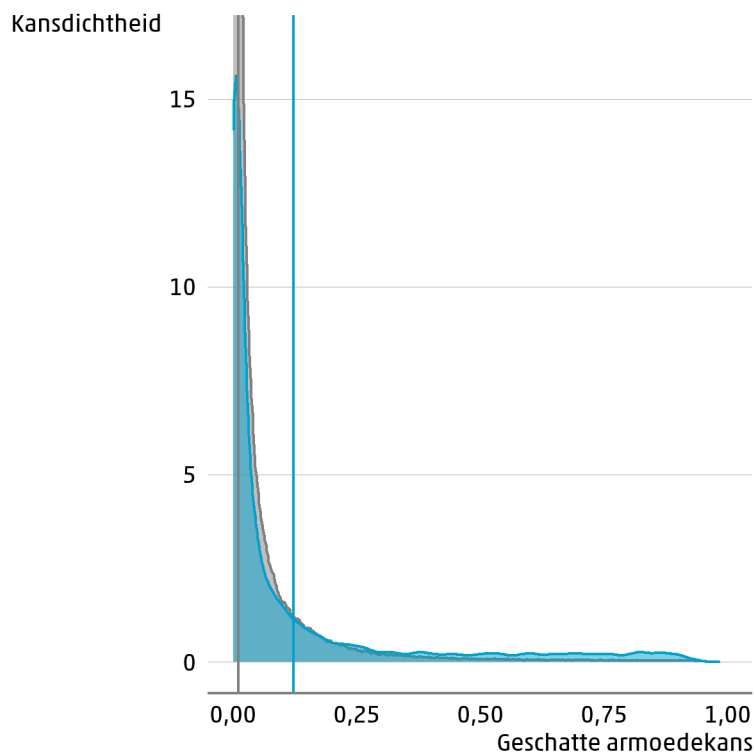
Onderstaande figuren tonen de verdeling van deze kenmerken in deze deelpopulatie (blauw) in vergelijking met de gehele niet-arme populatie (grijs). Merk op dat het opleidingsniveau deels op registraties en deels op enquêtes gebaseerd is, waardoor dit niet voor de gehele populatie bekend is. Hierdoor is de groep met een onbekend opleidingsniveau relatief groot. Ook is dit een selectieve groep. Zo is het opleidingsniveau relatief vaak onbekend voor ouderen en personen met een migratieachtergrond. Dit kan leiden tot vertekening in de opleidingsniveauverdeling. Daarom laten we dit buiten beschouwing bij het beschrijven van de populatiekenmerken. Mogelijk heeft dit ook invloed op de selectie van deze deelpopulatie, hoewel dit bij jongeren minder speelt. Wel kan dit er bijvoorbeeld toe leiden dat personen met een eerste generatie migratieachtergrond oververtegenwoordigd zijn.

Onderstaande figuren laten zien dat laagopgeleide jongeren relatief vaker van 1e of 2e generatie migratieachtergrond (generatie = 1 of generatie = 2) zijn, alleenstaand (plhh_alleenstaande = 1) zijn of ongehuwd partner zijn (zowel met kinderen: plhh_partn_mkind = 1 als zonder: plhh_partn_zkind = 1) en relatief minder vaak gehuwde partner zijn (zowel met kinderen: plhh_partn_huw_mkind = 1 als zonder kinderen plhh_partn_huw_zkind = 1).



Kans om arm te worden

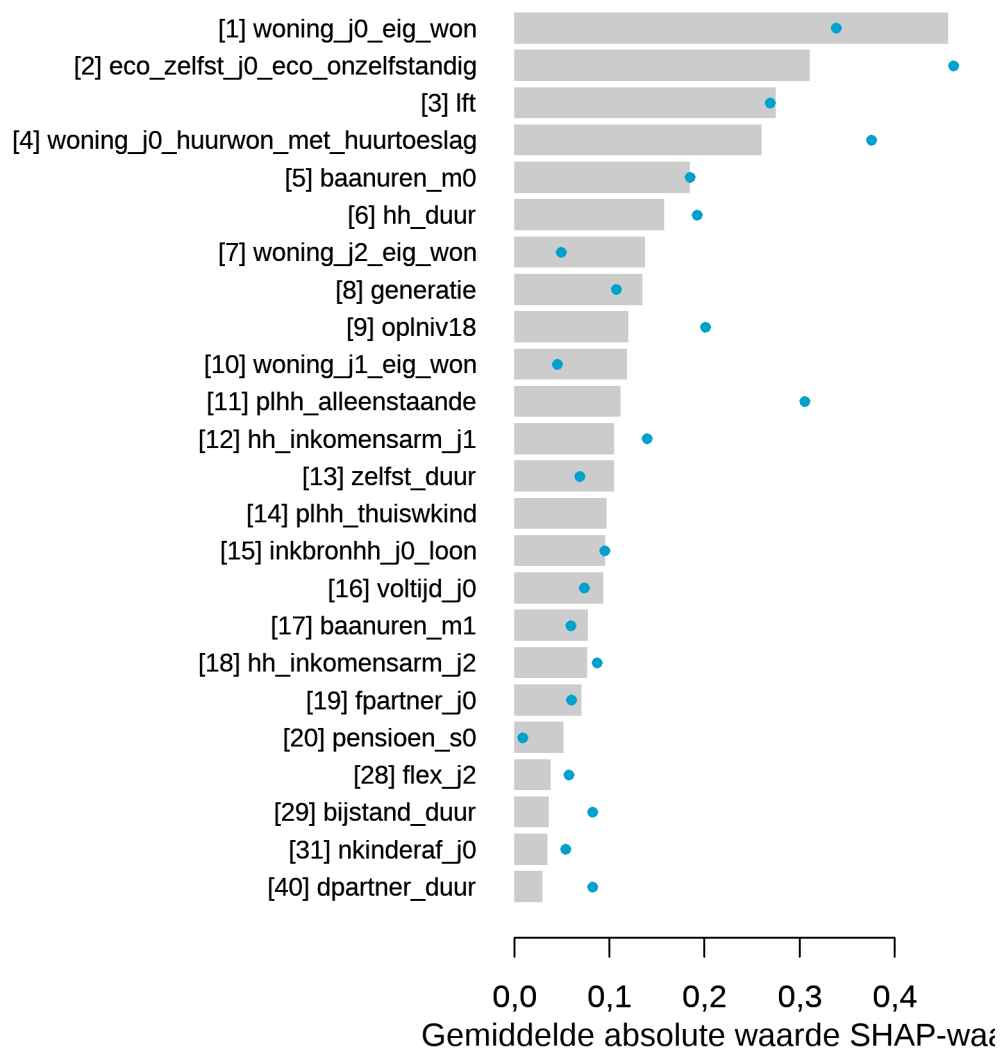
Onderstaande figuur toont de verdeling van de geschatte kans om arm te worden voor de niet-arme deelpopulatie laagopgeleide jongeren (blauw) en voor de gehele niet-arme populatie in Nederland (grijs). De verticale lijnen tonen de gemiddelde kans om arm te worden. De gemiddelde kans voor de laagopgeleide jongeren is met (12%) fors hoger dan die van de gehele Nederlandse niet-arme populatie (1,2%). Hoewel binnen de groep laagopgeleide jongeren de meesten net als in de gehele populatie een kleine kans hebben om arm te worden, is er ook een groep die een fors hoger risico loopt. Dit is te zien aan de staart van de verdeling die hoger ligt waardoor ook het gemiddelde in deze groep hoger ligt.



Belangrijkste kenmerken

Onderstaande figuur geeft de belangrijkste kenmerken weer op basis van SHAP-waarden. (Zie paragraaf 3.3 voor verdere toelichting.) De 20 belangrijkste kenmerken in de gehele niet-arme populatie (grijze balken) zijn aangevuld met de 20 belangrijkste kenmerken voor de deelpopulatie (blauwe punten). Bij kenmerken waarbij iedereen in de deelpopulatie dezelfde waarde heeft, ontbreekt de blauwe punt, aangezien dit kenmerk geen effect heeft binnen de deelpopulatie.

In onderstaande figuur is te zien dat de belangrijkste kenmerken voor deze deelpopulatie redelijk overeenkomen met de gehele arme populatie. (Zie paragraaf 4.1 voor een omschrijving van deze kenmerken.) Wel is te zien dat de blauwe punten voor een aantal kenmerken in meer of mindere mate afwijken van de grijze balken. Hieronder gaan we in op een aantal kenmerken die *belangrijker* zijn dan voor de Nederlandse niet-arme populatie, dus waarvoor de blauwe punt rechts van de grijze balk ligt. Dit is met name het geval voor het niet economisch zelfstandig zijn (`eco_zelfst_j0_eco_onzelfstandig = 1`), het wonen in een woning met huurtoeslag (`woning_j0_huurwon_met_huurtoeslag = 1`) en plaats in het huishouden alleenstaand (`plhh_alleenstaande = 1`). In de volgende figuren bekijken we hoe deze kenmerken de kans om arm te worden beïnvloeden.



Effect op kans om arm te worden

Onderstaande figuren tonen het *effect* op de kans om arm te worden en de *decompositie* van het effect. De staafdiagrammen tonen het effect van de 20 belangrijkste kenmerken op de geschatte kans om arm te worden voor de deelpopulatie (blauwe balken) en de gehele niet-arme populatie in Nederland (grijze balken). Een negatieve waarde op de y-as betekent dat personen met dit kenmerk een lagere kans dan gemiddeld hebben om arm te worden, een positieve waarde betekent dat personen een hogere kans dan gemiddeld hebben. De hoogte van de balk geeft de range van de kans weer voor 80% van de betreffende populatie. De breedte van de balk komt overeen met de relatieve grootte van de populatie met deze waarde. Een smalle balk betekent dus relatief weinig mensen en een brede balk relatief veel mensen. Zie paragraaf 3.3 voor verdere toelichting.

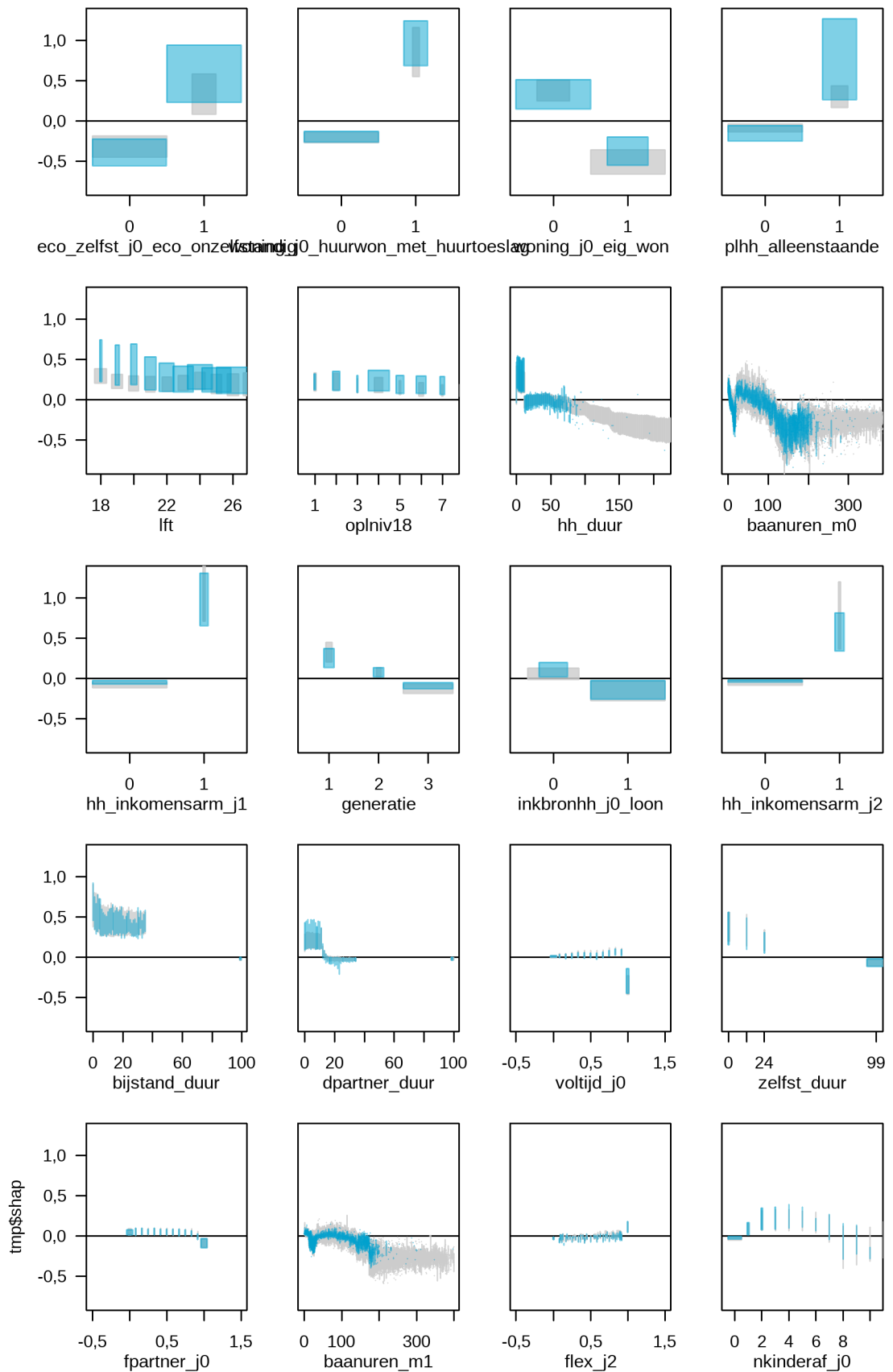
De tweede figuur toont de decompositie van het verschil in belang van de belangrijkste kenmerken. Het verschil in belang wordt opgesplitst in een component die het belang van de *verdeling* van het kenmerk weergeeft (horizontale as) en een component die het belang weergeeft van het *effect* dat het kenmerk heeft op de geschatte kans (verticale as). Zie paragraaf 3.3 voor verdere toelichting.

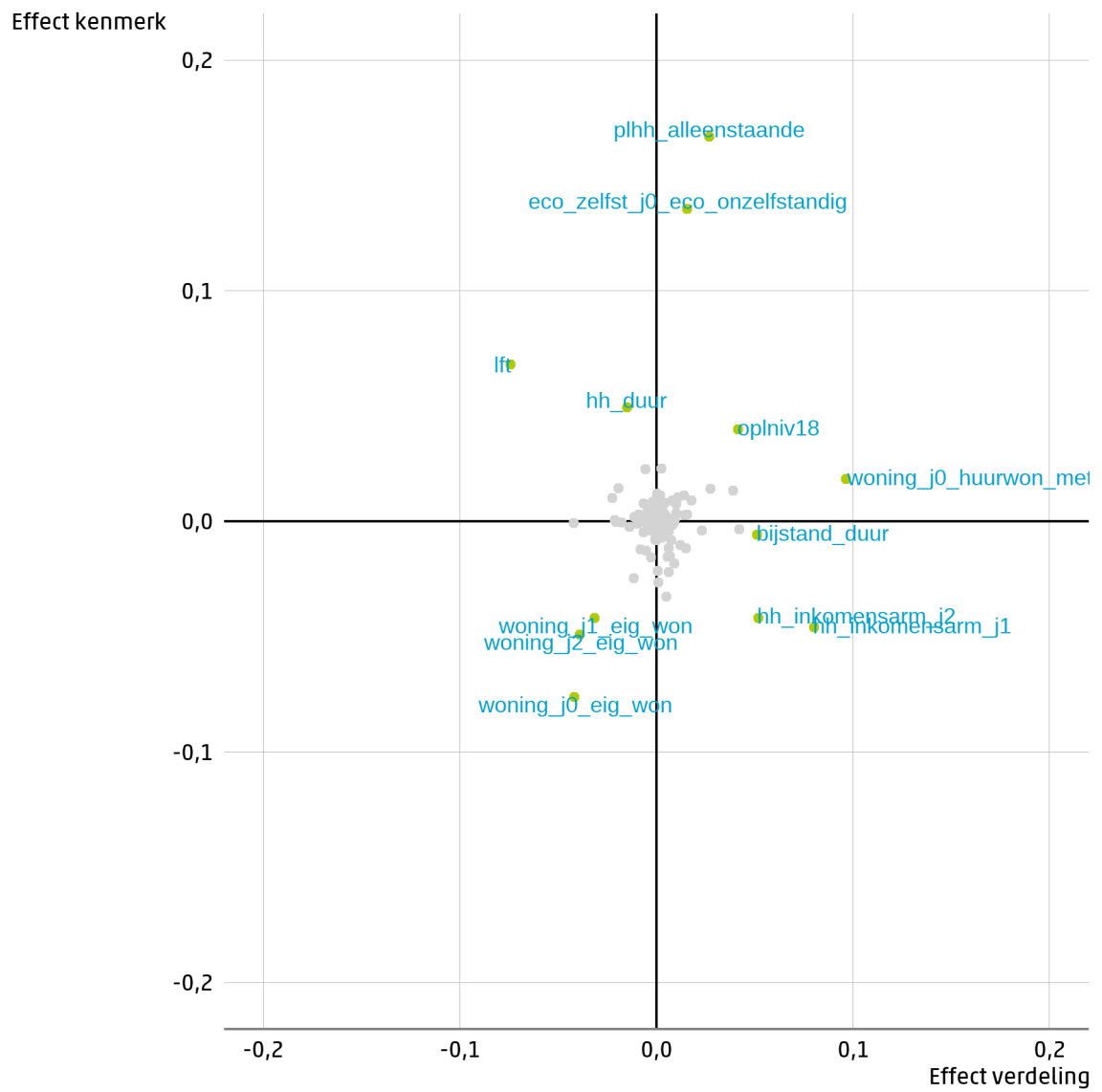
De decompositiefiguur laat zien dat het verschil in belang van alleenstaand (plhh_alleenstaande) met name verklaard wordt door het belang van een positief kenmerkeffect. In het staafdiagram is te zien dat

alleenstaanden in deze deelpopulatie een grotere kans hebben om arm te worden en niet-alleenstaanden juist een kleinere kans. De blauwe balken liggen verder van nul af. Daarnaast is deze deelpopulatie relatief vaker alleenstaand. De blauwe balk voor $p1hh_alleenstaande = 1$ is breder dan de grijze.

Voor het kenmerk economische onzelfstandigheid ($eco_zelfst_j0_eco_onzelfstandig$) toont de decompositiefiguur een vergelijkbaar effect; het effect van dit kenmerk is belangrijker dan de verdeling. De kans om in de armoede terecht te komen is—net als landelijk—groter bij personen die economisch onzelfstandig zijn, en kleiner voor personen die economisch zelfstandig zijn. In deze deelpopulatie is dit effect ook nog eens groter. De blauwe balken liggen namelijk verder van nul af dan de grijze. Daarnaast is in de decompositiefiguur ook in mindere mate een effect van de verdeling te zien. Dit is terug te zien in het staafdiagram; deze deelpopulatie is relatief vaker economisch onzelfstandig ($eco_zelfst_j0_eco_onzelfstandig = 1$) in vergelijking met de gehele niet-arme populatie. De blauwe balk is voor deze groep namelijk breder dan de grijze.

Verder is in de decompositiefiguur te zien dat het grotere belang van een huurwoning met huurtoeslag ($woning_j0_huurwon_met_huurtoeslag$) met name verklaard wordt door een positief verdelingseffect. Deze deelpopulatie woont namelijk vaker in een huurwoning met huurtoeslag. De blauwe balk voor $woning_j0_huurwon_met_huurtoeslag = 1$ is breder dan de grijze. Daarbij is het effect wat sterker dan voor de hele niet-arme populatie. De blauwe balk ligt wat hoger dan de grijze.





E.9 Niet-arme populatie - werkenden

Deze paragraaf zoomt in op de deelpopulatie niet-arme werkenden:

- hoe is deze deelpopulatie afgebakend,
- in hoeverre verschilt deze deelpopulatie qua kenmerken van de hele niet-arme populatie,
- hoe groot is de kans om arm te worden,
- in hoeverre verschillen de kenmerken die het belangrijkste zijn voor het schatten van de armoedekans van het landelijke beeld,
- welk effect hebben deze kenmerken op de kans om arm te worden.

Afbakening

- In referentiejaar niet in de armoede;
- In referentiejaar werkend: werknemer, directeur-groothouder, zelfstandig ondernemer, overig zelfstandige of meewerkend gezinslid.

Deze deelpopulatie bestaat uit personen die in het referentiejaar t niet arm en werkend zijn.

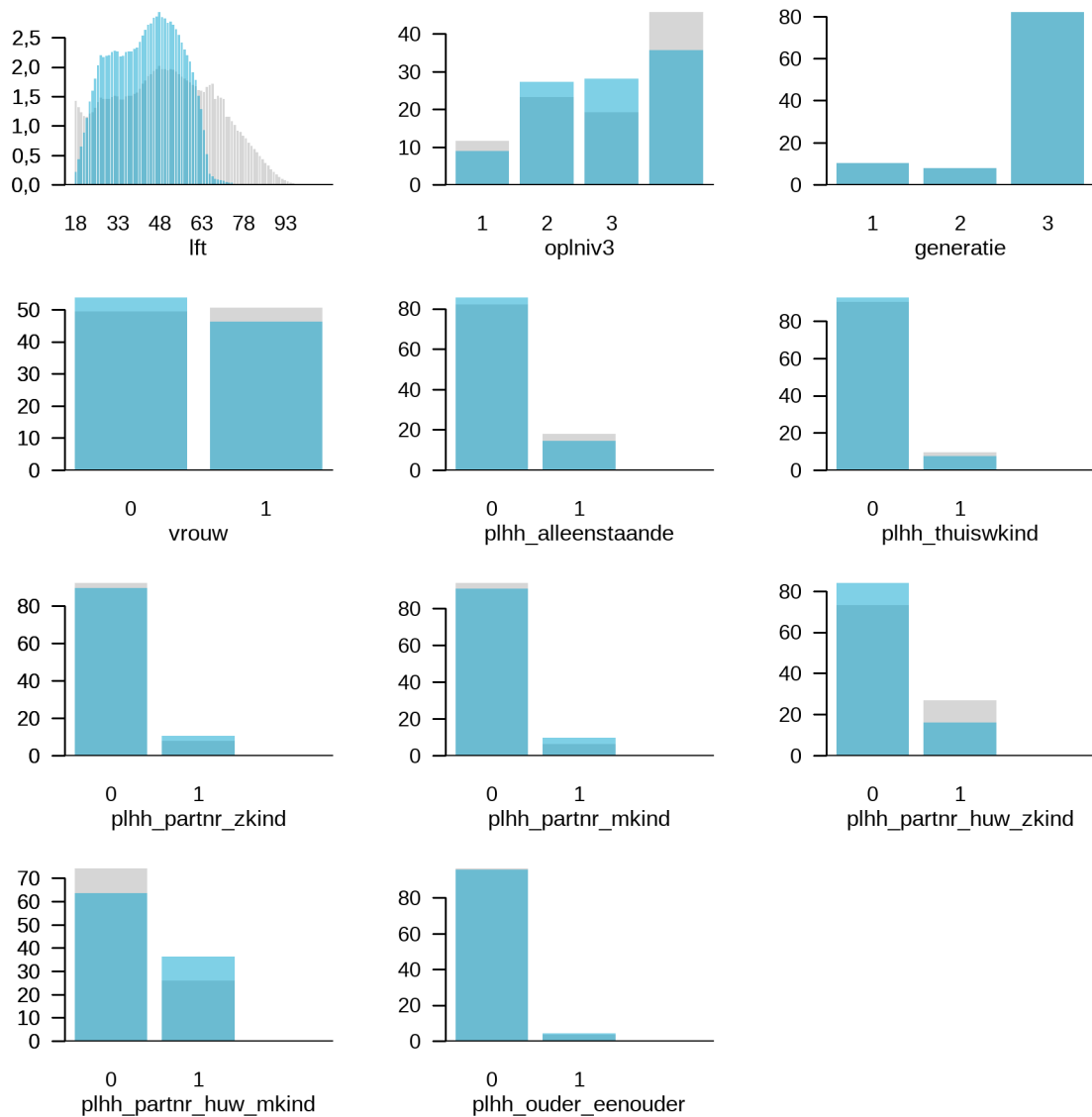
Kenmerken

Om beter zicht te krijgen op deze deelpopulatie hebben we gekeken naar een aantal achtergrondkenmerken:

- leeftijd ($1ft$)
- opleidingsniveau ($oplniv3$)
- generatie migratieachtergrond
 - eerste generatie migratieachtergrond ($generatie = 1$)
 - tweede generatie ($generatie = 2$)
 - Nederlandse achtergrond ($generatie = 3$)
- geslacht (vrouw: $vrouw = 1$)
- type huishouden
 - alleenstaande ($plhh_alleenstaande = 1$)
 - thuiswonend kind ($plhh_thuiswkind = 1$)
 - ongehuwd partner zonder kinderen ($plhh_partn_zkind = 1$)
 - ongehuwd partner met kinderen ($plhh_partn_mkind = 1$)
 - gehuwd partner met kinderen ($plhh_huw_zkind = 1$)
 - gehuwd partner zonder kinderen ($plhh_huw_mkind = 1$)
 - ouder in eenouderhuishouden ($plhh_ouder_eenouder = 1$)

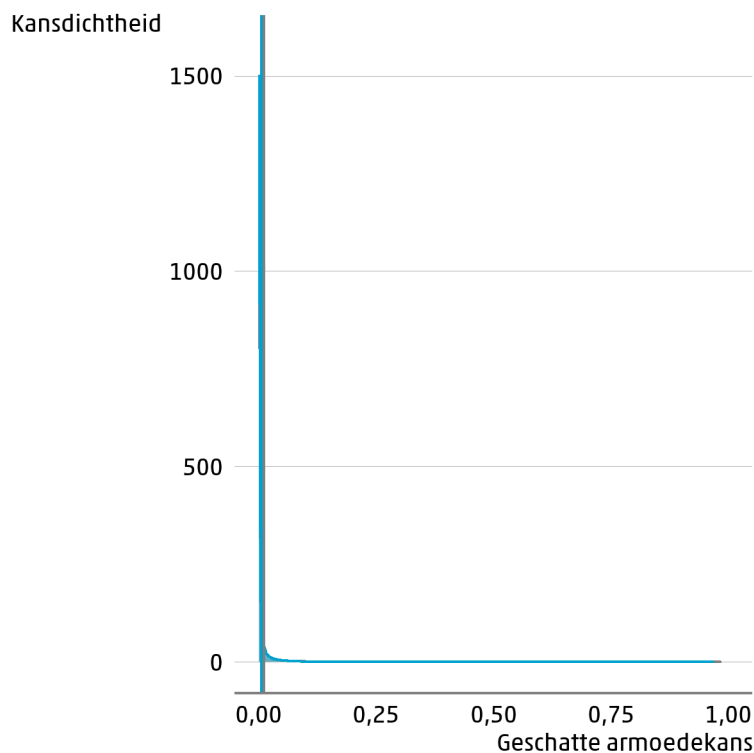
Onderstaande figuren tonen de verdeling van deze kenmerken in deze deelpopulatie (blauw) in vergelijking met de gehele niet-arme populatie (grijs). Merk op dat het opleidingsniveau deels op registraties en deels op enquêtes gebaseerd is, waardoor dit niet voor de gehele populatie bekend is. Hierdoor is de groep met een onbekend opleidingsniveau relatief groot. Ook is dit een selectieve groep. Zo is het opleidingsniveau relatief vaak onbekend voor ouderen en personen met een migratieachtergrond. Dit kan leiden tot vertekening in de opleidingsniveauverdeling. Daarom laten we dit buiten beschouwing bij het beschrijven van de populatiekenmerken.

Uit de figuren blijkt dat er in de deelpopulatie werkenden met name relatief meer personen tussen de ongeveer 20 en de 60 jaar zijn ($20 < 1ft < 60$), meer mannen ($vrouw = 0$) en relatief meer gehuwde partners met kinderen ($plhh_huw_mkind = 1$) en relatief minder gehuwde partners zonder kind ($plhh_huw_zkind = 1$).



Kans om arm te worden

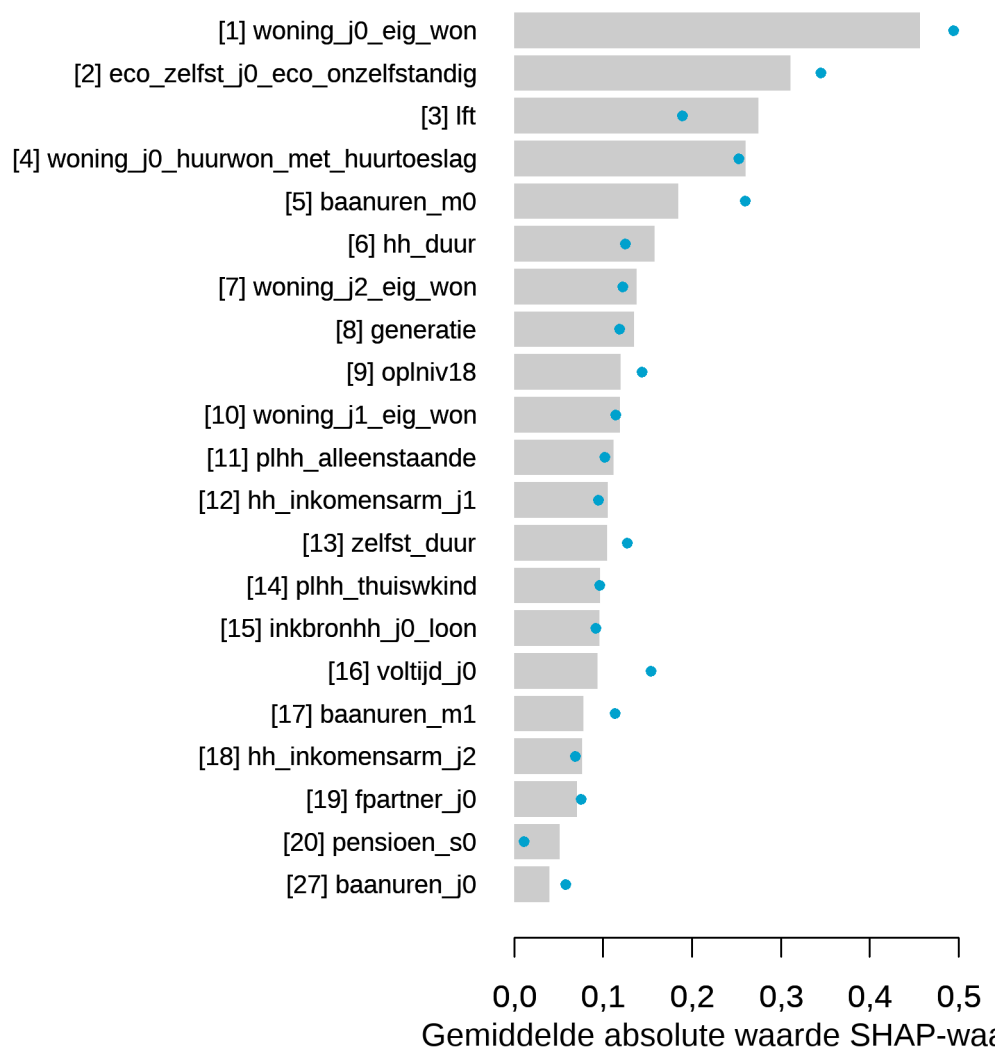
Onderstaande figuur toont de verdeling van de geschatte kans om arm te worden voor de niet-arme deelpopulatie wekrenden (blauw) en voor de gehele niet-arme populatie in Nederland (grijs). De verticale lijnen tonen de gemiddelde kans om arm te worden. De gemiddelde kans voor deze deelpopulatie (0,7%) is lager dan die van de gehele Nederlandse niet-arme populatie (1,2%).



Belangrijkste kenmerken

Onderstaande figuur geeft de belangrijkste kenmerken weer op basis van SHAP-waarden. (Zie paragraaf 3.3 voor verdere toelichting.) De 20 belangrijkste kenmerken in de gehele niet-arme populatie (grijze balken) zijn aangevuld met de 20 belangrijkste kenmerken voor de deelpopulatie (blauwe punten). Bij kenmerken waarbij iedereen in de deelpopulatie dezelfde waarde heeft, ontbreekt de blauwe punt, aangezien dit kenmerk geen effect heeft binnen de deelpopulatie.

In onderstaande figuur is te zien dat de belangrijkste kenmerken voor de niet-arme werkenden in grote mate overeenkomen met de gehele niet-arme populatie. (Zie paragraaf 4.1 voor een omschrijving van deze kenmerken.) Wel is te zien dat de blauwe punten voor een aantal kenmerken in meer of mindere mate afwijken van de grijze balken. Hieronder gaan we in op een aantal kenmerken die belangrijker zijn dan voor de Nederlandse niet-arme populatie, dus waarvoor de blauwe punt rechts van de grijze balk ligt. Zo is te zien dat voor deze deelpopulatie het aantal baanuren (`baanuren_m0`) en het voltijd werken (`voltijd_j0`) belangrijker zijn dan voor de niet-arme populatie. Dit is in lijn met eerder onderzoek waaruit blijkt dat risico op armoede kleiner is voor mensen die voltijd werken dan die deeltijd werken, en dat het risico groter is voor kleine deeltijdbanen dan voor substantiële deeltijdbanen (Van den Brakel and Otten, 2019). Ook eigenwoningbezit (`woning_j0_eig_won`) en economische zelfstandigheid (`eco_zelfst_j0_ec0_onzelfstandig`) zijn voor deze deelpopulatie wat belangrijker dan voor de niet-arme populatie. In de volgende figuren bekijken we hoe de kenmerken voltijd werk en baanuren de kans om arm te worden beïnvloeden.



Effect op kans om arm te worden

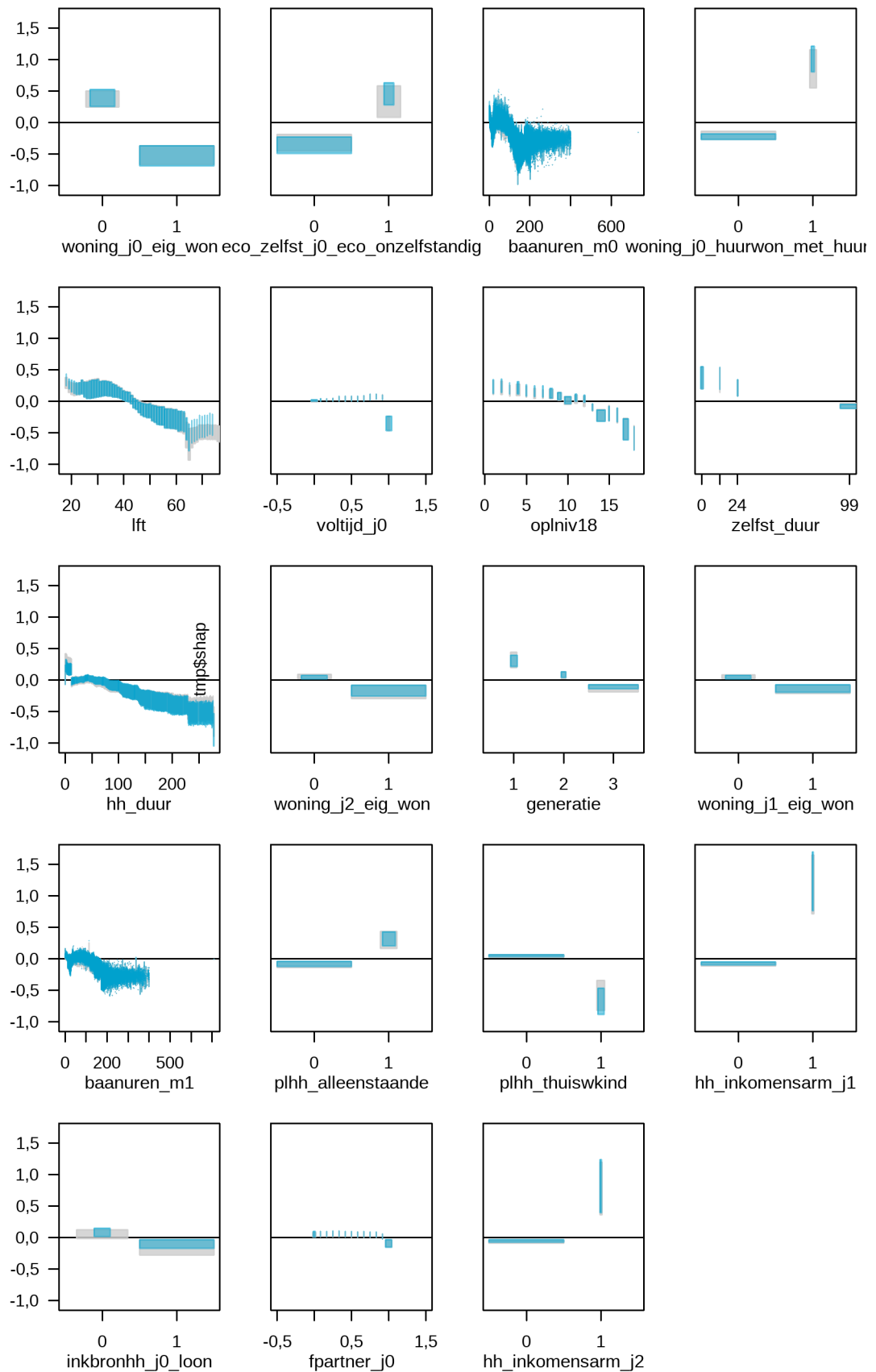
Onderstaande figuren tonen het *effect* op de kans om arm te worden en de *decompositie* van het effect. De staafdiagrammen tonen het effect van de 20 belangrijkste kenmerken op de geschatte kans om arm te worden voor de deelpopulatie (blauwe balken) en de gehele niet-arme populatie in Nederland (grijze balken). Een negatieve waarde op de y-as betekent dat personen met dit kenmerk een lagere kans dan gemiddeld hebben om arm te worden, een positieve waarde betekent dat personen een hogere kans dan gemiddeld hebben. De hoogte van de balk geeft de range van de kans weer voor 80% van de betreffende populatie. De breedte van de balk komt overeen met de relatieve grootte van de populatie met deze waarde. Een smalle balk betekent dus relatief weinig mensen en een brede balk relatief veel mensen. Zie paragraaf 3.3 voor verdere toelichting.

De tweede figuur toont de decompositie van het verschil in belang van de belangrijkste kenmerken. Het verschil in belang wordt opgesplitst in een component die het belang van de *verdeling* van het kenmerk weergeeft (horizontale as) en een component die het belang weergeeft van het *effect* dat het kenmerk heeft op de geschatte kans (verticale as). Zie paragraaf 3.3 voor verdere toelichting.

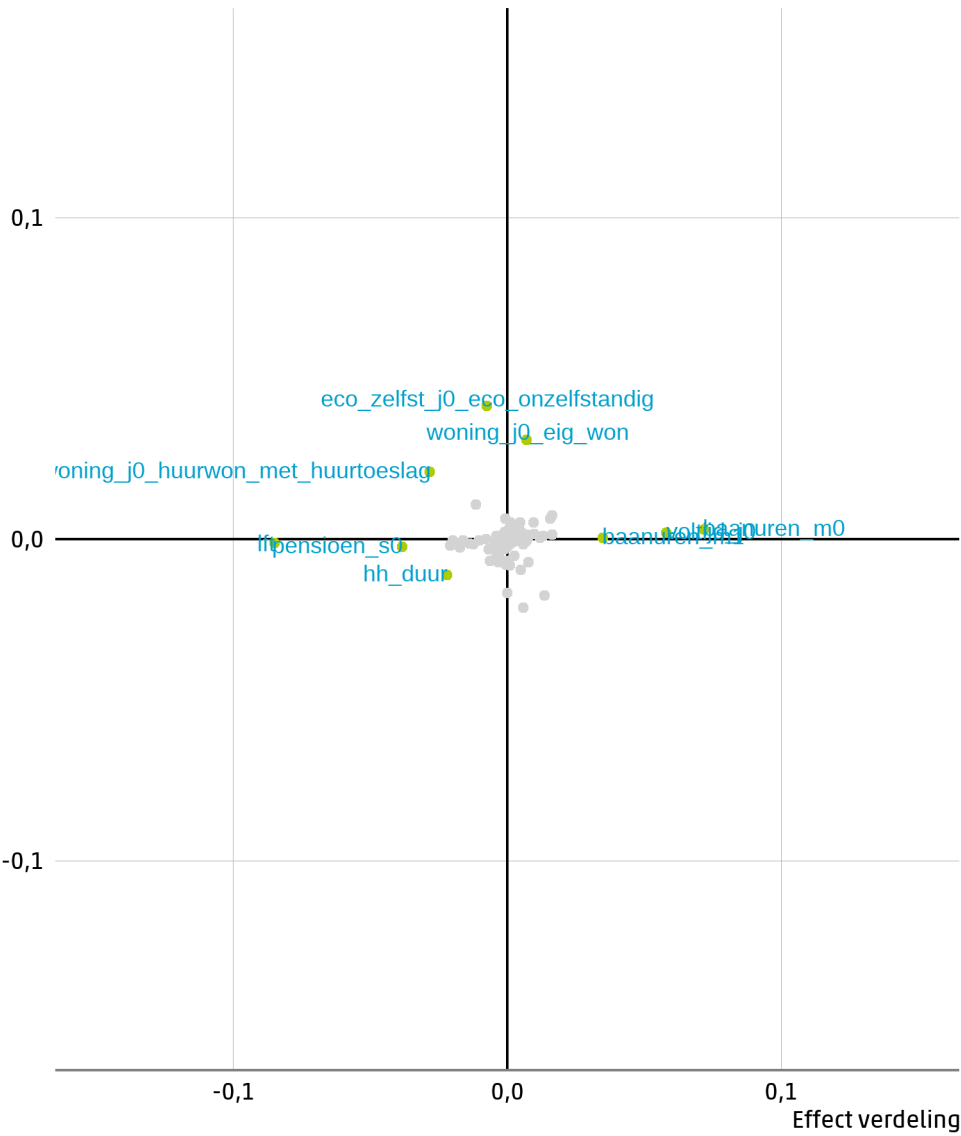
De decompositie figuur laat zien dat het recent aantal baanuren (baanuren_m0) en het voltijd werken (voltijd_j0) met name belangrijker zijn vanwege de verdeling van de deelpopulatie. Werkenden hebben

tenslotte meer baanuren en werken relatief vaker voltijd dan de gehele niet-arme populatie. In de staafdiagrammen is dit niet goed te zien, doordat beide kenmerken uit veel mogelijke waarden bestaan, waardoor de blauwe en grijze balken niet zichtbaar in breedte verschillen. Hierdoor is dus niet te zien in hoeverre bepaalde groepen oververtegenwoordigd zijn in de deelpopulatie.

Voor eigenwoningbezit (`woning_j0_eig_won`) en economische zelfstandigheid (`eco_zelfst_j0_ec0_onzelfstandig`) geldt het omgekeerde. Deze kenmerken zijn in deze deelpopulatie belangrijker dan in de gehele niet-arme populatie vanwege een positief kenmerkeffect. In de figuur is vooral te zien dat economisch onzelfstandigen in deze deelpopulatie een nog grotere armoedekans hebben.



Effect kenmerk



E.10 Niet-arme populatie - alleenstaande ouders

Deze paragraaf zoomt in op de deelpopulatie niet-arme alleenstaande ouders:

- hoe is deze deelpopulatie afgebakend,
- in hoeverre verschilt deze deelpopulatie qua kenmerken van de hele niet-arme populatie,
- hoe groot is de kans om arm te worden,
- in hoeverre verschillen de kenmerken die het belangrijkste zijn voor het schatten van de armoedekans van het landelijke beeld,
- welk effect hebben deze kenmerken op de kans om arm te worden.

Afbakening

- In referentiejaar niet in de armoede;
- In referentiejaar ouder in eenouderhuishouden.

Deze deelpopulatie betreft personen die in het referentiejaar t niet arm zijn en in een eenouderhuishouden wonen.

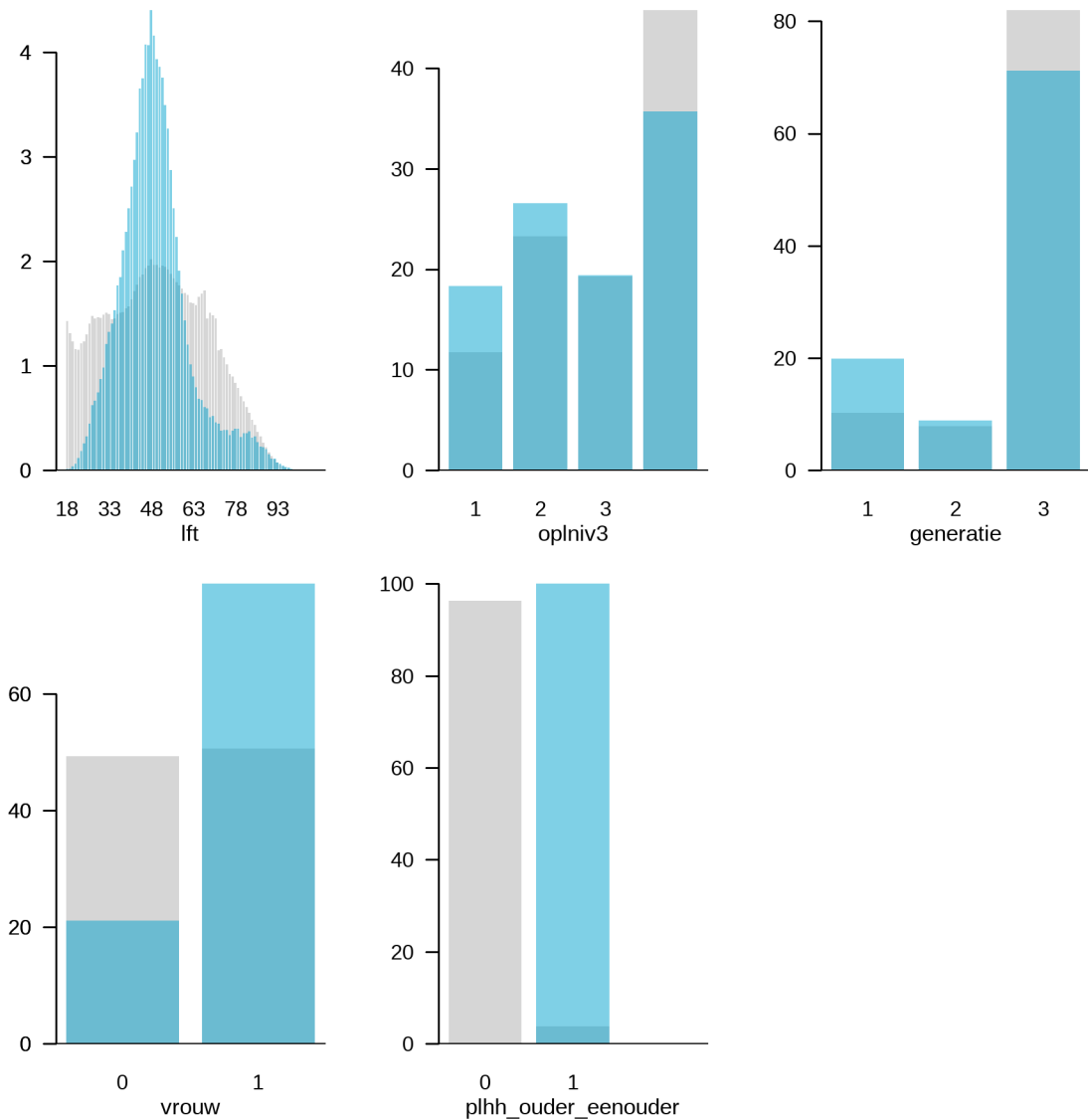
Kenmerken

Om beter zicht te krijgen op deze deelpopulatie hebben we gekeken naar een aantal achtergrondkenmerken:

- leeftijd (lft)
- opleidingsniveau ($oplniv3$)
- generatie migratieachtergrond
 - eerste generatie migratieachtergrond ($generatie = 1$)
 - tweede generatie ($generatie = 2$)
 - Nederlandse achtergrond ($generatie = 3$)
- geslacht (vrouw: $vrouw = 1$)
- type huishouden
 - alleenstaande ($plhh_alleenstaande = 1$)
 - thuiswonend kind ($plhh_thuiswkind = 1$)
 - ongehuwd partner zonder kinderen ($plhh_partn_zkind = 1$)
 - ongehuwd partner met kinderen ($plhh_partn_mkind = 1$)
 - gehuwd partner met kinderen ($plhh_huw_zkind = 1$)
 - gehuwd partner zonder kinderen ($plhh_huw_mkind = 1$)
 - ouder in eenouderhuishouden ($plhh_ouder_eenouder = 1$)

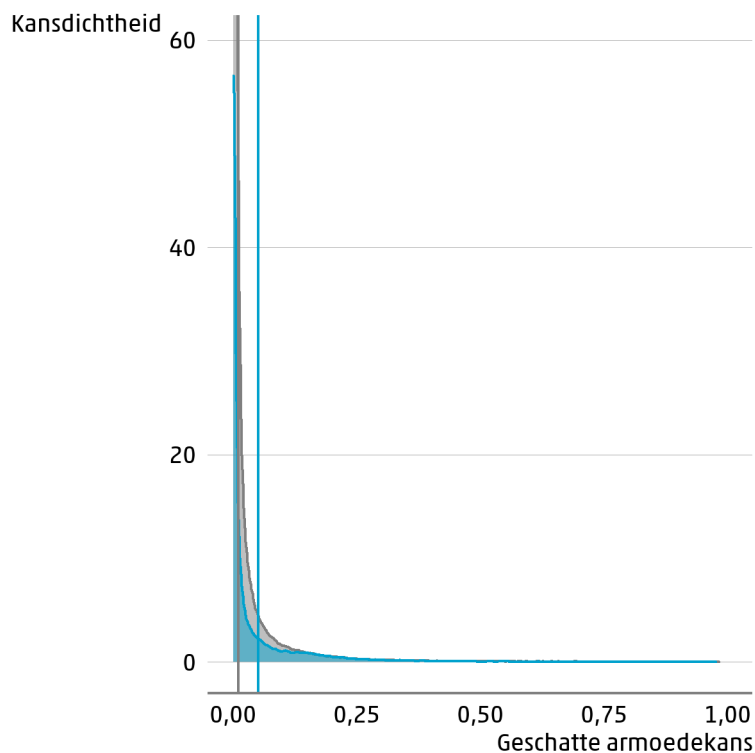
Onderstaande figuren tonen de verdeling van deze kenmerken in deze deelpopulatie (blauw) in vergelijking met de gehele niet-arme populatie (grijs). Merk op dat het opleidingsniveau deels op registraties en deels op enquêtes gebaseerd is, waardoor dit niet voor de gehele populatie bekend is. Hierdoor is de groep met een onbekend opleidingsniveau relatief groot. Ook is dit een selectieve groep. Zo is het opleidingsniveau relatief vaak onbekend voor ouderen en personen met een migratieachtergrond. Dit kan leiden tot vertekening in de opleidingsniveauverdeling. Daarom laten we dit buiten beschouwing bij het beschrijven van de populatiekenmerken.

Uit de figuren blijkt dat er in de deelpopulatie alleenstaande ouders met name relatief meer personen tussen de ongeveer 35 en de 60 jaar zijn ($35 < lft < 60$), personen met een eerste generatie migratieachtergrond ($generatie = 1$) en relatief meer vrouwen ($vrouw = 1$).



Kans om arm te worden

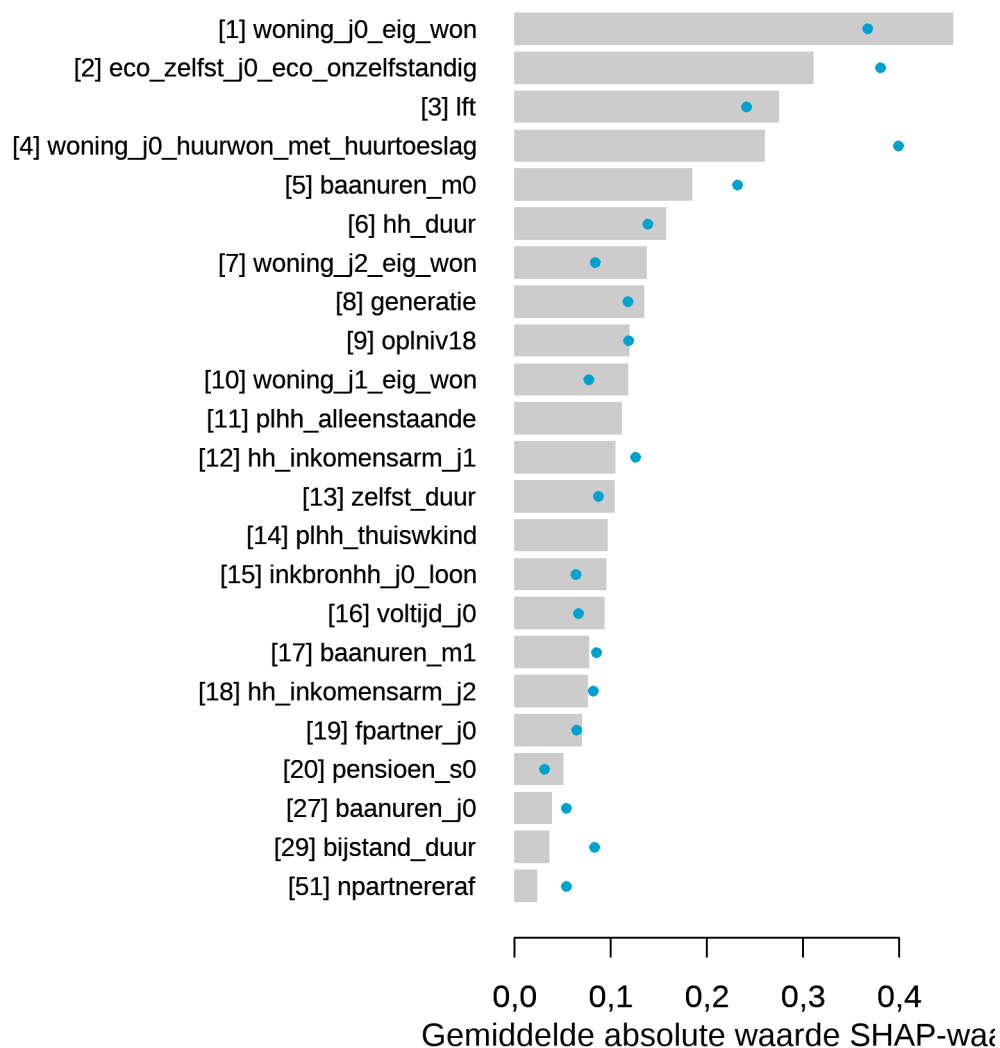
Onderstaande figuur toont de verdeling van de geschatte kans om arm te worden voor de niet-arme deelpopulatie alleenstaande ouders (blauw) en voor de gehele niet-arme populatie in Nederland (grijs). De verticale lijnen tonen de gemiddelde kans om arm te worden. Hieruit blijkt dat de gemiddelde kans om arm te worden voor alleenstaande ouders (5%) groter is dan die van de gehele Nederlandse niet-arme populatie (1,2%).



Belangrijkste kenmerken

Onderstaande figuur geeft de belangrijkste kenmerken weer op basis van SHAP-waarden. (Zie paragraaf 3.3 voor verdere toelichting.) De 20 belangrijkste kenmerken in de gehele niet-arme populatie (grijze balken) zijn aangevuld met de 20 belangrijkste kenmerken voor de deelpopulatie (blauwe punten). Bij kenmerken waarbij iedereen in de deelpopulatie dezelfde waarde heeft, ontbreekt de blauwe punt, aangezien dit kenmerk geen effect heeft binnen de deelpopulatie.

In onderstaande figuur is te zien dat de belangrijkste kenmerken voor niet-arme alleenstaande ouders in grote mate overeenkomen met de gehele niet-arme populatie. (Zie paragraaf 4.1 voor een omschrijving van deze kenmerken.) Wel is te zien dat de blauwe punten voor een aantal kenmerken in meer of mindere mate afwijken van de grijze balken. Hieronder gaan we in op een aantal kenmerken die belangrijker zijn dan voor de Nederlandse niet-arme populatie, dus waarvoor de blauwe punt rechts van de grijze balk ligt. Zo zijn onder andere het niet economisch zelfstandig zijn (`eco_zelfst_j0_eco_onzelfstandig` = 1) en het wonen in een woning met huurtoeslag (`woning_j0_huurwon_met_huurtoeslag` = 1) voor niet-arme alleenstaande ouders belangrijker dan voor de hele niet-arme populatie. In de volgende figuren bekijken we hoe deze kenmerken de kans om arm te worden beïnvloeden.



Effect op kans om arm te worden

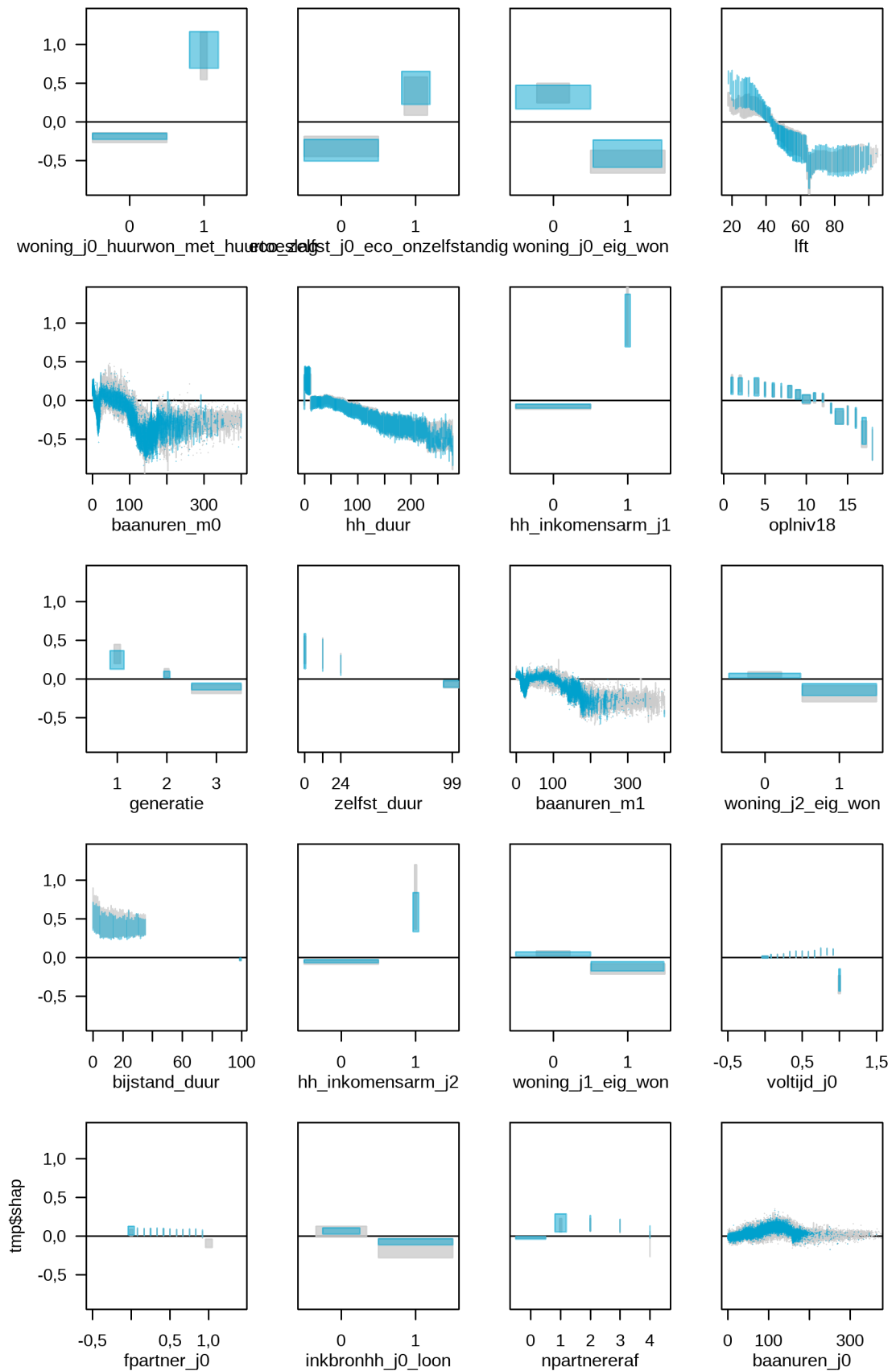
Onderstaande figuren tonen het *effect* op de kans om arm te worden en de *decompositie* van het effect. De staafdiagrammen tonen het effect van de 20 belangrijkste kenmerken op de geschatte kans om arm te worden voor de deelpopulatie (blauwe balken) en de gehele niet-arme populatie in Nederland (grijze balken). Een negatieve waarde op de y-as betekent dat personen met dit kenmerk een lagere kans dan gemiddeld hebben om arm te worden, een positieve waarde betekent dat personen een hogere kans dan gemiddeld hebben. De hoogte van de balk geeft de range van de kans weer voor 80% van de betreffende populatie. De breedte van de balk komt overeen met de relatieve grootte van de populatie met deze waarde. Een smalle balk betekent dus relatief weinig mensen en een brede balk relatief veel mensen. Zie paragraaf 3.3 voor verdere toelichting.

De tweede figuur toont de decompositie van het verschil in belang van de belangrijkste kenmerken. Het verschil in belang wordt opgesplitst in een component die het belang van de *verdeling* van het kenmerk weergeeft (horizontale as) en een component die het belang weergeeft van het *effect* dat het kenmerk heeft op de geschatte kans (verticale as). Zie paragraaf 3.3 voor verdere toelichting.

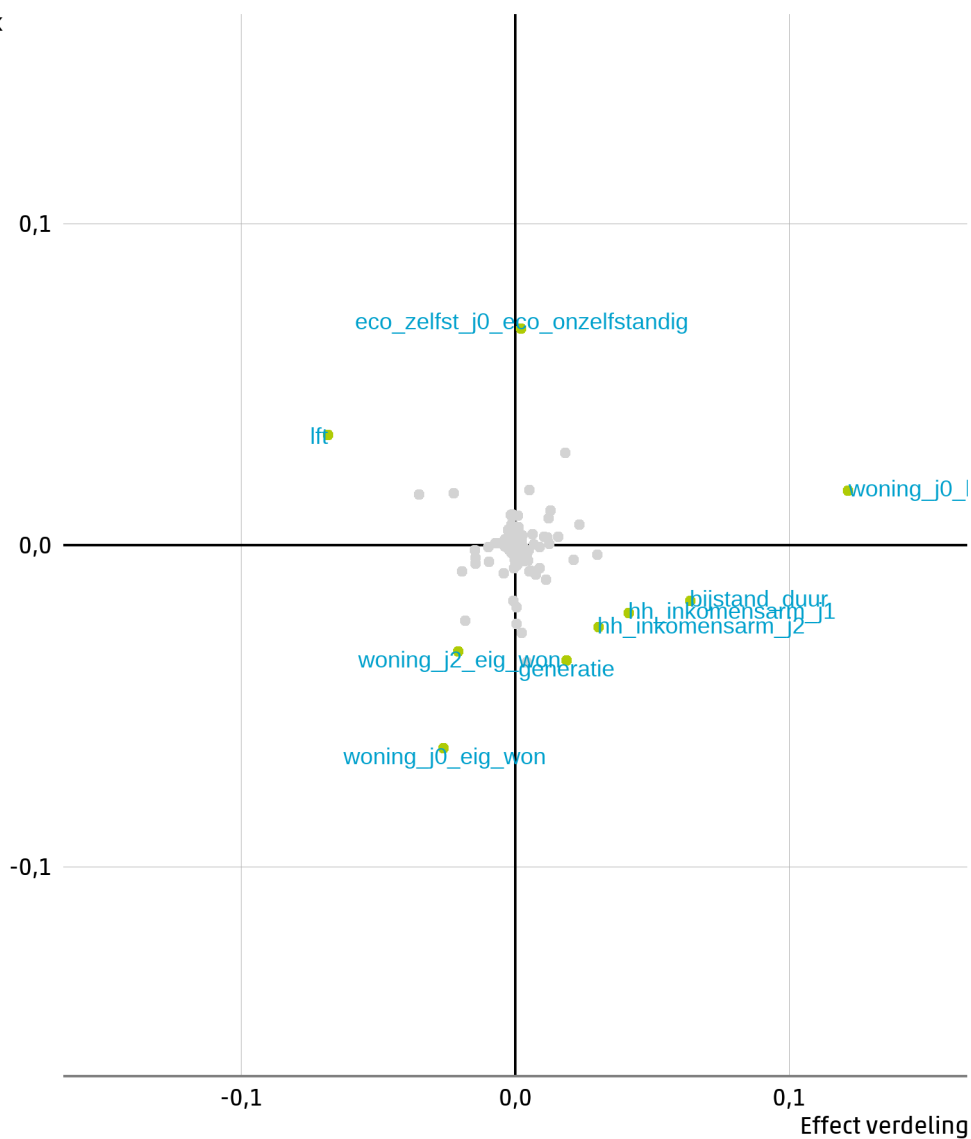
In onderstaande figuren is te zien dat het wonen in een huurwoning met huurtoeslag ($woning_j0_huurwon_met_huurtoeslag = 1$) vaker voorkomt bij alleenstaande ouders dan in de gehele

niet-arme populatie. Dit positieve verdelingseffect zorgt ervoor dat dit kenmerk belangrijker is voor deze deelpopulatie. De blauwe balk is voor deze groep namelijk breder dan de grijze. Daarbij heeft het wonen in een woning met huurtoeslag een iets hoger armoederisico dan in de gehele niet-arme populatie (de blauwe balk is iets hoger dan de grijze). Ook de decompositiefiguur laat dit positieve kenmerkeffect zien.

Daarnaast zien we in deze deelpopulatie dat personen die economisch onzelfstandig zijn ($\text{eco_zelfst_j0_eco_onzelfstandig} = 1$) een hogere kans hebben om arm te worden dan in de gehele niet-arme populatie. (De blauwe balken liggen namelijk wat verder van nul dan de grijze.) Dit komt ook terug in de decompositie figuur, waar vooral een verschil te zien is in het effect op de geschatte kans (afwijking in verticale richting).



Effect kenmerk



Referenties

- Burger, J. and Meertens, Q. (2020). The algorithm versus the chimps: On the minima of classifier performance metrics. In Cao, L., Kusters, W., and Lijffijt, J., editors, *Proceedings of the BNAIC/BeneLearn2020*, pages 38–55. Leiden University.
- CBS (2020). *Armoederisico 2019*. Centraal Bureau voor de Statistiek. <https://www.cbs.nl/nl-nl/nieuws/2020/49/armoederisico-bevolking-in-2019-een-fractie-lager>.
- CBS (2021a). *Risicofactoren voor transities in en uit armoede - samenvatting*. Centraal Bureau voor de Statistiek. <https://www.cbs.nl/-/media/innovatie/risicofactoren-armoede-samenvatting.pdf>.
- CBS (2021b). *Vermogen van huishoudens; huishoudenskenmerken, vermogensklassen*. Centraal Bureau voor de Statistiek. <https://opendata.cbs.nl/#/CBS/nl/dataset/83835NED/table?dl=53C00>.
- Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Krishnapuram, B. and Shah, M., editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. Association for Computing Machinery.
- Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., and Li, Y. (2021). *xgboost: Extreme Gradient Boosting*. R package version 1.3.2.1.
- De Waal, T. (2016). Obtaining numerically consistent estimates from a mix of administrative data and surveys. *Statistical Journal of the International Association for Official Statistics*, 32:231–243.
- EZK (2019). *Strategisch Actieplan voor Artificiële Intelligentie*. Ministerie van Economische Zaken en Klimaat. <https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/beleidsnotas/2019/10/08/strategisch-actieplan-voor-artificiele-intelligentie/Rapport+SAPAI.pdf>.
- Friedman, J. and Popescu, B. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3):916–954.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition.
- Liu, Y. and Just, A. (2021). *SHAPforxgboost: SHAP Plots for XGBoost*. R package version 0.1.0.
- Lundberg, S. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Proceedings of the Advances in Neural Information Processing Systems*. Neural Information Processing Systems Foundation.
- Molnar, C. (2019). *Interpretable Machine Learning*. <https://christophm.github.io/interpretable-ml-book/>.
- Powers, D. (2011). Evaluation: From precision, recall and f-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63.
- Therneau, T. and Atkinson, B. (2019). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-15.
- Van den Brakel, M., Arts, K., and Herbers, D. (2021). *Financiële gevolgen van verweeduwing*. Centraal Bureau voor de Statistiek. <https://www.cbs.nl/nl-nl/longread/statistische-trends/2021/financiële-gevolgen-van-verweeduwing>.

- Van den Brakel, M. and Otten, F. (2019). *Armoede en Sociale Uitsluiting 2019*. Centraal Bureau voor de Statistiek. <https://longreads.cbs.nl/armoede-en-sociale-uitsluiting-2019/>.
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.