Universiteit Leiden
The Netherlands

Opleiding Informatica & Economie

Het verbeteren van het proces rond de evaluatie van zorgaanbieders

Maurits de Groot

Supervisors:
Arno Knobbe (LIACS)
Dirk Meijer (LIACS)
Rob Konijn (Zilveren Kruis)

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

10/08/2018

Abstract

Het doel van dit onderzoek was om te kijken of er op een betere manier gebruik gemaakt kan worden van de beschikbare informatie om beslissingen te nemen in het controleproces van zorgaanbieders door zorgverzekeraar Zilveren Kruis. Om deze vraag te beantwoorden is gekeken naar verbeteringen op basis van het proces. Er is een nieuwe aanpak voorgesteld die gebruik maakt van lineaire regressie die naast de huidige methodiek een voorspelling doet over de verwachte uitkomsten van de controles.

Ook is voorgesteld een inschatting te maken van de tijd en de kosten van een controle. Met deze informatie kan er op een betere manier gestuurd worden in de volgorde van controleren. Verder is er gekeken naar intertemporele verbanden tussen deze controles, Hieruit is gebleken dat er een hoge kans is dat als er in het huidige jaar een risico is gedetecteerd dat dit risico in het verleden ook heeft plaatsgevonden.

Om dit onderzoek uit te voeren, is gebruik gemaakt van verschillende (samengestelde) datasets. De gebruikte gegevens zijn op te splitsen in gegevens over de geestelijke gezondheidszorg (GGZ) en gegevens over wijkverpleging. De gegevens over wijkverpleging zijn gebruikt als indicator voor een volledigere GGZ dataset. Dit omdat de datakwaliteit van de gegevens van de GGZ niet voldeed om de analyses over uit te voeren. Bij het maken van een voorspellingsmodel is de kwaliteit van de uitkomst afhankelijk van de kwaliteit van de gegevens die gebruikt worden om het model te maken. Dit betekent dat als er gegevens worden gebruikt die onvolledig zijn, zoals de gegevens van de GGZ, de voorspelling een lage nauwkeurigheid zal hebben.

Bij het gebruik van een grotere en volledigere dataset, zoals de dataset van de wijkverpleging, kwam naar boven dat de kwaliteit van de voorspelling hoger was. De nauwkeurigheid van de methode die gebruikt maakt van lineaire regressie is vergeleken met de huidige methodiek. Het toevoegen van de lineaire regressie geeft een toegevoegde nauwkeurigheid. Dit is te verklaren doordat lineaire regressie gebruik maakt van alle beschikbare informatie en een afweging maakt waarin gekeken wordt welke attributen bijdragen aan een betere voorspelling. Hierbij komt het voor dat er attributen worden meegenomen die op zichzelf bijna geen nauwkeurigheid bijdragen, alleen door meerdere van dit soort attributen samen te nemen, verhoogt de algemene nauwkeurigheid van de voorspelling.

Inhoudsopgave

Hoofdstuk 1

Introductie

In Nederland is iedereen van 18 jaar en ouder verplicht een basis zorgverzekering af te sluiten. Dat kan bij één van de negen zorgverzekeraars (concerns) die ieder op hun beurt meerdere labels onder zich hebben. De vier grootste concerns (Achmea, VGZ, CZ en Menzis) hebben 90% van de markt in handen [Koenraadt, 2016].

Wanneer een verzekerde zorg nodig heeft, gaat hij naar een zorgaanbieder. Deze aanbieder verleent de zorg. Heeft de verzekerde een naturapolis dan gaat de rekening rechtstreeks naar de zorgverzekeraar. Heeft de verzekerde een restitutiepolis, dan betaalt de verzekerde de rekening en declareert die vervolgens bij de zorgverzekeraar.

De kosten voor de gezondheidszorg in Nederland stijgen ieder jaar. In 2017 werd 97,5 miljard euro uitgegeven aan de gezondheidszorg. Dat is een stijging van 2,0 miljard ten opzichte van 2016 [CBS, 2018]. De stijging wordt veroorzaakt door een aantal factoren:

1. Vergrijzing: mensen worden ouder en ouderen vragen meer zorg.
2. Zorggebruik neemt toe: een toenemend aantal mensen maakt gebruik van zorg.
3. Hogere eisen: mensen gebruiken niet alleen vaker zorg, maar stellen er ook steeds hogere eisen aan.
4. Nieuwe technologie: de medische wereld boekt vooruitgang, waardoor steeds meer mensen kunnen worden geholpen en genezen. Ziekten waar men vroeger aan overleed, zijn nu goed te behandelen met nieuwe medicijnen en dure behandelmethodes. Door deze toenemende mogelijkheden leven mensen langer, maar stijgen de kosten ook weer.
5. Chronisch zieken: veel ziekten zijn niet te genezen, maar zijn chronisch geworden. Hierdoor worden patiënten levenslang behandeld, met alle bijkomende kosten.
6. Meer begeleiding: Nederland investeerde in de afgelopen decennia steeds meer in de begeleiding van mensen die moeilijk mee kunnen komen in de maatschappij, zoals ouderen en mensen met een

lichamelijke of verstandelijke beperking. Dat kost geld.

Het Rijksinstituut voor Volksgezondheid en Milieu (RIVM) spreekt in het Trendscenario Volksgezondheid Toekomst Verkenning 2018 [RIVM, 2018] de verwachting uit dat in 2040 de zorgkosten in Nederland zijn gestegen tot een bedrag van 174 miljard euro. Om de zorgpremie niet te hard te laten stijgen, zijn maatregelen nodig om de zorguitgaven in de hand te houden. Zo moet de zorgverzekeraar zorgvuldig kijken naar de gedeclareerde zorg, zodat alleen voor de zorg betaald wordt die daadwerkelijk geleverd is.

Omdat de zorg blijft vernieuwen, moeten zorgverzekeraars blijven verbeteren om hierop in te spelen. Ook het proces dat de rechtmatigheid van zorgdeclaraties controleert moet blijven verbeteren. De beschikbare hoeveelheid gegevens over declaraties neemt de komende periode alleen maar toe [Coffman and Odlyzko, 2002]. Hierdoor heeft een zorgverzekeraar meer middelen (data) tot zijn beschikking. Dit vraagt om slimmere methodes om die groeiende hoeveelheid gegevens te blijven verwerken.

Zorgverzekeraar Zilveren Kruis is onderdeel van Achmea, marktleider in zorgverzekeringen. Met ruim 3,4 miljoen verzekerden is Zilveren Kruis één van de grootste zorgverzekeraars van Nederland [Kruis, 2017]. Zilveren Kruis doet er alles aan om de toenemende zorgkosten te beheersen. Een van de manieren is het continu verbeteren van het proces dat de rechtmatigheid van zorgdeclaraties controleert.

In dit document wordt vanuit een data gedreven perspectief gekeken naar het interne proces binnen Zilveren Kruis. Dit om te kijken hoe Zilveren Kruis deze gegevens kan gebruiken om de zorgdeclaraties efficiënter te kunnen controleren zodat de zorgkosten beperkt te houden.

Hoofdstuk 2

Probleemstelling

Het is belangrijk binnen bedrijfsprocessen te blijven streven naar verbetering om te zorgen dat het bedrijf concurrerend blijft in de markt. Dat geldt voor ieder bedrijf, dus ook voor een zorgverzekeraar. Verder is het belangrijk om te zorgen dat nieuwe vormen van onrechtmatige zorgdeclaraties worden gedetecteerd. Deze onrechtmatigheden zorgen er voor dat de zorg duurder wordt zonder dat de kwaliteit ervan omhoog gaat [Kirlidog and Asuk, 2012]. Dat moet natuurlijk voorkomen worden.

In dit onderzoek wordt gekeken naar het proces waarmee zorgverzekeraar Zilveren Kruis declaraties van zorgaanbieders in de geestelijke gezondheidszorg (GGZ) controleert op rechtmatigheid. Hierbij ligt de focus op het maken van beslissingen binnen dit controleproces. Er wordt onderzocht welke ontwikkelingen zich hebben afgespeeld op het gebied van data driven decision making en hoe deze toepasbaar zijn binnen het huidige proces dat het Zilveren Kruis hanteert.

Dit wordt omvat in de volgende onderzoeksvraag:

> *"Hoe kan er beter gebruik worden gemaakt van data bij het maken van beslissingen binnen het controleproces van zorgaanbieders bij zorgverzekeraar Zilveren Kruis"*

De opbouw van het onderzoek is als volgt: eerst wordt er gekeken naar de context waarin het onderzoek zich bevindt. Hier wordt gekeken naar relevant onderzoek dat in het verleden is uitgevoerd, worden een aantal definities gegeven die belangrijk zijn voor de rest van het onderzoek, vervolgens wordt een beschrijving gegeven van het huidige proces.

Daarna volgt de analyse. Deze analyse wordt uitgevoerd door in gesprek te gaan met diverse actoren binnen het proces. Op deze manier wordt er een objectief beeld gevormd van de huidige situatie. Dit zal gebeuren door verbeteringen voor te stellen binnen het bestaande proces. Vervolgens worden deze bevindingen geëvalueerd, dit wordt gedaan door concrete aanbevelingen te doen en te reflecteren op de beperkingen van dit onderzoek.

Tot slot worden de bevindingen samengevat in de conclusie.

Het resultaat van het onderzoek bestaat uit een analyse van de huidige situatie gevolgd door aanbevelingen over verbeteringen voor de toekomst. Deze aanbevelingen kunnen betrekking hebben op wijzigingen op het huidige proces of een advies om nieuwe actie te ondernemen die nog niet in het proces is opgenomen.

Eventuele verbeteringen worden geformuleerd op basis van recente ontwikkelingen binnen de literatuur en binnen de bedrijfscontext.

Omdat de insteek van dit onderzoek ligt op beslissingen die gebaseerd worden op data wordt ook gekeken naar de huidige beschikbare gegevens. Deze worden vergeleken met de gegevens die nodig zijn voor verschillende methodieken die gebruikt kunnen worden en beter zouden werken dan de huidige situatie om op deze manier verbeteringen te vinden van het proces.

Hoofdstuk 3

Context

In dit hoofdstuk wordt er meer toelichting gegeven over de achtergrond van het onderzoek en de bedrijfscontext waarin het onderzoek wordt uitgevoerd. Hierbij wordt gekeken naar gerelateerd werk. Dit is belangrijk om het wetenschappelijke aspect van het onderzoek goed te begrijpen en de rol van verschillende datamining technieken binnen de zorgverzekeringencontext te plaatsen. Ook worden definities afgebakend. Tot slot wordt een beschrijving gegeven van de huidige situatie vanuit waar de verbetering plaats kan vinden.

3.1 Gerelateerd werk

De laatste jaren wordt steeds meer gebruik gemaakt van datamining technieken voor het nemen van beslissingen binnen een bedrijf [Cao and Zhang, 2008]. Tegenwoordig is informatie een steeds belangrijker bezit aan het worden van bedrijven [Redman, 2008]. Daarom wordt er in dit veld ook steeds meer onderzoek gedaan.

Zoals te lezen in het onderzoek van [Brynjolfsson et al., 2011]. In dit onderzoek wordt gekeken naar het effect van Data-Driven Decision Making op de prestaties van bedrijven. Uit dit onderzoek blijkt dat er een sterke associatie is tussen het gebruik van op data gebaseerde beslissingen en een verhoogde productiviteit.

Verder wordt onderzoek gedaan naar de toepassing van datamining technieken binnen zorgverzekeringen, bijvoorbeeld door [S. Viveros et al., 1996]. Dit onderzoeken toont aan dat het gebruik van deze technieken tot een verhoogde efficiëntie leidt bij het opsporen van onrechtmatige claims.

3.2 Definities

In dit rapport wordt een aantal definities binnen de context van het Zilveren Kruis en data-analyse gebruikt. Deze definities worden in dit hoofdstuk verder toegelicht.

3.2.1 Zorgaanbiedereigenschappen

Zorgaanbiedereigenschappen verwijzen naar de kenmerken van een zorgaanbieder. Deze kenmerken zijn per zorgaanbieder per tijdsperiode beschikbaar voor analyse. Voorbeelden van zorgaanbiedereigenschappen zijn "verblijfsduur per cliënt" of "diagnosehoofdgroep per cliënt". Deze eigenschappen kunnen zowel nominaal als numeriek zijn.

3.2.2 Risico's

Binnen de context van Zilveren Kruis is een risico:

"Een onderdeel waarop er een fout gemaakt kan worden binnen een zorgaanbieder"

De risico's waar het binnen de GGZ om gaat, zijn de volgende:

Verwijzing

Een bewijslast die een zorgaanbieder heeft waarin bewezen moet worden dat er een verwijzing heeft plaatsgevonden voorafgaand aan de behandeling. Dit moet vanuit het patiëntendossier gedaan worden.

Hoofdbehandelaarschap

Volgens de Wet marktordening gezondheidszorg of vanwege polisvoorwaarden wordt er een hoofdbehandelaar als bevoegd vermeld. Deze hoofdbehandelaar heeft in de diagnostiekfase direct patiëntgebonden tijd besteed aan de patiënt waarvoor wordt gedeclareerd.

Onverzekerde zorg

Alle behandelingen die niet voldoen aan de stand van wetenschap en praktijk vallen onder onverzekerde zorg. Dit zijn behandelingen waarvan niet is aangetoond dat deze een effect hebben en worden daardoor niet vergoed.

Juist gebruik verblijfsprestaties

De Nederlandse zorgautoriteit heeft in regeling NR/CU-570 verschillende types verblijfsprestaties gedefinieerd. Zo geld dat voor verblijfsprestaties E, F en G de patiënten doorgaans tijdens de duur van de behandeling in een kliniek verblijven.

Verblijfsdagen

DBC-spelregels 2013: Een verblijfsdag met overnachting kan alleen geregistreerd worden als de patiënt voor 20:00 uur is opgenomen (eerste opname) en 's nachts in de instelling verblijft. De dag van opname en de daarop volgende nacht gelden als één verblijfsdag.

Diagnose Behandeling Combinatie (DBC)

Op het moment dat een patiënt ergens last van heeft, wordt deze behandeling vaak door verschillende instanties afgehandeld. Meestal komt een patiënt eerst bij een specialist om vast te stellen waar deze patiënt last van heeft. Nadat deze diagnose is gesteld wordt de patiënt vaak doorverwezen naar een ziekenhuis om behandeld te worden. Soms komt het zelfs nog voor dat er nazorg nodig is.

Om bij te houden wat de kosten zijn van een bepaalde behandeling wordt er per diagnose een dossier aangemaakt waarin alle activiteiten gerelateerd aan deze behandeling worden bijgehouden.

3.3 Procesbeschrijving

Voordat verbeteringen in het bestaande proces aangebracht kunnen worden, is het essentieel het huidige proces in kaart te brengen. Omdat het onderzoek zich richt op beslissingen die op basis van gegevens gemaakt kunnen worden, zal de beschrijving van het proces ook vanuit dat perspectief zijn.

De beschrijving zal in verschillende delen opgedeeld worden. Elk deel vertegenwoordigt een stap in het proces waar gegevens verwerkt worden en er een keuze gemaakt wordt op basis van deze transformatie.

Het volledige overzicht van het proces is te vinden in Appendix B.

Het proces begint op het moment dat er risico's zijn gesignaleerd en eindigt als de controle uitgevoerd is. Dit betekent dat er geen aanbevelingen zijn over de methode waarop risico's gesignaleerd worden of over de uitspraak van de controle.

3.3.1 Zorgaanbiedereigenschappen naar risico's

Beschrijving

Om er achter te komen of er geen fouten worden gemaakt, moeten de risico's zoals beschreven in 3.2.2 meetbaar gemaakt worden. Het meetbaar maken van risico's gebeurt op verschillende manieren. Sommige risico's kunnen rechtstreeks uit de data gehaald worden. Zo moet er tussen sommige behandelingen een bepaalde tijd zitten. Dit is rechtstreeks af te lezen uit de data door te kijken of de tijd tussen deze specifieke behandelingen groot genoeg is. Deze risico's worden formele risico's genoemd. Controles hierop worden formele controles genoemd. Bij formele controles is direct vast te stellen of de declaratie goed of fout is.

Figuur 3.1: Zorgaanbiedereigenschappen naar risico's

Andere risico's zijn niet vast te stellen door naar één zorgaanbiedereigenschap te kijken. Voor deze risico's wordt gekeken naar meerdere zorgaanbiedereigenschappen. Wat deze eigenschappen zijn, wordt bepaald door de desbetreffende domeinexperts en specialisten. Van deze risico's is niet direct te bepalen of ze van toepassing zijn voor een bepaalde zorgaanbieder. Wel is het onderlinge verschil uit de data te halen door deze zorgaanbiedereigenschappen terug te brengen tot één meetwaarde.

Nadat de domeinexperts samen hebben bepaald hoe de meetwaarde tot stand wordt gebracht wordt er nog een populatiecorrectie uitgevoerd die per zorgaanbieder corrigeert op de type patiënten.

Voorbeeld

Ter illustratie van dit proces een voorbeeld van de fictieve risico's X en Y en zorgaanbieders A, B en C. Van risico X is bekend dat deze afhankelijk is van zorgaanbiedereigenschap P1, P2 en P3. Risico Y afhankelijk van P4 en P5. In dit voorbeeld blijft de populatiecorrectie achterwege.

Over zorgaanbieders A, B en C is het volgende bekend: (tabel 3.1)

Zorgaanbieder	P1	P2	P3	P4	P5
A	6	42	106,40	J	Cat1
B	7	42	96,50	J	Cat5
C	6	42	102,33	N	Cat1

Tabel 3.1: Voorbeeld - Gegevens zorgaanbieders

Voor iedere zorgaanbieder wordt de meetwaarde van de risico's X, Y bepaald. Dit wordt voor risico X bijvoorbeeld gedaan door $(P1 \cdot P2) - P3$ te nemen, voor risico Y gekeken of $P5 = Cat1$ en $P4 = J$. Dit resulteert in de volgende waardes voor de risico's (tabel 3.2)

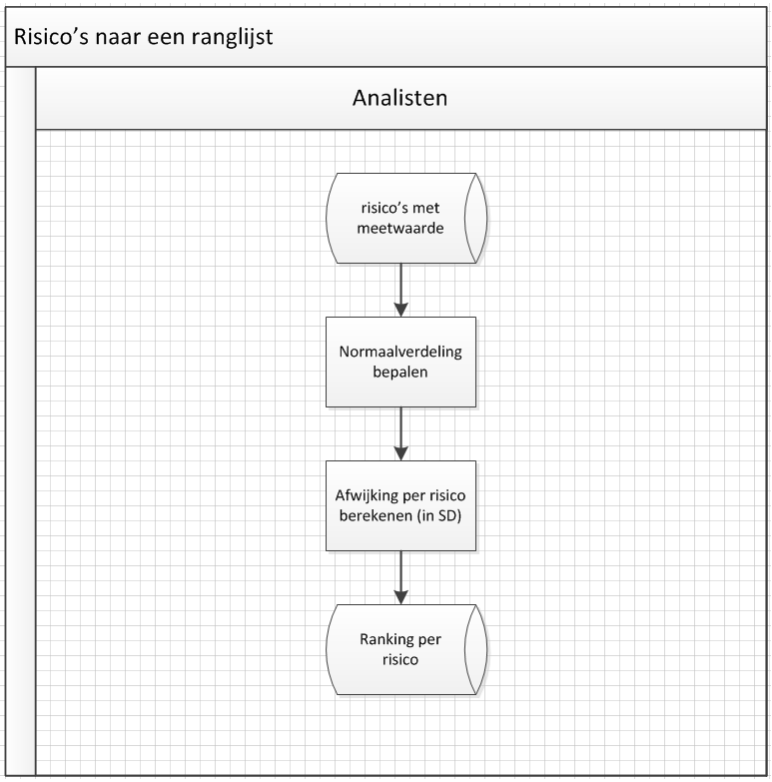
Zorgaanbieder	Meetwaarde X	Meetwaarde Y
A	145.60	1
B	155.50	0
C	149.67	0

Tabel 3.2: Voorbeeld - Meetwaardes X, Y

3.3.2 Risico's naar een ranglijst

Beschrijving

Wanneer alle risico's zijn vastgesteld en meetbaar zijn gemaakt, moet een prioritering aangebracht worden van alle zorgaanbieders. Dit omdat het niet mogelijk is alle zorgaanbieders te controleren. Dit gebeurt momenteel door te kijken naar afwijkend declaratiegedrag tussen zorgaanbieders. Voor alle zorgaanbieders wordt een verdeling vastgesteld voor ieder risico. Voor ieder risico wordt er een meetwaarden bijgehouden, deze specifieke waarde is representatief voor een bepaald risico.



Figuur 3.2: Risico's naar een ranglijst

Voor ieder van deze meetwaardes wordt een z-score opgesteld. Een z-score is een statistische maat die aangeeft hoeveel standaarddeviatie een waarde afwijkt van het gemiddelde. Een z-score van 0 betekend dat de waarde gelijk is aan het gemiddelde. een z-score van 1 geeft aan dat een waarde 1 standaard deviatie afwijkt van het gemiddelde. Op het moment dat een z-score hoger is krijgt deze een hoge prioritering. Dit resulteert in een ranglijst van zorgaanbieders voor ieder risico waarbij de hoogste z-scores bovenaan staan.

Voorbeeld

Om dit verder te verduidelijken, wordt het voorbeeld uit 3.3.1 uitgebreid. Aangenomen wordt dat er veel meer zorgaanbieders zijn dan zorgaanbieders A, B, C. In dit voorbeeld wordt alleen naar de risico's X, Y gekeken.

Voor risico X is een gemiddelde meetwaarde bekend van 150. Hiervoor geld dat $z\text{-}score(150) = 0$. Verder is bekend dat $z\text{-}score(157) = 1$. Hierdoor krijgen de zorgaanbieders A, B, C een z-score van respectievelijk -0.24, 0.25 en -0.14. Voor risico Y geld dat overal waar de meetwaarde 1 staat gecontroleerd moet worden.

In de vorm van een ranglijst ziet dat er voor de z-scores van risico X als volgt uit: (tabel 3.3)

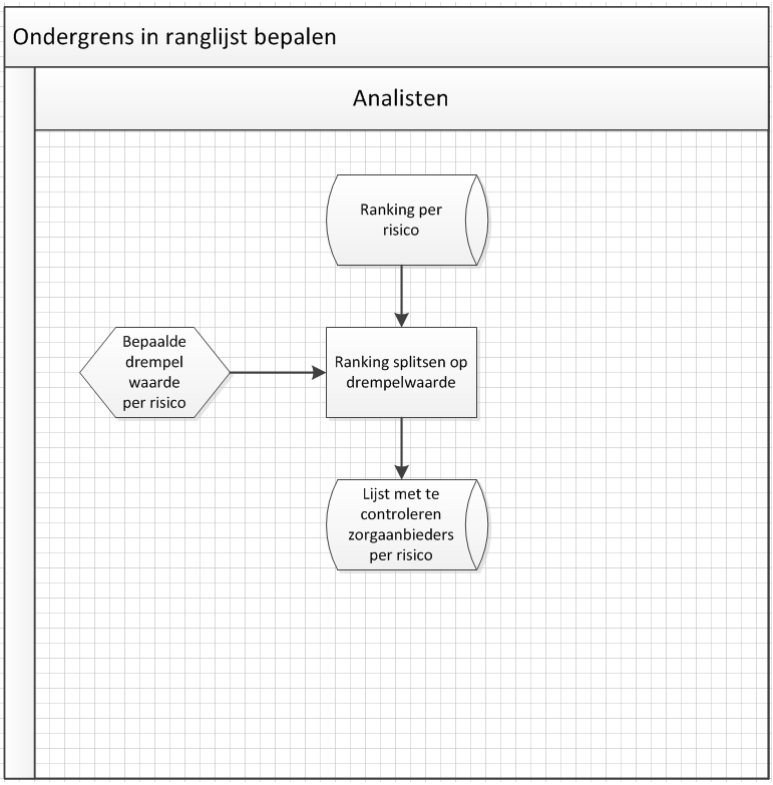
Zorgaanbieder	z-score X
G	0.38
E	0.28
B	0.25
D	0.15
F	-0.06
C	-0.14
A	-0.24

Tabel 3.3: Voorbeeld - Ranglijst

3.3.3 Ondergrens in ranglijst bepalen

Beschrijving

Voordat de ranglijsten per risico zijn gemaakt, wordt per risico een drempelwaarde bepaald waar de resultaten onder moeten zitten. Alle zorgaanbieders die boven deze vooraf bepaalde ondergrens zitten komen op een lijst om gecontroleerd te worden. Hierbij wordt voor de controle geen onderscheid gemaakt in het aantal zorgaanbieders.



Figuur 3.3: Ondergrens in ranglijst bepalen

Voorbeeld

Voor de risico's X, Y wordt nu per risico een ondergrens bepaald. Voor deze ondergrens wordt gekeken naar de ranking van risico X zoals beschreven in tabel 3.4. Voor dit risico is de ondergrens van een z-score van 0.25.

Deze waarde is per risico verschillend en wordt bepaald door de domeinexperts binnen Zilveren Kruis op basis van ervaring. Alle zorgaanbieders die een z-score hebben die groter is dan 0.25 moeten gecontroleerd worden.

In dit voorbeeld ziet de lijst er als volgt uit:

Zorgaanbieder	z-score(X)
G	0.38
E	0.28
B	0.25
D	0.15
F	-0.06
C	-0.14
A	-0.24

Tabel 3.4: Voorbeeld - Ranglijst met ondergrens

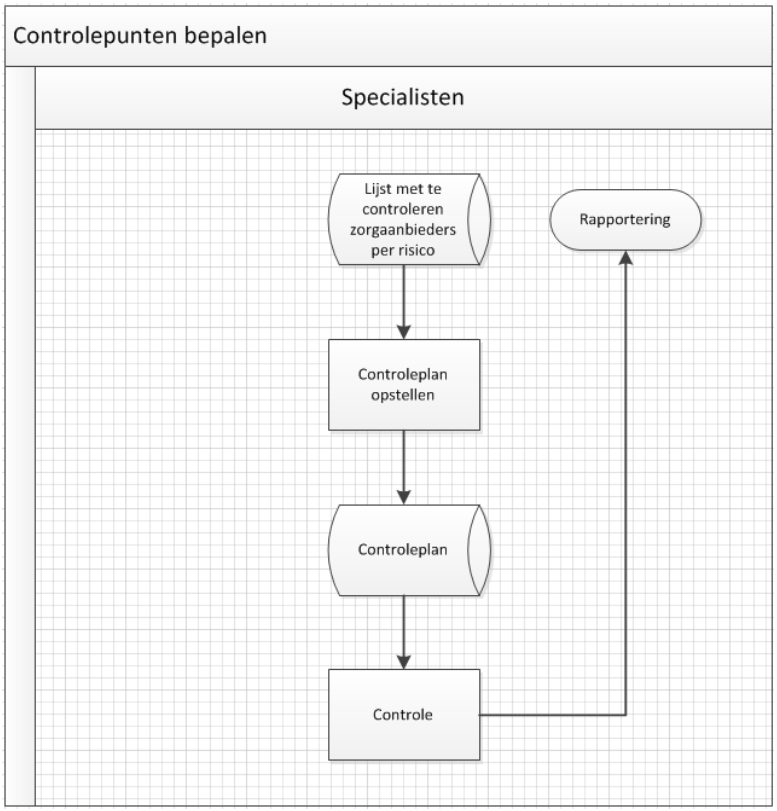
Hierbij moeten de zorgaanbieders G, E en B gecontroleerd worden op risico X. Dit proces wordt herhaald voor ieder risico waardoor er ook andere aanbieders op andere risico's gecontroleerd kunnen worden.

3.3.4 Controlepunten bepalen

Beschrijving

Nadat er voor ieder risico is bepaald welke zorgaanbieders er per risico gecontroleerd moeten worden wordt er per zorgaanbieder gecontroleerd op de verschillende risico's waar deze opvielen.

Op het moment van controle wordt een werkdocument bijgehouden waar per risico de terug te vorderen bedragen worden genoteerd. Van dit overzicht wordt alleen het totaal terug te vorderen bedrag opgeslagen in de rapportering. Dit zorgt er voor dat binnen de huidige situatie niet te bepalen is op welk risico een bedrag terug gevorderd wordt, dit probleem wordt later in detail besproken.



Figuur 3.4: Controlepunten bepalen

Voorbeeld

Voor dit voorbeeld wordt gekeken naar zorgaanbieder B. Van deze aanbieder is uit 3.3.3 bekend dat hij gecontroleerd gaat worden op risico X. Verder wordt aangenomen dat zorgaanbieder C ook verdacht is op risico W.

Dit betekent dat zorgaanbieder C op zowel risico X als risico W gecontroleerd gaat worden.

Hoofdstuk 4

Aanpak

4.1 Gebruikte software

Er is in dit onderzoek gekeken naar onderlinge verbanden tussen de resultaten van risico's. Daarvoor is gebruik gemaakt van drie softwarepakketten die voor dit doel beschikbaar zijn:

- Cortana [Leiden, 2017]: een datamining tool ontwikkeld door de Universiteit Leiden dat zich richt op het detecteren van subgroepen;
- SAS Enterprise Guide [SAS, 2018]: hiermee kan informatie verzameld en gekoppeld worden om datasets te genereren om te analyseren;
- Weka [Waikato, 2018]: hiermee kunnen op een snelle manier diverse modellen gebouwd worden, waarmee de data op een efficiënte manier geanalyseerd kunnen worden.

4.2 Data

In dit hoofdstuk wordt gekeken naar de beschikbare data voor dit onderzoek. Hierbij gaat het alleen om de informatie die gebruikt wordt binnen de GGZ. In dit hoofdstuk wordt gekeken hoe deze informatie verzameld wordt; hoe om wordt gegaan met fouten in de informatie, missende informatie en op wat voor manier deze gegevens verwerkt worden.

4.2.1 Data verzamelen

De informatie die is gebruikt in dit onderzoek is samengesteld uit verschillende plaatsen. Allereerst was informatie nodig om de voorspelling over te maken. Het is belangrijk om per zorgaanbieder op een zo laag

mogelijk niveau onrechtmatigheden te kunnen voorspellen.

De meest gedetailleerde informatie die op zorgaanbieders niveau beschikbaar is ziet er als volgt uit:

AGB_CODE	De AGB code van de zorgaanbieder
NAAM	De naam van de zorgaanbieder
JAAR	Het jaar van de controle
CONT_A	Uitkomst van de controle van risico A
CONT_B	Uitkomst van de controle van risico B
CONT_C	Uitkomst van de controle van risico C
RES_A	Resultaat in bedrag van controle A
RES_B	Resultaat in bedrag van controle B
RES_C	Resultaat in bedrag van controle C
Totaal	Totaal van de resultaten van alle controles

Tabel 4.1: Gegevens zorgaanbieders

De dataset bestaat uit drie delen. Het eerste deel (weergeven in het blauw) geeft identificerende informatie weer van de zorgaanbieder. Dit deel bestaat uit een unieke identificatiecode van de zorgaanbieder, de Algemeen GegevensBeheer-(AGB)code, de naam van de zorgaanbieder en het betreffende jaar.

In het tweede deel van de dataset (groene deel) staat of deze zorgaanbieder gecontroleerd is in het betreffende jaar. Hier zijn verschillende mogelijkheden:

- Er heeft geen controle plaatsgevonden voor een bepaald risico. In de dataset wordt dit weergegeven met "n.v.t.".
- Er is wel gecontroleerd maar er zijn geen onrechtmatigheden gevonden. In de dataset wordt dit weergegeven met "voldoende".
- Er is wel gecontroleerd en er zijn onrechtmatigheden gevonden. In de dataset wordt dit weergegeven met "onvoldoende".

Het laatste deel van de tabel (rode deel) bevat informatie over het resultaat van de controle. Als er geen controle heeft plaatsgevonden dan staat er n.v.t. in deze tabel. Op het moment dat er wel een controle heeft plaatsgevonden staat het resultaat van de controle in deze tabel. Dit resultaat is het bedrag in euro's dat is teruggevorderd als resultaat van het risico waarop gecontroleerd is. Tot slot staat het totaal teruggevorderde bedrag in deze tabel.

		Wel gecontroleerd	Niet gecontroleerd
Wel onrechtmatigheden gevonden	Controle	Onvoldoende	n.v.t.
	Resultaat	$BEDRAG > 0$	
Geen onrechtmatigheden gevonden	Controle	Voldoende	
	Resultaat	0	

Tabel 4.2: Voorwaarden werkdocument

Deze dataset wordt het **werkbestand** genoemd omdat dit de tussentijdse informatie is. Dit werkbestand bestaat uit een totaal van 81 zorgaanbieders waarbij minimaal 1 controle was uitgevoerd over een periode van twee jaar.

Deze informatie kan gebruikt worden als target voor verschillende analyses. Omdat dit alleen resultaten zijn, is meer informatie nodig om voorspellingen over te doen. Daarom wordt deze dataset aangevuld met een dataset vanuit de analyse-afdeling, de zogenaamde **zorgaanbiedersmatrix**. Deze data bevat beschrijvende informatie per zorgaanbieder. Deze data ziet er als volgt uit:

AGB_CODE	De AGB code van de zorgaanbieder
JAAR	Het jaar van de controle
SOM_DBC	Aantal verschillende behandeltrajecten
Klas_Risico_A	Resultaat (z-score) voor risico A
Klas_Risico_B	Resultaat (z-score) voor risico B
Klas_Risico_C	Resultaat (z-score) voor risico C
Score_Totaal	Totaal van de klassen van alle risico's

Tabel 4.3: Zorgaanbiedersmatrix

Ook deze dataset is verdeeld in drie delen. Het eerste deel wordt gebruikt om een zorgaanbieder te identificeren. Het tweede deel (groen) geeft aan hoeveel verschillende behandeltrajecten liepen in het desbetreffende jaar. Her derde deel geeft aan tot welke klasse een zorgaanbieder behoort. Dat werkt als volgt:

Voor iedere zorgaanbieder wordt een waarde bepaald die representatief is voor het risico (meer hierover in 3.3). Voor ieder risico wordt deze waarde vergeleken met die van vergelijkbare zorgaanbieders. Aan de hand van de afwijking op het gemiddelde wordt een ondergrens in standaard deviatie bepaald.

Als een zorgaanbieder als zelfstandige binnen een grotere praktijk werkt dan heeft de overkoepelende praktijk een eigen AGB code en heeft iedere zorgaanbieder een eigen AGB code. Bij het samenvoegen van de zorgaanbiedersmatrix met het werkbestand kwam het voor dat in werkbestand zowel de AGB code van de praktijk stond als van de individuele behandelaar. Ook kwam het voor dat er de code van de praktijk in het ene bestand stonden en van het individu in het andere bestand. Hierdoor waren de bruikbare rijen met informatie verminderd tot 69.

De laatste informatie die per zorgaanbieder bekend is, zijn de declaratiedata. Dit zijn gegevens die ontstaan omdat een zorgaanbieder declareert. Deze informatie ziet er als volgt uit:

Ook deze dataset is op te delen in drie stukken. Deze stukken zijn hier weer met kleuren aangegeven. Het eerste stuk (blauw) lijkt in eerste instantie hetzelfde als bij de zorgaanbiedersmatrix. Het enige verschil is dat

AGB_NR	De AGB nummer van de zorgaanbieder
JAAR	Het jaar van de controle
AANTL_VERZ	Het aantal patiënten minstens één ingediend declaratie
BEDRAG_DECL	Het totaal gedeclareerde bedrag
AANTL_VERG	Het aantal vergoedingen die zijn uitbetaald
PERC_AANTL_VERZ	Verdeling van het aantal verzekerde binnen een praktijk
PERC_BEDRAG_DECL	Verdeling van het gedeclareerde bedrag binnen een praktijk
PERC_AANTL_VERG	Verdeling van het aantal vergoedingen binnen een praktijk

Tabel 4.4: Declaratiedata

er in dit geval een AGB_NR staat in plaats van een AGB_CODE.

Een AGB_NUMMER is een verkorte versie van de AGB_CODE. Hierdoor was niet voldoende homogene informatie beschikbaar om alle AGB_CODES terug te vinden waardoor in de samengestelde dataset van het werkbestand, de risicomatrix en de declaratiedata 54 van de 81 zorgaanbieders terug te vinden zijn. Op deze problemen wordt dieper ingegaan in Sectie 5.1.3

Verder staat er in het groene blok van tabel 4.4 de volgende informatie:

- AANTL_VERZ: het aantal patiënten waarvoor er door de zorgaanbieder minstens één declaratie ingediend is;
- BEDRAG_DECL: het totaal gedeclareerde bedrag (zegt niets over de verdeling van dit bedrag over verschillende verzekerden);
- AANTL_VERG: het aantal vergoedingen die zijn uitbetaald.

In tabel 4.4 staat dat er 10 verzekerden zijn waarvoor 5 vergoedingen zijn geweest met een totaalbedrag van 1000 euro. Dit betekend niet dat 5 van de 10 verzekerden één vergoeding hebben laten opeisen. Het kan zijn dat meerdere vergoedingen zijn uitgekeerd als een verzekerde voor meerdere DBC's naar dezelfde zorgaanbieder is gegaan.

In het laatste (rode) deel staat informatie over de verhoudingen waarin de informatie uit het groene blokje is verdeeld binnen een praktijk. Dit wordt gedaan met de volgende attributen:

- PERC_AANTL_VERZ
- PERC_BEDRAG_DECL
- PERC_AANTL_VERG

Hierbij is PERC_AANTL_VERZ de verdeling van de verzekerde die naar die praktijk gaan. Dit is gedaan op de volgende manier:

$$PERC\_AANTL\_VERZ_{Huidigezorgaanbieder} = \frac{AANTL\_VERZ_{Huidigezorgaanbieder}}{\Sigma_n AANTL\_VERZ} \cdot 100\%$$

Waarbij n gelijk is aan het aantal zorgaanbieders binnen de praktijk. In andere woorden: PERC_AANTL_VERZ is het aandeel van de totale verzekerde patiënten die een zorgaanbieder behandelt binnen een praktijk.

Dit geld op de zelfde manier voor PERC_BEDRAG_DECL en PERC_AANTL_VERG. Hier is PERC_BEDRAG_DECL gedefinieerd als:

$$PERC\_BEDRAG\_DECL_{Huidigezorgaanbieder} = \frac{BEDRAG\_DECL_{Huidigezorgaanbieder}}{\Sigma_n BEDRAG\_DECL} \cdot 100\%$$

Waarbij n gelijk is aan het aantal zorgaanbieders binnen de praktijk. PERC_AANTL_VERG kan wiskundig gedefinieerd worden als:

$$PERC\_AANTL\_VERG_{Huidigezorgaanbieder} = \frac{AANTL\_VERG_{Huidigezorgaanbieder}}{\Sigma_n AANTL\_VERG} \cdot 100\%$$

Ook hier is n het aantal zorgaanbieders binnen de praktijk. Als er maar één zorgaanbieder in een praktijk zit (dit gebeurt als een zorgaanbieder niet in een groepspraktijk zit) dan zullen alle PERC variabele gelijk zijn aan 100.

4.2.2 Data voorbereiden

Op het moment dat alle gegevens zijn samengevoegd is er een dataset waarin de volgende attributen staan:

Soort	Aantal per aanbieder	Korte beschrijving
AGB_CODE	1	De identificerende sleutel van de aanbieder
CONT_RISICO	10	Per risico per jaar wordt aangegeven of hierop gecontroleerd is en wat de uitslag is Hierbij is CONT_A de controle op risico A
RES_RISICO	10	Per risico wordt aangegeven wat het teruggevorderde bedrag is. Hierbij is RES_A het resultaat van risico A
SOM_DBC	1	De totale hoeveelheid Diagnose Behandeling Combinaties.
KLASSEN	16	De scores per klasse waarin een zorgaanbieder is ingedeeld volgens de uitkomst van de spiegel methode.
AANTALLEN	3	Het totaal aantal vergoedingen, verzekerde en totaalbedrag uit de declaratiedata.
PERCENTAGES	3	De relatieve percentages van het totaal van een praktijk.
Totaal	**44**	

Tabel 4.5: Samengevoegde dataset

In deze dataset valt op dat er erg veel attributen (44) zijn voor het aantal rijen (54) waar informatie over is. Dit zorgt er voor dat de kans op overfitten erg groot is [Hall and Holmes, 2003]. Om dit tegen te gaan moet de informatie op een andere manier ingedeeld worden. Daarbij mag de AGB_CODE niet meegenomen worden als voorspellende variabele omdat dit een identificerende variabele is.

Om te zorgen dat het aantal attributen teruggedrongen wordt, kan nagedacht worden over de hoeveelheid toegevoegde informatie die een attribuut heeft op het moment dat andere attributen al worden meegenomen. Zo is te zien dat alle controles van risico's alleen zeggen of iets voldeed of niet, of dat hier geen informatie

over is. Een overzicht hiervan is te vinden in Tabel 4.2.

Hierin staat dat wanneer niet gecontroleerd op risico A de attributen CONT_A en RES_A allebei op "n.v.t." staan. De informatie die CONT_A en RES_A toevoegen is in deze situatie gelijk. Ook is te zien dat als er wel gecontroleerd is en er geen onrechtmatigheden gevonden zijn, de waardes van CONT_A en RES_A gelijk zijn aan respectievelijk "voldoende" en "0". Daarom levert het hebben van beide attributen in dit geval ook geen extra informatie.

Als er wel onrechtmatigheden zijn aangetroffen, staat CONT_A altijd op "onvoldoende" en zal RES_A altijd groter zijn dan 0. Ook hier levert het hebben van beide attributen niet meer informatie. Daarom kan CONT_RISICO uit de dataset gehaald worden. Gekozen wordt om CONT_RISICO uit de dataset te halen en niet RES_RISICO omdat RES_RISICO meerdere waardes kan hebben waar CONT_RISICO maar drie mogelijke waardes heeft.

Hierdoor bestaat de dataset niet meer uit 44, maar uit 34 attributen. Omdat de AGB_CODE niet meegenomen kan worden als attribuut, bestaat de dataset uiteindelijk uit 33 attributen.

Het is verder niet mogelijk om attributen te verwijderen zonder informatie te verliezen. Hierbij is nog niet vastgesteld welke informatie relevant is en welke niet. Het is wel een probleem dat "n.v.t." in de data staat. Deze waarde betekent dat er geen controle is geweest en dat daardoor ook onbekend is of er zich onrechtmatigheden afspelen voor dat risico. Het is niet wenselijk dat dit attribuut voorspellende waarde krijgt.

Als deze gegevens behandeld worden als waarde komt het voor dat de waarde "n.v.t." de nauwkeurigheid kan verhogen. In andere woorden, het missen van informatie draagt dan bij aan de voorspelling. Om dit probleem te verhelpen zijn alle gevallen waar "n.v.t." staat, vervangen door een missing value. Dit laat zich representeren door een vraagteken. Dit zorgt er ook voor dat alle attributen in de dataset nu bestaan uit numerieke waardes of missing values.

De nieuwe dataset zou er nu als volgt uit kunnen zien:

RES_1_2013[1]	...	RES_5_2013	RES_1_2014	...	RES_5_2014	Andere Informatie Zorgaanbieder
100	...	40	0	...	30	Andere Informatie Zorgaanbieder
?	...	45	?	...	25	Andere Informatie Zorgaanbieder
110	...	?	35	...	?	Andere Informatie Zorgaanbieder
90	...	?	78	...	?	Andere Informatie Zorgaanbieder

Tabel 4.6: Nieuwe dataset

In deze dataset komt het voor dat niet ieder risico is voorzien van een resultaat. Dit komt doordat niet op ieder risico is gecontroleerd bij iedere zorgaanbieder. Toch is het wenselijk een voorspelling te doen die iets

[1]Resultaat van de controle over risico 1 in 2013

zegt over een risico. Dit kan alleen op het moment dat dit risico een resultaat heeft. Daarom wordt deze grote dataset opgesplitst in vijf kleinere datasets.

In deze datasets worden alleen rijen uit de originele dataset opgenomen als deze geen missing values hebben in de target variabele. Als de dataset van het voorbeeld gesplitst zou worden per risico zouden de volgende datasets ontstaan:

RES_1_2013	...	RES_5_2013	RES_1_2014	...	RES_5_2014	Andere Informatie Zorgaanbieder
100	...	40	0	...	30	Andere Informatie Zorgaanbieder
110	...	?	35	...	?	Andere Informatie Zorgaanbieder
90	...	?	78	...	?	Andere Informatie Zorgaanbieder

Tabel 4.7: RES_1_DATASET

RES_1_2013	...	RES_5_2013	RES_1_2014	...	RES_5_2014	Andere Informatie Zorgaanbieder
100	...	40	0	...	30	Andere Informatie Zorgaanbieder
?	...	45	?	...	25	Andere Informatie Zorgaanbieder

Tabel 4.8: RES_5_DATASET

In deze datasets zullen de attributen waarover de voorspelling wordt gedaan altijd gevuld zijn. Hierdoor worden minder missing values meegenomen in de analyse. Tot slot wordt deze informatie opgeslagen in Attribute-Relation File Format (ARFF) formaat zodat deze informatie goed ingelezen kan worden door de verschillende datamining programma's zoals beschreven in 4.1.

4.2.3 Data verwerken

De manier waarop de gegevens verwerkt worden is afhankelijk van het doel van de analyse. In het hoofdstuk 5 worden verschillende analyses toegepast met verschillende doeleindes. Hier zal per analyse aangegeven worden op welke manier de informatie verwerkt wordt.

4.3 Procesinformatie

Voor de proces beschrijving uit 3.3 is op verschillende manieren informatie verzameld. Hierbij is eerst gekeken naar de beschikbare procestekeningen. Deze tekeningen waren op basis van taken ingericht en waren niet meer helemaal up-to-date. Om dit op te lossen zijn er verschillende interviews afgenomen met iedereen die een taak heeft in het proces.

Hierin werd gevraagd hoe bepaalde stappen in het proces nu ingericht waren. De gegeven antwoorden zijn

gevalideerd in een vervolginterview. Met deze informatie is een nieuwe procesbeschrijving opgesteld. In deze beschrijving staat niet de taak maar de informatie centraal. Hiervoor is gekozen omdat er daardoor vanuit gegevensperspectief beslissingen in kaart gebracht kunnen worden om deze vervolgens te verbeteren.

De aangepaste procestekening is weergegeven in Appendix B.

Hoofdstuk 5

Procesanalyse

Hieronder staan de procesgerelateerde verbeteringen beschreven. Deze verbeteringen hebben dezelfde indeling als Sectie 3.3. Hierdoor kan er een vergelijking worden gemaakt tussen de huidige situatie en de beschreven verbeteringen.

5.1 Zorgaanbiedereigenschappen naar risico's

Zoals te lezen in 3.3.2 wordt in de huidige situatie een score gegeven aan alle risico's waarop gecontroleerd wordt. Deze score wordt dan gebruikt als referentiewaarde om te kijken of de desbetreffende zorgaanbieder een relatief hoge (of lage) score heeft. Als deze score ver afwijkt wordt de zorgaanbieder gecontroleerd.

In de nieuwe situatie is het niet de bedoeling te kijken waar de grootste afwijking van deze score in zit maar is het doel te voorspellen wat het verwachte bedrag is dat te wijten is aan onrechtmatigheden. Deze voorspelling maakt het mogelijk in Sectie 5.2 om op verschillende manieren een prioritering mee te geven aan de te maken ranking.

Om dit op een systematische manier te doen, worden zes verschillende stappen doorlopen waarin de transitie van data (onverwerkte informatie) naar een verwacht terug te halen bedrag in euro per risico per zorgaanbieder gemaakt wordt. De volgende stappen worden doorlopen:

1. Domeinkennis en beperkingen
2. Dataselectie
3. Data pre-processing
4. Datamining
5. Interpretatie van de informatie

5.1.1 Domeinkennis en beperkingen

Wanneer een dataset beschikbaar is voor analyse zijn er diverse methodieken die gebruikt kunnen worden [Bramer, 2013]. Deze methodieken zijn in te delen in twee categorieën: black-box modellen en white-box modellen. Bij black-box modellen is het niet mogelijk op een gemakkelijke manier te kijken waarom een algoritme tot een bepaald antwoord is gekomen. Bij een white-box model is het wel mogelijk dit op een gemakkelijke manier te achterhalen.

Binnen de context van Zilveren Kruis is het belangrijk de voorspellingen te kunnen achterhalen omdat ze uitgelegd moeten kunnen worden aan de zorgaanbieders die gecontroleerd worden en om te voldoen aan eventuele regulatie die er nu is of in de toekomst zal ontstaan. Ook is het natuurlijk wenselijk vanuit het Zilveren Kruis om te weten welke attributen belangrijk zijn voor het controleren van de zorgaanbieders.

Een andere reden om een traceerbaar model te gebruiken is voor inzicht en validatie. Door te zien welke attributen belangrijk zijn binnen een model kan hier op andere manieren op gestuurd worden. Verder is er binnen Zilveren Kruis een grote hoeveelheid specialistische kennis. Door gebruik te maken van deze kennis zouden bijzonderheden binnen een model geïdentificeerd en gecorrigeerd kunnen worden.

Een deel van de kennis die er al binnen de analyse van Zilveren Kruis aanwezig is, is op dit moment verwerkt in het slim genereren van de scores. het is daarom verstandig deze menselijke kennis te gebruiken als aanvulling. Een model dat gebruik kan maken van de huidige analyse uitkomsten (en deze weegt naar relevantie) zou ervoor zorgen dat een methodiek gecreëerd wordt die de huidige werkwijze completeert.

Hiermee wordt een situatie bereikt waarin de meest relevante onderdelen voor de voorspelling worden geselecteerd om een zo nauwkeurig mogelijk resultaat te bereiken.

5.1.2 Dataselectie

De gegevens die voor dit onderdeel worden gebruikt, zijn de gegevens zoals beschreven in Sectie 4.2.1. Hier gaat het om de samengestelde dataset op zorgaanbiedersniveau. Hierbij wordt voorspeld wat het terug te vorderen bedrag per zorgaanbieder is in het jaar 2014. Dit wordt gedaan voor ieder risico apart in plaats van op een geaggregeerd niveau. Hierdoor wordt op risiconiveau inzicht verkregen over de zorgaanbieders.

5.1.3 Data pre-processing

Om te zorgen dat de gebruikte informatie klaar is om te analyseren, moet deze op een aantal factoren worden gecontroleerd. Binnen dit onderzoek wordt onderscheid gemaakt in vijf factoren die de kwaliteit van de data beïnvloeden.

1. Ambiguïteit in de data
2. Duplicaten in de data
3. Normalisatie van AGB codes
4. Compleetheid van de data
5. Onmogelijkheden in de data

De kwaliteit van de data moet zo hoog mogelijk zijn. Dit omdat de kwaliteit van de voorspelling gelimiteerd is door de kwaliteit van de data waarover de voorspelling wordt gemaakt. De kwaliteit van de data is hoog als al deze factoren geoptimaliseerd zijn. Om per factor aan te geven waar rekening mee gehouden moet worden, wordt hieronder iedere factor individueel behandeld.

Ambiguïteit in de data

Binnen een dataset mogen de data geen ambiguïteiten vertonen. Op het moment dat de data op verschillende manieren te interpreteren is zullen (1) verschillende individuen andere inzichten uit de analyse halen en (2) zal de data aan de hand van persoonlijke interpretatie ingevuld worden.

Zo worden in de dataset van Zilveren Kruis de begrippen “n.v.t.” en “0” door elkaar gebruikt. Dit wordt zichtbaar bij het terug gehaalde bedrag, dit is een waarde die met de hand wordt ingevuld in het werkdocument. Sommige werknemers binnen Zilveren Kruis zijn van mening dat als een zorgaanbieder geen geld terug hoeft te geven omdat er geen sprake is van onrechtmatigheden de terugvordering niet van toepassing is. Andere medewerkers binnen Zilveren Kruis zien dit als een terugvordering die gelijk is aan 0 euro.

Voor beide invullingen is wat te zeggen, alleen zorgt deze inconsequentie er wel voor dat er een verschil ontstaat binnen de data. Daarom is het belangrijk om dit op een uniforme manier in te vullen, omdat anders de voorspellende waarde van de gegevens omlaag gaat.

Om dit probleem op te lossen is er binnen deze dataset het volgende gedaan: alle waardes die niet van toepassing waren maar wel gecontroleerd zijn, hebben een terugvorderingswaarde van 0 gekregen. Alle waardes die niet van toepassing waren en niet gecontroleerd waren hebben een ‘?’ gekregen als waarde, dit omdat deze waarde er niet is (missing value).

Duplicaten in de data

Een duplicaat binnen de data kan doorgaans twee dingen betekenen. Het kan zijn dat dezelfde regel twee keer in de dataset staat, in dit geval wordt deze regel als twee keer verwerkt waardoor de zorgaanbieder in kwestie als twee keer zo belangrijk wordt verwerkt. Een andere mogelijkheid is dat de gegevens gefragmenteerd opgeslagen zijn. Hierdoor wordt de voorspellende waarde van de gegevens lager zonder dat dat nodig zou zijn omdat de gegevens wel allemaal aanwezig zijn. Om dit visueler te maken staat in tabel 5.1 een voorbeeld. Binnen de geselecteerde dataset van zilveren kruis waren overigens geen duplicaten.

Unieke sleutel	Attribuut 1	attribuut 2
ABC123	100	Yes
ABC123	100	Yes
ABC124	?	No
ABC124	50	?
ABC125	73	No

(a) Met duplicaten

→

Unieke sleutel	Attribuut 1	attribuut 2
ABC123	100	Yes
ABC124	50	No
ABC125	73	No

(b) Zonder duplicaten

Tabel 5.1: Duplicaten in de data

Normalisatie van AGB codes

Wanneer verschillende datasets samengevoegd worden, is het belangrijk dat de unieke identificatie van een zorginstelling door alle datasets heen het zelfde is. Als dit niet volgens de zelfde unieke sleutel gebeurt dan is het vaak niet meer mogelijk om op grote schaal verschillende datasets samen te voegen.

Daarom moet één sleutel worden afgesproken voor alle datasets. Ook als deze sleutel om wat voor reden dan ook verandert, moet deze sleutel, bijvoorbeeld met behulp van een conversietabel, om te zetten zijn naar de sleutel uit het verleden. Op dit moment is dit helaas niet het geval binnen Zilveren Kruis. Zo wordt zowel met "AGB codes" als met "AGB nummers" gewerkte. Ook zijn de "AGB codes" en "AGB nummers" niet op een eenduidige manier om te zetten.

In de meeste gevallen is het zo dat een AGB_NR hetzelfde is als de eerste 8 karakters van een AGB_CODE, helaas gaat dit niet voor alle gevallen op. Het gevolg hiervan is dat koppelingen tussen verschillende informatiebronnen niet gemaakt kunnen worden.

Compleetheid van de data

Een dataset is compleet als er binnen deze dataset geen waardes ontbreken. Op het moment dat er waardes ontbreken, zijn er verschillende mogelijkheden [Grzymala-Busse and Hu, 2000]. Veel gebruikte methodes zijn het verwijderen van de rij waarin een waarde mist, het combineren van onvolledige rijen en het gemiddelde

uit de dataset nemen als invulling voor de missende waarde.

Ook zijn er algoritmes die zelf beslissen wat de meest efficiënte methode is om met ontbrekende waardes om te gaan. Het nadeel van het invullen van de ontbrekende waardes is dan een aanname wordt gemaakt over de gegevens. Wat de invloed van deze aannames is, is afhankelijk van de hoeveelheid missende waardes in de gegevens.

Kijkend naar het risico hoofdbehandelaarschap wordt het aandeel missende waardes binnen die dataset als volgt bepaald:

$$percentagemissingvalues = \frac{aantalmissingvalues}{aantalattributen \cdot aantalzorgaanbieders} \cdot 100\%$$

Bij het toepassen van de formule op de dataset van Zilveren Kruis blijkt dat 6.5% van de gegevens van het risico hoofdbehandelaarschap ontbreekt. Wanneer deze waardes ingevuld worden op een van de hierboven beschreven methodieken, wordt een aannamen gemaakt over 6.5% van de dataset. Omdat dit een relatief hoog percentage is om aannames over te doen, wordt dit niet gedaan.

Onmogelijkheden in de data

Wanneer een zorgaanbieder als voldoende door de controle komt maar er toch een bedrag wordt teruggevorderd, is sprake van een onmogelijkheid in de data. In dit geval zou deze zorgaanbieder of ten onrechte als voldoende gemarkeerd worden of is er ten onrechte een bedrag teruggevorderd. Hoewel dit niet in de bekeken dataset voorkomt, is het wel belangrijk om hier bewust van te zijn en dit te controleren.

5.1.4 Datamining

Als de gegevens klaar zijn gemaakt voor gebruik kan een voorspellende techniek toegepast worden. Hierbij moet rekening gehouden worden met de beperkingen uit Sectie 5.1.1. Bij het voorspelen van een waarde voor de onrechtmatigheden per jaar per risico zullen niet alle attributen even veel voorspellende waarde hebben. Om te bepalen welke attributen relevant zijn voor de voorspelling wordt voorafgaand aan de voorspelling bepaald welke attributen de meeste relevantie hebben voor het voorspellen.

Dit fenomeen staat bekend als “Attribute selection”, hiervoor bestaan verschillende manieren om dit te doen, gespecialiseerd op verschillende taken [Hall and Holmes, 2003]. In dit onderzoek moet een waarde voorspeld worden. Dit type probleem wordt ook wel een regressie probleem genoemd.

Om de meeste relevante attributen te bepalen voor een regressieprobleem wordt vaak gekozen om (1) een gulzige methode [Caruana and Freitag, 1994] (greedy method) toe te passen of (2) gebruik te maken van het M5 algoritme [Quinlan, 1992]. Omdat het M5 algoritme gebruik maakt van een gulzig algoritme, wordt

hieronder een beschrijving gegeven van een implementatie van het M5 algoritme voor attribuut selectie.

Stel dat een dataset bestaat uit 4 voorspellende attributen; A, B, C en D. Uit deze set moeten de meest efficiënte attributen gehaald worden. Iedere mogelijke combinatie van attributen heeft een bijbehorende nauwkeurigheid van de voorspelling. De hoeveelheid verschillende combinaties is te berekenen met de formule: 2^k waarbij k het aantal attributen binnen de dataset is. In het eerder genoemde voorbeeld zijn dus $2^4 = 16$ verschillende mogelijkheden. Deze mogelijkheden zijn grafische weergeven in Figuur 5.1.

Binnen het voorbeeld is het nog te doen om alle 16 gevallen door te rekenen en daaruit de beste voorspelling te bepalen. In de echte dataset gaat het over 33 attributen, dit betekent dat $2^{33} = 8.589.934.592$ verschillende modellen met elkaar vergeleken zouden moeten worden. Er van uitgaande dat een model in een honderdste van een seconde opgesteld zou kunnen worden dan zou dit alsnog meer dan 163 jaar duren. Daarom wordt een gulzige zoekmethode binnen het M5 algoritme gebruikt om te selecteren welke uitkomsten met elkaar vergeleken worden.

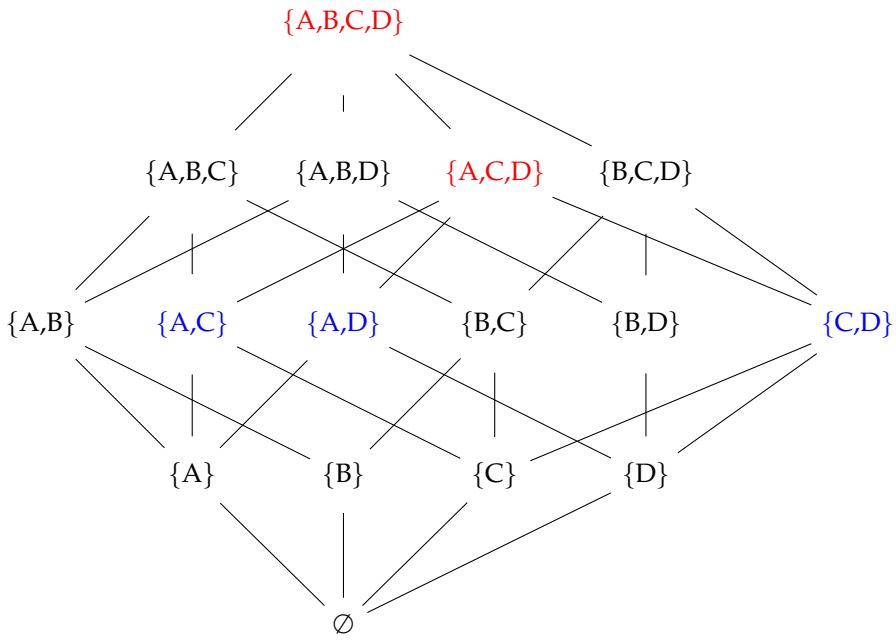
Daarom wordt niet iedere optie doorgerekend maar wordt een selectie gemaakt binnen deze mogelijke uitkomstruimte. Daartoe wordt eerst een score berekend voor de dataset met alle attributen. Hoe deze score wordt berekend wordt later in dit hoofdstuk besproken. Als deze score is berekend voor de dataset met k attributen wordt deze vergeleken met de scores van alle mogelijke datasets met $k-1$ attributen. In dit voorbeeld de dataset met k attributen {A,B,C,D} en zijn de mogelijke opties van datasets met $k-1$ attributen {A,B,C}, {A,B,D}, {A,C,D} en {B,C,D}. Hier geldt dat een lagere score correspondeert met een betere attribuut selectie.

Voor deze 4 attribuut selecties wordt degene met de laagste score genomen als beste optie. Als de dataset met de laagste score meer attributen heeft dan de samenstelling van attributen waarmee deze vergeleken wordt stopt de selectie en wordt de samenstelling met deze lage score gekozen. Mocht een van de combinaties van attributen met minder attributen een lagere score hebben dan wordt deze dataset gekozen om verder te zoeken.

Volgens het voorbeeld betekent dit dat als de set {A,B,C,D} een lagere score zou hebben dan {A,B,C}, {A,B,D}, {A,C,D} of {B,C,D} set {A,B,C,D} gekozen zou worden. Stel dat dit in het voorbeeld niet het geval is. In het voorbeeld heeft de set {A,C,D} de laagste score. Dit betekent dat vanuit de set {A,C,D} verder gekeken gaat worden naar mogelijkheden met minder attributen. Hier gaat het alleen om mogelijkheden die een subset zijn van {A,C,D}, deze zijn in het blauw aangegeven in Figuur 5.1.

Op dit moment wordt de vergelijking gemaakt tussen de voorlopige beste keuze en alle subsets daarvan. Hierbij wordt alleen gekeken naar de subsets die precies 1 attribuut minder bevatten. Na deze scores berekend te hebben, worden deze weer met elkaar vergeleken om de laagste score te vinden. Gesteld dat in het voorbeeld de set {A,C,D} een lagere score heeft dan {A,C}, {A,D} en {C,D}. Dit betekent dat de attributen A, C en D

gekozen worden om te gebruiken binnen het model.

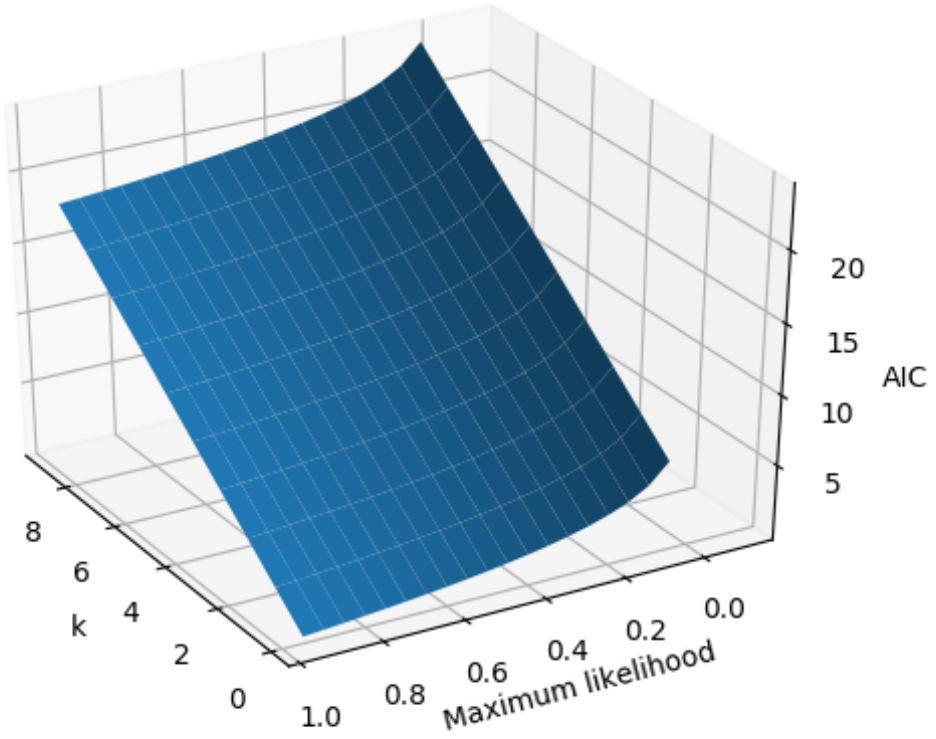


Figuur 5.1: Zoekruimte attribuut selectie

De scores kunnen op verschillende manieren bepaald worden. In dit onderzoek wordt een methode toegepast die is gebaseerd op de Akaike information criterion (AIC) [Burnham and Anderson, 2004] De score bij deze methode wordt als volgt bepaald:

$$AIC = 2k - 2\ln(\hat{L})$$

Hierbij is k het aantal attributen binnen een geteste set en $\hat{L}$ de likelihood [Akaike, 1987] gegeven deze attributen. De $\hat{L}$ is een maat die aangeeft hoe goed het berekende voorspelmodel bij de geobserveerde data past. Iedere combinatie van attributen heeft een verschillend voorspelmodel en daarmee ook een verschillende $\hat{L}$. Deze waarde wordt bepaald door de uitkomst van het regressiemodel te vergelijken met de gegevens waarmee dit model gemaakt is. Hoe beter de gegevens het voorspelmodel representeren hoe hoger de $\hat{L}$. Een grafische weergave van dit verband is te vinden in Figuur 5.2. Op basis van een voorspellend model wordt de score bepaald.



Figuur 5.2: Akaike information criterion

Nadat de optimale attributen zijn gevonden worden deze gebruikt om het voorspellingsmodel te maken. Dit wordt gedaan met lineaire regressie [Neter et al., 1996] door de gekozen attributen te nemen uit de selectie. Het voorbeeld van hierboven volgend zou dat betekenen dat er naar de attributen A, C en D wordt gekeken. Met behulp van deze attributen wordt geprobeerd om een zo een nauwkeurig mogelijke voorspelling op te stellen van een target.

In het geval van de echte dataset is het target de kosten van een risico per zorgaanbieder. In het theoretische voorbeeld werd het target T genoemd. Dit geeft voor het voorbeeld een model dat er als volgt uit ziet:

$$T = x_1 A + x_2 C + x_3 D + Const$$

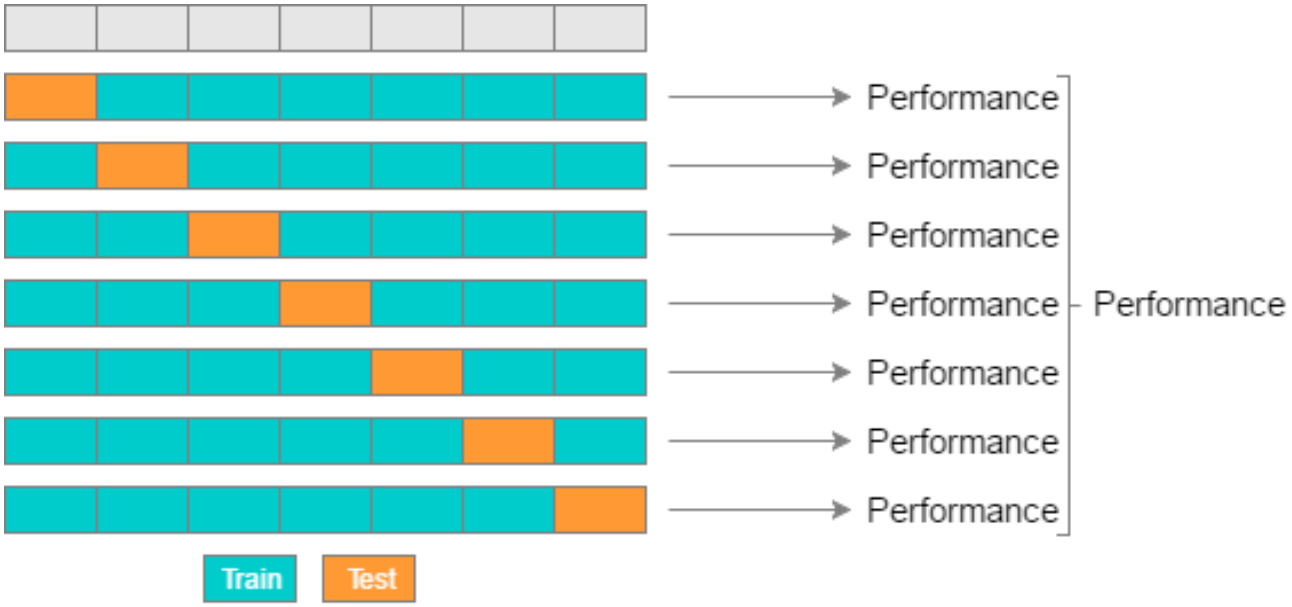
Hierbij is T het target attribuut en zijn A,C en D de attributen met de sterkste voorspellende waarde volgens de attribuut selectie zoals hierboven beschreven. $Const$ staat voor een constante waarde die wordt gebruikt als gemiddelde waarde die verwacht wordt wanneer geen van de attributen aanwezig is.

Vanaf nu wordt gekeken naar het resultaat van deze methode op de datasets zoals beschreven in Sectie 4.2.1. Deze analyse wordt uitvoerd op de dataset die zich richt op het risico hoofdbehandelaarsschap. Binnen deze

dataset wordt getracht op basis van de beschikbare gegevens het bedrag te voorspellen of per zorgaanbieder onrechtmatig gedeclareerd is.

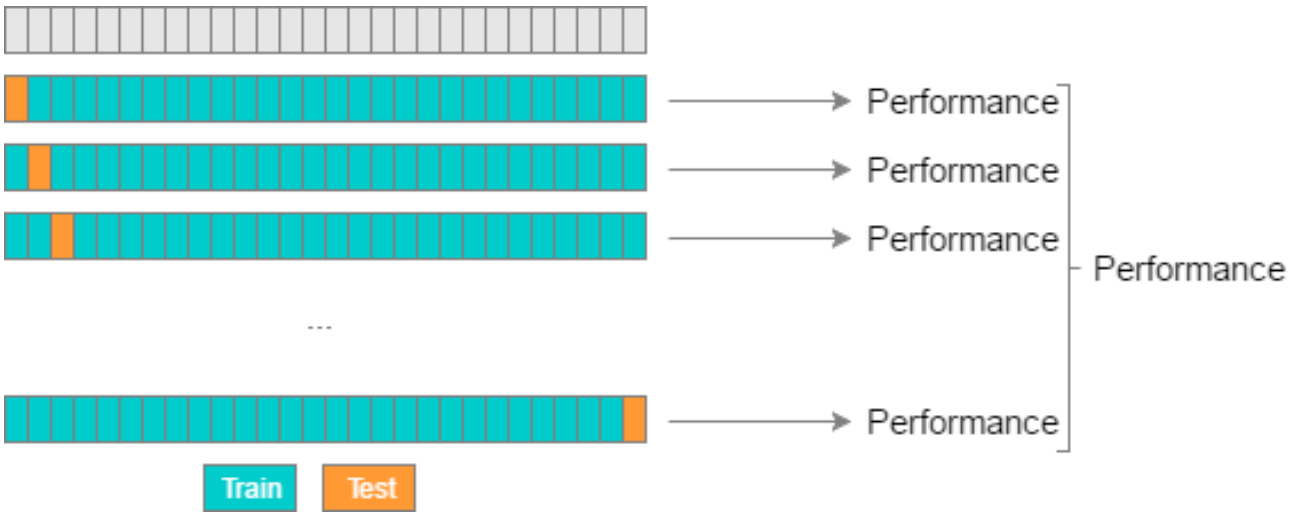
Naast de hierboven beschreven methodieken worden de resultaten gevalideerd met "leave-one-out cross-validatie" [Refaeilzadeh et al., 2016]. Om dit concept goed te begrijpen wordt eerst uitgelicht wat "k-fold cross-validatie" [Kohavi et al., 1995] is. Bij k-fold cross-validatie wordt de beschikbare dataset opgedeeld in k gelijken delen. Hierbij wordt steeds één deel achterwege gelaten, zodat er later een berekend model mee getoetst kan worden. Dit deel wordt we de test data genoemd. De overig $k-1$ delen worden samengevoegd en hierop wordt vervolgens het model getraind. Dit is nadat de attribute selection is uitgevoerd.

Hieruit volgt een model door middel van lineaire regressie, welke vervolgens getoetst wordt op de test data. Dit proces wordt k keer herhaald, waarbij steeds één deel als test data fungeert. Een visuele representatie staat in figuur 5.3. Van ieder deel wordt bijgehouden hoe goed het model het doet, van deze scores wordt het gemiddelde genomen. Dit is de uiteindelijke score van het model.



Figuur 5.3: k-fold cross validatie [Bonada, 2017a]

Door op deze manier de voorspelling te valideren, krijg je een realistischere weergave van de gemaakte fout. Het aantal splitsingen waarin een dataset opgedeeld kan worden is ten hoogste gelijk aan het aantal rijen binnen deze dataset. Dit omdat een rij in een dataset een atomair gegeven is, een model kan niet getest worden op een halve rij informatie. Deze speciale situatie is leave-one-out cross-validatie, hierbij wordt 1 rij genomen om het model op te testen en de rest van de gegevens worden gebruikt om het model te trainen (zie figuur 5.4). Dit is in dit geval mogelijk omdat hier sprake is van een kleine dataset. Als de dataset groter wordt, betekent dit dat er meer modellen moeten worden getraind, een actie die veel rekenkracht kost. Daarom wordt in veel situaties gekozen om 10-fold cross-validatie toe te passen.



Figuur 5.4: leave one out cross validatie [Bonada, 2017b]

De resultaten van de analyse zoals beschreven in 5.1.4 zijn de volgende:

===Linear Regression Model===

$$Res\_Hoofdbehandelaarschap\_2014 =$$
$$2.2121 \cdot Res\_Verwijzing\_een + 0.1097 \cdot Res\_Hoofdbehandelaarschap\_een - 0.9705 \cdot Res\_Verblijfsdagen\_een -$$
$$0.011 \cdot som\_bdrg + 110.5458 \cdot som\_dbc + 98599.6217 \cdot klas\_klinamb - 15513.2579 \cdot klas\_up\_vrblfzw -$$
$$27891.5266 \cdot klas\_famtraj - 11193.1243 \cdot klas\_ndl\_serie + 8409.2691 \cdot klas\_indirect + 10564.676 \cdot klas\_kindertijd -$$
$$3375.6442 \cdot score\_tot - 20972.8138 \cdot dummy\_sort - 72.9629 \cdot ranking - 17.5169 \cdot aantal\_verg - 3106.9667 \cdot$$
$$perc\_bedrag\_decl + 3762.2743 \cdot perc\_aantal\_verz + 0.0171 \cdot bedrag\_decl - 6577.87 \cdot LN\_bedrag\_decl + 51062.7852$$

=== Summary ===

Correlation coefficient	0.0205
Mean absolute error	33287.9562
Root mean squared error	57540.8882
Relative absolute error	198.7763%
Root relative squared error	193.0006%
Total Number of Instances	54

Tabel 5.2: data summary

Het resultaat is opgedeeld in twee delen. In het eerste deel staat het lineaire regressie model met bijbehorende waardes voor de gegeven dataset. In het tweede deel staan beschrijvende statistieken over deze voorspelling. Hieronder volgt een uitleg van de beschrijvende statistieken uit de *summary*. In sectie 5.1.5 wordt inhoudelijk ingegaan op deze waardes.

Correlation coefficient

De correlatie coëfficiënt verwijst hier naar Pearson's correlatie coëfficiënt [Benesty et al., 2009]. Dit getal is een waarde tussen de -1 en 1 en geeft aan hoe sterk de correlatie is tussen de lijn zoals getekend door het lineaire regressiemodel en de daadwerkelijke waardes. Een waarde van 1 betekent dat er een perfecte correlatie is tussen de voorspelling en de realiteit. Een waarde van -1 betekent dat de voorspelling het tegenovergestelde voorspelt van de realiteit. Een waarde van 0 betekent dat de voorspelling en de realiteit niet gecorreleerd zijn.

Mean absolute error

De mean absolute error (MAE) [Willmott and Matsuura, 2005a] geeft aan hoe groot het gemiddelde verschil is tussen de voorspelde waarde en de echte waarde. Hierbij wordt geen rekening gehouden of deze fout positief of negatief is. Alleen de afstand tot de echte waarde wordt gebruikt voor de MAE.

Root mean squared error

De root mean squared error (RMSE) [Hyndman and Koehler, 2006] heeft veel weg van de MAE. Het verschil tussen deze twee methodes is dat de RMSE het kwadraat van de gemiddelde afwijking aangeeft. Dit is voornamelijk nuttig wanneer het maken van grote fouten erger is dan is dan het maken van fouten in het algemeen.

Relative absolute error

De relative absolute error (RAE) [Willmott and Matsuura, 2005b] beschrijft in welke mate de MAE afwijkt van de absolute variatie binnen de gegeven dataset. In dit geval wordt de absolute variatie gegeven door $\frac{\Sigma|\bar{\theta} - \theta_i|}{N}$ waarbij N de hoeveelheid rijen binnen de gegeven dataset is. Er geldt specifiek dat:

$$RAE = \frac{MAE}{\frac{\Sigma|\bar{\theta} - \theta_i|}{N}}$$

Root relative squared error

De Root relative squared error (RRSE) [Willmott and Matsuura, 2005b] beschrijft in welke mate RMSE afwijkt van de standaard deviatie (SD) voor de gegeven dataset. Er geldt specifiek dat:

$$RRSE = \frac{RMSE}{SD} \cdot \sqrt{\frac{N-1}{N}}$$

Total Number of Instances

Hiermee wordt het aantal gevallen aangegeven waarover de analyse in uitgevoegd. Dit is het aantal rijen binnen de dataset.

5.1.5 Interpretatie van de informatie

Bij de resultaten van de bovenstaande analyse (tabel 5.2) vallen een aantal dingen op. De correlation coefficient van 0.0205 geeft aan dat de voorspelde waarde niet nauwkeurig is. Dit is waarschijnlijk door de lage data kwaliteit en lage hoeveelheid voorbeelden. Verder is de RAE hoger dan 100%, dit betekent dat de afwijking groter is dan de grootte van de dataset. Als deze waarde hoger is dan 100%, dan is de voorspelling vaak niet van een hoge kwaliteit. Een andere verklaring is dat er informatie is verdwenen binnen de zorgaanbiedersmatrix door alleen een 0 of 1 neer te zetten bij bepaalde delen.

Concluderend is de kwaliteit van de gegevens te laag om er een goede voorspelling mee te maken. Om te beoordelen of het gebruiken van een regressie methode zou werken bij het maken van een voorspelling voor een terugvordering binnen Zilveren Kruis, wordt gekeken naar een completere dataset vanuit de wijkverpleging. Deze dataset gaat over een andere zorgsoort, maar bevat vergelijkbare gegevens en wordt daarom beschouwd als een goede indicator voor een completere dataset van de GGZ, mocht deze later ontstaan.

Deze dataset bestaat uit 212 rijen en 22 kolommen. Een volledige beschrijving van de wijkverplegingdataset staat in Appendix A.

Een data-analyse op de wijkverplegingsdataset volgens de zelfde methodiek als de hierboven beschreven analyse op de gegevens van de GGZ toont de volgende resultaten. De voorspellende waarde van het re-

Correlation coefficient	0.3818
Mean absolute error	36236.6152
Root mean squared error	69161.9748
Relative absolute error	89.4559%
Root relative squared error	114.1769%
Total Number of Instances	212

Tabel 5.3: summary wijkverpleging

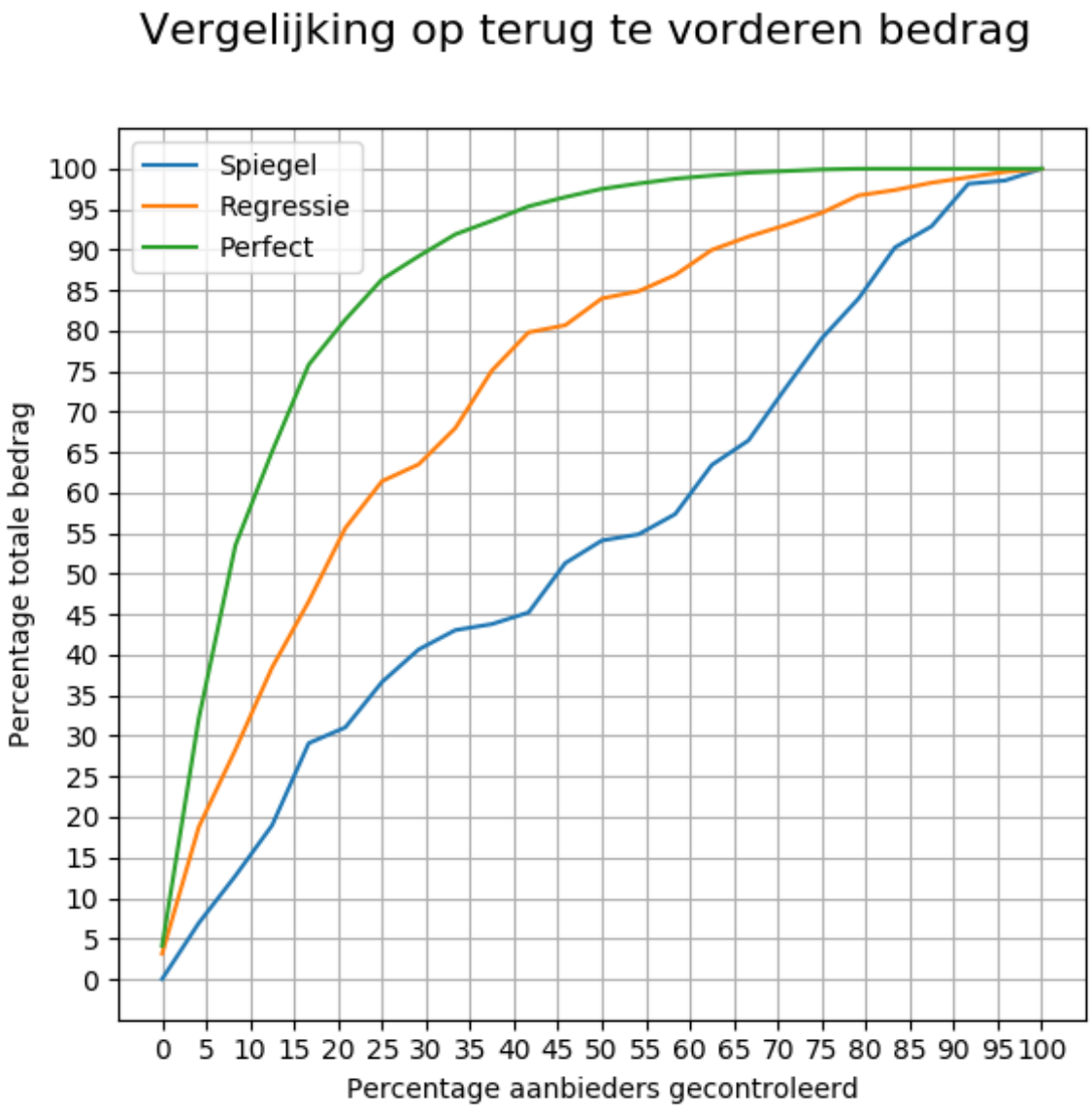
gressiemodel verbetert als er meer gegevens beschikbaar zijn. Dit zegt alleen nog niets over het verschil in prestaties tussen de huidige (spiegel) methodiek en de voorgelegde methode in dit onderzoek.

Om deze twee methodieken op een eerlijke manier te vergelijken zal er voor allebei een ranking gemaakt moeten worden die met elkaar vergeleken kunnen worden. Dit kan vanuit verschillende perspectieven gedaan worden. Dit onderzoek richt zich op zowel het teruggevorderde bedrag als het wel of niet detecteren van een

onrechtmatigheid.

Om deze vergelijking te maken moet er een ranking ontstaan volgens de spiegelmethode, dit is de methode zoals beschreven in Sectie 3.3. Hiervoor is in overleg gekozen om het attribuuft 'uur_p_client' te gebruiken om op te spiegelen. Dit is het attribuut dat volgens de analisten binnen Zilveren Kruis het meest representatief is voor de dataset. Deze wordt vervolgens gesorteerd van grootste afwijking naar laagste.

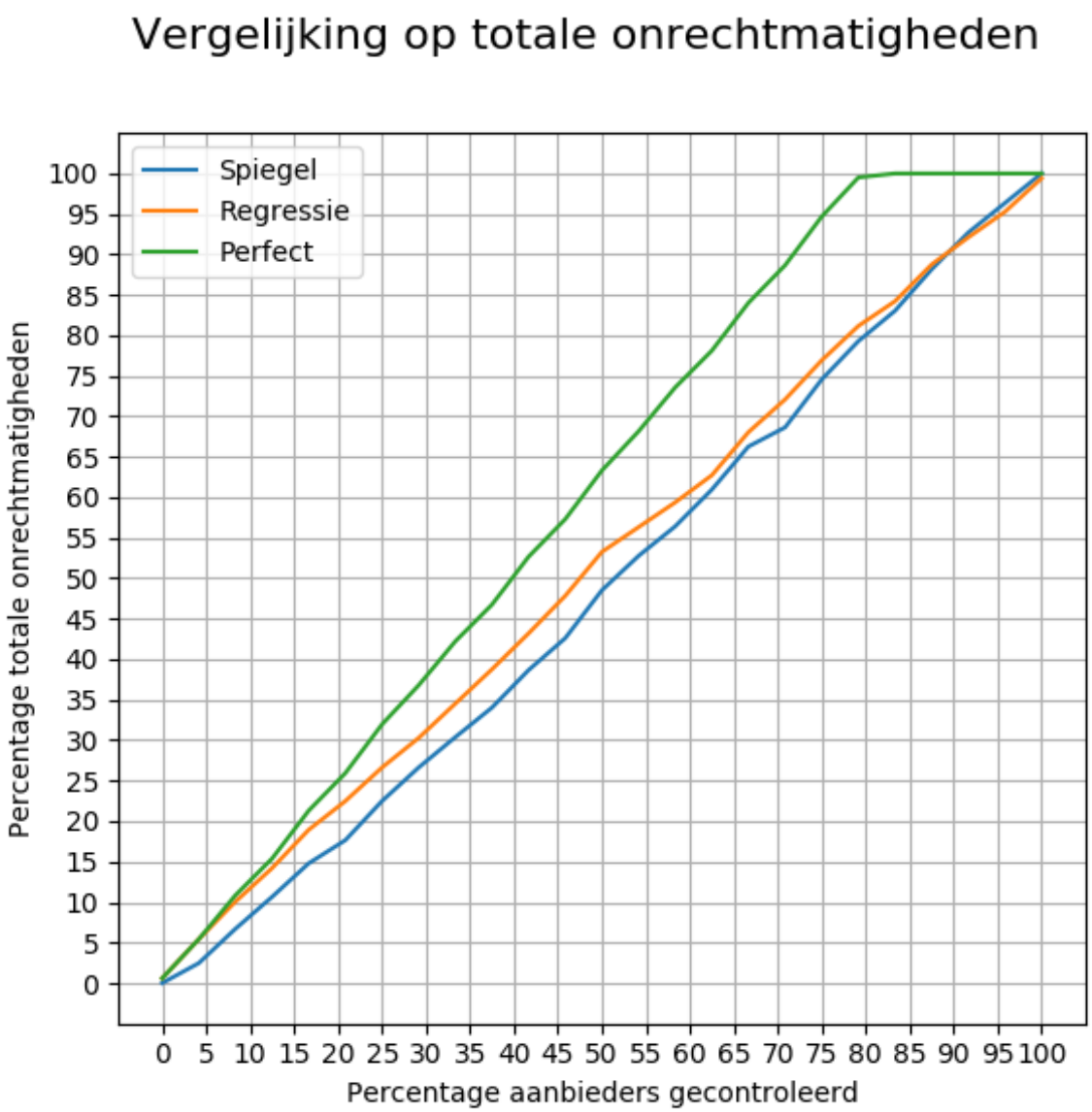
Voor de vergelijking op basis van bedrag wordt er binnen het regressiemodel de grootste voorspelde terugvordering het hoogst gerankt en de laagste voorspelde terugvordering als laagste gerankt. Verder wordt er nog gekeken naar de perfecte ranking. Hierbij wordt gekeken naar een situatie waarin de echte bedragen al bekend zijn. Dit resulteert in de volgende Figuur 5.5



Figuur 5.5: Vergelijking op terug te vorderen bedrag

In Figuur 5.5 wordt het percentage van het totale bedrag weergeven op de y-as en de procentuele aanbieders gecontroleerd op de x-as. Dit figuur laat zien dat onder de gemaakte aannames het toevoegen van een regressie-element aan de huidige analyse zorgt voor een verbetering in terugvordering. Hierin is te zien dat er op 50% van de gecontroleerde aanbieders onder de spiegel methode 55% van het totale bedrag is teruggevorderd. Bij dezelfde hoeveelheid is er met de regressie methode 85% van het totale bedrag teruggevorderd.

Ook kan er een vergelijking gemaakt worden in nauwkeurigheid van voorspelling. Hierin wordt gekeken wanneer alle onrechtmatigheden gevonden zijn. Hoe sneller alle gevallen zijn gevonden waar een terugvordering van toepassing is hoe nauwkeuriger de voorspelmethodiek. Dit resulteert in figuur 5.6.



Figuur 5.6: Vergelijking op totale onrechtmatigheden

Ook in dit figuur is te zien dat het gebruik maken van regressie bijdraagt aan het resultaat. Hier is te zien dat de voorspelling op een nauwkeurigere manier gemaakt wordt. Te zien is dat er in 80% van de zorgaanbieders

een terugvordering van toepassing is. Als 50% van de zorgaanbieders gecontroleerd zijn dan is er volgens de spiegel methodiek 49% van de onrechtmatigheden gevonden. Bij regressie wordt hier 54% terug gevonden en in de perfecte situatie 63%.

5.1.6 Toepassing en verbeterpunten

Binnen deze analyse zijn er een aantal factoren waarmee rekening gehouden moet worden. Ten eerste zijn de gemaakte vergelijkingen op basis van aannames. Dit was de enige manier om de vergelijking op een zo eerlijk mogelijke manier te maken.

In de toekomst zal er meer informatie beschikbaar zijn. Als die goed wordt opgeslagen dan kan deze informatie meegenomen worden in de beschreven analyse en zal de nauwkeurigheid hiervan toenemen. Dit geld zowel voor de huidige methode als eventuele toevoegingen hieraan. Hierdoor zal het ook als een iteratief proces behandeld moeten worden.

5.2 Risico's naar een ranglijst

Het maken van een ranglijst wordt volgens de huidige methodiek gedaan op basis van afwijking. Dit zorgt er impliciet voor dat het doel van de ranking het vinden van de grootste afwijkende zorgaanbieders is. Hoewel dit vanuit een analytisch perspectief vaak interessant is, komt het niet per definitie overeen met bedrijfsmatige en strategische belangen. Door het kijken naar de afwijkingen komen kleine bedrijven waar relatief veel variatie in zit relatief hoog te staan op de ranking. Hierbij gaat het vaak ook om relatief kleine bedragen.

Om te zorgen dat er vanuit verschillende belangen volgens een data gebaseerde manier beslissingen gemaakt kunnen worden, moeten deze belangen in kaart gebracht worden. Verschillende perspectieven kunnen gebruikt worden om te ranken:

- Kans dat er onrechtmatigheden gevonden worden
- Terug te vorderen bedrag
- Kosten van de controle
- Tijd die de controle kost

Bij de dataset van Zilveren Kruis is nu alleen een ranking mogelijk op basis van kans (afwijking). De uitkomst van het beschreven regressiemodel is een voorspeld terug te vorderen bedrag en een kans. Deze manieren van ranken zijn eerder gebruikt om de figuren te maken in Sectie 5.1.5.

Verder is het relevant te weten wat een bepaalde controle kost. Door alleen te kijken naar de verwachte opbrengst wordt een onvolledig beeld geschetst van een bepaalde actie. Ook de benodigde tijd van een

controle kan invloed hebben op de beslissing van de controles.

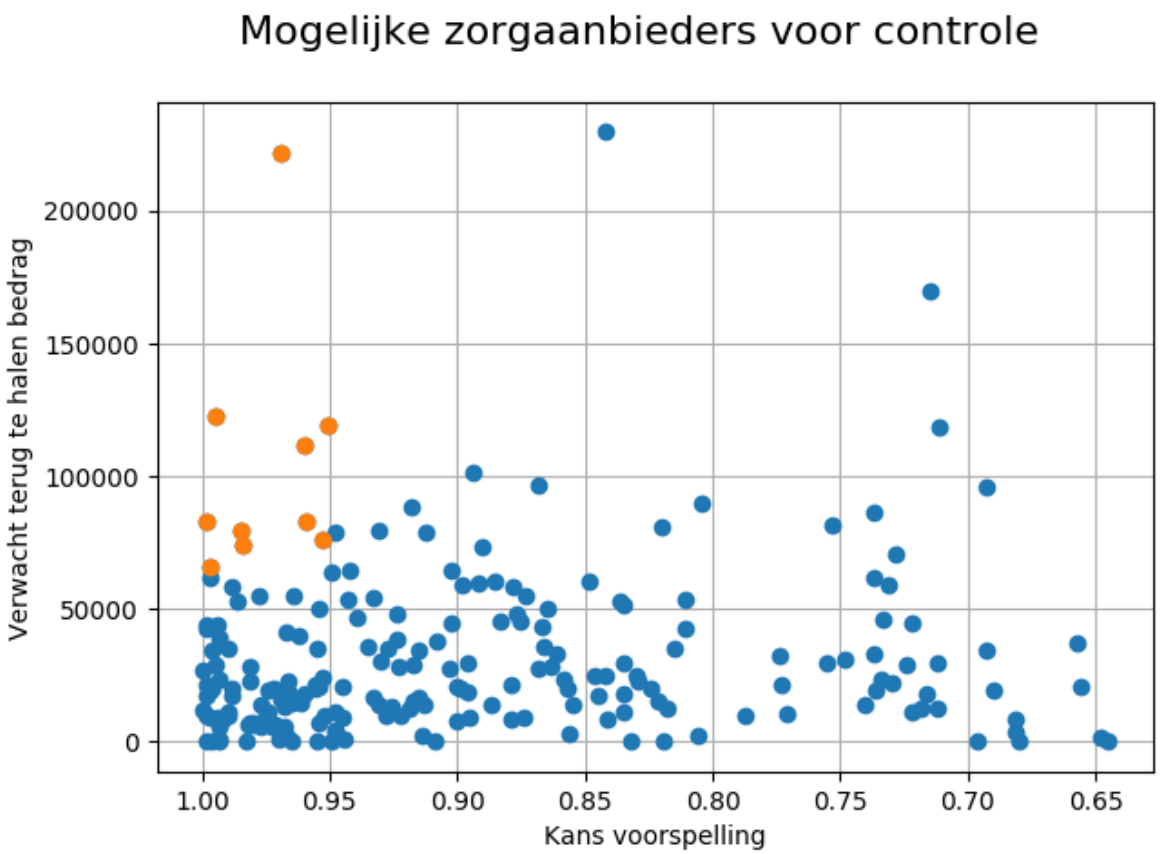
Deze vier mogelijkheden zijn in te delen in wat er nu gebeurt, wat er in de nabije toekomst kan verbeteren en wat er in de verdere toekomst verbeterd kan worden. Nu wordt er alleen op basis van kans gerangschikt. Bij het gebruik van regressie met de huidige gegevens kan op basis van kans en verwachte opbrengsten gerankt worden. Door het bijhouden van meer gegevens over kosten en tijd kan een voorspelling gemaakt worden van deze factoren zodat deze meegenomen kunnen worden voor het maken van de strategische beslissingen.

5.3 Ondergrens in ranglijst bepalen

Op het moment dat ranglijsten zijn opgesteld, moet een keuze gemaakt worden welke zorgaanbieders gecontroleerd moeten worden. Welke afwegingen hierbij belangrijk worden gevonden is een management beslissing. Het kan voorkomen dat er gestuurd wordt op kans, verwacht terug te halen bedrag, netto terug te halen bedrag (verwacht - kosten) of een zo hoog mogelijke controle-efficiëntie. Bijvoorbeeld door de verwachte tijd en het verwachte netto terug te halen bedrag te combineren.

Dit zijn allemaal strategieën waarop gestuurd kan worden als de betreffende ranglijsten opgesteld zijn. Door van meerdere ranglijsten te gebruiken is het mogelijk om prioriteit te verdelen naar meerdere factoren. Dit kan gebeuren door bijvoorbeeld een weging mee te geven aan de relatieve positie binnen de ranglijst.

Dit kan ook gebeuren door niet de relatieve posities te nemen maar de absolute waardes binnen de ranglijsten. Hierdoor kunnen regels opgesteld worden zoals: maak een ranking van iedereen met een kans van 95+% en controleer de 10 zorgaanbieders met de hoogste verwachte terugvordering mits deze hoger zijn dan 50.000 euro. Een visuele representatie hiervan is te zien in Figuur 5.7.



Figuur 5.7: Mogelijke zorgaanbieders voor controle

Het figuur laat zien dat er verschillende zorgaanbieders zijn met dezelfde verwachte terugvordering. Wanneer enkel naar deze terugvordering word gekeken wordt een zorgaanbieder met 70% kans en een verwachte terugvordering van 35.000 euro even hoog gescoord als een zorgaanbieder met 97% kans en een verwachte terugvordering van 35.000 euro.

Dit kan verder gegeneraliseerd worden door te stellen dat van alle punten met gelijke hoogte de meest linker geprefereerd moet worden, omdat altijd sprake is van gelijke terugvordering met verschil in kans. Dit zelfde geldt ook voor punten die boven elkaar liggen. Hierbij is het hoogste punt altijd geprefereerd omdat deze een hogere verwachte terugvordering heeft met dezelfde kans.

Bij het kiezen van zorgaanbieders moet rekening gehouden worden met deze principes door te kijken of er niet een zorgaanbieder beschikbaar is die objectief beter is dan de geselecteerde aanbieder.

5.4 Controlepunten bepalen

Nadat bepaald is welke punten gecontroleerd moeten worden, kan gekeken worden of er een verband bestaat tussen de punten die sowieso gecontroleerd gaan worden en andere mogelijke risico's. Om dit te onderzoeken

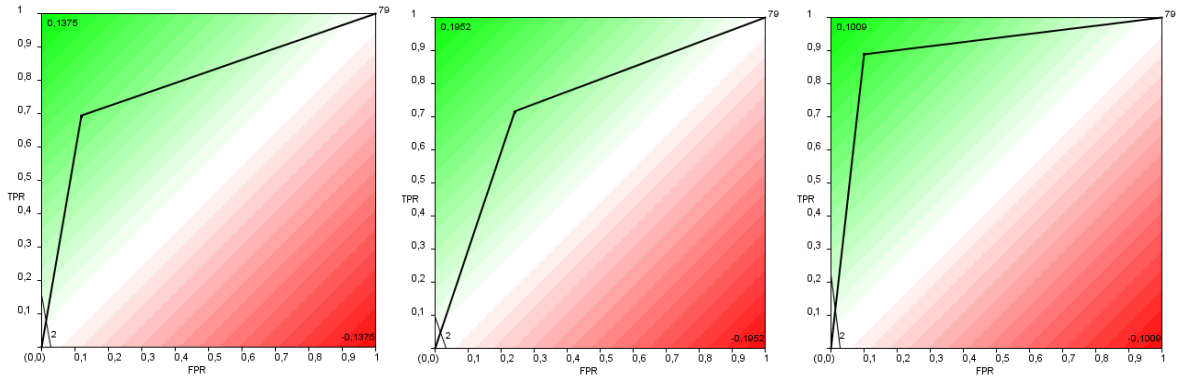
wordt gekeken naar de informatie binnen het werkbestand zoals beschreven in Sectie 4.2.1. Hier wordt niet gekeken naar het bedrag wat voorspeld wordt maar of er een bedrag teruggevorderd is per risico per jaar en of er correlaties bestaan tussen bestaande risico's per jaar. Dit wordt gedaan door te kijken of - gebruikmakend van Cortana [Meeng and Knobbe, 2011] - subgroepen gemaakt kunnen worden binnen de risico's.

Binnen Cortana wordt getracht subgroepen te vinden die de risico's uit het tweede jaar representeren. Dit wordt gedaan volgens de maatstaaf Weighted Relative Accuracy [Lavrač et al., 1999] (WRAcc). Hierbij wordt uitgegaan van een 99% betrouwbaarheidsinterval. Binnen dit betrouwbaarheidsinterval zijn verbanden gevonden voor de risico's verwijzingen, hoofdbehandelaarschap en verblijfsdagen

Wat hier opvalt als we kijken naar welke subgroepen die de beste representaties zijn van de risico's in 2014 dit de risico's uit 2013 zijn op de zelfde gebieden. Dit staat weergeven in Tabel 5.4. In Figuur 5.8 is de bijbehorende ROC curve [Bradley, 1997] te vinden.

Doel subgroep	FPR[1]	TPR[2]	Probability	P-waarde	WRAcc	Conditie
Verwijzing2014	0.121	0.692	0,529	$4.440 \cdot 10^{-16}$	0,4	Verwijzing2013 = '1'
Hoofdbehandelaarschap2014	0.241	0.714	0,517	$1.119 \cdot 10^{-8}$	0,33	Hoofdbehandelaarschap2013 = '1'
Verblijfsdagen2014	0.1	0.888	0,533	$5.286 \cdot 10^{-11}$	0,55	Verblijfsdagen2013 = '1'

Tabel 5.4: Intertemporele verbanden $n + 1$



(a) Verwijzingen (b) Hoofdbehandelaarschap (c) Verblijfsdagen

Figuur 5.8: ROC curve per subgroep

Hieruit kunnen we concluderen dat er sprake is van correlaties over intertemporele risico's. Dit geeft aan dat binnen de geanalyseerde gegevens de beste indicator van het volgende jaar het verleden is. Dit is te verklaren door te bedenken dat het terughalen van onrechtmatigheden vaak op een later tijdstip wordt gedaan, hier kan tot vier jaar overheen gaan. Daardoor zijn de zorgaanbieders in 2014 nog niet aangesproken op de onrechtmatigheden uit 2013. Hierdoor is er ook geen stimulans om het declaratiegedrag te veranderen. Mensen die administratieve fouten maken zullen waarschijnlijk dezelfde fouten blijven maken omdat zij op dit tijdstip

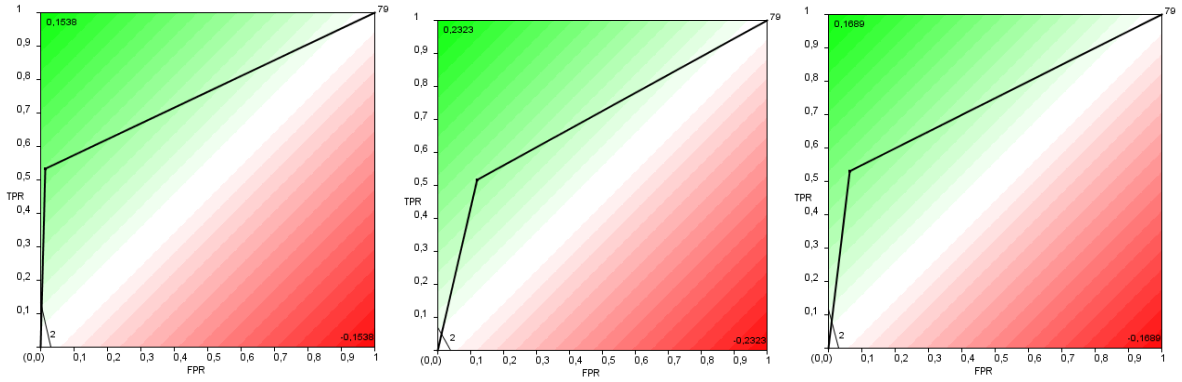
[1]False Positive Rate
[2]True Positive Rate

niet weten dat er een fout is gemaakt.

Vanuit het oogpunt van de controle is niet alleen interessant om te weten dat er een verband is tussen jaar n en jaar $n+1$. Het is ook interessant om te kijken wat er gebeurt als dit verband omgedraaid wordt en gekeken wordt naar de invloed die het jaar n heeft op het jaar $n-1$. Als hiertussen een verband bestaat is het mogelijk om controles met terugwerkende kracht te verantwoorden. De resultaten hiervan staan in Tabel 5.5 en Figuur 5.9.

Doel subgroep	FPR	TPR	Probability	P-waarde	WRAcc	Conditie
Verblijfsdagen2013	0.015	0.533	0,888	$1.913 \cdot 10^{-8}$	0,37	Verblijfsdagen2014 = '1'
Hoofdbehandelaarschap2013	0.12	0.517	0,714	$1.305 \cdot 10^{-8}$	0,28	Hoofdbehandelaarschap2014 = '1'
Verwijzing2013	0.064	0.529	0,692	$4.283 \cdot 10^{-7}$	0,33	Verwijzing2014 = '1'

Tabel 5.5: Intertemporele verbanden $n-1$



(a) Verblijfsdagen (b) Hoofdbehandelaarschap (c) Verwijzingen

Figuur 5.9: ROC curve per subgroep

Hierin is te zien dat er een grotere waarschijnlijkheid is dat er onrechtmatigheden optreden voor een bepaald risico in jaar $n-1$ als er een onrechtmatigheid is geconstateerd in jaar n.

Hoofdstuk 6

Evaluatie

In dit hoofdstuk worden concrete aanbevelingen gedaan. Ook wordt stilgestaan bij de verbeterpunten binnen het beschreven proces. Verder wordt er gereflecteerd in de discussie.

6.1 Aanbevelingen

In deze sectie worden er kansen binnen Zilveren Kruis geadresseerd in combinatie met een mogelijke verbetering. Hierbij gaat het deels om verdere uitwerking van eerder genoemde mogelijkheden en deels om niet eerder benoemde gevallen.

6.1.1 Alleen negatieve voorbeelden

De gegevens die nu worden bijgehouden en geanalyseerd bestaan voornamelijk uit voorbeelden van zorgaanbieders waar een onrechtmatigheid gevonden is. Dit geeft voor een analyse veel voorbeelden van negatieve gevallen. Er zijn weinig voorbeelden van zorgaanbieders die alles goed doen.

Machine learning algoritme zijn goed in het vinden van het onderscheid tussen zorgaanbieder met rechtmatige en onrechtmatige declaraties. Hiervoor moeten er wel voorbeelden beschikbaar zijn van beide gevallen.

Om dit te verbeteren is het noodzaak meer gegevens op te slaan over zorgaanbieders waar geen onrechtmatigheden zijn gedetecteerd. Dit kan zowel door verkennende controles uit te voeren waarbij gegevens van willekeurige zorgaanbieders worden verzameld, als ook door meer resultaten op te slaan van zorgaanbieders waar alles goed was bij het controleren.

Mocht alles op orde zijn als uitkomst van bijvoorbeeld een zelfcontrole van een zorgaanbieder dan is het

verstandig om alsnog de gegevens van deze zorgaanbieder toe te voegen als voorbeeld van een terugvordering van 0 euro.

6.1.2 In de toekomst te verzamelen gegevens

Zoals benoemd in 5.4 blijft de hoeveelheid beschikbare informatie toenemen. Dit betekent echter niet dat deze toename automatisch doorgevoerd wordt naar de analyses die worden uitgevoerd. Hiervoor zal op een actieve manier meer gegevens vergaard moeten worden. Omdat het niet altijd vooraf te zeggen is welke informatie van een zorgaanbieder de meeste voorspellende waarde heeft, is het verstandig om te kijken welke kwalitatieve gegevens op een zo geautomatiseerd mogelijke methode verzameld kunnen worden.

Alle nieuwe informatie die toegevoegd wordt aan een voorspellingsmodel draagt potentieel bij aan de nauwkeurigheid van dit model. Toch zijn er een aantal gegevens die direct een bijdrage kunnen leveren op andere onderdelen van het detectieproces. Dit zijn de kosten per controle en het aantal uren dat deze controle gekost heeft.

Met deze informatie kan een regressiemodel een inschatting maken van de kosten en doorlooptijd van toekomstige controles. Als deze informatie nauwkeurig geschat kan worden, is dit te gebruiken om hiermee nieuwe ranglijsten te maken. Het hebben van deze ranglijsten stelt Zilveren Kruis in staat om dit mee te nemen als prioriteit bij het kiezen van de te controleren zorgaanbieders.

Hiermee kan er gerichter gestuurd worden op strategische en bedrijfskundige doelen. Ook stelt dit Zilveren Kruis in staat om onderscheid te maken tussen twee zorgaanbieders waarvan met verschillende kans hetzelfde terug te vorderen bedrag is voorspeld.

6.1.3 Datakwaliteit

De kwaliteit van de gegevens is bepalend voor de kwaliteit van de resultaten. Op het moment dat de kwaliteit van de input laag is zal de kwaliteit van de output ook laag zijn. Daarom wordt stil gestaan bij verschillende punten waar Zilveren Kruis aan kan denken bij het verbeteren van de kwaliteit van de gegevens. Ze worden hieronder per gebied uitgelicht.

Duplicaten

Het hebben van duplicaten binnen de gegevens zorgt ervoor dat de entiteit die meerdere malen voorkomt verkeerd verwerkt wordt. Wanneer er twee keer exact dezelfde informatie staat, wordt deze informatie meerdere malen meegenomen in een analyse waardoor deze informatie als waardevoller meegenomen wordt. Ook kan de informatie op verschillende plaatsen onvolledig genoteerd staan. Dan wordt deze informatie

meerdere malen half verwerkt.

Beide opties zorgen ervoor dat de gegevens niet correct verwerkt worden. Dit zorgt er indirect voor dat het resultaat ook in kwaliteit achteruit gaat. Daarom is het noodzaak om hierop te controleren.

Consistentie

Bij het uitvragen van medewerker om het veld “terug te vorderen bedrag” in te vullen kan het voorkomen dat er een bedrag staat ingevuld. Ook komt het voor dat medewerkers “o”, “-”, “nvt”, “ ” of “n.v.t.” invullen. Met al deze antwoorden wordt bedoeld dat er geen geld teruggevorderd moet worden. Al deze antwoorden worden door een voorspelmodel als verschillende situaties opgevat.

Omdat meerdere notaties worden gebruikt voor hetzelfde zal een minder nauwkeurige voorspelling gemaakt kunnen worden. Dit effect treedt ook op wanneer een spelfout gemaakt wordt binnen een dataset of inconsistent gebruik wordt gemaakt van punten en komma‘s.

Om te voorkomen dat gelijkwaardige waardes op verschillende manieren geïnterpreteerd worden, is het verstandig afspraken te maken over de notatie. Ook is het verstandig om waar mogelijk te forceren dat gebruikers alleen maar invoer kunnen geven die een bepaald patroon volgt. Zo zou bij het veld “terug te vorderen bedrag” alleen een numerieke waarde worden geaccepteerd en geen lege waardes.

Compleetheid

Werken met gegevens die niet compleet zijn, betekent het mislopen van kennis. Dit is op zichzelf al een slechte situatie vanuit het oogpunt van datakwaliteit. Om de ontbrekende gegevens alsnog in te vullen wordt vaak een statistische schatting gemaakt van de meest waarschijnlijke waarde op die plaats. Deze invulling is een aanname die ervoor zorgt dat de ingevulde waarde een andere voorspellende waarde krijgt. Dit komt de nauwkeurigheid van de algehele voorspelling niet ten goede.

Ook komt het voor dat een missende waarde wordt behandeld als een aparte waarde. Als dit het geval is dan heeft het ontbreken van een waarde een voorspellende waarde die niet gewenst is. Dit stelt voorspellingen in staat om nauwkeuriger te zijn op het moment dat er minder informatie beschikbaar is, omdat er dan geen toeval in de dataset zit. Hierbij wordt uitgegaan dat het missen van een waarde geen voorspellende waarde heeft.

Omdat beide situaties niet bijdragen aan een goede voorspelling moeten deze vermeden worden. Dit kan door te controleren of alles is ingevuld bij het invoeren van gegevens.

Onmogelijkheden

Verder is het belangrijk dat de gegevens geen combinaties mogen bevatten die in de realiteit niet kunnen voorkomen. Bijvoorbeeld als bij controle als uitkomst staat dat deze voldoende was maar er alsnog een bedrag teruggevorderd is.

Om deze onmogelijkheden er uit te halen is domeinkennis nodig. Dit gaan alleen als er samen met de desbetreffende domeinspecialist wordt gekeken naar mogelijke combinaties van invoer. Vervolgens kan hier nog op gecontroleerd worden bij het invoeren van gegevens.

6.1.4 Detailniveau van gegevensopslag

Zorgaanbieders kunnen binnen een kliniek vallen, controles van meerdere zorgaanbieders kunnen in een dossier verwerkt worden en teruggevorderde bedragen kunnen als risicototaal opgeslagen worden. Bij al deze groeperingen kan ervoor gekozen worden om gegevens op verschillende niveaus op te slaan.

Op het moment dat de informatie van alle zorgaanbieders binnen een kliniek als totaal wordt gezien is het niet meer mogelijk om naar een individuele zorgaanbieder binnen die kliniek te kijken. Op het moment dat naar alle individuele zorgaanbieders binnen die kliniek gekeken wordt, is het alsnog mogelijk om de gegevens van de kliniek te aggregeren.

Dit zelfde principe geldt voor het samenvoegen van controles binnen een dossier of het samenvoegen van verschillende risico's bij een terugvordering. Op het moment dat data gegroepeerd word opgeslagen gaat er informatie verloren. Daarom is het aan te raden om gegevens op een zo laag mogelijk niveau op te slaan. Het samenvoegen van deze gegevens is altijd nog mogelijk.

6.1.5 Fragmentatie gegevensopslag

Binnen Zilveren Kruis worden gegevens op verschillende plaatsen opgeslagen. Vaak worden tussenresultaten, teruggevorderde bedragen op risiconiveau, alleen opgeslagen in een Excel-document op verschillende locaties. Door de verspreiding van informatie is het niet mogelijk om alle gegevens die er zijn mee te nemen in een analyse. Dit heeft tot gevolg dat er niet optimaal gebruik gemaakt wordt van deze gegevens.

Toch is het goed om gebruik te maken van deze gegevens. Het is daarom te overwegen om de tussentijdse informatie niet gefragmenteerd te laten verdwijnen in diverse bestanden maar ze op een centrale plaats op te slaan.

6.1.6 Impactrealisatie bij gegevensinvoer

Het invullen van verschillende gegevens kan overkomen als een administratieve last die geen bijdrage levert aan de dagelijkse werkzaamheden. Werknemers die deze gegevens bijhouden hebben vaak geen overzicht van de impact die dit later in het proces heeft. Als de werknemers die deze gegevens moeten invullen zich realiseren wat het gevolg van hun werkzaamheden is zou dit kunnen bijdragen aan het begrip en de zorgvuldigheid van het invullen. Hierom is het verstandig om een terugkoppeling te geven waarin de waarde van deze gegevens worden benadrukt.

6.1.7 Intertemporele foutgevoeligheid

In hoofdstuk 5.4 is vastgesteld dat als er een onrechtmatigheid van een risico in jaar n optreedt dit een vergrote kans heeft om in jaar $n+1$ op te treden. Hierbij moet meegenomen worden dat er in dit geval nog geen correctie heeft opgetreden over jaar n. Dit betekent dat zorgaanbieders een vergrote kans hebben om vergelijkbaar gedrag te vertonen over de jaren heen. Nadat er wordt gecorrigeerd voor dit gedrag zal pas een impuls ontstaan om dit gedrag te veranderen.

Deze kennis is op verschillende manieren toe te passen. Als er een onrechtmatigheid is gevonden voor een specifiek risico en hier nog geen terugvordering over is uitgevoerd, is het verstandig om dit risico het volgende jaar mee te nemen met een controle. Hierbij is wel het probleem dat er in deze situatie er vanuit wordt gegaan dat de zelfde zorgaanbieder twee opeenvolgende jaren gecontroleerd wordt.

Een andere methode is om op het moment dat een onrechtmatigheid voor een bepaald risico geconstateerd wordt, dit risico voor voorgaande jaren ook te controleren. Deze methode maakt het mogelijk correcties over meerdere jaren uit te voeren in één controle. De kans dat er een onrechtmatigheid optreedt in het jaar $n-1$ is groter dan dat deze optreed in het jaar $n+1$ als er in jaar n een onrechtmatigheid is opgetreden.

Daarom is het aan te raden te kijken naar voorgaande jaren als hierin een onrechtmatigheid is opgetreden. Hierbij gaat het uitsluitend om het risico waarover de onrechtmatigheid wordt gevonden.

6.1.8 Continuous improvement

Het maken van een model op de huidige gegevens zorgt voor de beste voorspelling voor het huidige moment. Alleen door continue verandering in wetgeving binnen de zorg, beschikbare informatie en beschikbare voorspelmodellen is het noodzakelijk om te blijven verbeteren. Nieuwe risico's moeten gesignaleerd blijven worden op basis van veranderende regulaties en nieuwe frauduleuze constructies.

Nieuwe manieren van onrechtmatigheden vragen om nieuwe manieren van detectie. Het is belangrijk om te

blijven onderzoeken welke informatie beschikbaar is en hoe deze informatie verwerkt kan worden. Op het moment dat je je vandaag niet voorbereidt op morgen ben je alleen klaar voor gisteren.

6.1.9 Implementatie

Het implementeren van deze adviezen is een complexe taak. Toch zijn er binnen Zilveren Kruis een aantal kansen die hiervoor gebruikt kunnen worden. Dit omdat er op relatief korte termijn een vervanging van het informatieopslagsysteem aan komt. Dit geeft de kans om gegevens op een centrale plaats op te slaan.

Verder zou het verstandig zijn om mensen aan te stellen om de kwaliteit van de gegevens te waarborgen. Deze zouden aan de kwalitatieve kant verbeteringen kunnen doorvoeren door standaarden over het opslaan van gegevens op te stellen en te waarborgen.

Verandering gaat in kleine stappen. Het is niet realistisch om de gang van zaken te veranderen op een korte termijn. Toch kunnen er grote stappen gemaakt worden door kleine aanpassingen te doen. Waar een analist nu gebruik maakt van Sherlock (voorheen SAS FF) om de voorgeprogrammeerde spiegelanalyse-code, een macro, aan te roepen kan de kwaliteit van de voorspelling verhoogd worden door een andere macro aan te roepen. Het gebruik van verschillende macro's met toenemende complexiteit zal over de tijd heen toenemen met het begrip over deze macro's.

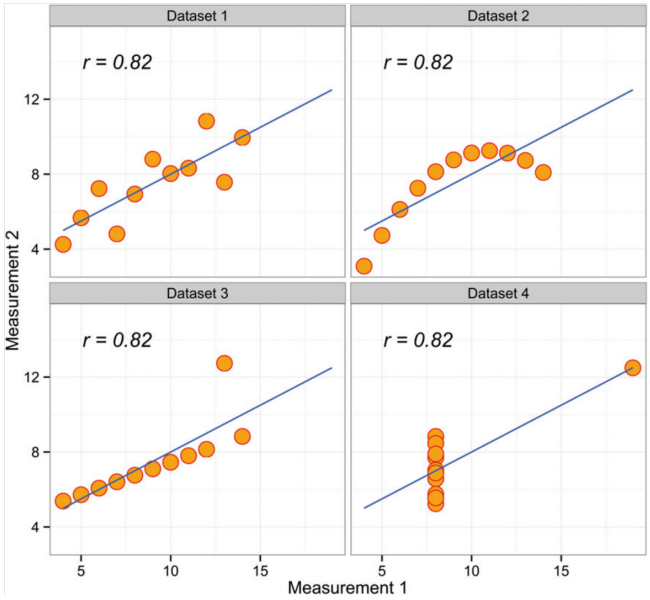
Het is niet verstandig om dit direct over alle verschillende afdelingen door te voeren. Door gebruik te maken van een pilot zorgt men ervoor dat het succes van verschillende methodieken bewezen kan worden. Door het behalen van een vroegtijdig succes zullen mensen gemotiveerder zijn om volgens een andere werkwijze aan de slag te gaan. Een voorbeeld hiervan is het controleren van voortgaande jaren op het moment dat een risico gedetecteerd is. Dit is relatief makkelijk door te voeren en het resultaat hiervan is direct zichtbaar.

6.2 Discussie

In dit verslag wordt regressie gebruikt als voorspellingsmethode. Door gebruik te maken van deze methodiek kunnen voorspellingen gemaakt worden die een numerieke waarde opleveren. Dit wordt voorgesteld als oplossing voor het probleem dat in dit document optreedt. Toch is het goed stil te blijven staan dat er binnen voorspelmodellen geen model is dat voor iedere dataset altijd het beste resultaat geeft. Dit fenomeen wordt beschreven als de *no free lunch theorem* [Wolpert and Macready, 1997].

De voorspelling die uit lineaire regressie komt geeft het verband aan binnen een dataset. Het kan echter voorkomen dat dezelfde verbanden worden gevonden in verschillende datasets. Voor het menselijk oog is het duidelijk dat de vier datasets zoals weergeven in figuur 6.1 verschillend zijn. Als er echter gekeken wordt naar de statistische kenmerken, zoals gemiddelde etc, van deze datasets dan zijn deze zo goed als identiek.

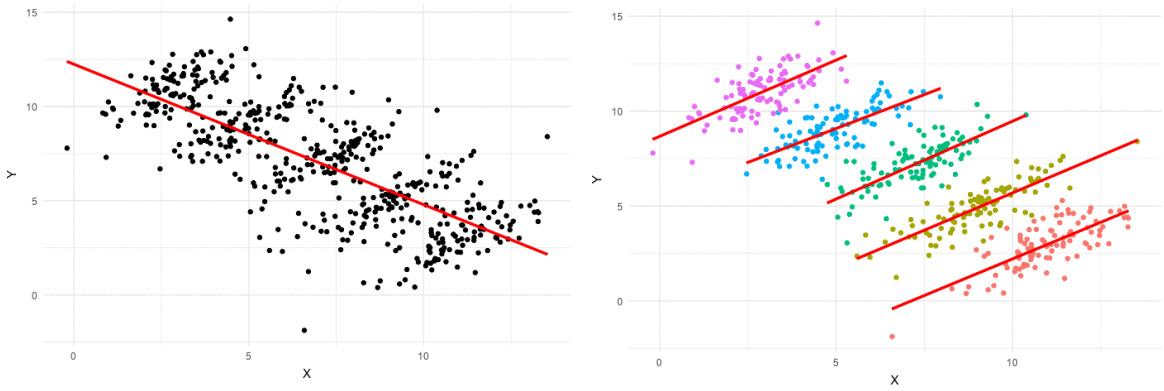
Daardoor wordt er ook dezelfde schatting gemaakt binnen deze datasets. Dit beroemde voorbeeld wordt *Anscombe's quartet* [Anscombe, 1973] genoemd. Dit benadrukt het belang van datavisualisatie bij het gebruik van lineaire regressie.



Figuur 6.1: Anscombe's quartet [Hickey et al., 2015]

Bij het gebruik van lineaire regressie is het goed in het achterhoofd te houden dat lineaire regressie een voorspelling geeft over volledige datasets. Hierbij kan het voorkomen dat de geanalyseerde dataset een verzameling is van kleinere datasets. Denk hierbij aan het groeperen van alle risico's tot een terug te vorderen bedrag.

Het resultaat van een dergelijke samenvoeging kan zijn dat het zogeheten *Simpson's Paradox* [Simpson, 1951] effect optreedt. Hierin is het verband binnen de gegevens tegengesteld aan het verband van de samengevoegde gegevens. Een visuele representatie hiervan staat in figuur 6.2. Hierin is te zien in het zwart dat de totale



(a) Totale dataset (b) verdeeld in subgroepen

Figuur 6.2: Simpson's Paradox [svwiki, 2017]

correlatie negatief verloopt. Als hier echter gekeken wordt naar de verschillende subgroepen waar deze dataset

uit is opgebouwd, wordt duidelijk dat het in werkelijkheid om een verzameling positieve verbanden gaat. Ondanks het bestaan van deze effecten is het gebruikmaken van regressie een verbetering ten opzichte van de huidige methode.

Hoofdstuk 7

Conclusie

Het onrechtmatigheden-detectieproces waarnaar is gekeken bestaat uit verschillende onderdelen. Binnen dit onderzoek is een verbeterde methode voorgesteld die is op te delen in drie hoofddelen: het rangschikken van de mogelijke zorgaanbieders, het bepalen van de drempelwaarde en verbanden tussen de resultaten van deze controles.

Het rangschikken van deze zorgaanbieders wordt gedaan op basis van lineaire regressie. Hierbij wordt er een inschatting gemaakt van het terug te vorderen bedrag. Als dit voor iedere zorgaanbieder is gedaan, worden ze op volgorde van hoog naar laag gezet om er een ranglijst van te maken.

Het maken van deze ranking is anders dan het kijken naar de grootte van de afwijking volgens de spiegelmethodiek. De grootste verschillen tussen deze twee is dat (1) lineaire regressie van meer attributen gebruikmaakt om een voorspelling te doen. En (2) er gebruik wordt gemaakt van een supervised methode ten opzichte van een unsupervised methode. De attributen die hiervoor gebruikt worden, kunnen ook bestaan uit de uitkomsten van de spiegelanalyse. Hierdoor wordt er meer informatie gebruikt bij het voorspellen, hetgeen resulteert in een nauwkeurigere beslissing.

Het volgende onderdeel is de bepaling van de drempelwaarde voor de opgestelde ranglijst. Dit is onder de huidige werkwijze een statisch bepaalde waarde. Deze waarde houdt geen rekening met factoren zoals grootte van de zorgaanbieder of het verwachte bedrag. Hierdoor komt het vaak voor dat relatief kleine zorgaanbieders een grote afwijking hebben en relatief lage terug te vorderen bedragen.

Omdat deze situatie niet optimaal is, wordt er in dit onderzoek voorgesteld om deze beslissing te maken op basis van strategische en bedrijfskundige afwegingen. Door bij te houden hoeveel tijd en geld een controle kost en oplevert kan er op meerdere factoren gestuurd worden. Zo is het mogelijk om te sturen op een zo groot mogelijk terug te vorderen bedrag zolang er tijd is voor controle. Of te kiezen voor de zorgaanbieders

die het hoogste terug te vorderen bedrag per verwachte aantal uur werk hebben.

Tot slot is er gekeken naar het verband tussen verschillende uitkomsten van controles. Hier is gevonden dat er een verband bestaat tussen de uitkomst van de huidige controle en de uitkomst van de controle over hetzelfde risico in voorgaande jaren. Hiervoor geldt dat als een zorgaanbieder in het huidige jaar een fout heeft gemaakt over een risico, deze fout waarschijnlijk in voorgaande jaren ook is gemaakt.

Bibliografie

[Akaike, 1987] Akaike, H. (1987). Factor analysis and aic. In *Selected Papers of Hirotugu Akaike*, pages 371–386. Springer.

[Anscombe, 1973] Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, 27(1):17.

[Benesty et al., 2009] Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). *Pearson Correlation Coefficient*, pages 1–4. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Bonada, 2017a] Bonada, E. (2017a). Cross-validation strategies - kfold cross-validatie.

[Bonada, 2017b] Bonada, E. (2017b). Cross-validation strategies - leave-one-out cross-validatie.

[Bradley, 1997] Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145 – 1159.

[Bramer, 2013] Bramer, M. (2013). *Principles of data mining*. Springer Publishing Company, Incorporated.

[Brynjolfsson et al., 2011] Brynjolfsson, E., Hitt, L. M., and Kim, H. H. (2011). Strength in numbers: How does data-driven decisionmaking affect firm performance?

[Burnham and Anderson, 2004] Burnham, K. P. and Anderson, D. R., editors (2004). *Model Selection and Multimodel Inference*. Springer New York.

[Cao and Zhang, 2008] Cao, L. and Zhang, C. (2008). Domain driven data mining. In *Data Mining and Knowledge Discovery Technologies*, pages 196–223. IGI Global.

[Caruana and Freitag, 1994] Caruana, R. and Freitag, D. (1994). Greedy attribute selection. In Cohen, W. W. and Hirsh, H., editors, *Machine Learning Proceedings 1994*, pages 28 – 36. Morgan Kaufmann, San Francisco (CA).

[CBS, 2018] CBS (2018). Zorguitgaven; kerncijfers. `https://opendata.cbs.nl/statline/#/CBS/nl/dataset/84047NED/line?ts=1530442110020`.

[Coffman and Odlyzko, 2002] Coffman, K. G. and Odlyzko, A. M. (2002). *Internet Growth: Is There a "Moore's Law" for Data Traffic?*, pages 47–93. Springer US, Boston, MA.

[Grzymala-Busse and Hu, 2000] Grzymala-Busse, J. W. and Hu, M. (2000). A comparison of several approaches to missing attribute values in data mining. In *International Conference on Rough Sets and Current Trends in Computing*, pages 378–385. Springer.

[Hall and Holmes, 2003] Hall, M. A. and Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data Engineering*, 15(6):1437–1447.

[Hickey et al., 2015] Hickey, G., Dunning, J., Seifert, B., Sodeck, G., J Carr, M., Burger, H., and Beyersdorf, F. (2015). Statistical and data reporting guidelines for the european journal of cardio-thoracic surgery and the interactive cardiovascular and thoracic surgery. 48.

[Hyndman and Koehler, 2006] Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679 – 688.

[Kirlidog and Asuk, 2012] Kirlidog, M. and Asuk, C. (2012). A fraud detection approach with data mining in health insurance. *Procedia - Social and Behavioral Sciences*, 62:989 – 994. World Conference on Business, Economics and Management (BEM-2012), May 46 2012, Antalya, Turkey.

[Koenraadt, 2016] Koenraadt, B. (2016). Marktverdeling zorgverzekeraars in nederland (infographic). https://www.zorgwijzer.nl/zorgverzekering-2017/marktverdeling-zorgverzekeraars-infographic.

[Kohavi et al., 1995] Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada.

[Kruis, 2017] Kruis, Z. (2017). Stijging aantal verzekerden zilveren kruis. `https://nieuws.zilverenkruis.nl/stijging-aantal-verzekerden-zilveren-kruis/`.

[Lavrač et al., 1999] Lavrač, N., Flach, P., and Zupan, B. (1999). Rule evaluation measures: A unifying view. In Džeroski, S. and Flach, P., editors, *Inductive Logic Programming*, pages 174–185, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Leiden, 2017] Leiden, U. (2017). Cortana subgroup discovery.

[Meeng and Knobbe, 2011] Meeng, M. and Knobbe, A. (2011). Flexible enrichment with cortana–software demo. In *Proceedings of BeneLearn*, pages 117–119.

[Neter et al., 1996] Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996). *Applied linear statistical models*, volume 4. Irwin Chicago.

[Quinlan, 1992] Quinlan, J. R. (1992). Learning with continuous classes. pages 343–348. World Scientific.

[Redman, 2008] Redman, T. C. (2008). *Data driven: profiting from your most important business asset*. Harvard Business Press.

[Refaeilzadeh et al., 2016] Refaeilzadeh, P., Tang, L., and Liu, H. (2016). *Cross-Validation*, pages 1–7. Springer New York, New York, NY.

[RIVM, 2018] RIVM (2018). Zorguitgaven - hoe ontwikkelen zich de zorguitgaven in de toekomst? https://www.vtv2018.nl/zorguitgaven.

[S. Viveros et al., 1996] S. Viveros, M., P. Nearhos, J., and Rothman, M. (1996). Applying data mining techniques to a health insurance information system. pages 286–294.

[SAS, 2018] SAS (2018). Sas enterprise guide.

[Simpson, 1951] Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 13(2):238–241.

[svwiki, 2017] svwiki, P. (2017). Simpsons paradox.

[Waikato, 2018] Waikato, U. (2018). Weka 3: Data mining software in java.

[Willmott and Matsuura, 2005a] Willmott, C. J. and Matsuura, K. (2005a). Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate Research*, 30(1):79–82.

[Willmott and Matsuura, 2005b] Willmott, C. J. and Matsuura, K. (2005b). Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, 30(1):79–82.

[Wolpert and Macready, 1997] Wolpert, D. H. and Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82.

Appendix A

Beschrijving dataset wijkverpleging

In deze appendix wordt de gebruikte dataset die vanuit wijkverpleging is verkregen beschreven. De verkregen dataset voldoet aan de kwaliteitseisen die beschreven zijn in Sectie 5.1.3. In deze appendix wordt gekeken naar de opbouw van de dataset, vervolgens wordt gekeken naar overbodige attributen binnen deze dataset.

De verkregen dataset bestaat uit de volgende attributen:

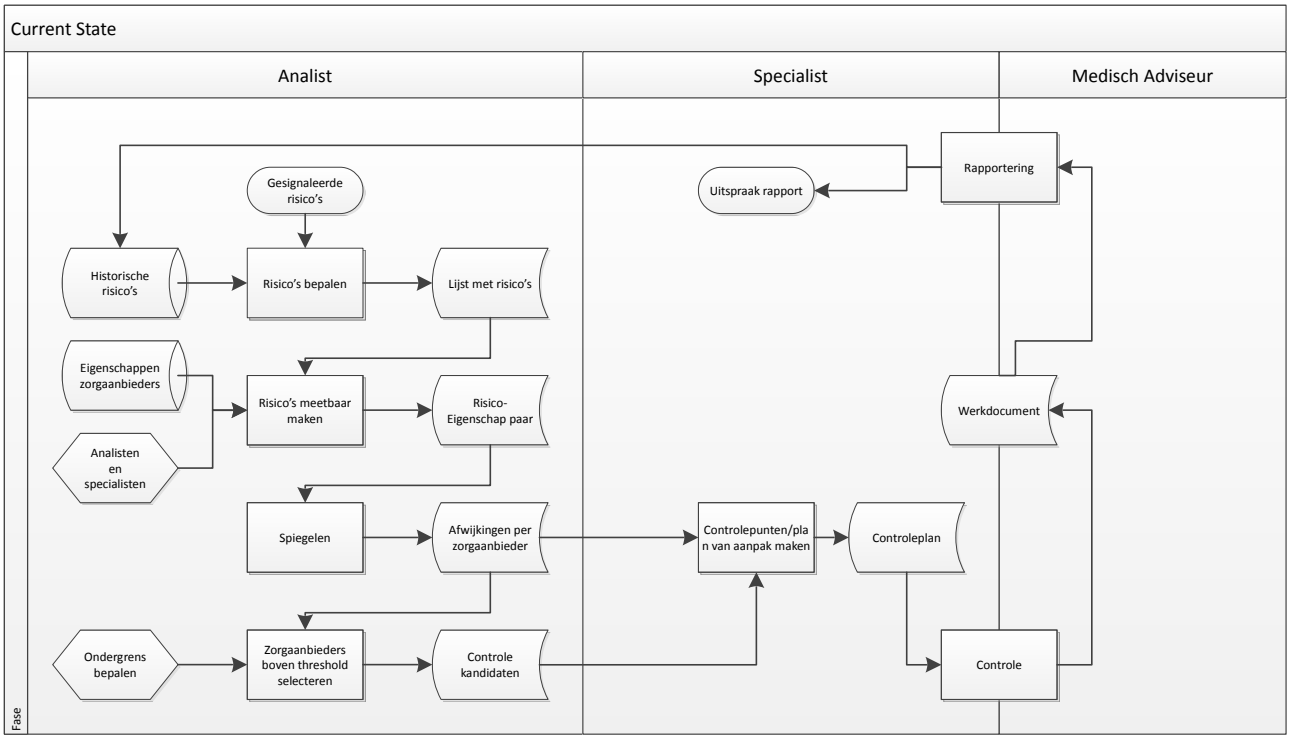
Naam	Type	Beschrijving
antl_vzd	Numeric	Aantal verzekerde
kosten_p_clnt	Numeric	Kosten per cliënt
antl_uur	Numeric	Aantal uur
uur_p_clnt	Numeric	Aantal uur per cliënt
antl_uur_PV	Numeric	Aantal uur persoonlijke verzorging
antl_uur_VP	Numeric	Aantal uur verpleging
verh_vp_pv	Numeric	Verhouding verpleging / persoonlijke verzorging
gem_antl_dagen	Numeric	Gemiddeld aantal dagen
gem_lft	Numeric	Gemiddelde leeftijd
perc_man	Numeric	Percentage man
perc_vrouw	Numeric	Percentage vrouw
perc_jong	Numeric	Percentage jong
perc_kind	Numeric	Percentage kind
perc_overl	Numeric	Percentage overledenen
gem_dagen_tot_decl	Numeric	Gemiddeld aantal dagen tot declaratie
antl_notas	Numeric	Aantal nota's
antl_regels_p_nota	Numeric	Aantal regels per nota
antl_unieke_regels	Numeric	Aantal unieke regels
antl_regels_restitutie	Numeric	Aantal regels restitutie
perc_restitutie	Numeric	Percentage restitutie
bedrag_teruggevorderd	{TRUE, FALSE}	Is er een bedrag teruggevorderd
target	Numeric	Grote van het teruggevorderde bedrag

Tabel A.1: Beschrijving dataset wijkverpleging

Hoewel deze data kwalitatief goed is zijn er toch een aantal onnodige attributen te vinden. Het attribuut 'bedrag_teruggevorderd' is een deel informatie die pas bekend is nadat de controle is uitgevoerd. Hierom zou deze niet meegenomen kunnen worden in een voorspelling. Verder valt het nog op dat de attributen 'perc_man' en 'perc_vrouw' perfect negatief gecorreleerd zijn ($R = -1$). Dit betekend dat het toevoegen van 'perc_man' geen informatie toevoegt als 'perc_vrouw' al aanwezig is en andersom.

Appendix B

Procesbeschrijving



Figuur B.1: Procesbeschrijving