Inhoud

1 Inleiding **2**

2 Methodiek **2**
2.1 Gevoelige attributen 3
2.2 Metrieken 4

3 Data verzamelen en opschonen **5**
3.1 Data collectie (pre-pilot) 6
3.2 Data opschonen 6

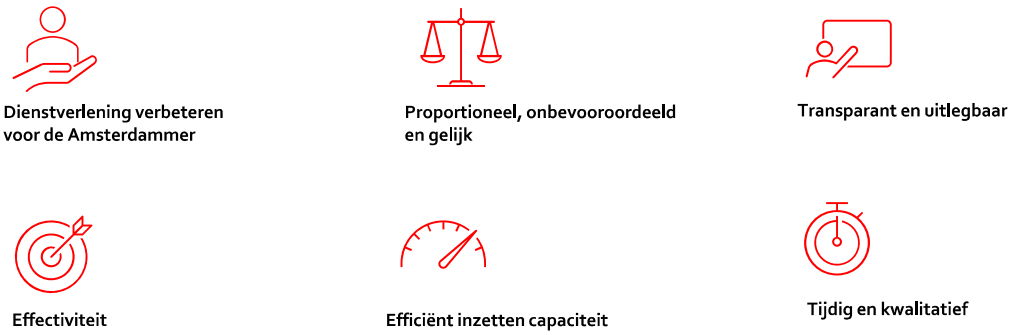
4 Analyse en herweging model **7**
4.1 Analyse 7
4.2 Eerste resultaten 8
4.3 Herweging model 10

5 Resultaten **11**
5.1 Directe bias analyse 11
5.2 Indirecte bias analyse 13
5.3 Conclusie 14

Annex 1 **15**
Lijst westerse landen 15
Lijst niet-westerse landen 15
Reden features niet geselecteerd voor indirecte bias analyse 16

1 Inleiding

Het WPI algoritme wordt ontwikkeld om de afdeling Handhaving te ondersteunen bij het inschatten van de onderzoekswaardigheid van aanvragen levensonderhoud. De ontwikkeling van het WPI algoritme is gedaan met een zesvoudig doel (Figuur 1) in gedachten. Het model wordt ingezet aan de kop van het proces, en voorsorteert of een aanvraag verder onderzocht dient te worden, of niet. Een medewerker voert hierna altijd een vooronderzoek uit om te bepalen of de aanvraag daadwerkelijk onderzoekswaardig is alvorens het naar handhaving wordt doorgezet. Dit model zal dus geen geautomatiseerde beslissingen nemen.



Figuur 1: Doelstellingen WPI algoritme

Het is voor de Gemeente Amsterdam van belang dat de algoritmes die gebruikt worden in de stad niet onbedoeld bias hebben, en aanvragers vaker onterecht selecteert op basis van bijvoorbeeld gender, leeftijd of migratieachtergrond. Dit soort gevoelige attributen worden niet gebruikt door het model, maar het is mogelijk dat andere variabelen die door het model gebruikt worden toch indirect verband houden met dit soort attributen. Dit noemen we proxy variabelen.

Een van de technieken die we gebruiken om er zeker van te zijn dat dit algoritme op een eerlijke, proportionele, onbevooroordeelde en gelijkwaardige manier te werk gaat, is een bias analyse. In dit document zullen we toelichten welke methodiek we hiervoor gebruiken, en wat de uitkomsten hiervan zijn. Als het model eenmaal in gebruik is genomen, zal de bias analyse regelmatig herhaald worden om er zeker van te zijn dat het model vrij blijft van onwenselijke verschillen.

De bias analyse is slechts een van de tools die we tot onze beschikking hebben in de ontwikkeling van eerlijke algoritmes. Voor meer informatie over *fairness* heeft de Gemeente Amsterdam het Fairness Handbook[1] ontwikkelt.

2 Methodiek

Tijdens de bias analyse zullen we vergelijken hoe verschillende groepen scoren op een vooraf geselecteerde bias metriek. We zullen bijvoorbeeld vergelijken hoe mannen scoren versus vrouwen, en mensen met een Nederlandse versus niet Nederlandse nationaliteit. We zullen zowel kijken naar bias in het model als bias in het huidige proces: immers niet alleen een algoritme, maar ook mensen kunnen beïnvloed worden door impliciete bias.

[1] Op dit moment is het Fairness Handbook enkel beschikbaar in het Engels. Aan de Nederlandse vertaling wordt gewerkt.

2.1 Gevoelige attributen

We voeren een directe en indirecte bias analyse uit voor het WPI algoritme. De directe bias analyse is gebaseerd op data uit de BRP, deels verkregen via Socrates[2] en deels direct uit de BRP. We kunnen hierdoor onderstaande gevoelige attributen door middel van een directe bias analyse analyseren.[3] De groeperingen zijn gemaakt op basis van logica in de statistische verdeling van de data; business logica verkregen uit de werkgroep; en de beschikbaarheid van voldoende data binnen elke groep.

Attribuut	Bevoordeelde groep	Benadeelde groep
Leeftijd	30- 40- 50-	30+ 40+ 50+
Geslacht	Man	Vrouw
Geboorteland	Nederland Westers land[4]	Niet-Nederland Niet-westers land[4]
Nationaliteit	Nederlands Westers [4]	Niet-Nederlands Niet-westers [4]

Voor de indirecte bias analyse onderzoeken we in hoeverre bepaalde features indirect bias zouden kunnen veroorzaken. Hierbij kijken we naar een selectie van vijf features. Deze features zijn geselecteerd op basis van 1) een verwachte link tussen de feature en een gevoelig attribuut[5]; en 2) voldoende data beschikbaarheid om een betrouwbare analyse te kunnen doen.

Feature	Attribuut waarvoor proxy	Bevoordeelde groep	Benadeelde groep
Eerder LO ontvangen	Socio-economische status Migratieachtergrond	Nee	Ja
Eerder LO aangevraagd	Socio-economische status Migratieachtergrond	Nee	Ja
Dagen sinds einde dienst	Socio-economische status Migratieachtergrond	Meer dan een jaar Meer dan 2 maanden	Minder dan een jaar Minder dan 2 maanden

[2] Socrates is het zaaksysteem dat gebruikt wordt door de inkomensconsulenten.

[3] §Ook data over de burgerlijke staat is beschikbaar. Helaas is de groep die ofwel een geregistreerd partnerschap heeft, ofwel getrouwd is te klein om een betrouwbare data analyse op uit te voeren. Tijdens de pilotfase van het model zal aanvullende data verzameld worden, en zal deze analyse uitgevoerd kunnen worden.

[4] Voor de indeling westers; niet-westers baseren we ons op de indeling zoals gemaakt door het CBS. In 2022 is het CBS gestopt met het gebruik van deze indeling, en gebruikt in plaats hiervan een indeling op basis van werelddelen en veelvoorkomende immigratielanden (zie https://www.cbs.nl/nl-nl/longread/statistische-trends/2022/nieuwe-indeling-bevolking-naar-herkomst). Voor deze bias analyse hebben wij echter onvoldoende aanvragers met een migratieachtergrond om een dergelijke meer toegespitste analyse te maken. Toch is het aannemelijk dat iemand met bijvoorbeeld een Belgische migratieachtergrond minder risico heeft om een negatieve bias te ervaren dan iemand met bijvoorbeeld een Syrische migratieachtergrond. Om dit met beperkte data inzichtelijk te maken hebben we ervoor gekozen de oude westers; niet-westers indeling van het CBS te gebruiken. Met deze groepering zijn de groepen namelijk wel groot genoeg om de analyse te kunnen doen.

Annex 1 bevat een overzicht van welke landen wel en niet als westers worden gelabeld

[5] Deze link is gebaseerd op de ervaringen van de business, niet op wetenschappelijk beschikbare informatie.

Dagen sinds verhuizing	Sociale klasse Socio-economische status Migratieachtergrond	Meer dan een jaar	Minder dan een jaar
Aantal adressen	Sociale klasse	Eén adres	Meer dan één adres

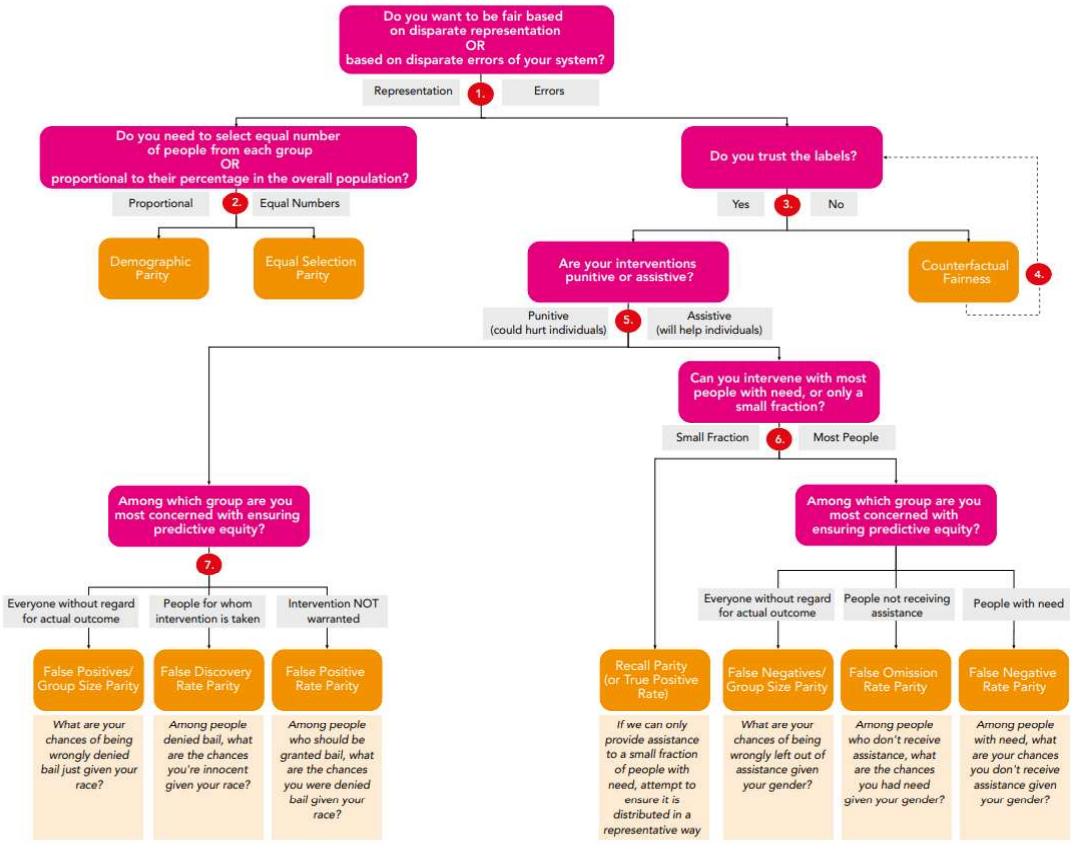
Annex 1 bevat een toelichting bij alle features die niet geselecteerd zijn voor de indirecte bias analyse.

2.2 Metrieken

Om de juiste metriek te selecteren voor de bias analyse maken we gebruik van de Fairness tree (Figuur 2). We doorlopen de boom voor het WPI algoritme als volgt:

- Het model moet eerlijk zijn op basis van fouten (*disparate errors*), vanwege ons doel om proportioneel te handhaven: omdat een handhavingsonderzoek een last is voor de burger, moet dit alleen gedaan worden wanneer het echt nodig is.
- Hoewel er altijd een risico is op valsnegatieven wanneer we het hebben over handhaving, hebben we over het algemeen vertrouwen in onze labels omdat de handhavingsonderzoeken met veel zorg worden uitgevoerd en het besluit onderbouwd wordt genomen.
- Onze interventies zijn 'straffend' (*punitive*), want een interventie door handhaving kan als belastend worden ervaren door de aanvrager.

Zo eindigt de route linksonder. Dit betekent dat we drie mogelijke metrieken over hebben (Figuur 3).



Figuur 2: Fairness tree ontwikkeld door Aequitas

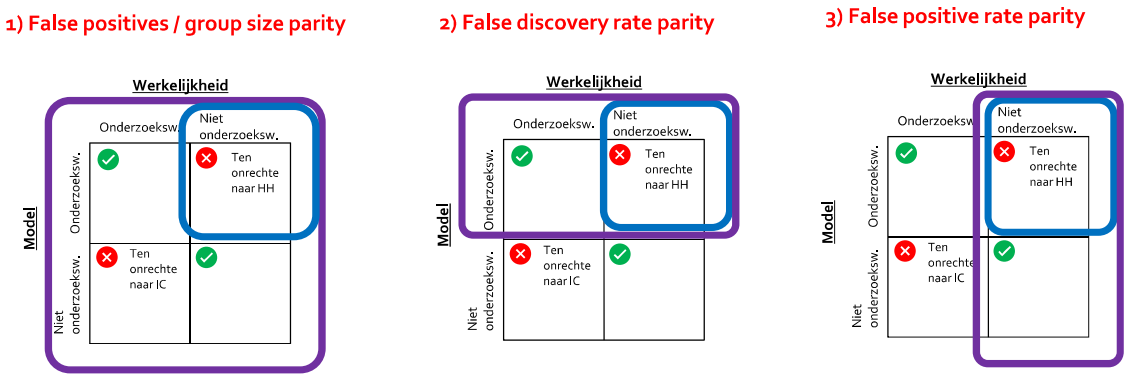
De eerste metriek, *false positives / group size parity*, is uiterst geschikt voor deze analyse. De metriek is te interpreteren als: *wat is de kans dat een willekeurig persoon die een uitkering aanvraagt ten onrechte wordt onderzocht op basis van zijn/haar gender, achtergrond of andere sensitieve attributen?*

De tweede metriek, *false discovery rate parity*, valt af voor onze analyse. De metriek is te interpreteren als: *wat is de kans dat een aanvrager die door het model is geselecteerd als onderzoekswaardig, in werkelijkheid niet onderzoekswaardig is?* Deze metriek is naar ons inzicht vooral gericht op hoe goed het model werkt –

iets dat we al analyseren door middel van de hit-rate – en op hoe efficiënt de handhavingscapaciteit wordt ingezet.

De derde metriek, *false positive rate parity*, kan niet door ons geanalyseerd worden. Tijdens de pre-pilot zullen we enkel aanvragen onderzoeken die door het model als onderzoekswaardig worden aangemerkt. We kunnen daardoor geen onderscheid maken tussen de onderste twee vakken in Figuur 3.

Tijdens de bias analyse zal gebruik gemaakt worden van de eerste metriek, *false positives / group size parity*.

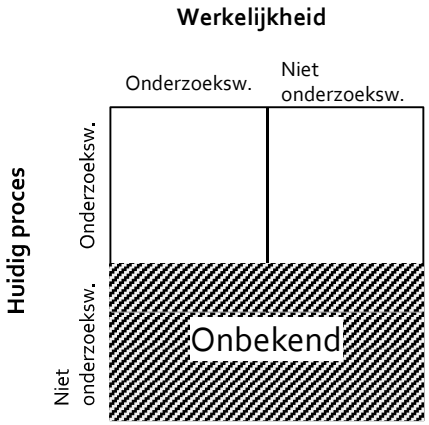


Figuur 3: Drie mogelijke metrieken voor de bias analyse. De uitkomst van de metriek is het blauw omlijnde datapunt gedeeld door de paars omlijnde datapunten.

3 Data verzamelen en opschonen

Om een representatieve bias analyse uit te kunnen voeren, is het startpunt een representatieve, gelabelde dataset. Dit houdt in dat we van een representatief datasample hebben, waarvan we weten of de aanvraag in werkelijkheid onderzoekswaardig is.

De oorspronkelijke trainingsset is enkel gebaseerd op de aanvragen die in het huidige proces als onderzoekswaardig worden geclassificeerd. Deze set heeft dus geen labels voor data die in het huidige proces als niet onderzoekswaardig worden geclassificeerd (Figuur 4). In het huidige proces wordt 7% van de aanvragen geselecteerd door de Business Rules of Inkomensconsulent en doorgestuurd naar handhaving. Dit betekent dat voor 93% van de aanvragen geen label beschikbaar is.



Figuur 4: Confusion matrix met data beschikbaarheid huidig proces

3.1 Data collectie (pre-pilot)

Om de datakwaliteit en -kwantiteit te vergroten doen we een pre-pilot. In deze pre-pilot zullen drie handhavingsspecialisten een sample van aanvragen die in het huidige proces niet als onderzoekswaardig zijn aangemerkt, maar door het model wel als onderzoekswaardig worden aangemerkt, alsnog onderzoeken en labelen. Het gaat hierbij puur om onderzoek. Er zit hier geen rechtsgevolg of andere gevolgen aan, ook niet als de handhaver iets onderzoekswaardig vindt. Daarnaast is het onderzoek puur gebaseerd op dossieronderzoek, en vindt dus geen klantcontact plaats.

Deze pre-pilot heeft betrekking op aanvragen van 26 april 2021 t/m 28 maart 2022.
Van de aanvragen die het model selecteert en die nog niet eerder door handhaving onderzocht zijn is zo'n 69% reeds afgewezen door de inkomensconsulent. Omdat we vooraf verwachten dat de kans groot is dat deze aanvragen ook door de handhavingsspecialisten zullen worden afgewezen, en om de tijd van de handhavingsspecialisten zo goed mogelijk te benutten limiteren we het aantal van dit soort aanvragen die worden meegenomen in de pre-pilot.

Bij de start van de pre-pilot ontvangen de handhavingsspecialisten elk een individueel template. Op dit template vinden zij een lijst met aanvragen, met voor elke aanvraag het administratienummer[6] en de vijf belangrijkste features waardoor de aanvraag geselecteerd is door het model. De handhavingsspecialisten voeren dossieronderzoek uit in Sherlock, Focus, BRP en Suwinet, en boeken de aanvraag af in het template op een van de afboekcodes. Indien gewenst kunnen ze een toelichting bij hun keuze typen.

3.2 Data opschonen

Nadat de resultaten door de handhavingsspecialisten zijn ingestuurd, worden ze samengevoegd in een Excel document. Aanvragen waarbij een administratienummers twee aanvraagnummers heeft worden verwijderd, omdat niet te achterhalen is naar welke aanvraag de handhavingsspecialisten hebben gekeken (m.u.v.. één aanvraag na waar dit wel bekend was). Ook aanvragen waarbij de handhavingsspecialist geen score heeft toegekend en aanvragen gedaan door een Amsterdammer behorende tot een Bijzondere Doelgroep worden verwijderd. Van de 328 aanvragen die op de oorspronkelijke templates stonden zijn hierna nog 279 aanvragen over.

Via het document *20220517 Aanvragen om te onderzoeken in prepilot* hebben we de aanvraagnummers gematcht op naam handhavingsspecialist en administratienummer. Voor administratienummers die niet in deze lijst voorkomen omdat de risicoscore in het model onder de 0.63 zakte na aanpassing van het model was geen match. Voor deze aanvragen is het aanvraagnummer gematcht op administratienummer in de lijst *20220509 Aanvragen om te onderzoeken in prepilot*.

De labels bij de aanvragen worden versimpeld. De dertien afboekcodes worden teruggebracht naar twee opties: *onderzoekswaardig* en *niet onderzoekswaardig*. De gebruikte mapping is weergegeven in Tabel 1.

Afboekcode	Onderzoekswaardigheid
Afwijzing: middelen	Onderzoekswaardig
Afwijzing: voorliggende voorziening	Onderzoekswaardig
Afwijzing: voorliggende voorziening Zelfstandigen	Onderzoekswaardig

[6] Voor een zestal aanvragen die toebehoren aan drie Amsterdammers is het ook het dienstnummer meegenomen, zodat de aanvragen individueel te onderzoeken zijn.

Afwijzing: woon/leefsituatie	Onderzoekswaardig
Geen wijziging	Niet onderzoekswaardig[7]
Geen wijziging: Rechtvaardigheid	Niet onderzoekswaardig[8]
Niet onderzoekswaardig	N/A
Onjuiste opvoer	N/A
Wijziging: middelen	Onderzoekswaardig
Wijziging: middelen (Rechtvaardigheid)	N/A
Wijziging: woon/leefsituatie	Onderzoekswaardig
Wijziging: woon/leefsituatie (Rechtvaardigheid)	Onderzoekswaardig
Aanvullend onderzoek vereist namelijk:	Onderzoekswaardig[9]

Tabel 1: Onderzoekswaardigheid mapping afboekcodes. De N/A codes zijn codes die beschikbaar waren voor de deelnemende handhavingsspecialisten, maar waarvan zij geen gebruik gemaakt hebben.

4 Analyse en herweging model

In dit hoofdstuk beschrijven we de wijze waarop de verzamelde data geanalyseerd is; de eerste resultaten van de bias analyse; en de vervolgstappen die zijn genomen op basis van deze eerste resultaten.

4.1 Analyse

De dataset die we gebruiken voor het analyseren van bias in het model bestaat uit de volgende groepen:

1. Aanvragen die door zowel het model als het huidige proces als onderzoekswaardig worden gelabeld. Omdat deze aanvragen al in het huidige proces door een handhavingsspecialist beoordeeld worden, zijn deze vragen niet meegenomen in de pre-pilot. De labels (onderzoekswaardig / niet onderzoekswaardig) zijn beschikbaar vanuit het huidige proces.
2. Aanvragen die wel door het model zijn geselecteerd, maar niet in het huidige proces. Deze aanvragen splitsen we op in twee subgroepen:
    a. Aanvragen die in het huidige proces door de Inkomensconsulent zijn afgewezen;
    b. Aanvragen die in het huidige proces door de Inkomensconsulent zijn toegewezen.
3. Aanvragen die wel zijn onderzocht in het huidige proces, maar niet door het model zijn geselecteerd. De labels (onderzoekswaardig / niet onderzoekswaardig) zijn beschikbaar vanuit het huidige proces.

De metrieken voor het model worden berekend op groepen 1 en 2, de metrieken voor het huidige proces op groepen 1 en 3.

Zoals beschreven in sectie 3.1 hebben de handhavingsspecialisten geen representatief sample ontvangen van de aanvragen die door het model als onderzoekswaardig geselecteerd zijn, omdat we meer geïnteresseerd zijn in aanvragen die nog niet zijn afgewezen. Daardoor is enig handmatig werk vereist

[7] Uitzondering zijn veertien aanvragen die allen door dezelfde handhaver beoordeeld zijn, en waarbij uit de toelichting bleek dat de aanvraag toch onderzoekswaardig is.
[8] Dit gaat slechts om één aanvraag. In het algemeen nemen we *Geen wijziging: Rechtvaardigheid* wel mee als onderzoekswaardig, maar in dit geval heeft de ICer ook zonder tussenkomst handhaving maatwerk verleent.
[9] Alle aanvragen met afboekcode *Aanvullend onderzoek vereist namelijk:* zijn individueel gecontroleerd op onderzoekswaardig d.m.v. het bekijken van de toelichting.

zodat de uitkomsten van de bias analyse representatief is. Het is nodig om de resultaten van de prepilot (groep 2 hierboven) zo te wegen dat een representatieve verhouding ontstaat tussen aanvragen die door de inkomensconsulent worden toegekend en afgewezen. Dit gebeurt in twee stappen:

1. Per gevoelig attribuut en voor iedere combinatie van toegewezen/afgewezen en bevoordeeld/benadeeld rekenen we een multiplier uit tussen het aantal aanvragen met die combinatie in de gehele dataset en het aantal aanvragen met die combinatie in de pre-pilot. Per gevoelig attribuut krijgen we dus vier multipliers.
2. Het aantal valspositieven wat in de pre-pilot is gevonden voor ieder gevoelig attribuut en combinatie van toegewezen/afgewezen en bevoordeeld/benadeeld schalen we met behulp van de berekende multipliers. Hiervoor wordt het aantal valspositieven voor die combinatie vermenigvuldigd met de betreffende multiplier.

Hieruit volgt het aantal valspositieven per afgewezen/toegewezen en bevoordeelde/benadeelde groep, als het sample voor de pre-pilot representatief zou zijn geweest voor de echte verhoudingen tussen afgewezen en toegewezen aanvragen. Door dit te combineren met de rest van de cijfers, kunnen de metrieken worden berekend:

1. Voor zowel het model als het proces en zowel de bevoordeelde als de benadeelde groepen tellen we de opgeschaalde aantallen valspositieven bij elkaar op. Voor het model zijn dit de geschaalde aantallen zoals hierboven beschreven, plus de aantallen onder de aanvragen die zowel door model als proces zijn geselecteerd (groep 1). Voor het huidige proces zijn dit de aantallen onder de aanvragen die alleen in het proces zijn onderzocht (groep 3), plus de aantallen onder de aanvragen die zowel door model als proces zijn geselecteerd (groep 1).
2. Met de totale aantallen valspositieven, geschaald waar nodig, kan uiteindelijk de *false positives / group size* uitgerekend worden. Hiervoor wordt het aantal valspositieven voor de bevoordeelde en benadeelde groep gedeeld door het totale aantal aanvragen in de betreffende groep.
3. Tot slot vergelijken we de bevoordeelde en benadeelde groep met elkaar door de metrieken van beide groepen van elkaar af te trekken om de zogenoemde *difference* te verkrijgen, of door elkaar te delen voor de *ratio*.

4.2 Eerste resultaten

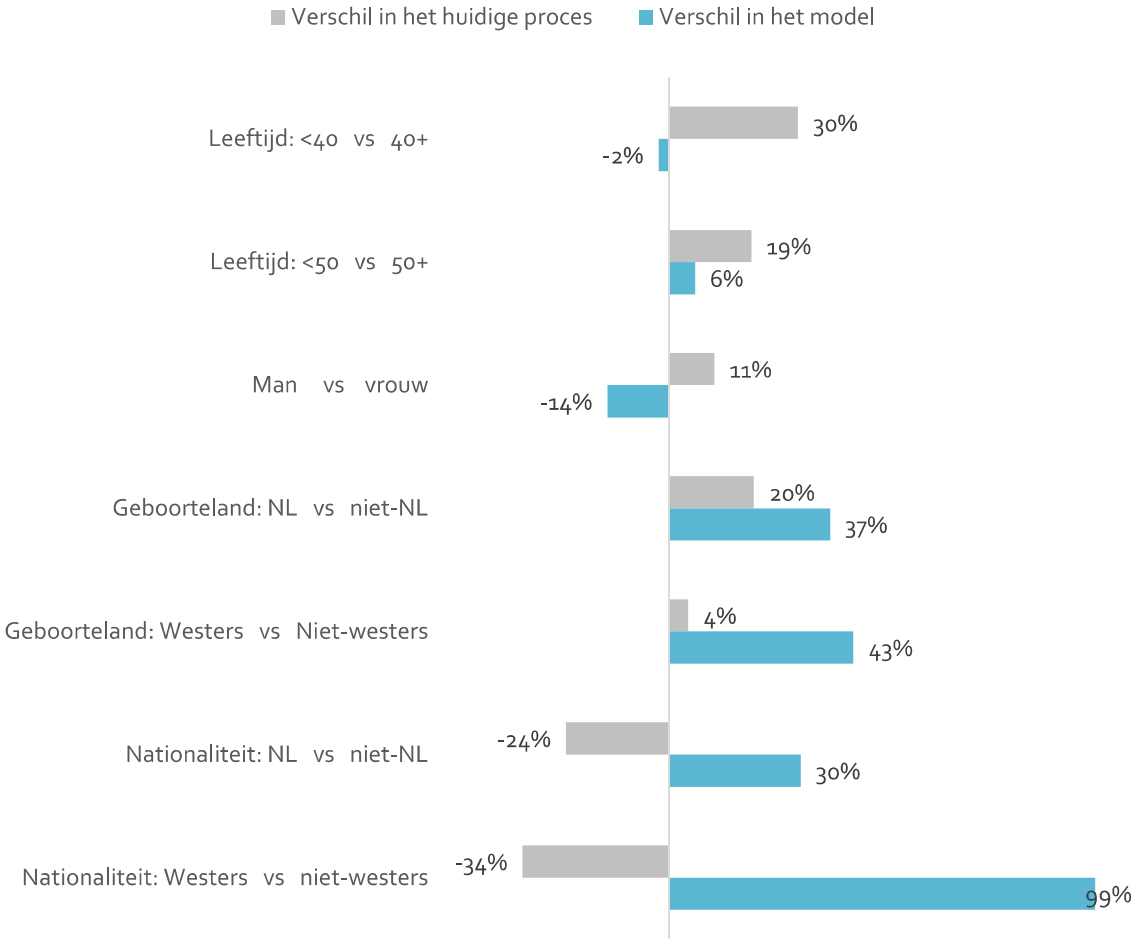
Op basis van bovenstaande analyse komen we tot de resultaten afgebeeld in Figuur 5. In deze bias analyse kijken we naar het verschil in de kans dat een persoon van de ene of de andere groep ten onrechte door het model geselecteerd wordt. Het is hierbij belangrijk om op te merken dat het verschil niet direct gelijk staat aan een bias. Als je twee groepen vergelijkt, is er vrijwel altijd een klein verschil door toeval. We hebben dit besproken met een groep specialisten die bij het ontwikkelen van het Fairness Handbook betrokken waren, van onder andere het Civic AI lab, Utrecht Data School en Eindhoven University of Technology. Tijdens deze gesprekken en naar aanleiding van aanvullend onderzoek hebben we geconcludeerd dat er op dit moment geen algemeen geaccepteerde manier is om te bepalen bij welk percentage sprake is van bias. Het vakgebied van bias is jong, en literatuur is beperkt. We kijken daarom samen met de business naar deze resultaten, en bepalen samen wanneer er volgens onze interpretatie sprake is van bias.

Allereerst kijken we naar leeftijd. In het huidige proces is de kans dat een aanvrager van 40+ jaar ten onrechte door handhaving wordt onderzocht 30% groter dan van iemand die jonger dan 40 jaar is. Het model daarentegen selecteert 2% vaker ten onrechte iemand van jonger dan 40 jaar dan iemand van ouder 40 jaar. Hierbij zorgt het model dus dat het verschil tussen de groepen afneemt, en het selectieproces dus gelijkwaardiger wordt. Bij de 50+ / 50- groep zien we een soortgelijk resultaat: hoewel zowel in het huidige proces als bij gebruik van het model de kans iets groter is dat iemand van meer dan

50 jaar oud ten onrechte geselecteerd wordt, neemt die kans af van 19% naar 6% bij gebruik van het model. Op leeftijd stellen wij vast dat er in het model geen bias zit.

De kans dat een vrouw in het huidige proces ten onrechte geselecteerd wordt is 11% groter dan bij een man. Met het model is de kans dat een man ten onrechte geselecteerd wordt juist 14% groter. Hoewel het verschil in het model iets groter is dan in het huidige proces is het nog altijd dusdanig laag dat we geen bias vaststellen.

Bij geboorteland en nationaliteit verandert dit echter. Hierbij zien we dat het model over het algemeen een stuk vaker mensen die niet in Nederland geboren zijn, of niet de Nederlandse nationaliteit hebben ten onrechte selecteert. Als we een westers en niet-westers geboorteland en een westerse of niet-westerse nationaliteit met elkaar vergelijken is dit verschil nog groter. Gezien de grootte van dit verschil en het feit dat het model hier significant slechter presteert dan het huidige proces, stellen we hier een bias vast. In 4.3 lichten we toe op welke manier we het model aanpassen om deze bias drastisch te verminderen.



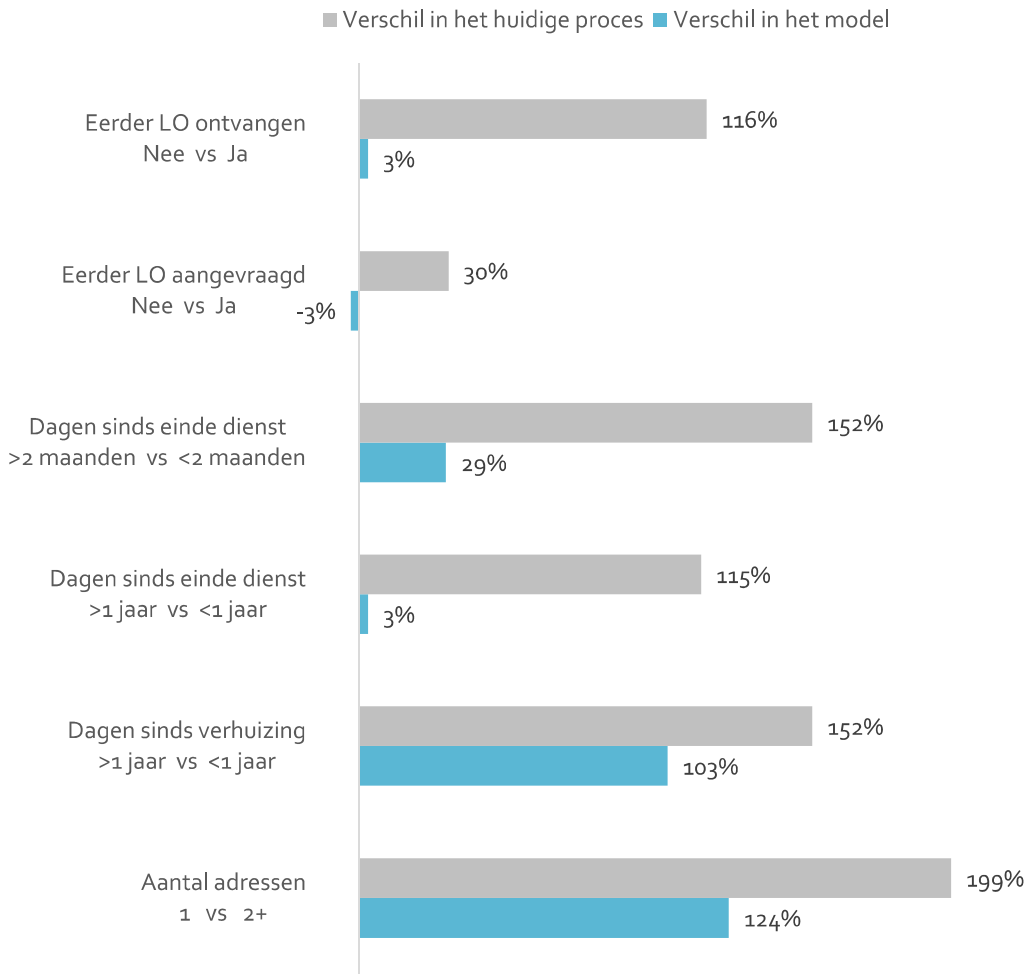
*Figuur 5: Eerste resultaten directe bias analyse. Een positief resultaat betekent dat een aanvrager uit de benadeelde groep (rechts) x% **meer** kans heeft om onterecht onderzocht te worden; een negatief resultaat betekent dat een aanvrager uit de benadeelde groep (rechts) x% **minder** kans heeft om onterecht onderzocht te worden.*

Naast directe bias bekijken we ook mogelijke indirecte bias in model. De eerste resultaten hiervan zijn weergegeven in Figuur 6. Bij de indirecte bias is de tolerantie voor een verschil wat hoger. We kijken hierbij immers naar de features waaruit het model bestaat. Het model bestaat uit vijftien features, die allen een hoge voorspellende waarde hebben. Als een feature een hoge voorspellende waarde heeft, worden aanvragen dus op basis van deze features geselecteerd. Het is dan ook logisch dat deze aanvragen die voldoen aan deze features ook vaker ten onrechte geselecteerd worden. Ook is voor bias op leeftijd,

geboorteland/nationaliteit en gender al rechtstreeks gecontroleerd in de directe bias analyse. We weten daardoor dat een eventueel verschil in de indirecte bias analyse niet veroorzaakt wordt door een van deze attributen.

We zien dat in de eerste vier features in het figuur (*Eerder LO ontvangen*; *Eerder LO aangevraagd*; *Dagen sinds einde dienst*) het verschil bij gebruik van het model kleiner is dan in het huidige proces. Daarnaast is het verschil bij deze features relatief laag, en zien we hier dus geen bias.

Dagen sinds verhuizing en aantal adressen hebben beide een zeer hoge voorspellende waarde. We weten uit gesprekken met de handhavingsspecialisten dat recente verhuizingen en het ingeschreven staan op meerdere adressen ook in het huidige proces een zeer duidelijk signaal voor mogelijke onrechtmatigheid is. Het is dan ook niet verrassend dat het verschil hier relatief groot is. Wel zien we dat het model hier beter presteert dan het huidige proces. Ook hier wordt geen bias gevonden.



*Figuur 6: Eerste resultaten indirecte bias analyse. Een positief resultaat betekent dat een aanvrager uit de benadeelde groep (rechts) xx% **meer** kans heeft om onterecht onderzocht te worden; een negatief resultaat betekent dat een aanvrager uit de benadeelde groep (rechts) xx% **minder** kans heeft om onterecht onderzocht te worden.*

4.3 Herweging model

Om de aangetroffen bias op basis van geboorteland en nationaliteit te corrigeren gebruiken we de herwegingsmethode. Bij de herwegingsmethode[10] geef je gewichten aan je trainingsdata om te corrigeren voor scheve verhoudingen in de dataset en de daaruit voortvloeiende ongewenste effecten in de resultaten.

In dit model is de voornaamste bias die tegen aanvragers met een niet-westerse nationaliteit. We passen daarom de trainingsdata aan om aanvragers met een niet-westerse nationaliteit wiens aanvraag onrechtmatig is een relatief lager gewicht te geven, en aanvragers met een niet-westers nationaliteit wiens aanvraag rechtmatig is relatief een hoger gewicht te geven. Ook geven we aanvragers met een westerse nationaliteit wiens aanvraag rechtmatig is een relatief lager gewicht, en aanvragers met een westerse nationaliteit wiens aanvraag onrechtmatig is een relatief hoger gewicht. De exacte gewichten worden op wiskundige wijze bepaald[10].

Met de nieuwe gewichten wordt het model opnieuw getraind en vervolgens wordt de bias analyse opnieuw uitgevoerd. Dit doen we enerzijds om te controleren of de interventie om bias te verminderen succesvol is geweest, en anderzijds om te kijken of niet onbedoeld andere bias in het model is geïntroduceerd door de herweging van het model. De eindresultaten van de bias analyse staan beschreven in Hoofdstuk 5. Na herweging op basis van westerse versus niet-westerse nationaliteit is ook het verschil in Nederlandse versus niet-Nederlandse nationaliteit en in geboorteland sterk afgenomen. Dit is niet verrassend, aangezien al deze attributen sterk samenhangen met het attribuut waarop herwogen is. Er is geen bias op andere attributen ontstaan. Ook zien we dat het model nog steeds even goed in staat is om onderzoekswaardige aanvragen te herkennen na deze herweging – de effectiviteit van het model lijdt dus niet onder deze herweging.

5 Resultaten

In dit hoofdstuk worden de eindresultaten van de bias analyse beschreven. Om ervoor te zorgen dat dit hoofdstuk zelfstandig gelezen kan worden zit er wat herhaling in van eerdere hoofdstukken (met name hoofdstuk 4). In deze bias analyse kijken we naar het verschil in de kans dat een persoon van de ene of de andere groep ten onrechte door het model geselecteerd wordt. Het is hierbij belangrijk om op te merken dat het verschil niet direct gelijk staat aan een bias. Als je twee groepen vergelijkt, is er vrijwel altijd een klein verschil door toeval. We hebben dit besproken met een groep specialisten die bij het ontwikkelen van het Fairness Handbook betrokken waren, van onder andere het Civic AI lab, Utrecht Data School en Eindhoven University of Technology. Tijdens deze gesprekken en naar aanleiding van aanvullend onderzoek hebben we geconcludeerd dat er op dit moment geen algemeen geaccepteerde manier is om te bepalen bij welk percentage sprake is van bias. Het vakgebied van bias is jong, en literatuur is beperkt. We kijken daarom samen met de business naar deze resultaten, en bepalen samen wanneer er volgens onze interpretatie sprake is van bias.

5.1 Directe bias analyse

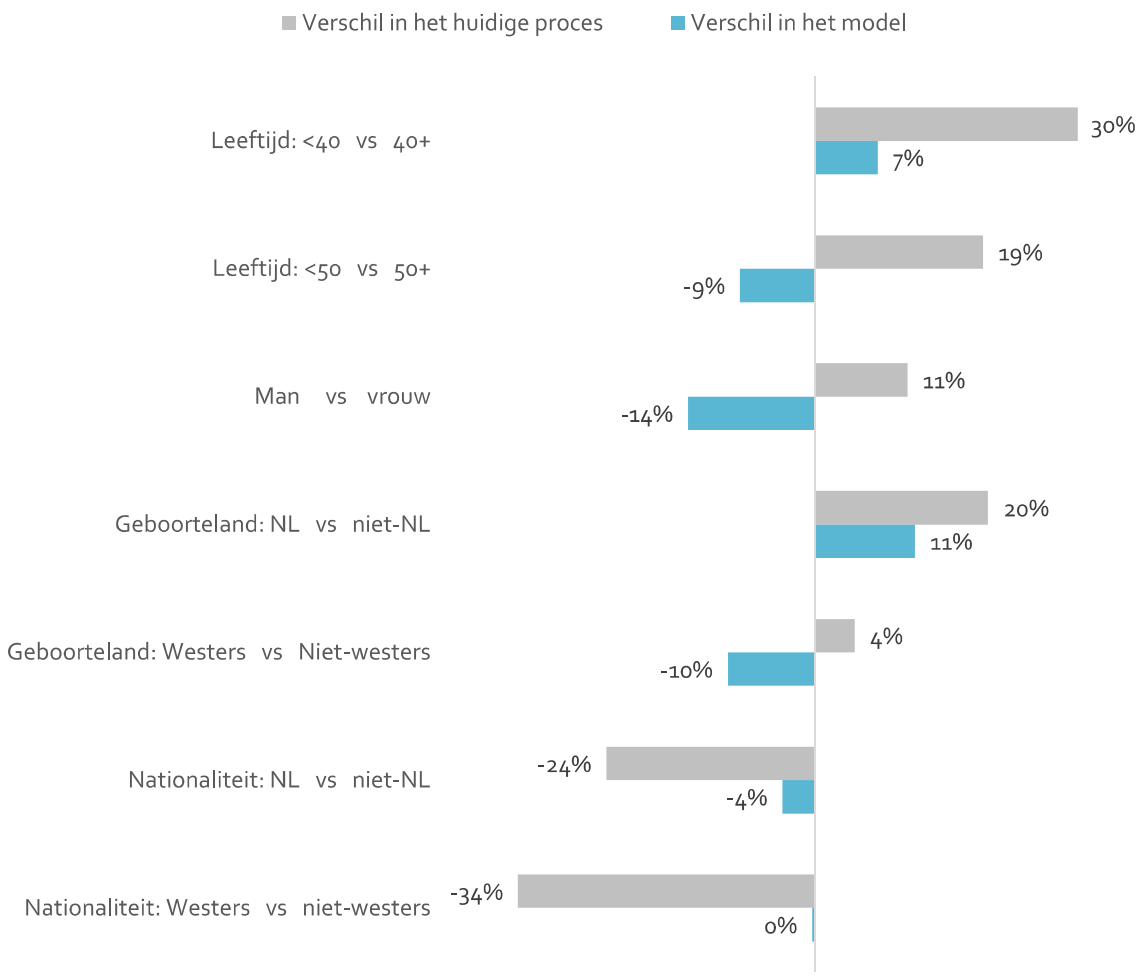
Figuur 7 geeft het eindresultaat van de directe bias analyse weer.

We kijken eerst naar eventuele bias op basis van leeftijd. In het huidige proces is de kans dat een aanvrager van 40+ jaar ten onrechte door handhaving wordt onderzocht 30% groter dan van iemand die jonger dan 40 jaar is. In het model daalt dit verschil naar 7%. Hierbij zorgt het model dus dat het verschil tussen de groepen afneemt, en het selectieproces gelijkwaardiger wordt. Daarnaast is het in het huidige proces de kans dat iemand van 50+ jaar oud ten onrechte geselecteerd wordt 19% groter dan voor een aanvrager van jonger dan 50. In het model daarentegen is de kans dat iemand van jonger dan 50 jaar ten onrechte geselecteerd is 9% groter. Op verzoek van de business controleren we achteraf ook of de groep jonger dan 30 jaar bias ervaart, en dit blijkt niet het geval te zijn. Op leeftijd stellen wij vast dat er in het model geen bias zit.

[10] Zie F. Kamiran and T. Calders, "Data Preprocessing Techniques for Classification without Discrimination," Knowledge and Information Systems, 2012.

De kans dat een vrouw in het huidige proces ten onrechte geselecteerd wordt is 11% groter dan bij een man. Met het model is de kans dat een man ten onrechte geselecteerd wordt juist 14% groter. Hoewel het verschil in het model iets groter is dan in het huidige proces is het nog altijd dusdanig laag dat we geen bias vaststellen.

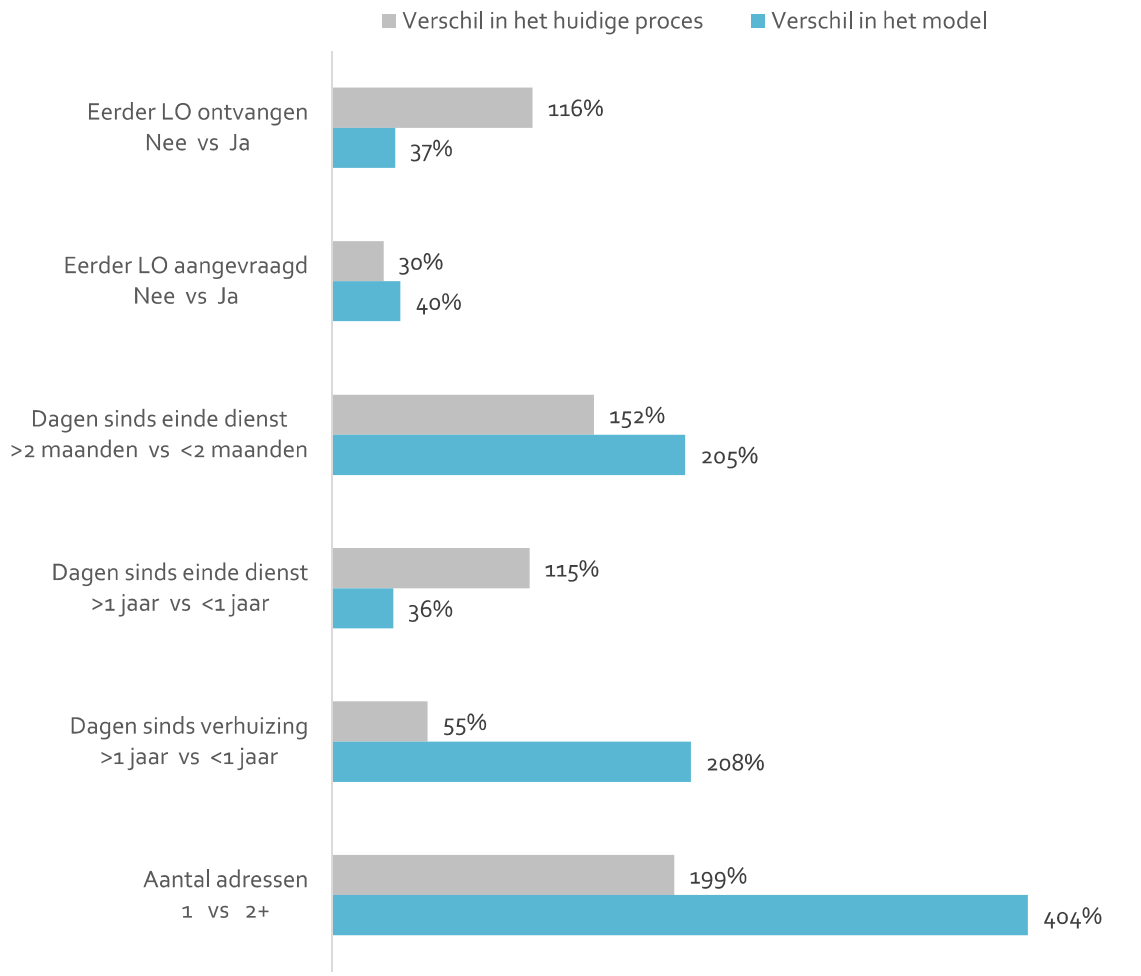
Ook op geboorteland en nationaliteit is het verschil nooit groter dan 11%. Daarnaast daalt in de meeste gevallen het verschil in het model ten opzichte van het huidige proces. Ook op nationaliteit en geboorteland stellen we geen bias vast.



*Figuur 7: Eindresultaat directe bias analyse. Een positief resultaat betekent dat een aanvrager uit de benadeelde groep (rechts) xx% **meer** kans heeft om onterecht onderzocht te worden; een negatief resultaat betekent dat een aanvrager uit de benadeelde groep (rechts) xx% **minder** kans heeft om onterecht onderzocht te worden.*

5.2 Indirecte bias analyse

Naast directe bias bekijken we ook mogelijke indirecte bias in model. De resultaten hiervan zijn weergegeven in Figuur 8. Bij de indirecte bias analyse is de tolerantie voor een verschil wat hoger. We kijken hierbij immers naar de features waaruit het model bestaat. Het model bestaat uit vijftien features, die allen een hoge voorspellende waarde hebben. Als een feature een hoge voorspellende waarde heeft, worden aanvragen met deze features vaker geselecteerd. Het is dan ook logisch dat deze aanvragen die voldoen aan deze features ook vaker ten onrechte geselecteerd worden. Ook is voor bias op leeftijd, geboorteland/nationaliteit en gender al rechtstreeks gecontroleerd in de directe bias analyse. We weten daardoor dat een eventueel verschil in de indirecte bias analyse niet veroorzaakt wordt door een van deze attributen.



*Figuur 8: Eindresultaat indirecte bias analyse. Een positief resultaat betekent dat een aanvrager uit de benadeelde groep (rechts) xx% **meer** kans heeft om onterecht onderzocht te worden; een negatief resultaat betekent dat een aanvrager uit de benadeelde groep (rechts) xx% **minder** kans heeft om onterecht onderzocht te worden.*

De eerste twee features, *eerder levensonderhoud (LO) ontvangen*; en *eerder levensonderhoud (LO) aangevraagd*, hebben een relatief laag verschil. Aanvragers die eerder LO hebben ontvangen of aangevraagd hebben respectievelijk een 37% en 40% hogere kans om door het model ten onrechte geselecteerd te worden. In het huidige proces is dit 116% en 30% respectievelijk. Omdat de tolerantie voor verschil in de indirecte bias analyse wat hoger is, stellen we hier geen bias vast.

Als een aanvrager een dienst bij de gemeente minder dan twee maanden geleden beëindigd heeft, is de kans dat hij of zij door het model ten onrechte geselecteerd wordt 205% groter. Dit is niet verrassend,

want een recente beëindigde dienst is een belangrijk signaal voor mogelijke onrechtmatigheid. In het huidige proces is deze kans 152% groter. Als we echter verder uitzoomen en kijken hoe dit verloopt een jaar nadat de laatste dienst is beëindigd, dan zien we dat dit verschil in het huidige proces zakt naar 152%, en in het model naar 36%. Het is dus niet zo dat aanvragers die in het verleden een dienst hebben afgenomen bij de gemeente hier lange tijd later nog nadelen van ondervinden bij het aanvragen van levensonderhoud. Er wordt dan ook geen bias vastgesteld.

Tot slot kijken we naar dagen sinds verhuizing en aantal adressen. Deze features hebben beiden een zeer hoge voorspellende waarde. We weten uit gesprekken met de handhavingsspecialisten dat recente verhuizingen en het ingeschreven staan op meerdere adressen ook in het huidige proces een zeer duidelijk signaal voor mogelijke onrechtmatigheid is. Het is dan ook niet verrassend dat het verschil hier relatief groot is. Daarnaast weten we dankzij de directe bias analyse dat dit verschil in ieder geval niet veroorzaakt wordt door het wel/niet hebben van een migratieachtergrond. Wel zien we dat het model hier minder scherp presteert dan het huidige proces. Echter omdat het hier gaat over de indirecte bias analyse op features met een hoge voorspellende waarde, beschouwen we dit niet als een bias.

5.3 Conclusie

Na herweging van het model is in overleg met de business geen bias aangetroffen in zowel de directe als indirecte bias analyse. Het is echter van belang om op bias te blijven controleren indien het model geïmplementeerd wordt. Meer informatie over op welke momenten de bias analyse herhaald zal worden in het beheerproces is te vinden in het beheerplan.

Annex 1

Lijst westerse landen

Argentinië	Duitsland	Indonesië	Nederlandse Antillen	Slowakije
Australië	Estland	Italië	Nieuw-Zeeland	Spanje
België	Finland	Japan	Noorwegen	Tsjechië
Bondsrepubliek Duitsland	Frankrijk	Kroatië	Oostenrijk	Tsjecho-Slowakije
Bulgarije	Griekenland	Letland	Polen	USA
Canada	Groot-Brittannië	Litouwen	Portugal	Zweden
Canarische Eilanden	Hongarije	Luxemburg	Roemenië	Zwitserland
Cyprus				
Denemarken	Ijsland	Nederland	Slovenië	

Lijst niet-westerse landen

Afghanistan	Djibouti	Italiaans-Somaliland	Mozambique	Somalië
Albanië	Dominicaanse Republiek	Ivoorkust	Myanmar	Sovjet-Unie
Algerije	Ecuador	Jamaica	Namibië	Sri Lanka
Angola	Egypte	Jemen	Nieuw-Guinea	Suriname
Armenië	El Salvador	Joegoslavië	Nepal	Syrië
Aruba	Equatoriaal-Guinea	Jordanië	Nicaragua	Taiwan
Azerbeidzjan	Eritrea	Kaapverdië	Niger	Tanzania
Bangladesh	Ethiopië	Kameroen	Nigeria	Thailand
Belarus	Filipijnen	Kazachstan	Noord-Jemen	Tibet
Bhutan	Frans West-Afrika	Kenya	Noord-Korea	Togo
Bolivia	Frans-Guyana	Kirgizië	Oekraïne	Trinidad en Tobago
Bosnië-Herzegovina	Gambia	Koeweit	Oman	Tunesië
Brazilië	Georgië	Korea	Opper-Volta	Turkije
Brits-Guyana	Ghana	Libanon	Pakistan	Uganda
Burkina Faso	Goudkust	Liberia	Panama	Uruguay
Burma	Guadeloupe	Libië	Peru	Venezuela

Burundi	Guatemala	Macedonië	Puerto Rico	Verenigde Arabische Emiraten
Ceylon	Guinee	Malakka	Qatar	Verenigde Arabische Republiek
Chili	Guinee-Bissau	Malawi	Rhodesië	Vietnam
China	Guyana	Maleisië	Rusland	Zambia
Colombia	Haïti	Mali	Rwanda	Zimbabwe
Congo	Honduras	Marokko	Réunion	Zuid-Afrika
Congo-Kinshasa	Hongkong	Mauritanië	Saoedi-Arabië	Zuid-Jemen
Costa Rica	India	Mauritius	Senegal	Zuid-Korea
Cuba	Irak	Mexico	Sierra Leone	
Dahomey	Iran	Moldavië	Singapore	
DRC	Israël	Mongolië	Soedan	

Reden features niet geselecteerd voor indirecte bias analyse

Feature	Feature is een expliciete beleidsregel voor afwijzing	Moeilijk om direct verband te leggen met gevoelig attribuut	Beperkte data beschikbaar
Aantal afspraken geen contact		X	X
Aantal afspraken no show		X	X
Adres in Amsterdam	X		
Gemiddelde percentage maatregel		X	X
Inkomen onbekend		X	
Medebewoner		X	
Partner			X
Percentage deelnames gestart		X	
Totaal bruto inkomen	X		
Totaal vermogen	X		

Feature is een expliciete beleidsregel voor afwijzing
De reden om dit soort features niet te analyseren is dat ze expliciet deel uitmaken van het beleid, wat legitimeert dat ernaar gekeken wordt.

Moeilijk om direct verband te leggen met gevoelig attribuut

De reden om dit soort features niet te analyseren is dat, als er een verschil gevonden wordt op de feature, het zo goed als onmogelijk is om dit te koppelen aan een gevoelig attribuut. Daardoor kan niet bepaald worden of er iets aan het model aangepast moet worden of niet.

Beperkte data beschikbaar

De reden om dit soort features niet te analyseren is dat voor een betrouwbare analyse de groepen die met elkaar worden vergeleken voldoende groot moeten zijn, anders worden er mogelijk conclusies getrokken op basis van te weinig, toevallige data.