Technische documentatie voorspelmodel bijstandsfraude

GEMEENTE NISSEWAARD

TOTTA DATA LAB

CONNECTING THE DOTS

Inhoudsopgave

1 Introductie

Totta data lab heeft een algoritme ontwikkeld voor de gemeente Nissewaard. Dit algoritme brengt patronen in data in kaart die duiden op een verhoogd risico op onrechtmatigheden. Op basis van deze patronen voorspelt het algoritme voor elke cliënt de kans op onrechtmatig gedrag. In dit document is uitgelegd welke data gebruikt wordt en hoe dit algoritme tot stand is gekomen.

1.1 Data

Alle databronnen die verwerkt zijn in het model, zijn in overleg en na goedkeuring van de gemeente Nissewaard gebruikt. Van huidige en voormalige bijstandsgerechtigden in de gemeente Nissewaard zijn de uitkeringsgegevens en de achtergrondkenmerken samengevoegd met gegevens omtrent:

- ☐ kinderen;
- ☐ vakantieperioden;
- ☐ rechten, plichten en ontheffingen;
- ☐ maatregelen;
- ☐ contactmomenten;
- ☐ samenloopsignalen;
- ☐ participatietrajecten;
- ☐ bijzondere situaties;
- ☐ Liaan: fraude onderzoeken;

Aan deze set gegevens is een label toegevoegd dat aangeeft of een bijstandsgerechtigde in het verleden onrechtmatigheden heeft gepleegd. Dit label is toegekend wanneer bijvoorbeeld inkomsten, vermogen of samenwonen verzwegen zijn. Deze onrechtmatigheden hebben niet altijd geleid tot het stopzetten van de bijstandsuitkering, maar soms ook tot het verlagen van de uitkering of tot een andere wijziging in de uitkering.

Voor het vinden van een robuust algoritme is deze data set willekeurig verdeeld in een training dataset (75% van de data) en een test dataset (25%). De training dataset is gebruikt om aan de hand van verschillende modeleringstechnieken patronen te herkennen in de data. De test dataset is gebruikt om te valideren of de gevonden patronen robuust genoeg zijn om de data te kunnen beschrijven. Met robuust bedoelen we dat de voorspellingen van het model weer net zo goed zijn wanneer het model voorspellingen doet op een nieuwe dataset. Wanneer het model in onze test 50% fraude correct voorspelt, maar bij een nieuwe dataset maar 20% correct voorspelt, dan is het geen robuust model. Bij het valideren van de verschillende modeleringstechnieken is de kwaliteit van een voorspelling vergeleken met een model dat een willekeurige voorspelling maakt. Daarnaast is er gekeken naar de 10 gevallen met de hoogste voorspelde kans op onrechtmatigheid en vervolgens is vergeleken hoeveel gevallen daarvan daadwerkelijk onrechtmatig gedrag vertonen. Een combinatie van de geteste modelleertechnieken resulteert in een eindmodel. De technieken dragen niet allemaal evenveel bij aan het eindmodel, maar de bijdrage van de technieken is afhankelijk van de prestaties gedurende de tests. Des te meer fraude gevallen in de top 10 correct zijn voorspelt in de tests, des te meer de techniek zal meewegen in het eindmodel. Met het uiteindelijke model zijn vervolgens voorspellingen gemaakt op de voorspel dataset. In de voorspel dataset zitten alleen personen die op dat huidige moment in de uitkering zaten.

1.2 Algoritme

Om tot een algoritme te komen dat robuuste voorspellingen maakt, is de kwaliteit van de volgende modeleringstechnieken getest: Random forest, Neurale netwerken, Lasso modeling en MARSplines. Van deze 4 technieken bleek een combinatie van de Random forest en MARSplines het beste om de data te beschrijven. Er is voor en combinatie van twee modellen gekozen: Random forest en MARSplines. Dit is om ervoor te zorgen dat het algoritme op zowel de huidig bekende gegevens als de toekomstige gegevens van nieuwe bijstandsgerechtigden een robuuste voorspelling kan maken. De Neurale netwerken en lasso modellen voorspelden niet beter dan wanner je random zou aanwijzen wie wel of niet fraude pleegt. Om die reden zijn de neurale netwerken en lasso modellen in dit geval niet van toegevoegde waarde. De MARSplines techniek en het Random forest daarentegen, voorspelde beter dan random. Om die reden is er voor een combinatie van deze twee modellen gekozen. Dit betekent dat 60% van de voorspellingen gebaseerd is op het Random forest en 40% van de voorspellingen gebaseerd is op MARSplines. Door twee technieken te combineren, is het risico weggenomen dat één model alleen goed kan voorspellen tijdens het testen van de modellen, maar in de toekomst niet goed presteert.

Daarnaast is het algoritme verbeterd door personen waarbij een onderzoek naar onrechtmatigheid heeft geleid tot een besparing ook een onrechtmatigheidslabel toe te kennen. Uit een test waarbij de modelperformance is vergeleken van een model met en zonder het meenemen van de besparing als onrechtmatigheid, bleek het model met besparing beter te presteren dan het model zonder besparing. In de test voorspelde het model mèt besparing gemiddeld 6 van de 10 hoogste fraude kansen correct en het model zonder besparing voorspelde er gemiddeld 5 van de 10 correct.

Algoritmes zoals bovenstaande worden vaak gebruikt voor het behalen van een nauwkeurige voorspelling, maar vertellen de gebruiker (te) weinig over het waarom. Algoritmes zijn weinig transparant over het waarom van bijvoorbeeld een risicoscore. Traditionele statische methoden geven daarentegen meer inzicht in het waarom, maar zijn minder geschikt om een nauwkeurige voorspelling te doen en tevens introduceren dit type modellen veel bias gericht op variabelen met een hoge correlatie naar de uitkomst. Uiteraard willen we een goede voorspelkracht behouden en bias zoveel mogelijk mitigeren. Totta data lab en Nissewaard hebben daarom in 2020 een nieuwe aanpak geïntroduceerd, welke transparantie combineert met het behouden van voorspelkracht en zonder extra bias. Hiervoor wordt gebruik gemaakt van 'Open A.I.'. Naast een risicoscore levert deze verzameling aan nieuwe modelleertechnieken namelijk ook de mogelijkheid om op individuele niveau terug te geven welke (combinatie van) variabelen het meest hebben bijgedragen aan de hoogte van de risicoscore. De methode die de ontwikkelaars van Totta data lab hiervoor inzetten heet Rulefit.

2 Uitleg van het project

Totta data lab heeft voor de gemeente Nissewaard een algoritme ontwikkeld dat de kans op fraude in de bijstand voorspelt voor elke bijstandsgerechtigde. Dit algoritme brengt patronen in data in kaart die duiden op een verhoogd risico op fraude. Het algoritme genereert een kansberekening op onrechtmatigheden op cliëntniveau. Op basis van deze kansberekeningen voert de gemeente onrechtmatigheidsonderzoeken uit. Tot op heden blijkt deze aanpak succesvol, aangezien er in 2018 en 2019 meerdere onrechtmatigheden zijn gevonden.

In dit document wordt uitgelegd hoe het algoritme werkt. Eerst zal er gekeken worden naar de data die gebruikt is voor het ontwikkelen van een voorspellend model. Vervolgens

zal worden uitgelegd hoe deze data verwerkt en geprepareerd is. Tenslotte laten we zien hoe is onderzocht welke techniek de data het beste kan beschrijven zodat uiteindelijk een robuuste voorspelling gemaakt kan worden.

2.1 Definitie van een algoritme

Voordat uitgelegd gaat worden hoe het algoritme van Nissewaard werkt, bespreken we eerst de definitie van een algoritme. Een algoritme is een eindige reeks handelingen die vanuit een begintoestand naar een beoogd doel leidt. Een algoritme is vergelijkbaar met een recept waarbij eieren, meel en melk de begintoestand zijn. Na een reeks van handelingen worden deze ingrediënten omgevormd tot een pannenkoek. Voor het algoritme van Nissewaard zijn de gegevens van bijstandsgerechtigden gebruikt als begintoestand. Deze gegevens zijn na verschillende handelingen, gebruik makende van modeleertechnieken, omgevormd tot een kans op onrechtmatigheden bij bijstandsgerechtigden.

3 De data

Voor het voorspellen van onrechtmatigheden bij bijstandsgerechtigden binnen de gemeente Nissewaard zijn verschillende databronnen gebruikt. In dit hoofdstuk wordt uitgelegd welke gegevens nodig zijn om een voorspelling te maken. Alle databronnen verwerkt in het model, zijn in overleg en na goedkeuring van de gemeente Nissewaard gebruikt.

3.1 Brondata

De data die wij nodig hebben bestaat uit gegevens omtrent:

1. Uitkeringen en cliëntgegevens
2. Kinderen
3. SHTABHIS kinderen
4. Vakantieperioden
5. Rechten, plichten en ontheffingen
6. Participatietrajecten
7. Contacten
8. Bijzondere situaties
9. Maatregelen
10. Samenloopsignalen
11. Liaan: gegevens omtrent onrechtmatigheidsonderzoeken

Niet alle aangeleverde gegevens zijn uiteindelijk meegenomen in het model. Dit komt omdat de we alleen een goed werkend model kunnen ontwikkelen op data met voldoende kwaliteit van voldoende kwantiteit.

3.2 Bevooroordeeld model

Een bevooroordeeld model ontstaat wanneer het model gestuurd wordt door alleen vroegere controles. Hierdoor zouden personen die al eerder gecontroleerd zijn, weer opnieuw gecontroleerd worden. Om dit te voorkomen, nemen we de resultaten van de nieuwe fraude controles mee in het model. Hierdoor leert het model nog beter welke patronen wel een indicatie van fraude zijn en welke niet. Naast het meenemen van de nieuwe controle resultaten elk kwartaal, voorkomen we de bias ook met combinaties van modellen. Verschillende modelleertechnieken zoals Random forest of neurale netwerken detecteren fraude patronen op een andere manier. In deze technieken wordt fraude op een andere manier getriggerd. Door de voorspellingen van meerdere technieken te combineren, zorgen we verschillende fraude patronen in kaart hebben. Hiermee

voorkomen we ook deels dat dezelfde personen opnieuw naar boven komen met de hoogste kans op fraude. De keuze voor de combinatie van modellen is sinds 2020 de Rulefit methode, omdat deze techniek transparantie over de uitkomst mogelijk maakt. Als laatste, ontwikkelen we alleen modellen op basis van data die van voldoende kwaliteit en kwantiteit is. Wat dit precies inhoudt lees je in het volgende hoofdstuk.

3.3 Data kwaliteit & kwantiteit

Wanneer de data niet van voldoende kwaliteit en kwantiteit is, zal het model niet goed presteren. Je krijgt dan vergelijkbare voorspellingen met het willekeurig aanwijzen wie fraude pleegt.

3.3.1 Vullingsgraad

De data moet voldoende gevuld zijn. Wanneer variabelen voor 80% missende waarden hebben, dan is dit niet werkbaar. Dit schetst geen goed beeld van de werkelijke situatie want je weet maar van 20% van de personen wat de situatie is. We hanteren voor dit algoritme een minimale vulling van 50%. Dit baseren we op de ervaringen die we hebben op gedaan bij het ontwikkelen van eerdere voorspelmodellen.

3.3.2 Variatie

De data moet voldoende variatie laten zien. Wanneer een variabele alleen gevuld is met dezelfde categorie, dan is er geen variatie. Stel dat iedereen in je dataset een man is, dan kan het algoritme geen onderscheid maken op basis van geslacht. Omdat geslacht nooit verschilt, is dit geen bruikbare variabele voor het algoritme. Daarnaast mag een variabele ook niet te veel variatie hebben. Stel dat de variabele 'reden van bijstandsbeëindiging' voor iedereen anders zou zijn. Dan zou het algoritme veel te weinig voorbeelden hebben van elke situatie om een duidelijk patroon te herkennen. Er moeten dus genoeg voorbeelden zijn van elke categorie om van te leren. Om die reden gebruiken wij alleen categorische variabelen waarbij er in elke categorie minimaal ongeveer 5% van de bijstandscliënten voorkomt.

3.3.3 Betrouwbaarheid

De data moet voldoende betrouwbaar zijn om mee te kunnen werken. Zo is het belangrijk dat wijzigingen in gegevens consequent worden geüpdatet. Wanneer de gegevens niet consequent zijn geüpdatet, kun je er niet van op aan dat de status van een cliënt op dat moment klopt. Wanneer de status niet klopt, kun je geen goed beeld schetsen van de werkelijke situatie. Dit is wel nodig om een betrouwbare voorspelling te kunnen maken. Naast het consequent updaten van de gegevens is het ook belangrijk dat iedereen op dezelfde manier categorieën van variabelen vult. Wanneer niet iedereen het gedrag van een klant hetzelfde beoordeelt, wordt er ook geen betrouwbaar beeld geschetst.

3.3.4 Paaseieren

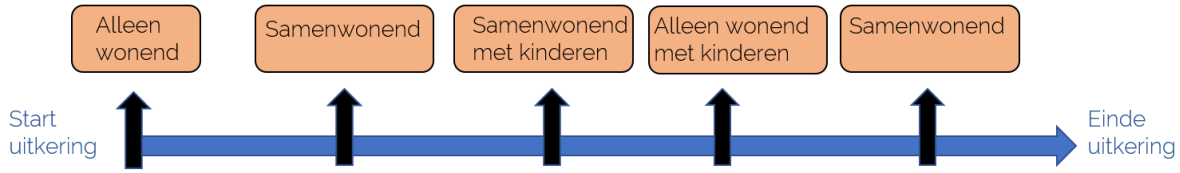
Het kan gebeuren dat er in variabelen indicaties worden gegeven of iemand gecontroleerd is op fraude of dat iemand een onrechtmatigheid heeft begaan. In principe staan deze gegevens in de fraude controle tabel of in een debiteuren tabel, maar soms kunnen ze ook in andere tabellen voorkomen.. Dit soort variabelen nemen we nooit mee in het model. Als je dit namelijk wel doet, dan vertel je eigenlijk van tevoren aan het model dat iemand onrechtmatig gedrag heeft gepleegd. Dat is dan geen voorspelling meer, want je weet al van tevoren dat er sprake is van onrechtmatigheid. Soms is het niet meteen duidelijk dat een variabele onrechtmatigheid indiceert. We nemen zo'n variabele dan wel mee in het model, maar komen er dan tijdens het testen van het model achter dat het model alle fraudegevallen correct voorspelt. Als dat gebeurd, dan weten we dat het mode van tevoren al wist wie er fraude heeft gepleegd en dan weten we dat er een variabele is

die deze informatie aan het model geeft. Deze variabele noemen we dan een paasei. We moeten dan namelijk gaan zoeken, welke variabele precies de fraude informatie doorgeeft aan het model.

3.3.5 Historische data

Wanneer er geen veranderingen van gedrag zijn geregistreerd over de tijdlijn (Figuur 1) van de uitkering is het moeilijk om individuele gedragspatronen te herkennen waarop onrechtmatigheden voorspelt kunnen worden. Stel dat iedereen hetzelfde gedrag zou vertonen aan het begin van de uitkering en aan het einde van de uitkering, dan zouden er niet zoveel gegevens zijn om op te bepalen waarom de een wel onrechtmatigheden pleegt en de andere niet. Wanneer gedrag verandert dan is het mogelijk om een bepaald omslagpunt gedurende de uitkering aan te wijzen dat een indicatie voor onrechtmatigheid zou kunnen zijn.

Figuur 1 - Voorbeeld van historische data:



3.3.6 Databronnen en koppelsleutels

Om met de data te kunnen werken moeten er koppelsleutels aanwezig zijn waarop de verschillende databronnen aan elkaar gekoppeld kunnen worden. Ook voor de koppelsleutels geldt dat ze voldoende gevuld moeten zijn. Soms kan het zijn dat er een koppelsleutel ontbreekt, om dan de data te kunnen koppelen, moet het mogelijk zijn om zelf een koppelsleutel te creëren. Wanneer koppelsleutels ontbreken, kan het gedrag wat is vastgelegd, niet gekoppeld worden aan de cliënt waar het om draait. Wanneer dit voor te veel databronnen het geval is, kan er geen individueel voorspelmodel gemaakt worden.

3.3.7 Mutatiegegevens

Mutatiegegevens moeten te herleiden zijn. Stel dat er in de tabel kinderen alleen de huidige status vermeldt staat, dan moet deze aangepast worden zodat er ook historische gegevens gebruikt kunnen worden. Hiervoor kan dan een mutatietabel voor kinderen bestaan waarin alle wijzigingen per cliënt zijn vastgelegd. Om te kunnen herleiden welke status bij welke cliënt hoort en bij welk kind, is er een mutatiecode nodig. Wanneer er in zo'n geval geen mutatiecode aanwezig is, heb je niks aan de gegevens omdat je dan geen historische gegevens hebt. Je kunt dan alleen de cliënten met elkaar vergelijken op het huidige moment. Dit geeft geen duidelijk beeld van de werkelijkheid. Om een goed model te maken, moet de hele tijdlijn van een uitkering in kaart zijn gebracht

3.3.8 Aantal bruikbare variabelen

Er moeten voldoende bruikbare variabelen aanwezig zijn om een voorspelmodel te kunnen maken. Stel dat er maar 3 variabelen gevuld zijn, met waar weinig variatie, dan is het onwaarschijnlijk dat daar een goede voorspelling uitkomt.

3.3.9 Aantal unieke cliëntnummers

Er zijn voldoende unieke cliëntnummers nodig om een goede voorspelling te kunnen maken. Zo kun je met maar 100 of 1000 cliëntnummers niet alle patronen terugvinden en

minder goed generaliseren. Uit ervaring hanteren we ongeveer een minimum van 3000 cliëntnummers voor een robuust model.

3.3.10 Aantal gevallen van onrechtmatigheid

Er zijn voldoende gevallen nodig waarbij onrechtmatigheid is vastgesteld zodat het algoritme genoeg voorbeeld heeft om op te leren. Voor het voorspellen van onrechtmatigheid in de bijstand hanteren wij een minimum van 50 onrechtmatigheidsgevallen.

3.3.11 Privacy

De data moet gepseudonimiseerd zijn. Met pseudonimiseren worden persoonsgegevens getransformeerd in een dataset die niet meer direct herleidbaar is tot een persoon. De burgerservicenummers van personen mag nooit gebruikt worden. Voor deze cliëntnummers zal de kans op onrechtmatigheden berekend worden. Wanneer gevoelige gegevens toch zijn aangeleverd, worden ze direct verwijderd. De gemeente wordt op de hoogte gesteld en een datalek wordt geregistreerd.

3.4 De gebruikte data

In de onderstaande tabel is elke kolom als databron weergegeven met in elke rij de variabelen die in deze bron voorkomen. Alle blauw gekleurde cellen geven de variabelen weer die uiteindelijk in het model zijn gebruikt. De andere cellen zijn variabelen die niet zijn meegenomen omdat ze niet voldoende kwaliteit met voldoende kwantiteit hebben. Daarnaast kunnen bepaalde gegevens ook niet zijn meegenomen omdat ze niet relevant zijn voor het voorspellen van onrechtmatigheid in de bijstand,

Uitkeringen en cliëntgegevens	Kinderen	SHTABHIS kinderen	Vakantieperioden	Rechten, plichten en ontheffingen
Dossiernummer	Cliëntnummer	Cliëntnummer	Cliëntnummer	Cliëntnummer
Cliëntnummer	Volgnummer	Entiteit	Ingangsdatum	Code recht / plicht
Code regeling	Geboortedatum	Gegeven	Einddatum	Omschrijving recht / plicht
Regeling	Ten laste komend	Key 1	Indicatie cliënt / partner / beiden	Begindatum
Code groep	Uitwonend	Key 2	Toelichting	Einddatum
Groep	Inwonend	Key 3		Code reden einde ontheffing
Datum registratie	Overlijdensdatum kind	Key 4		Omschrijving code reden einde ontheffing
Soort uitkering		Datum mutatie		
Omschrijving soort uitkering		Oude waarde		
Startdatum periodiek algemeen		Nieuwe waarde		
Einddatum periodiek algemeen		Aanvraagnummer		
Indicatie aard bijstand		Gemeentecode		
Aard bijstand				
Leefvorm				
Omschrijving leefvorm				
Cliëntnummer partner				
Gemeentecode				
Cliënttype				
Omschrijving cliënttype				
Oorzaak bijstand				
Omschrijving oorzaak bijstan				
Reden beëindiging bijstand				
Omschrijving reden beëindiging bijstand				
Geboortedatum				
Geslacht				
Aantal unieke cliëntnummers				

Tabel 1 - Variabelen per tabel:

Participatietrajecten	Contacten	Bijzondere situaties	Maatregelen	Samenloopsignalen	Liaan
Cliëntnummer	Cliëntnummer	Cliëntnummer	cliëntnummer	Cliëntnummer	Cliëntnummer
Trajectnummer	Nummer van het contact	Code bijzondere situaties	Dossiernummer	Dossiernummer	ID zaak
Code soort traject	Status van het contact	Omschrijving bijzondere situaties	Soort	SSID	ID persoon
Soort traject	Datum registratie	Volgnummer	Aanvraagnummer	Wijziging datum signaal	Zaaknummer
Begindatum	Afhandelingsdatum	Ingangsdatum	Code regeling	Gemeentecode GBA	Ingang
Einddatum	Nummer van de externe instantie	Vervaldatum	Categorie	Bron	Uit
Type traject	Nummer van de informatiesoort		Omschrijving	Samenloopcode	Gemeente
Omschrijving type traject	Omschrijving van de informatiesoort		Datum besluit	Aanvangsdatum	Naam
Gemeentecode	Aktiecode		Indicatie recidive	Einddatum	Datink
Gemeente	Nummer van de informatierepliek		Datum registratie	Creatie datum signaal	Datuit
Dossiernummer	Omschrijving 1		Begindatum periode van het gedrag dat tot een maatregel geleid heeft	Significante wijzigingsdatum	Fraude
Datum akkoord	Code wijze binnenkomst		Einddatum periode van het gedrag dat tot een maatregel geleid heeft		Reden onderzoek
Code status	Code regeling				Resultaat
Status	Regeling				
Datum status					ID conclusie
Indicatie afzien deelname					Conclusie
Code reden beëindiging					ID
Reden beëindiging					Nummer
					Woonplaats
					Geboortedatum
					Geslacht

4 Data verwerken

4.1 Definitie van onrechtmatigheden

Aan de hand van onrechtmatigheden uit het verleden leert het model herkennen wat mogelijke onrechtmatigheden in de toekomst zijn. Van tevoren is vastgesteld wanneer iemand onrechtmatigheden heeft gepleegd. Deze onrechtmatigheden zijn meestal vastgesteld na onderzoek van deze uitkeringsgerechtigde. Wanneer blijkt dat er bij een uitkeringsgerechtigde sprake is van onjuiste adresopgave of wanneer inkomsten, vermogen of samenwonen worden verzwegen, is er onrechtmatig gedrag vastgesteld. De uitkering van de uitkeringsgerechtigde wordt na zo'n incident stopgezet, aangepast, voortgezet of blijft ongewijzigd.

Voor de gemeente Nissewaard is er nog een extra onderdeel aan de definitie van onrechtmatigheid toegevoegd. Uit tests blijkt dat het model hierdoor een hoger percentage correct voorspelt uit de top 10 cliënten met de hoogste kans op onrechtmatig gedrag. De toevoeging houdt in dat een onrechtmatigheidsonderzoek dat tot een besparing leidt ook is opgenomen als onrechtmatig gedrag in het algoritme.

4.2 Data preparatie en opschoning

Voordat een model getraind kan worden, moet de data geprepareerd zijn. Het preparen en opschonen van de data verloopt via de volgende stappen:

1. Brontabellen werkbaar maken;
2. Training en voorspel datasets maken en de peilmoment selectie;
3. Onrechtmatigheden definiëren in de data;
4. Basis dataframe en voorspel niveau;
5. Samenvoegen tabellen tot 1 dataset en creëren van variabelen;
6. Incomplete, dubbele of missende data corrigeren;

4.2.1 Brontabellen werkbaar maken

In deze stap laden we de data in en vertalen we de bron tabellen naar tabellen waarmee het model kan werken. Dit betekent dat bijvoorbeeld namen worden veranderd van tabellen en variabelen zodat er geen vreemde tekens meer in voorkomen, De modelleer technieken kunnen namelijk niet omgaan met vreemde tekens.

4.2.2 Training en voorspel datasets maken en de peilmoment selectie

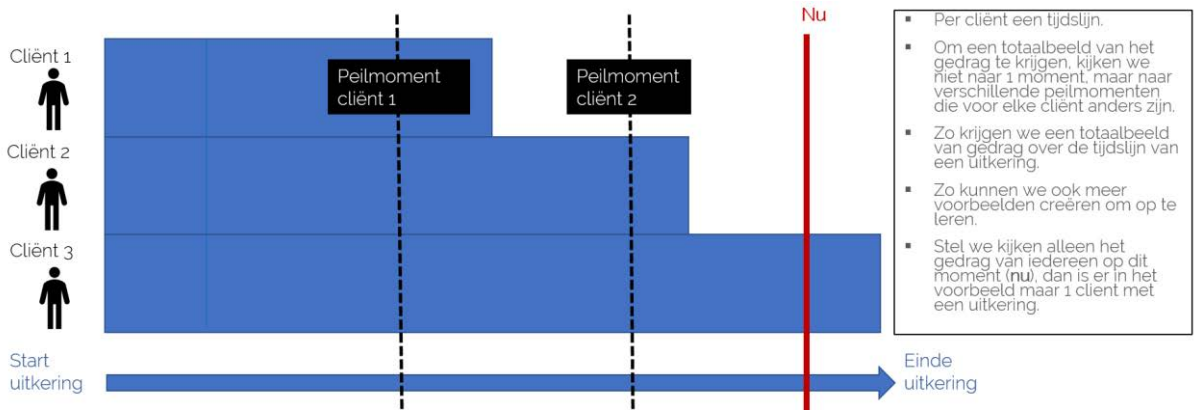
Om het model te leren welke patronen over het algemeen voorkomen bij onrechtmatigheden en welke niet, gaan we het model trainen. Het trainen van het model betekent eigenlijk dat we het model voorbeelden geven van onrechtmatig gedrag en dat het model gaat proberen om deze voorbeelden zelf in de data te vinden. Daarna testen we hoe goed het model instaat is om andere gevallen van fraude te vinden, die het model nog niet eerder heeft gezien. Wanneer we het model in de praktijk gaan gebruiken, willen we namelijk ook dat het model in staat is om nieuwe fraude gevallen te vinden.

Om dit te kunnen doen is de data opgesplitst in een dataset om op te trainen en een dataset om op te testen, de training dataset, en een dataset waarop de uiteindelijke voorspellingen zijn gedaan, de voorspel dataset. Eerst is train dataset gebruikt om het model op te trainen en op te testen totdat het uiteindelijke model is ontwikkeld. Daarna is de voorspel dataset gebruikt om de uiteindelijke voorspellingen op te doen. In deze dataset is per cliënt een willekeurig moment binnen de uitkering gekozen waarop naar het gedrag van deze cliënt is gekeken. Dit moment is het peilmoment (Figuur 2)**.** Alle informatie uit andere bronnen voor elke cliënt is ook geselecteerd op dit peilmoment, dus informatie

na het peilmoment wordt niet meegenomen. Door op deze willekeurige peilmomenten naar het gedrag van de cliënten te kijken, is ervoor gezorgd dat je op elk moment in de uitkering een goede voorspelling kan maken. Je zorgt namelijk dat op verschillende punten over de tijdslijn van de uitkering, gedrag in kaart is gebracht. Ook zorg je dat je de informatie van alle personen in de uitkering kan gebruiken. Stel dat je alleen maar naar het gedrag zou kijken op 2018-01-01, dan mis je een heleboel uitkeringen die al zijn gestopt voor 2018-01-01 of die nog moeten beginnen. Je mist dus een heleboel voorbeelden waarop het model kan leren. Dus door met een peilmoment te werken, kun je de informatie van iedereen meenemen en op verschillende momenten over de tijdslijn van de uitkering. Op die manier krijgt het model zo'n realistisch mogelijk beeld van situatie en heeft het zoveel mogelijk situaties om op te leren.

Binnen de voorspel dataset is het peilmoment de huidige datum. Het is namelijk de bedoeling om op het huidige moment een voorspelling te doen om onrechtmatig gedrag op te sporen. Je hoeft dan ook alleen maar voor de mensen die op dit moment in de uitkering zitten een voorspelling te doen. Om die reden kun je wel de huidige datum als peilmoment gebruiken. Voor de uiteindelijk voorspelling gebruik je dan een model dat heeft geleerd op alle voorbeelden van alle uitkeringen uit de training dataset. Dat model gaat dan met alles wat het heeft geleerd opzoek naar onrechtmatig gedrag in de voorspel dataset, de mensen die op dit moment een uitkering hebben. Het model geeft dan alle personen uit de voorspel dataset een kans op onrechtmatigheid.

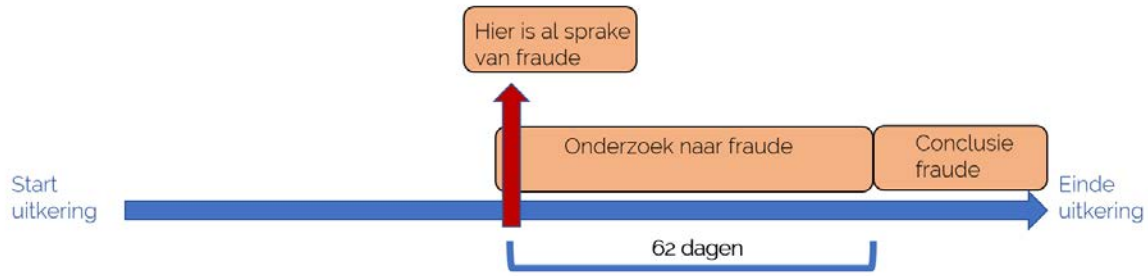
Figuur 2 - Weergave van het peilmoment:



4.2.3 Onrechtmatigheid definiëren in de data

De onrechtmatigheid is gedefinieerd in de data door een label van onrechtmatigheid te hangen aan alle cliënten die na een onrechtmatigheidsonderzoek een conclusie van onrechtmatig gedrag hebben gekregen of waarbij het onderzoek tot een besparing heeft geleid. Vervolgens is het peilmoment van cliënten met onrechtmatig gedrag veranderd naar het moment 62 dagen voordat het onrechtmatigheidsonderzoek is gestart (Figuur 3). Bij gemeente Nissewaard zit er gemiddeld 62 dagen tijd tussen de start van het onderzoek en het concluderen van onrechtmatig gedrag. Dus het onrechtmatige gedrag vind al eerder plaats dan dat de conclusie is getrokken dat dit inderdaad onrechtmatig gedrag is. Door op dit moment te trainen, wordt het meest betrouwbare moment van onrechtmatig gedrag meegenomen in het model.

Figuur 3 - Weergave van tijdslijn onrechtmatig gedrag:



4.2.4 Basis dataframe en voorspel niveau

De basis dataframe is een tabel met maar 1 uniek cliëntnummer per uitkering. Er is dus maar 1 regel per cliëntnummer (Tabel 2). Deze unieke cliëntnummers zijn nodig om een correcte koppeling te kunnen maken op het peilmoment. Door deze basisstructuur neer te zetten, is het mogelijk om een individuele voorspelling te maken, met genoeg onderscheidend vermogen voor het model, waarbij gelijk gecorrigeerd wordt voor eventuele tijdsinvloeden (wat vroeger was, hoeft nu niet meer zo te zijn), terwijl je wel wilt leren van de historie. Stel je zou de informatie niet samenvatten per persoon, maar je zou meerdere regels per persoon laten staan, dan zou het model minder goed kunnen voorspellen. Het model ziet namelijk elke regel als een voorbeeld. De voorbeelden die een samenvatting weergeven tot aan het peilmoment zijn veel informatiever dan de regels die alleen de informatie op het peilmoment weergeven. Tevens kan het dan voorkomen dat verschillende cases voor het model teveel op elkaar lijken, terwijl het model juist op zoek is naar het onderscheid, het verschil, van de verschillende cases.

Het is belangrijk dat we per persoon een voorspelling doen omdat de Nissewaard per persoon fraude onderzoeken uitvoert. De gemeente Nissewaard kan nu bijvoorbeeld de 10 personen met de hoogste kans op fraude selecteren en op deze personen een fraude onderzoek uitvoeren. Wanneer de gemeente willekeurig personen zou selecteren om te onderzoeken, vinden ze minder personen die daadwerkelijk fraude plegen. Door op deze manier te werken, kan de gemeente veel meer fraude opsporen dan voorheen. In het onderstaande figuur zie je een voorbeeld het voorspelniveau. Je ziet per cliëntnummer maar 1 regel met daarin informatie samengevat (zoals 'Aantal keer in de bijstand geweest').

Tabel 2 - Weergave van het voorspel niveau:

Cliëntnummer	Peilmoment	Aantal keer in de bijstand geweest	Onrechtmatigheid
1	2018-01-01	0	1
2	2018-10-01	0	0
3	2017-02-15	3	1

4.2.5 Samenvoegen tabellen tot 1 dataset & creëren van variabelen

Alle databronnen zijn gekoppeld aan de basis dataframe op cliënt niveau of dossier niveau en per databron zijn er variabelen gecreëerd. Per cliëntnummer is als het ware een samenvatting gegeven van het gedrag tot aan/op het peilmoment. Variabelen die bijvoorbeeld zijn gemaakt:

- Leeftijd; dit is de leeftijd op het peilmoment en die is berekend op basis van de geboortedatum.
- Aantal dossiers; dit is het aantal uitkeringsdossiers dat iemand heeft gehad tot aan het peilmoment.

4.2.6 Incomplete, dubbele of missende data corrigeren

Om met een modelleer techniek te kunnen werken, moeten alle missende gegevens gecorrigeerd zijn. Zo zijn voor alle numerieke waarden de missende gegevens vervangen met 0. Binnen categorische variabelen zijn de missende gegevens vervangen met 'geen'. Alle categorieën van categorische variabelen zijn meegenomen als aparte variabelen (dummies). Dit betekent dat iemand die in een bepaalde categorie valt, binnen de nieuwe 'dummy' variabele een 1-waarde krijgt en wanneer iemand erbuiten valt krijgt die een 0-waarde (Tabel 3). Voor elke categorie van de categorische variabele wordt er zo'n dummy variabele aangemaakt. Dit is gedaan omdat de modelleertechnieken niet kunnen werken met categorische waarden, maar wel met 1-en en 0-en.

Tabel 3 - Weergave van een categorische variabele en de dummy varianten ervan:

Cliëntnummer	Leefvorm	Leefvorm Samenwonend	Leefvorm Samenwonend met kinderen	Leefvorm Alleen wonend
1	Samenwonend	1	0	0
2	Samenwonend met kinderen	0	1	0
3	Alleen wonend	0	0	1

5 De modelontwikkeling

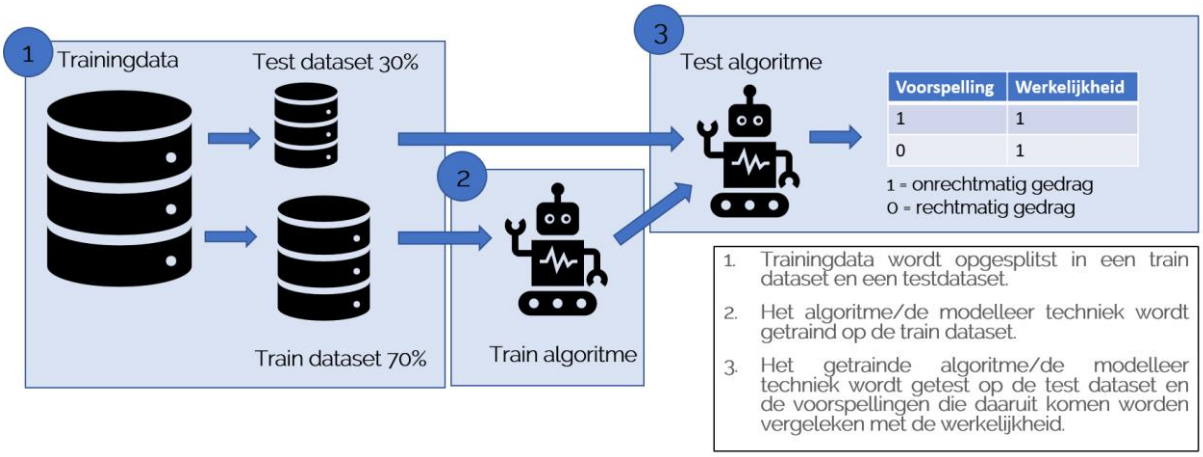
In de vorige hoofdstukken is besproken welke data verzameling en data verwerking nodig is voor het maken van een algoritme dat onrechtmatigheden binnen bijstandsuitkeringen voorspelt. In dit hoofdstuk wordt het keuzeproces beschreven voorafgaande aan het uiteindelijke model. Ook wordt uitgelegd welke modelleringstechnieken zijn gebruikt.

5.1 Aanpak modelkeuze

Om te bepalen welk model of samenstelling van modellen het beste werkt is de kwaliteit van de modellen getest. Dus eerst zijn de modellen getraind en daarna zijn getest op kwaliteit. Het trainen van het model betekent dat het model leert welke patronen over het algemeen voorkomen bij onrechtmatigheden en welke niet. We geven het model voorbeelden van onrechtmatig gedrag en dan gaat het model proberen om deze voorbeelden zelf in de data te vinden. Daarna testen we hoe goed het model instaat is om andere gevallen van fraude te vinden, die het model nog niet eerder heeft gezien.

Wanneer we het model in de praktijk gaan gebruiken, willen we namelijk ook dat het model in staat is om nieuwe fraude gevallen te vinden. Om dit te kunnen doen is de trainingdata uit het vorige hoofdstuk is willekeurig verdeeld in een train dataset (75%) en een test dataset (25%). Met deze datasets zijn de modellen getraind en is er een voorspelling gemaakt (Figuur 4). Met train dataset leert het model fraude patronen herkennen en met de test dataset is de kwaliteit getest. Tijdens een test wordt er een voorspelling gedaan op de test dataset met het model dat was getraind op de train dataset. Deze voorspelling is vergeleken met de werkelijke gevallen van bijstandsonrechtmatigheden. Hierbij is gekeken naar de 10 gevallen met de hoogste voorspelde kans op onrechtmatigheid en vervolgens is vergeleken hoeveel gevallen daarvan daadwerkelijk onrechtmatig gedrag vertonen. Het model of combinatie van modellen met de hoogste verhouding correct voorspelde onrechtmatigheden in de top 10 is uiteindelijk geselecteerd en in de praktijk getest en gebruikt door Nissewaard.

Figuur 4 - Weergave van het trainen en testen van het algoritme:



5.2 Uitleg over toegepaste modellen en technieken

Voor het voorspellen van onrechtmatigheden bij bijstandsuitkeringen is de data getraind en daarna getest met verschillende typen algoritmes. Al deze technieken kunnen in de data patronen te herkennen en hiermee regels opstellen die een label toekennen aan nieuwe gevallen van onrechtmatig gedrag.

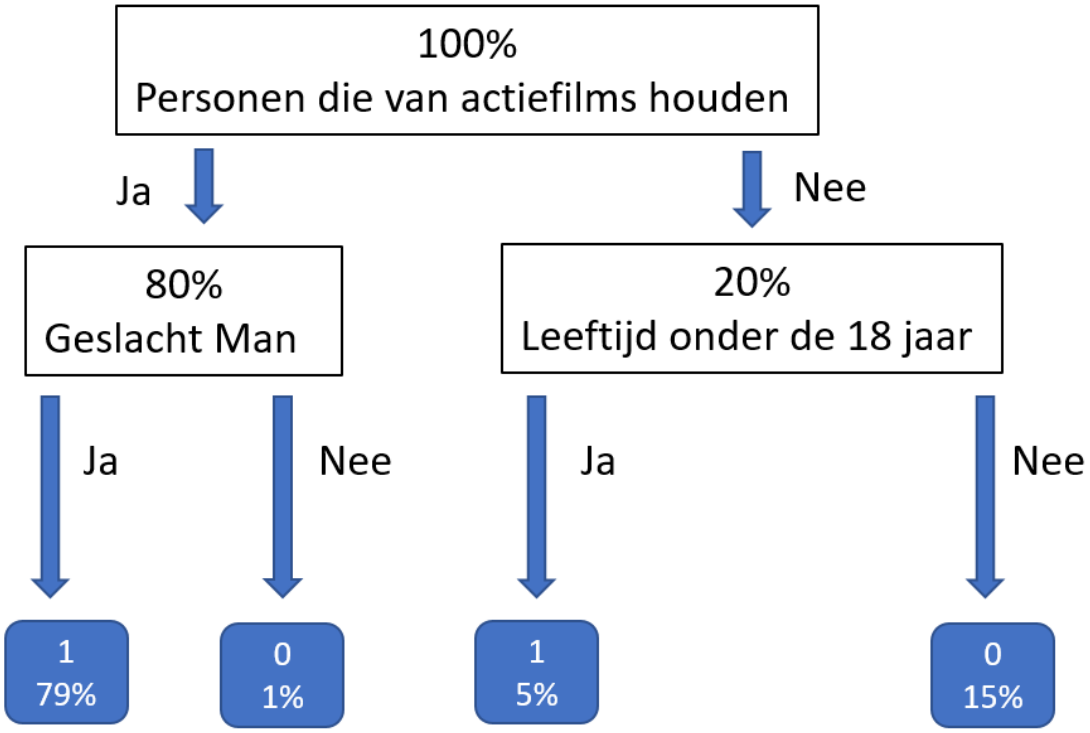
Totta data lab en Nissewaard hebben daarom in 2020 een nieuwe aanpak geïntroduceerd, welke transparantie combineert met het behouden van voorspelkracht en zonder extra bias. Hiervoor wordt de verzamel techniek 'Open A.I.' gebruikt. Deze technieken maken het namelijk mogelijk om op individuele niveau terug te geven welke (combinatie van) variabelen het meest hebben bijgedragen aan de hoogte van de risicoscore. De methode die wordt toegepast is Rulefit. Deze techniek maakt gebruikt van de Random forest tecnhiek en Lasso regressie. In onderstaande lichten we eerst deze technieken toe om de werking van Rulefit vervolgens te specificeren.

5.2.1 Random forest

Random forest is een techniek waarbij het model op de data wordt getraind aan de hand van beslisbomen (Figuur 5). Dus het model leert op basis van voorbeelden van onrechtmatig gedrag om zelf onrechtmatig gedrag te herkennen. Een beslisboom is een overzicht van alle mogelijke uitkomsten van een reeks gerelateerde keuzes. Aan de hand van een beslisboom kunnen nieuwe gevallen geclassificeerd worden als onrechtmatig of

rechtmatig. Om uit te leggen hoe een beslisboom eruit ziet, nemen we een voorbeeld waarbij we in kaart te brengen wie er naar de film 'Captain Marvel 'gaat. In het onderstaande voorbeeld zie je daar een visualisatie van:

Figuur 5 - Welke groepen mensen gaan er naar 'Captain Marvel'?

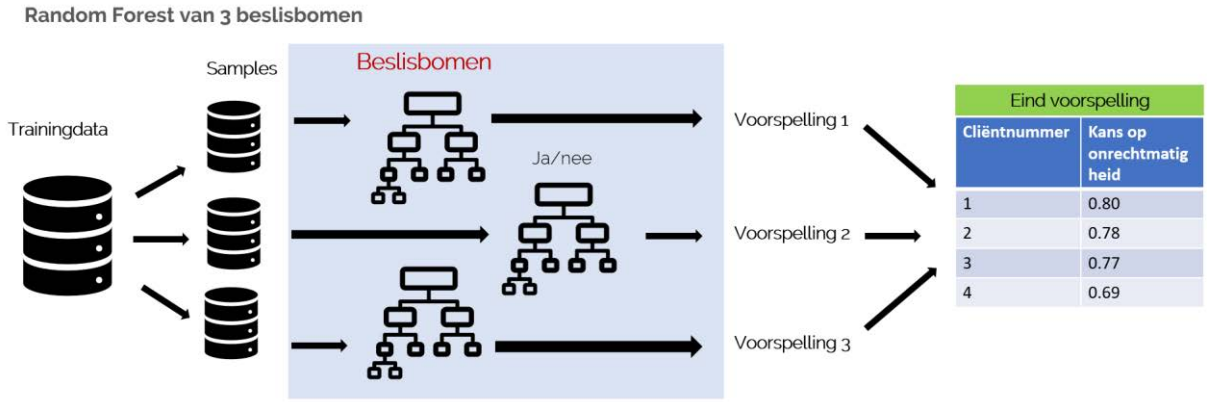


We zien hier dan de beslisboom 3 variabelen gebruikt om in te delen wie er wel en wie er naar 'Captain Marvel' gaat.: houdt een persoon van actiefilms, het geslacht en de leeftijd. De beslisboom heeft 2 groepen gevonden die naar 'Captain Marvel' gaan. In de eerste groep zit 79% van de personen uit de data en dit zijn allemaal mannen die van actiefilms houden. De andere groep bestaat uit personen die allemaal niet van actiefilms houden en jonger dan 18 jaar zijn. Deze groep bedraagt 5% van de personen in de dataset. Een beslisboom gebruikt dus een aantal variabelen om groepen te vinden die wel of niet in een bepaalde klasse vallen.

Wanneer de Random forest techniek gebruikt wordt voor het trainen van het model en voorspellen met het model, hebben we het over een 'bos' aan beslisbomen (Figuur 6). Het aantal beslisbomen dat in een Random forest model gebruikt wordt, kan gevarieerd worden in het model Meer bomen is altijd beter, omdat het naar een optimale verdeling convergeert. Echter, we gebruiken minder bomen, omdat dit simpelweg sneller is.. Door het aantal bomen in de tests te variëren vinden we het optimale model. Dit model voorspeld zo veel mogelijk gevallen correct, maar is ook snel in het herkennen van patronen in de data. Het Random forest selecteert willekeurig een aantal variabelen en maakt met deze variabelen een beslisboom. Op deze manier kan er geen bias ontstaan waarbij bepaalde variabelen steeds geselecteerd worden om een patroon op te herkennen.. De kwaliteit van de verschillende beslisbomen wordt gemeten en geeft aan elke boom een gewicht, waarbij bomen die beter voorspellen een hoger gewicht krijgen dan minder goed voorspellende beslisbomen. Het uiteindelijk model zal nieuwe gevallen classificeren aan de hand van al deze beslisbomen en hun gewichten. Dit bos van

beslisbomen is beter dan een enkele grote beslisboom omdat een grote beslisboom veel te specifiek zou worden voor een dataset. Die beslisboom zou dan niet meer goed werken op een nieuwe dataset. Door heel veel bomen te combinen, die op willekeurige variabelen zijn gebaseerd, kan er een algemeen patroon herkend worden. Dit patroon is dan ook te vinden in een nieuwe dataset.

Figuur 6 - Weergave van de Random forest techniek:



5.2.2 Lasso

De lasso modeling techniek is een uitbreiding op de lineaire regressie analyse. Bij een regressie analyse wordt een afhankelijke variabele beschreven door een som van onafhankelijke variabelen vermenigvuldigd met een gewicht plus een restterm. Denk bijvoorbeeld aan de prijs van fruit, die is afhankelijk van het seizoen en van de markt. In dit geval hebben we het over onrechtmatig gedrag en dat is bijvoorbeeld afhankelijk van variabelen zoals leeftijd en aantal dossiers.. Bij een lineaire regressie analyse worden gewichten gezocht zodat het verschil tussen de voorspelling van onrechtmatigheid en de daadwerkelijke waarde zo klein mogelijk is. De gewichten geven aan in hoeverre een variabele invloed heeft op de uiteindelijke uitkomst.

Lasso modeling is een techniek die de voorspel kracht probeert te verbeteren door alleen de variabelen een gewicht toe te kennen die het model verbeteren. De overige variabelen zullen een gewicht van 0 krijgen en hierdoor niet gebruikt worden voor het classificeren van nieuwe gevallen. Op deze manier selecteert de techniek zelf de belangrijkste variabelen, waardoor je enigszins kunt corrigeren voor overfitting. Overfitting betekent dat het model te specifieke patronen heeft geleerd gedurende training, en daardoor niet goed voorspelt op nieuwe data. Uiteindelijk moet het model namelijk de algemene patronen herkennen die ook in de toekomstige data voorkomen. Dit zijn dus patronen die niet individueel zijn, maar die bij meerdere personen voorkomen. Gedragspatronen zijn immers nagenoeg niet gelijk. Verder wordt het model ook getest met verschillende lasso gewichten. We testen dit om dezelfde reden als het testen van het aantal bomen bij een Random forest. Namelijk, het model moet optimaal presteren, de patroonherkenning mag niet te specifiek zijn, maar ook niet te onspecifiek.

5.2.3. Rulefit (Open A.I.)

De methode Rulefit maakt gebruik van de technieken Random forest en Lasso regressie.

Er wordt eerst een Random forest gecreëerd. Omdat elke beslisboom binnen de Random forest verschillend is, zijn ook de beslisregels in elke beslisboom verschillend. Eén beslisregel is één pad in de beslisboom. De beslissingen in een pad vormen de

beste split tussen rechtmatig en onrechtmatig gedrag op basis van historische data. Voor alle beslisbomen binnen de Random forest worden de beslisregels opgeslagen. Er ontstaat een lijst met verschillende beslisregels. . De beslisregels van een Random Forest zijn bijvoorbeeld:

- Beslisregel 1: Personen die van actiefilms houden
- Beslisregel 2: Personen die niet van actiefilms houden
- Beslisregel 3: Personen die van actiefilms houden & Man
- Beslisregel 4: Personen die van actiefilms houden & Vrouw
- Beslisregel 5: Personen die niet van actiefilms houden & Leeftijd onder de 18 jaar
- Beslisregel 6: Personen die niet van actiefilms houden & Leeftijd ouder dan 18 jaar
- Beslisregel 7: Personen die van komediefilms houden
- Beslisregel 8: Personen die niet van komediefilms houden
- Beslisregel 9: Personen die van komediefilms houden & Man
- Beslisregel 10: Personen die van komediefilms houden & Vrouw
- Beslisregel 11: Personen die niet van komediefilms houden & Leeftijd onder de 18 jaar
- Beslisregel 12: Personen die niet van komediefilms houden & Leeftijd ouder dan 18 jaar

Vervolgens wordt er voor elke client gekeken aan welke beslisregels de client voldoet. Uiteindelijk moet het model de algemene patronen herkennen die gelden voor iemand die onrechtmatigheid pleegt en iemand die dat niet doet. Dit zijn dus patronen die niet individueel zijn, maar die bij meerdere personen voorkomen. Om ervoor te zorgen dat alleen de beslisregels overblijven die belangrijk zijn voor het algemene patroon wordt een Lasso regressie toegepast. Beslisregels met een sterk verband tot (on)rechtmatigheden krijgen in de Lasso regressie een hogere weging. Er wordt per client beoordeeld welke beslisregels belangrijk zijn om tot de uiteindelijke voorspelling te komen. Als een client bijvoorbeeld aan beslisregel 3 voldoet, is de kans groter dat hij van Captain Marvel houdt dan iemand die aan beslisregel 6 voldoet. Lasso regressie probeert beslisregels te vinden die duiden op een hogere kans op onrechtmatigheid.

Per client wordt een risicovoorspelling berekend aan de hand van het verband, de weging en de gedragsvariabelen van de client. De Rulefit methode geeft door het gebruik te maken van Lasso regressie ook inzicht in het belang van de beslisregels voor elke client om tot deze risicovoorspelling te komen.

Er zijn verschillende parameters die aangepast kunnen worden binnen Rulefit. Ten eerste kan het aantal bomen in de Random forest worden gevarieerd. Als er meer beslisbomen in een Random forest zitten ontstaan er meer verschillende beslisregels. Maar het wordt ook moeilijker voor de Lasso regressie om de belangrijkste beslisregels eruit te halen. Er wordt dus gezocht naar het optimale aantal beslisbomen in een Random forest. Ten tweede kan het aantal variabelen in een beslisregel worden gevarieerd. Met meer variabelen in een beslisregel is het moeilijker de beslisregels te interpreteren. Maar te weinig variabelen in een beslisregel zou een te algemeen beeld geven waardoor het moeilijker wordt voor het model om cliënten te identificeren met een hogere kans op het plegen van onrechtmatigheid. Transparantie met behoud van een kwalititatief goede voorspelling is de belangrijkste motivatie voor het gebruik van de Rulefit methode, daarom is er gekozen om alleen te kijken naar hoe het model presteert met maximaal 4 of 5 variabelen per beslisregel. Uit wetenschappelijk onderzoek door Totta data lab is naar voren gekomen dat het gebruiken van 4 of 5 variabelen per beslisregel dezelfde

voorspelkracht geeft als 'black-box' technieken en de resultaten tegelijkertijd voldoende interpreteerbaar zijn.

5.2.4 Hoe ziet de uitkomst eruit?

In onderstaande tabel wordt een voorbeeld gegeven van hoe de resultaten van Rulefit er in de praktijk uitzien bij een oplevering. In deze weergave worden voor iedere client de drie belangrijkste beslisregels weergegeven. Deze beslisregels zijn voor de individuele cliënt belangrijk om de hoogte van de kans op onrechtmatigheid per client te verklaren. Deze beslisregels verschillen daarom per client. Hieronder ziet u een voorbeeld van deze weergave:

Clientnummer	Risico	Individuele regel 1 (positief)	Individuele regel 2 (positief)	Individuele regel 3 (positief)
1	0.80	- duur uitkering > 15 jaar - aantal kinderen > 5 - aantal trajecten < 2.5 - leefvorm: Alleenstaande	- leeftijd > 29 - duur uitkering > 15 jaar - man - oorzaak uitkering: einde uitkering werkloosheid	- leefvorm: alleenstaande - aantal trajecten < 2.5 - burgerlijke staat: ongehuwd - reden beëindiging: wijziging uitkering norm
2	0.77	- leeftijd > 29 - duur uitkering > 15 jaar - man - oorzaak uitkering: einde uitkering werkloosheid	- leeftijd > 29 - leeftijd start uitkering > 25 - oorzaak uitkering: andere oorzaak - burgerlijke staat: gehuwd	- duur uitkering > 15 jaar - aantal kinderen > 5 - aantal trajecten < 2.5 - leefvorm: Alleenstaande
3	0.75	- leeftijd > 29 - duur uitkering > 15 jaar - man - oorzaak uitkering: einde uitkering werkloosheid	- leefvorm: alleenstaande - aantal trajecten < 2.5 - burgerlijke staat: ongehuwd - reden beëindiging: wijziging uitkering norm	- duur uitkering < 5 jaar - leeftijd start uitkering > 25 - leefvorm: alleenstaande - vakantiedagen > 15

Zodra we naar de belangrijke beslisregels bij de risicoscore van client 1 kijken zien we dat de client vooral hoog scoort, omdat deze client:

- langer dan 15 jaar een uitkering heeft;
- meer dan 5 kinderen heeft;
- minder dan 2,5 trajecten volgt en
- alleenstaand is

Maar dit is niet het enige wat een significante bijdrage heeft geleverd aan deze score. Bij de tweede belangrijke regel zien we dat ook bijdraagt aan de risicoscore dat de client een man is van ouder dan 29 jaar en in de uitkering terecht gekomen is na een werkloosheidsuitkering.

Tot slot speelt bij de hoogte van de risicoscore bij deze man dat maximaal 2 trajecten heeft doorlopen en een vorige bijstandsuitkering is beëindigd door een wijziging in de norm van de uitkering.

6 Resultaten

Nu beschreven is welke data en welke technieken worden ingezet voor de kwartaal, bespreken we welke parameters zijn gebruikt om onrechtmatigheden in de bijstand te voorspellen.

6.1 Hoe werkt de modelbeoordeling

Om de kwaliteit van modellen te beoordelen worden de correcte en incorrecte voorspellingen met elkaar vergeleken. Dit is goed te vergelijken met ROC curves (Figuur 9). In een ROC curve worden de correcte voorspellingen afgezet tegen de incorrecte voorspellingen. De correcte voorspelling geeft aan dat het model onrechtmatigheid voorspelt waar dit ook daadwerkelijk het geval is. Bij de incorrecte voorspelling is ook onrechtmatigheid voorspeld, maar is dit in werkelijkheid niet het geval. De ROC curve is verder ook vergeleken met een curve van een model dat een willekeurige voorspelling geeft. Hoe groter de afstand tussen de ROC curve en de willekeurige curve, des te beter het model kan voorspellen. Daarbij is voornamelijk van belang dat er een groot verschil te zien is links, aan het begin, van de geplotte curves (zie de figuren hieronder). We willen namelijk vooral de gevallen die volgens het model de hoogste fraude kansen hebben, correct voorspellen.

De ROC curve is gebaseerd op de confusion matrix (Figuur 8). Naast de ROC curve kijken we ook naar de confusion matrix van de top 10 gevallen met de hoogste fraude kansen. Een confusion matrix geeft de volgende zaken weer:

- Daadwerkelijke aantal fraude gevallen
- Daadwerkelijk aantal gevallen zonder fraude
- Voorspelt aantal fraude gevallen
- Voorspelt aantal gevallen zonder fraude

Aan de hand van deze aantallen kan het percentage correct voorspelde fraude gevallen berekend worden. We willen uiteindelijk een model waarbij zoveel mogelijk gevallen binnen de top 10 hoogste fraude kansen correct voorspeld wordt. In de afbeelding hieronder zien we dat **3 van de 10 gevallen correct voorspeld** zijn. Dus 30% van de personen in de top 10 met de hoogste fraude kans is in dit voorbeeld correct voorspeld.

Figuur 7 – Confusion matrix

		Daadwerkelijke aantallen	
Voorspelde aantallen		Wel fraude	Geen fraude
	Wel fraude	3	2
	Geen fraude	3	2

Deze afbeelding is een voorbeeld van een confusion matrix voor de top 10 hoogste fraude kansen. Je ziet dat bij elkaar opgeteld de getallen in de matrix 10 zijn. In totaal zijn er daadwerkelijk 8 fraude gevallen in de top 10 aanwezig. Er zijn 5 gevallen van fraude voorspeld in de top 10. Van het aantal voorspelde fraude gevallen zijn er 3 daadwerkelijk fraudeur en 2 zijn geen fraudeur. Het model voorspelt dat 3 van de daadwerkelijk fraude gevallen, geen fraude gevallen zijn.

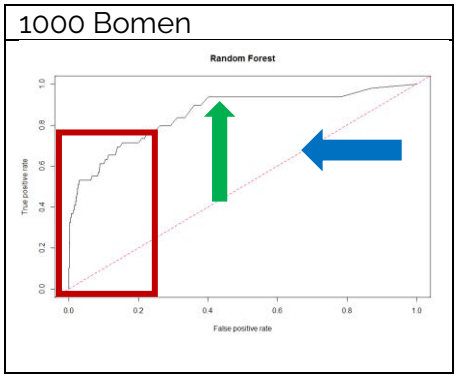
6.2 Modelleringsfase – assemblages, losse modellen

Een robuust model, of combinatie van modellen, zorgt ervoor dat er zowel bij het testen op bekende (in sample) als op onbekende (out of sample) data vergelijkbare resultaten uitkomen. Wanneer het algoritme is getest op bekende data, dan is het algoritme getest op dezelfde data als waarop het algoritme getraind is. Wanneer het model getest is op onbekende data dan is het algoritme getest op andere data dan waarop het getraind is. Het gaat hier dus om testdata die nog helemaal nieuw is voor het algoritme. Wanneer een model goed in sample kan voorspellen en minder goed out of sample kan voorspellen wordt dit 'overfitten' genoemd. Om ervoor te zorgen dat een model ook goed out of sample kan voorspelen wordt niet altijd het beste model gekozen. Een 'simpeler' model waar de kwaliteit van de voorspellingen bijna net zo goed zijn als het 'perfecte' model zal een betere voorspeller zijn voor nieuwe data. Door meerdere modellen te combineren kan overfitting worden tegengegaan. Op deze manier voorkom je dus een overfitting of underfitting in je voorspellingen.

Hieronder worden de resultaten van het Random forest model en de Rulefit techniek met elkaar vergeleken,

Random forest:

Figuur 8 – ROC curve Random forest



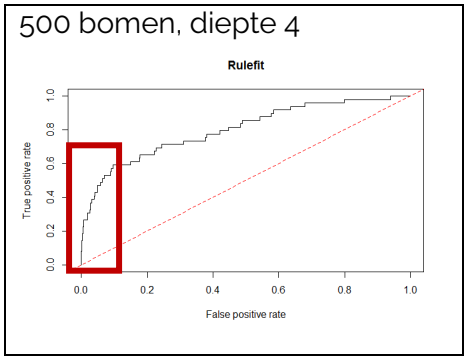
Uitleg van de grafieken: Hiernaast is in Figuur 8 de ROC curve van de Random forest techniek weergegeven. De curve toont een stabiel model, die vele malen beter is dan een willekeurige selectie en presteert het best in de hoogst voorspelde kansen op onrechtmatigheid (rood). Tabel 4 geeft aan dat van de top 10 hoogst voorspelde risico's 8 onrechtmatigheden ook daadwerkelijk zijn bewezen.

Tabel 4:: confusion matrix Random forest (top 10)

		Daadwerkelijke aantallen	
Voorspelde aantallen		Wel fraude	Geen Fraude
	Wel fraude	8	2

Rulefit:

Figuur 9 – ROC curve Rulefit



	Daadwerkelijke aantallen		
Voorspelde aantallen		Wel fraude	Geen Fraude
	Wel fraude	7	3

Tabel 5: confusion matrix Rulefit (top 10)

Uitleg van de grafieken: Hiernaast is in Figuur 9 de ROC curve van de Rulefit techniek weergegeven. Ook deze curve toont een stabiel model, die vele malen beter is dan een willekeurige selectie. Bij de Rulefit zien we vooral de beste modelprestaties in de hoogst voorspelde kansen op onrechtmatigheid (rood). Tabel 5 geeft aan dat van de top 10 hoogst voorspelde risico's 7 onrechtmatigheden ook daadwerkelijk zijn bewezen.

We zien dat de Rulefit methode nagenoeg dezelfde voorspelkracht houdt als de Random forest. We verkiezen echter de Rulefit methode omdat dit ons inzicht geeft in het waarom van de voorspelling op individueel niveau.

7. Proces van voorspellen en leren

Bekend is welke data gebruikt wordt, hoe deze geprepareerd is en welk technieken gebruikt zijn voor het vinden van een model dat zo goed mogelijke voorspellingen kan maken. De volgende stap is het maken van voorspellingen voor personen die op dit moment in de bijstand zitten. Om deze voorspelling te maken zijn per oplevering de volgende stappen uitgevoerd:

- De meest recente data van bijstandsgerechtigden is aangeleverd via een twee-factor beveiligde SFTP verbinding aangeleverd door Nissewaard.
- Deze data bevat ook de bijstandsgerechtigden die bij een eerdere oplevering zijn onderzocht door de gemeente (de top 10). Voor deze personen is het dan bekend of ze wel of niet fraude hebben gepleegd.
- Het model wordt vervolgens getraind op de data, inclusief de nieuwe gevonden fraude cases. Op deze manier leert het model nog beter fraude voorspellen.
- Met dit nieuwe getrainde model is een voorspelling gemaakt op basis van de voorspel dataset oftewel de nieuw aangeleverde data, waarin alleen cliënten voorkomen die momenteel een uitkering hebben.
- De voorspellingen zijn vervolgens aangeleverd aan Nissewaard via de twee-factor beveiligde SFTP verbinding.
- De gemeente Nissewaard gaat vervolgens de bijstandsgerechtigden uit de top 10 onderzoeken op fraude.
- De opleveringen worden gemiddeld een keer per kwartaal gedaan.

Voor elke lopende bijstandsgerechtigden is een kans berekend voor mogelijke onrechtmatigheden. Deze zijn van hoog naar laag gesorteerd en zijn samen met de gepseudonimiseerde cliëntnummers met de gemeente Nissewaard gedeeld om nader onderzocht te worden.