Universiteit Leiden
Opleiding Informatica & Economie

Toepassing van data mining bij het vinden van

onrechtmatige declaraties in de farmaceutische sector

Naam:	Jasper van Nijhuis
Datum:	19/08/2016
1e begeleider:	Cor Veenman (LIACS)
2e begeleider:	Siegfried Nijssen (LIACS)
3e begeleider:	Rob Konijn (Zilveren Kruis)

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Toepassing van data mining bij het vinden van onrechtmatige declaraties in de farmaceutische sector

Jasper van Nijhuis

Abstract

Om declaratiefraude binnen de farmaceutische sector tegen te gaan voert zorgverzekeraar Zilveren Kruis, onderdeel van Achmea, controles uit bij zorgverleners. Uit deze controles komen mogelijk onrechtmatige declaraties welke vervolgens teruggevorderd worden. Een groot deel van de controles zijn arbeidsintensief, waardoor er voor deze controles een selectie gemaakt dient te worden van de zorgverleners die onderzocht gaan worden. Om deze selectie met data mining te bepalen, worden er op basis van (gedrags)gegevens van zorgverleners twee voorspellingsmodellen gemaakt die voor deze zorgverleners de kans op terugvorderingen geven, waarbij het terugvorderingsbedrag binnen gespecificeerde grenzen ligt. Het eerste model doet een uitspraak over de kans op terugvorderingen, met als bedrag aan terugvorderingen €500 tot €1500. Het tweede model doet een uitspraak over de kans op terugvorderingen, met als bedrag aan terugvorderingen €1500 en hoger. Voor beide modellen geldt dat een random forest classifier het beste de modellen kan bouwen, gemeten in AUC. Het eerste model behaald een performance van 0,82 AUC en het tweede model 0,91 AUC. Daarnaast wordt duidelijk dat de splitsing in twee modellen een gemiddelde performance afname in AUC van 4% te wege brengt, ten opzichte van een voorspellingsmodel waarbij alleen de kans op terugvorderingen een rol speelt.

Inhoudsopgave

Hoofdstuk 1

Introductie

Data mining is niet meer alleen een wetenschappelijk onderzoeksgebied, maar ook een veelbelovend concept in het bedrijfsleven. Het kan namelijk inzichten en voorspellingen leveren, die voorheen verborgen en onbekend waren. Dit kan grote (financiële) voordelen met zich mee brengen. Zo ook heeft zorgverzekeraar Zilveren Kruis, onderdeel van Achmea, interesse in de toepassing van data mining binnen haar bedrijf. Ondanks de contracten die Zilveren Kruis met zorgverleners uit de farmaceutische sector heeft, omtrent welke zorg en hoe die zorg vanuit de zorgverzekering van de patiënt vergoed wordt, worden er alsnog jaarlijks voor miljoenen euro's declaraties onrechtmatig ingediend en uitbetaald. Het is dan ook niet gek dat Zilveren Kruis een volledige controle afdeling heeft die deze onrechtmatige declaraties probeert op te sporen en terug te vorderen. Het is voor Zilveren Kruis echter niet mogelijk om elke gecontracteerde zorgverlener individueel in de gaten te houden, dus moet er een selectie gemaakt worden die bepaald welke zorgverleners onderzocht gaan worden. Dit onderzoek bekijkt hoe deze selectie, met behulp van data mining in combinatie met domeinkennis, het beste te bepalen is. Hierbij wordt gezocht naar zorgverleners met een zo groot mogelijke kans op terugvordering en een zo hoog mogelijk verwacht terug te vorderen bedrag. De onderzoeksvraag luidt dan ook als volgt:

Hoe (goed) kan een voorspellingsmodel, met features gebasseerd op domeinkennis, een ranglijst maken van zorgverleners, waarbij de kans op terugvordering van declaraties en het verwachtte bedrag daarvan van belang zijn?

Dit onderzoek is gebasseerd op het CRISP-DM model [12], waarin het data mining process in zes fases opgedeeld is: *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation* en *Deployment*. CRISP-DM is in dit onderzoek terug te zien als het beschrijven van de (business) context, het prepareren van de data, inzicht verwerven in de data, het bouwen en evalueren van voorspellingsmodellen en als laatst het beschrijven van het gebruik van de modellen in de business.

Eerst wordt de context beschreven, welke de belangrijkste begrippen uit dit onderzoek toelicht. Vervolgens wordt de aanpak uitgelegd, welke bestaat uit het samenstellen van de dataset, het verwerven van inzicht in de data en het bouwen en valideren van voorspellingsmodellen. Dit wordt gevolgd door de resultaten van de data-analyse en de voorspellingsmodellen. Als laatst wordt de conclusie besproken en worden er aanbevelingen gedaan voor vervolgonderzoek.

Hoofdstuk 2

Context

Dit hoofdstuk licht de belangrijkste concepten en begrippen toe welke relevant zijn voor dit onderzoek. Dit hoofdstuk vervult de rol van *Business Understanding* volgens het CRISP-DM model. Het bespreekt als eerste waar binnen Zilveren Kruis dit onderzoek plaatsvindt en wat de rollen van de verschillende betrokken afdelingen zijn. Vervolgens omschrijft het wat de farmaceutische sector binnen Nederland is en licht het het begrip *onrechtmatige declaraties* binnen deze sector toe. Hierna volgt een toelichting van het beleid hiertegen door Zilveren Kruis. Als laatste worden de voor dit onderzoek relevante aspecten van data mining uitgelegd.

2.1 Zilveren Kruis

Met enkele miljoenen klanten en een premieomzet van ongeveer 19,9 miljard euro (2015) is Achmea de grootste verzekeraar in Nederland. Ze biedt verzekeringen aan onder verschillende merken, zoals Centraal Beheer Achmea, Interpolis, FBTO, Avéro, InShared en Zilveren Kruis. Zilveren Kruis verstrekt zorgverzekeringen en helpt klanten bij het vinden en krijgen van de juiste zorgbehandeling. Binnen Zilveren Kruis richt Het Kenniscentrum zich op het datamanagement van het bedrijf en ontsluit en koppelt ze externe data bronnen aan het interne systeem. Daarnaast maakt ze data en kennis beschikbaar en bruikbaar voor eindgebruikers. Denk hierbij aan laagdrempelige tools welke inzicht geven in data, zoals web-based dashboards. Als laatst analyseert ze de door Zilveren Kruis verzamelde data. Deze analyses zijn hoofdzakelijk gericht op controle vraagstukken, zoals mogelijke fraude, maar variëren ook tot het vergaren van nieuwe inzichten waarmee Zilveren Kruis haar beleid kan optimaliseren. Dit onderzoek, *Toepassing van data mining bij het vinden van onrechtmatige declaraties in de farmaceutische sector*, is uitgevoerd binnen Het Kenniscentrum bij het team Business Intelligence Analyse & Rapportage.

2.2 De Nederlandse Zorgautoriteit

De Nederlandse Zorgautoriteit, NZa, houdt toezicht op de zorgmarkt in Nederland. Haar doel is het bevorderen van de toegankelijkheid, betaalbaarheid en kwaliteit van de Nederlandse gezondheidszorg. De NZa is een zelfstandig bestuursorgaan, wat inhoudt dat het een individuele organisatie is welke overheidstaken uitvoert onder toezicht van een ministerie. In dit geval onder toezicht van het Ministerie van Volksgezondheid, Welzijn en Sport.

2.3 Farmacie

Met farmacie wordt de extramurale farmaceutische zorg bedoeld. Dit betreft het verstrekken van UR-geneesmiddelen en het verstrekken van informatie en begeleiding rond deze geneesmiddelen, buiten een instelling of ziekenhuis om. In praktijk uit dit zich tot een bezoek van een patiënt aan een apotheek. Een UR-geneesmiddel is een geneesmiddel welke uitsluitend verkrijgbaar is bij de apotheek op recept. Dit in tegenstelling tot een zelfzorggeneesmiddel welke zonder recept ook bij de apotheek en drogist verkrijgbaar is.

Farmacie wordt door verschillende partijen aangeboden:

Zelfstandige apotheken Individuele apotheken, niet aangesloten bij een branche organisatie
Ketenapotheken Franchise nemende apotheken, aangesloten bij een branche organisatie
Poliklinische apotheken Externe apotheken gevestigd in ziekenhuizen
Internetapotheken Apotheken welke hun diensten en producten online aanbieden en op afstand leveren
Dienstapotheken Apotheken die alleen tijdens de avond, nacht en in het weekend open zijn
Huisartsen met vergunning tot verstrekken van farmaceutische zorg

De farmacie kent 13 prestatiebekostigingen welke beschreven zijn door De Nederlandse Zorgautoriteit, NZa (zie hoofdstuk 2.2 *De Nederlandse Zorgautoriteit*) in *Beleidsregel prestatiebeschrijvingen voor farmaceutische zorg 2015* [3]. Daarnaast zijn er sinds 2012 vrije tarieven in de farmacie, waardoor er naast de prestatiebekostigingen ook marge behaald kan worden op de inkoop-verkoop van geneesmiddelen. De prestatiebekostigingen zoals beschreven door de NZa zijn in drie groepen te verdelen:

- Terhandstelling van een UR-geneesmiddel en bijbehorende zorg. De kosten van deze zorg kunnen een toeslag bevatten in bepaalde gevallen:
  - Bij de eerste keer van uitgifte van het geneesmiddel
  - Zorg welke in het weekend of 's nachts verleend is
  - Bij geneesmiddelen die de apotheker zelf moet bereiden (magistrale bereiding)
  - Bij geneesmiddelen die in meerdere eenheden geleverd worden
- Advies, voorlichting en begeleiding bij medicijn gebruik, zelfzorg, ziekenhuis/polikliniek bezoek en reizen
- Facultatieve prestaties: prestatiebeschrijving overeengekomen tussen zorgverzekeraar en zorgverlener, welke afwijkt van de door de NZa vastgestelde prestaties

Ongeveer de helft van de prestaties behoren tot de verzekerde zorg in het basispakket. Zorgverzekeraars dienen deze zorg in contracten met zorgverleners vast te leggen voor de bij hen aangesloten verzekerden. Over de overige prestaties kunnen verzekeraars selectiever afspraken maken met de zorgverleners. Voor patiënten worden UR-geneesmiddelen in zijn geheel vergoed, met uitzondering van het eigen risico. Daarnaast geldt voor een aantal van deze geneesmiddelen een eigen bijdrage van de patiënt.

2.3.1 Onrechtmatige declaraties

Binnen de farmacie bestaan een aantal vormen van onrechtmatig declareren. Dit houdt in dat zorgverleners declaraties indienen waarmee zij vergoedingen krijgen op een onrechtmatige manier. De NZa heeft deze onrechtmatige manieren uitgebreid beschreven in *Rapport Onderzoek zorgfraude* [2], paragraaf 7 *Extramurale Farmaceutische zorg*. Onderstaande is een opsomming van *tabel 7.1* tot en met *tabel 7.7* uit *Rapport Onderzoek zorgfraude*:

Overbehandeling Het onnodig vaak voorschrijven van een geneesmiddel

Een apotheekhoudende huisarts schrijft bovengemiddeld veel geneesmiddelen voor. Zijn patiënten komen het geneesmiddel vervolgens ook weer bij hem halen.

Onderbehandeling Minder eenheden afleveren dan nodig

Een apotheker levert een patiënt minder eenheden van het geneesmiddel dan voorgeschreven door een arts of overeengekomen met een zorgverzekeraar. De patiënt moet hierdoor nog eens terugkomen en de apotheker kan hierdoor nogmaals een prestatie in rekening brengen. Op het moment dat de patiënt onnodig terugkomt, is er niet langer sprake van onderbehandeling, maar van opknippen.

Spookzorg Het in rekening brengen van zorg die niet of deels geleverd is

Een patiënt komt bij de apotheek voor zijn geneesmiddelen. De apotheek brengt het uitgeven van de geneesmiddelen en de geneesmiddelen zelf in rekening. Vervolgens dient hij deze factuur elke week in bij de verzekeraar terwijl de patiënt niet meer is geweest;

Een patiënt komt bij de apotheek om zijn geneesmiddelen op te halen. De apotheker brengt een toeslag voor magistrale bereiding in rekening terwijl hij het geneesmiddel niet zelf (of onnodig) bereid heeft.

Upcoding Het declareren van duurdere behandelingen

Een patiënt haalt een doosje met 20 pillen bij de apotheker. De apotheker factureert echter een doosje met 40 pillen bij de verzekeraar.

Dubbele bekostiging U-bocht constructie

Een vrouw komt in het ziekenhuis voor het laten plaatsen van een spiraaltje. De behandelend arts vraagt de vrouw om eerst zelf een spiraaltje aan te schaffen via de apotheek. Vervolgens plaatst de arts dit spiraaltje. Zowel de apotheek als de arts/het ziekenhuis ontvangt nu een vergoeding voor het spiraaltje.

Andere prestatie Het in rekening brengen van een andere prestatie om een vergoeding te krijgen

Een patiënt komt bij de apotheker om Viagra te halen. Viagra behoort niet tot het verzekerde pakket en de patiënt dient dit dan ook zelf af te rekenen. Hij vraagt echter om het geneesmiddel als pijnstiller in rekening te brengen zodat het wel wordt vergoed. De apotheker doet dit vervolgens.

De factuur voor de geleverde diensten en/of producten gaat vanuit de zorgverlener direct naar de zorgverzekeraar. Daardoor kan een patiënt niet makkelijk zien of de zorg die bij zijn zorgverzekeraar in rekening is gebracht daadwerkelijk op de gedeclareerde manier geleverd is. Mocht de patiënt wel de factuur zien, dan kan hij alsnog niet altijd bepalen of de gefactureerde diensten en/of producten daadwerkelijk geleverd zijn. Een patiënt weet ondanks de factuur bijvoorbeeld niet of een apotheker daadwerkelijk een medicijn zelf bereid heeft of niet. Daarnaast wordt vaak aangenomen dat huisartsen het juiste voorschrijven en daardoor wordt over- en/of onderbehandeling door de patiënt niet opgemerkt. Kortom, de patiënt heeft beperkt inzicht in het declaratie process waardoor het vinden van onrechtmatige declaraties compleet in handen ligt van de zorgverzekeraar.

2.4 Onrechtmatige declaraties binnen Zilveren Kruis

Om onnodige kosten te voorkomen, moeten onrechtmatig gedeclareerde bedragen zoveel mogelijk teruggevorderd worden. De controles die binnen Zilveren Kruis hiervoor uitgevoerd worden, zijn in twee groepen te verdelen: Formele Controle en Materiële Controle.

2.4.1 Formele Controle

Formele Controles bestaan uit hard-coded checks welke op elke declaratie toegepast worden. Dit zijn voorgeprogrammeerde, automatische controles. Er is een zogenaamde *risicomatrix* waarin voor elke zorgsector een aantal controles staan die op elke declaratie binnen een bepaalde sector toegepast kunnen worden. Een voorbeeld hiervan is een check die controleert of het gedeclareerde bedrag correct is, op basis van het gedeclareerde product.

2.4.2 Materiële Controle

Materiële Controles zijn fysieke controles waarbij de administratie en declaraties van een zorgverlener op meerdere onderdelen onderzocht worden. Deze controles kunnen niet geautomatiseerd worden, omdat er veel verschil bestaat tussen de manieren waarop zorgverleners hun administratie bijhouden en omdat de controles voornamelijk op locatie bij de zorgverlener uitgevoerd moeten worden. Alleen dan is er toegang tot de benodigde data. Doordat deze controles arbeidsintensief en zodoende uitgebreid zijn dat het financieel niet mogelijk is om elke zorgverlener te onderzoeken, dient er een selectie gemaakt te worden. Hoewel er bij zowel Materiële Controles als Formele Controles onrechmatige declaraties gevonden worden, wordt er vaker fraude geconstateerd tijdens een Materiële Controle, aangezien deze controles uitgebreider zijn en toegang hebben tot meer data.

2.5 Data Mining

Data mining is simpel gezegd het zoeken naar verbanden in data. Dit kan handmatig, maar gezien de omvang en complexiteit van de te onderzoeken data worden algoritmen gebruikt om de verbanden te vinden. Een verzameling van data bestaat bijvoorbeeld uit de registratie van praktijkgegevens, of resultaten van eerder onderzoek. Dit onderzoek focust zich op het zoeken van verbanden met als doel het maken van voorspellingen, waarbij er een classificatie toegekend dient te worden met twee mogelijke waarden: wel terugvordering en geen terugvordering. Er zijn verschillende technieken om deze voorspelling te maken. Een techniek die belangrijk is voor dit onderzoek is random forest. Deze techniek wordt nader uitgelegd in hoofdstuk 2.5.2 *Random forest*. Volledige uitwerking van alle verschillende soorten data mining doelen en technieken is te vinden in *Data Mining: Practical Machine Learning Tools and Techniques* [14].

2.5.1 (Un)supervised learning

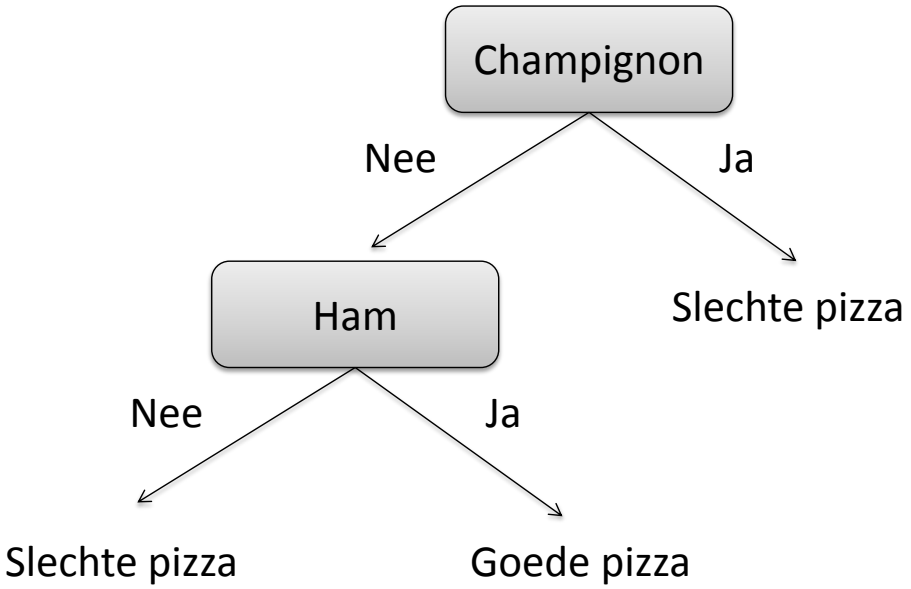
Elke classificatie taak is te bestempelen als unsupervised- of supervised learning probleem. *Unsupervised learning* houdt in dat de data ongelabeld is, oftewel dat er geen gevallen zijn waarvan geleerd kan worden. Een voorbeeld hiervan is een algoritme welke in een ongelabelde zoekt naar samenhangende groepen (clusters), zoals opvallende groepen zorgverleners met overeenkomende eigenschappen. *Supervised learning* houdt in dat de data gelabeld is, oftewel dat er data is waarop getraind kan worden, waarbij al aangegeven is wat het te bouwen model op zou moeten leveren. Een voorbeeld hiervan is een algoritme welke op basis van zorgverleners, waarvan bekend is of ze onrechtmatige declaraties ingediend hebben, voor nieuwe zorgverleners kan voorspellen of ze onrechtmatig declaraties in zullen dienen of niet. In dit onderzoek wordt dus een supervised learning probleem behandeld. Donalek laat in *Supervised and Unsupervised Learning* [1] zien welke algoritmen binnen welke *learning* vorm passen.

2.5.2 Random forest

Een random forest is een verzameling decision trees die op een bepaalde manier met elkaar samenwerken en tevens een supervised learning techniek. Om dit te beschrijven wordt eerst uitgelegd hoe een decision tree werkt.

Decicion tree

Een decision tree is een classificatie methode waarvan de output een boomvormige structuur heeft. Zie ter illustratie figuur 2.1. Een pizza heeft twee features: champignon en ham. Beide features zijn binair en kunnen dus de waarde *ja* en *nee* aannemen. Op basis van deze decision tree kan elke pizza geclassificeerd worden tot een goede pizza of slechte pizza, op basis van de aanwezigheid van champignon en ham.



Figuur 2.1: Decision tree om een pizza te classificeren: is de pizza goed of slecht?

Om een decision tree te bouwen is er een training set nodig. De training set welke gebruikt is voor figuur 2.1, zou er uit kunnen zien als tabel 2.1.

Champignon	**Ham**	**Kwaliteit**
Ja	Ja	Slecht
Ja	Nee	Slecht
Nee	Nee	Slecht
Nee	Ja	Goed

Tabel 2.1: De training set van figuur 2.1

De decision tree op basis van de data uit tabel 2.1 is erg vanzelfsprekend. Precies die en alleen die vier gevallen worden gemodelleerd. Het zou niet uitmaken of de eerste split op champignon of op ham plaatsvindt. Dit blijkt echter in de praktijk anders. Sommige features hebben een sterker verband met het gene dat voorspeld dient te worden, de target, dan andere features. Decision trees gebruiken maten zoals *Information Gain* om te berekenen welke features het sterkste verband hebben met de target. Op deze features wordt als eerst gesplitst, zodat er met een zo klein mogelijke boom een classificatie gemaakt kan worden. Uitgebreide uitleg en toepassingen van decision trees zijn te vinden in *Data Mining with Decision Trees: Theory and Applications* [10].

Random forest

Zoals gezegd, is een random forest een verzameling decision trees die op een bepaalde manier met elkaar samenwerken. Een random forest genereert op basis van de training set een aantal subsets, welke op basis van willekeur (random) een aantal gevallen uit de training set bevatten. Voor al deze random subsets wordt er een decision tree gegenereerd. Het genereren van de decision trees in een random forest kent één verschil met het genereren van normale decision trees. Bij de decision trees in een random forest wordt er bij het berekenen van de feature waar het beste op gesplitst kan worden een random subset aan features geëvalueerd, in plaats van alle features. Als alle decision trees gegenereerd zijn, kunnen nieuwe gevallen geclassificeerd worden door middel van het nieuwe geval door alle decision trees te laten classificeren. Al deze classificaties worden daarna volgens een bepaalde weging opgeteld tot een uitkomst. Een simpel voorbeeld is een random forest met vijf decision trees, welke allemaal *0* of *1* classificeren. Als er drie stuks een *0* classificeren en twee stuks een *1*, dan is de uitkomst *0*, uitgaande van een wegingsfunctie die alleen kijkt naar de classificatie die het meeste voorkomt.

Hoofdstuk 3

Aanpak

Dit hoofdstuk beschrijft de aanpak en de overwegingen welke leiden tot de uiteindelijke voorspellingsmodellen.

De eerste stap is het bepalen van de target. Wat moet er uiteindelijk voorspeld worden? De tweede stap is het verzamelen van de benodigde data uit verschillende databronnen. Deze data wordt samengevoegd, gefilterd en verrijkt door afgeleide features toe te voegen welke volgen uit gesprekken met specialisten op het gebied van terugvorderingen en onrechtmatige declaraties. Om inzicht te krijgen in de data worden er aan aantal analyses uitgevoerd waarmee de te bouwen modellen verklaard kunnen worden. De geprepareerde dataset dient als input voor een classificatie algoritme welke op basis van de gegeven dataset modellen bouwt waarmee zowel de kans op terugvorderingen voorspeld kan worden als het terug te vorderen bedrag. De laatste stap is het valideren van de modellen.

3.1 Gebruikte software

Om verwijzingen naar de gebruikte software binnen dit hoofdstuk te verduidelijken wordt de gebruikte software eerst beschreven. Binnen het Business Intelligence Analyse & Rapportage team, het team waarbinnen dit onderzoek plaatsvindt, wordt hoofdzakelijk SAS gebruikt als dataverwerking software. SAS biedt onder andere mogelijkheid tot samenvoegen, filteren en verrijken van data. Dit zijn dan ook de redenen waarom voor dit onderzoek SAS wordt gebruikt als dataverwerking software. Hoewel SAS ook data mining functionaliteit biedt, wordt bij dit onderzoek hiervoor *Weka 3: Data Mining Software in Java (http://www.cs.waikato.ac.nz/ml/weka/)* gebruikt omdat deze software meer data mining algoritmen biedt en in de huidige onderzoeksopzet gemakkelijker op snelle en eenvoudige wijze modellen kan bouwen.

3.2 Target

Zoals in hoofdstuk 2.4 *Onrechtmatige declaraties binnen Zilveren Kruis* besproken is, streeft Zilveren Kruis ernaar om onrechtmatige declaraties terug te vorderen. Dit doet ze door zowel Formele- als Materiële Controles uit te voeren. Uit hoofstuk 2.4 *Onrechtmatige declaraties binnen Zilveren Kruis* is te leren dat de Materiële Controles niet te automatiseren zijn. Er moet daarom om financiële redenen een selectie gemaakt worden, die bepaald welke zorgverleners er onderzocht gaan worden. Om deze selectie te bepalen zal de target gaan over terugvorderingen door Materiële Controles.

De afdeling Naleving & Controle selecteert normaliter de te onderzoeken zorgverleners, voert de controles uit en verwerkt de controles. Hiermee is deze afdeling de belangrijkste stakeholder. Aan de hand van gesprekken met specialisten en de manager van de afdeling Naleving & Controle blijkt dat er zowel interesse is in de kans dat er bij een zorgverlener een of meerdere terugvorderingen plaats gaan vinden als in het mogelijk terug te vorderen bedrag. Daarnaast is er ook vermeld dat het aantal terugvorderingen interessant is. Het blijkt echter dat in de database waarin de onrechmatige declaraties opgeslagen staan (zie hoofdstuk 3.3 *Data verzamelen*, ROC) er voor Materiële Controles niet wordt bijgehouden om hoeveel declaraties het gaat, maar dat er bulkboekingen gemaakt worden over bepaalde perioden. Hierdoor is het onmogelijk om een uitspraak te doen over het aantal onrechtmatige declaraties en daarmee het aantal terug te vorderen declaraties. Dit leidt tot het besluit dat er twee targets zijn:

- Kans op terugvordering bij een zorgverlener door Materiële Controle
- Terug te vorderen bedrag bij een zorgverlener door Materiële Controle

3.3 Data verzamelen

Binnen Zilveren Kruis zijn er een drietal interne databronnen waarin data over zorgverleners, dan wel data over declaratiegedrag staat:

- MIAZ (*Management Info Achmea Zorg*) Dimensie Zorgrelatie: basis gegevens van zorgverleners zoals NAW en type zorgverlener.
- MIAZ Farmacie Infomart: alle credit en debet boekingen op declaratieniveau binnen de farmaceutische sector. Dit betreffen de declaratie uitbetalingen en terugvorderingen vanuit Formele Controles, en terugbetalingen op eigen initiatief van zorgverlener.

De twee MIAZ databronnen zijn tabellen welke in de zogenaamde MIAZ database staan. MIAZ is een centrale database waarin de meeste data omtrent zorgverleners en declaraties binnen Zilveren Kruis opgeslagen staan. Naast de centrale MIAZ database is er ook de ROC database:

– ROC: terugvorderingen door de afdelingen Materiële Controle, Formele Controle en Speciale Zaken (wettelijk bewezen fraude). Let op, de naam *ROC* heeft niks te maken met het begrip *receiver operating characteristic*, vaak afgekort als ROC, welke uitgebreid besproken wordt in *ROC Graphs: Notes and Practical Considerations for Data Mining Researchers* [4].

Alle interne databronnen bevatten data omtrent zorgverleners die een contract hebben met Zilveren Kruis. Naast de interne bronnen is er ook een interessante externe bron, genaamd het AGB-register. Dit register wordt onderhouden door het bedrijf Vektis.

– AGB-register: vergelijkbare gegevens als MIAZ Dimensie Zorgrelatie, met als uitbreiding relaties tussen zorgverleners, vestigingen en eigenaren van vestigingen. Deze databron bevat alle geregistreerde zorgverleners in Nederland.

3.3.1 Dataset per kalenderjaar

Om de modellen te bouwen, valideren en testen wordt er data gebruikt uit 2012 en 2013. Data uit de jaren 2014 en nieuwer is nog niet compleet in de zin dat er nog fraude- en terugvorderingsonderzoeken lopen waarvan de uitslag nog niet bekend is. In de jaren voor 2012 is er significant minder data geregistreerd dan in latere jaren, waardoor het niet mogelijk is om deze nuttig te gebruiken. De data uit 2012 wordt gebruikt voor de training en validatie van de modellen. De data uit 2013 dient als test set. Deze verdeling wordt toegelicht in hoofdstuk 3.7 *Modellen valideren*. Zowel de training/validatie- als de test set bevatten dus data uit precies één kalenderjaar. Dit sluit goed aan bij het feit dat Materiële Controle aan het eind van elk kalenderjaar controles uitvoert over dat afgelopen kalenderjaar en vervolgens de terugvorderingen maakt, welke vaak zelfs op bulkniveau zijn. Hiermee maakt ze dus vaak per zorgverlener maar enkele terugvorderingen, welke gebaseerd zijn op meerdere onrechtmatige declaraties uit het gehele afgelopen kalenderjaar. Dit in tegenstelling tot Formele Controle welke per declaratie terugvorderingen maakt. De aanpak om de datasets per kalenderjaar in te delen zorgt er ook voor dat features op declaratieniveau gemakkelijk geaggregeerd kunnen worden naar zorgverlener niveau. Bijvoorbeeld het aantal declaraties van een zorgverlener kan zo per kalenderjaar gesommeerd worden en als feature worden toegekent aan de zorgverlener.

3.4 Data prepareren

De verzamelde data moet gefilterd worden om onbruikbare features te verwijderen, verrijkt worden met nieuwe features om impliciete informatie in de data welke mogelijk waardevol is expliciet uit te drukken zodat het door het classificatie algoritme gebruikt kan worden, uit verschillende data bronnen samengevoegd worden en als laatst moeten missende waarden geschat en ingevuld worden. Het is belangrijk om in acht te

nemen dat het model een voorspelling moet doen over zorgverleners en dat daarom de geprepareerde dataset op zorgverlener niveau moet zijn. Het prepareren van de data kan gezien worden als het *Data Preparation* onderdeel uit het CRISP-DM model. Hierbij wordt de ruwe data omgezet naar een dataset, welke bruikbaar is voor het classificatie algoritme.

3.4.1 Data filteren en verrijken

Het filteren en verrijken van de data is eigenlijk één proces waarbij de twee acties door elkaar heen gebeuren. Dit komt doordat er twee soorten data zijn, wat ook weer leidt tot verschillende soorten features:

- MIAZ Dimensie Zorgrelatie en het AGB-register bevatten data welke reeds op zorgverlener niveau is.
- MIAZ Farmacie Infomart en ROC bevatten data die op declaratieniveau is.

Aan de hand van gesprekken met specialisten en managers van de afdelingen Naleving & Controle en Speciale Zaken zijn naast de targets ook een aantal mogelijk voorspellende features van zorgverleners bedacht. Deze features zijn te vinden in bijlage 6.1 *Features uit gesprekken met stakeholders* en zijn onder te verdelen in vier groepen:

- Features welke al daadwerkelijk in een van de databronnen op zorgverlener niveau te vinden zijn. Een voorbeeld is het type zorgverlener.
- Features welke direct afgeleid kunnen worden uit features die al in een van de databronnen op zorgverlener niveau aanwezig zijn. Een voorbeeld is het postcodegebied uit de volledige alfanumerieke postcode.
- Features welke indirect, met behulp van een externe databron, afgeleid kunnen worden uit features die al in een van de databronnen op zorgverlener niveau aanwezig zijn. Een voorbeeld is het inwoneraantal uit het gemeentenummer waarin de zorgverlener actief is, met behulp van gegevens van het CBS.
- Features welke door het aggregeren van data op declaratieniveau terecht komen op zorgverlener niveau. Een voorbeeld hiervan is het aantal declaraties per zorgverlener.

Zoals eerder vermeld moet de geprepareerde dataset op zorgverlener niveau zijn. De MIAZ Dimensie Zorgrelatie en het AGB-register kunnen dus direct gebruikt worden, terwijl de MIAZ Farmacie Infomart en ROC geaggregeerd moeten worden naar zorgverlener niveau. Dit proces heeft betrekking op het filteren en selecteren van features waaruit waardevolle informatie afgeleid kan worden en daarmee heeft het dus ook betrekking op het verrijken van de geprepareerde dataset, doordat er nieuwe features gegenereerd worden dankzij aggregatie. Wanneer alle data op zorgverlener niveau is kunnen de andere drie soort features geselecteerd en afgeleid worden.

In verband met de tijdsbeperking van dit onderzoek zijn niet alle bedachte features zoals beschreven in bijlage 6.1 *Features uit gesprekken met stakeholders* daadwerkelijk geïmplementeerd in de dataset, maar is er een selectie gemaakt van de door Naleving & Controle belangrijkst geachte features. Deze lijst, inclusief afkortingen gebruikt later in dit onderzoek, is te vinden in bijlage 6.2 *Geïmplementeerde features*

3.4.2 Data samenvoegen

Om het voorspellingsmodel te kunnen bouwen is het essentieel om uiteindelijk een enkele dataset over te houden waarin alle relevante data staat. Bij zowel de geaggregeerde data als bij de data welke geselecteerd is uit databronnen welke reeds op zorgverlener niveau zijn wordt ook altijd het interne Zilveren Kruis relatienummer en/of het AGB nummer bewaard. Dankzij het feit dat de MIAZ Dimensie Zorgrelatie databron beide nummers bevat kunnen alle datasets aan elkaar gekoppeld worden.

3.4.3 Missende waarden

Om een dataset over te houden welke door elk classificatie algoritme gebruikt kan worden om een model te bouwen is het van belang dat de dataset geen missende waarden meer bevat. Om deze te elimineren worden drie veelgebruikte methoden besproken in *Data Mining Concepts, Models, Methods and Algorithms* [8]:

- Elke regel met een of meerdere missende waarden verwijderen
- Missende waarden vervangen door de modus of het gemiddelde van de feature
- Een voorspellingsmodel gebruiken om de missende waarden te voorspellen

Een aanvulling hierop is omschreven in *Using Classification and Regression Trees (CART) in SAS® Enterprise Miner™ For Applications in Public Health* [5].

- Een extra feature toevoegen die de aan-, en daarmee afwezigheid, van een andere feature aangeeft

De eerste drie de methoden hebben als gevolg dat er een dataset overblijft zonder missende waarden. De eerste en simpelste methode zorgt echter voor verlies van informatie en is daardoor alleen effectief op datasets met een klein percentage gevallen met missende waarden. De training set zoals omschreven in hoofdstuk 3.3.1 *Dataset per kalenderjaar* bevat 463 van de 3580 rijen, dus ongeveer 13%, met missende waarden en maakt hiermee deze methode niet geschikt. Om het onderzoek beperkt te houden is vervolgens besloten om de tweede methode te gebruiken, waarbij missende waarden vervangen worden door de modus in het geval van een nominale feature en het gemiddelde in het geval van een continue feature. Deze keuze wordt ook gehanteerd bij ander onderzoek in de volksgezondheid sector, zoals in *Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks* [13]. Om vervolgens de data zo te prepareren dat het duidelijk

is dat bepaalde waarden geschat zijn, wordt ook de vierde methode toegepast. Dit is belangrijk, omdat het feit dat een waarde mist juist waardevol kan zijn. Bij het declaratie process van geneesmiddelen door apothekers kan er bijvoorbeeld een verband bestaan tussen het niet invullen van de naam van de patiënt waar het geneesmiddel aan uitgegeven is en het feit dat de apotheker frequent onrechtmatige declaraties indient van geneesmiddelen welke nooit uitgegeven zijn.

3.5 Inzicht in de data

Om de stakeholders meer inzicht in de data te geven en om prominente features te vinden, worden er analyses van de data gemaakt. Deze analyses vervullen de rol van *Data Understanding* volgens het CRISP-DM model. Ten eerste wordt er gekeken naar de correlatie tussen verklarende features. Daarnaast wordt voor beide targets, "terugvordering vanuit Materiële Controle" en "bedrag teruggevorderd door Materiële Controle", de correlatie uitgerekend ten opzichte van de verklarende features. Als laatst wordt er voor elke feature een histogram gemaakt, waarmee voor elke waarde binnen die feature het percentage zorgverleners mét terugvorderingen wordt gegeven. Deze histogrammen geven inzicht in de eigenschappen welke zorgverleners met terugvorderingen vooral hebben. Nominale features krijgen in de histogram per mogelijke waarde een bin. Voor continue features worden tien bins uitgerekend met een zo gelijk mogelijk aantal zorgverleners per bin. Stel er zijn 1000 zorgverleners, dan worden de grenzen van de bins dus zo berekend dat er in elke bin 100 zorgverleners zitten. Dit voorkomt dat er door uitschieters in de data bins zijn, tussen de "normale" waarden en de uitschieters in, waar helemaal geen zorgverleners in zitten. Er zijn echter wel continue features waarvoor minder dan tien bins beschikbaar zijn. Dit komt doordat er voor bepaalde features waarden zijn welke heel veel voorkomen, zoals nul. Stel er zijn 1000 zorgverleners, waarvan 200 voor deze feature waarde nul hebben, dan bevatten de eerste twee bins alleen maar zorgverleners met waarde nul. Deze bins worden samengevoegd tot één bin.

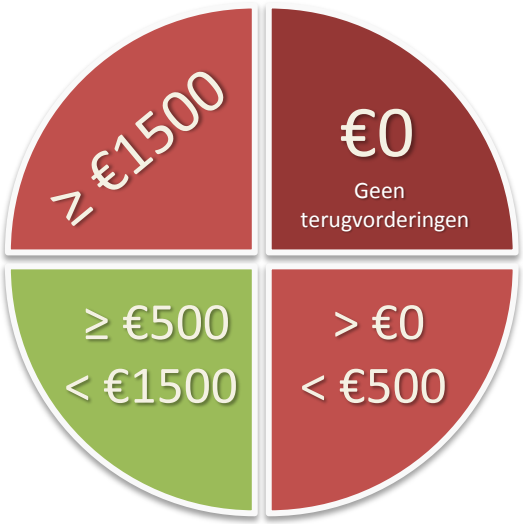
3.6 Modellen bouwen

De volgende stap in het process is *Modeling*, zoals het CRISP-DM model aangeeft. Gezien het feit dat er twee targets zijn worden er meerdere modellen gebouwd. Het is daarbij wenselijk dat er een samenhang is in de voorspelling tussen de kans op en het bedrag van terugvordering en dus niet simpelweg twee losse, op zichzelf staande modellen waarbij er een de kans voorspelt en de ander het bedrag. Volgens de stakeholders is de kans de belangrijkste target en zou het fijn zijn als er een indicatie van het bedrag gegeven kan worden.

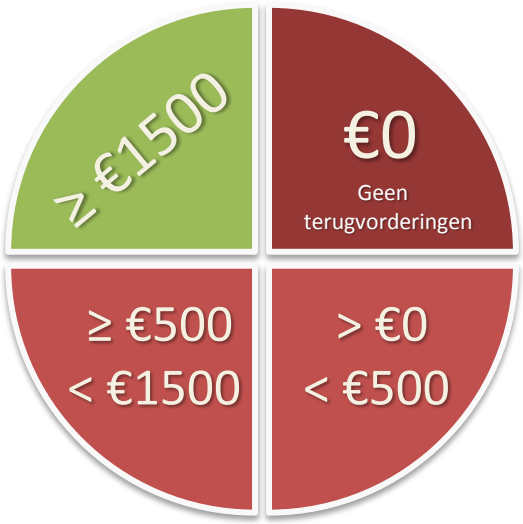
3.6.1 Implementatie van de twee targets

Het feit dat de kans op terugvordering de belangrijkste target is biedt een oplossing waarmee er een samenhang ontstaat tussen de kans op en het bedrag van terugvordering. De oplossing begint bij het bouwen van modellen op basis van verschillende datasets. Ter herinnering: elke rij in een dataset representeert een zorgverlener en zoals hoofdstuk 3.3.1 *Dataset per kalenderjaar* aangeeft zijn de datasets gebasseerd op een kalenderjaar aan data. Voor elke dataset geldt vervolgens dat de target aangeeft of er, per zorgverlener, terugvorderingen zijn geweest van declaraties uit het afgelopen kalenderjaar. Door hier een classificatie model op te bouwen en deze aan het eind van elk kalenderjaar toe te passen op de zorgverleners die dat kalenderjaar declaraties ingedient hebben, kan dus voorspeld worden hoe groot de kans is dat er bij die zorgverleners declaraties uit dat kalenderjaar teruggevorderd kunnen worden. Dit is exact in lijn met de werkwijze van Materiële Controle. Zij bepalen aan het eind van elk kalenderjaar welke zorgverleners er onderzocht gaan worden op onrechtmatige declaraties gedurende het afgelopen kalenderjaar, welke vervolgens teruggevorderd zullen worden.

De implementatie van de voorspelling van het verwachtte terug te vorderen bedrag gebeurd door per dataset grenzen te bepalen, waarbinnen het teruggevorderde bedrag dient te vallen. De gevallen binnen deze grenzen zijn targets en alle andere gevallen zijn non-targets. Dan rest nog de vraag hoe deze grenzen bepaald zijn. Om het onderzoek beperkt te houden is besloten om twee modellen te bouwen. Om de modellen zo goed mogelijk aan te laten sluiten bij de wensen van de eindgebruikers van Naleving & Controle, de afdeling welke onder andere Materiële Controles als taak heeft, zijn de grenzen in overleg met hen bepaald. Naleving & Controle heeft bepaald geen actie te ondernemen bij een verwacht terug te vorderen bedrag lager dan €500. Dit is dus de ondergrens. Er bleek echter geen duidelijke voorkeur te zijn voor de tweede grens. In de gehele training set van 3579 zorgverleners zijn er 821 targets, oftewel gevallen met terugvorderingen. Dit zijn er boven de ondergrens van €500 nog 783. Om de kwaliteit van de modellen te stimuleren is gekozen om de tweede grens zo te kiezen dat er op basis van de training set in beide modellen ongeveer evenveel targets zitten. Hierdoor is de balans tussen targets en non-targets in beide modellen ongeveer gelijk en wordt niet een van de twee modellen mogelijk meer benadeelt dan de ander, door een tekort aan targets. De mediaan van het geval met een bedrag tussen de €500 en €∞ ligt op €1375,91. Afgerond naar €1500 blijft er een verdeling over van 405 targets in de eerste dataset en 378 targets in tweede dataset. Beide datasets bevatten dus alle 3579 zorgverleners, maar per dataset is een andere groep zorgverleners de target. Zie ter illustratie firguur 3.1, 3.2 en tabel 3.2



Figuur 3.1: Dataset 1 heeft als **target** zorgverleners met een bedrag aan terugvorderingen vanaf €500 tot €1500 en als **non-target** alle zorgverleners met een ander bedrag aan terugvorderingen.



Figuur 3.2: Dataset 2 heeft als **target** zorgverleners met een bedrag aan terugvorderingen vanaf €1500 en als **non-target** alle zorgverleners met een ander bedrag aan terugvorderingen.

Tabel 3.1: Alternatieve representatie van figuur 3.1 en 3.2. Per dataset geeft een groen vakje de **target** aan en een rood vakje de **non-target**, waarbij elk vakje een groep zorgverleners representeert met een bedrag aan terugvorderingen binnen bepaalde grenzen.

Naast de modellen die rekening houden met het bedrag aan terugvorderingen, zal er ook een model gebouwd waarbij alle gevallen met terugvorderingen boven de €500 de target zijn. Tabel 3.2 laat zien hoe deze dataset zich representeert in de notatie zoals gebruikt in tabel 3.1. Aan de hand van dit model kan de performance verandering onderzocht worden van de toevoeging van de tweede target, het bedrag aan terugvorderingen, ten opzichte van een model waarbij alleen kans op terugvordering een rol zou spelen.

Tabel 3.2: Dataset waar de target onafhankelijk is van het bedrag aan terugvorderingen. Een groen vakje geeft de **target** aan en een rood vakje de **non-target**, waarbij elk vakje een groep zorgverleners representeert met een bedrag aan terugvorderingen binnen bepaalde grenzen.

3.6.2 Gewogen datasets

Zoals duidelijk wordt in hoofdstuk 3.6.1 *Implementatie van de twee targets* zijn de datasets ongebalanceerd. Met 405 targets in de ene dataset en 378 targets in de andere dataset, ten opzichte van 3579 gevallen totaal in beide datasets, zijn er in beide datasets net iets meer dan 10% targets. Dit kan een probleem opleveren bij sommige classificatie algoritmen. Deze algoritmen nijgen namelijk overwelmd te worden door de klasse welke in de meerderheid is, de non-targets in dit geval, waardoor de andere klasse genegeert wordt. Dit gebeurdt omdat sommige algoritmen hun performance proberen te optimaliseren over het absolute aantal gevallen in de dataset [9]. De Classbalancer in Weka (zie hoofdstuk 3.1 *Gebruikte software*) biedt hier echter een oplossing voor. Deze optie kent gewichten toe aan de gevallen, zodat elke klasse (wel terugvordering, geen terugvordering) dezelfde totale weging heeft. Daarnaast blijft de totale som van alle (gewogen) gevallen gelijk. Dit houdt in dat er bijvoorbeeld bij een dataset met 300 targets en 700 non-targets gewichten toe worden gekend waardoor de totale weging van zowel de targets als non-targets op 500 komt. Dit zorg er tevens voor dat het totaal gewicht *2 * 500 = 1000* wordt, wat gelijk is aan *700 + 300 = 1000*. Elk algoritme in Weka heeft zijn eigen implementatie hoe het vervolgens omgaat met de gewichten.

3.6.3 Algoritmen

Om per dataset het beste classificatie algoritme te vinden worden er verschillende algoritmen uit Weka (zie hoofdstuk 3.1 *Gebruikte software*) toegepast om de modellen te bouwen, hieronder vermeld als "Type: Weka naam (toelichting)":

- Naive Bayes: NaiveBayes
- Logistic Regression: SimpleLogistic
- Support Vector Machine: LibSVM
- K-nearest neighbours: IBk
- Decision Table: DecisionTable
- Decision Tree: J48 (Weka's C4.5 implementatie)
- Decision Forest: RandomForest

3.6.4 Parameter tuning

Om de modellen te optimaliseren, worden voor elke dataset de parameters van de gebruikte algoritmen aangepast. Zo zou voor een beslisboom bijvoorbeeld een maximale diepte ingesteld kunnen worden en voor een nearest neighbour algoritme de manier veranderd kunnen worden waarop de afstand tot buren berekend wordt. Om het effectief maar simpel te houden wordt dit op een "greedy" manier gedaan: de

eerste parameter wordt aangepast tot er een maximale verbetering in performance van het model plaats vindt. In dit geval een verbetering in AUC, meer daarover in hoofdstuk 3.7.1 *Performance maat*. Mocht er geen verbetering gevonden worden behoudt de parameter zijn initiële waarde. Vervolgens wordt er geprobeerd die situatie te verbeteren door de volgende parameter aan te passen. Dit herhaalt zich tot alle parameters mogelijk zijn aangepast. Naast deze gretige, handmatige aanpak bestaan er ook geautomatiseerde manieren van parameter tuning. Een voorbeeld is *grid-search*. Hierbij wordt er voor meerdere parameters een bereik en stap-waarde ingesteld, waarna alle mogelijke combinaties geprobeerd worden. De combinatie waarmee de performance van het model maximaal is wordt vervolgens gebruikt. [11].

3.7 Modellen valideren

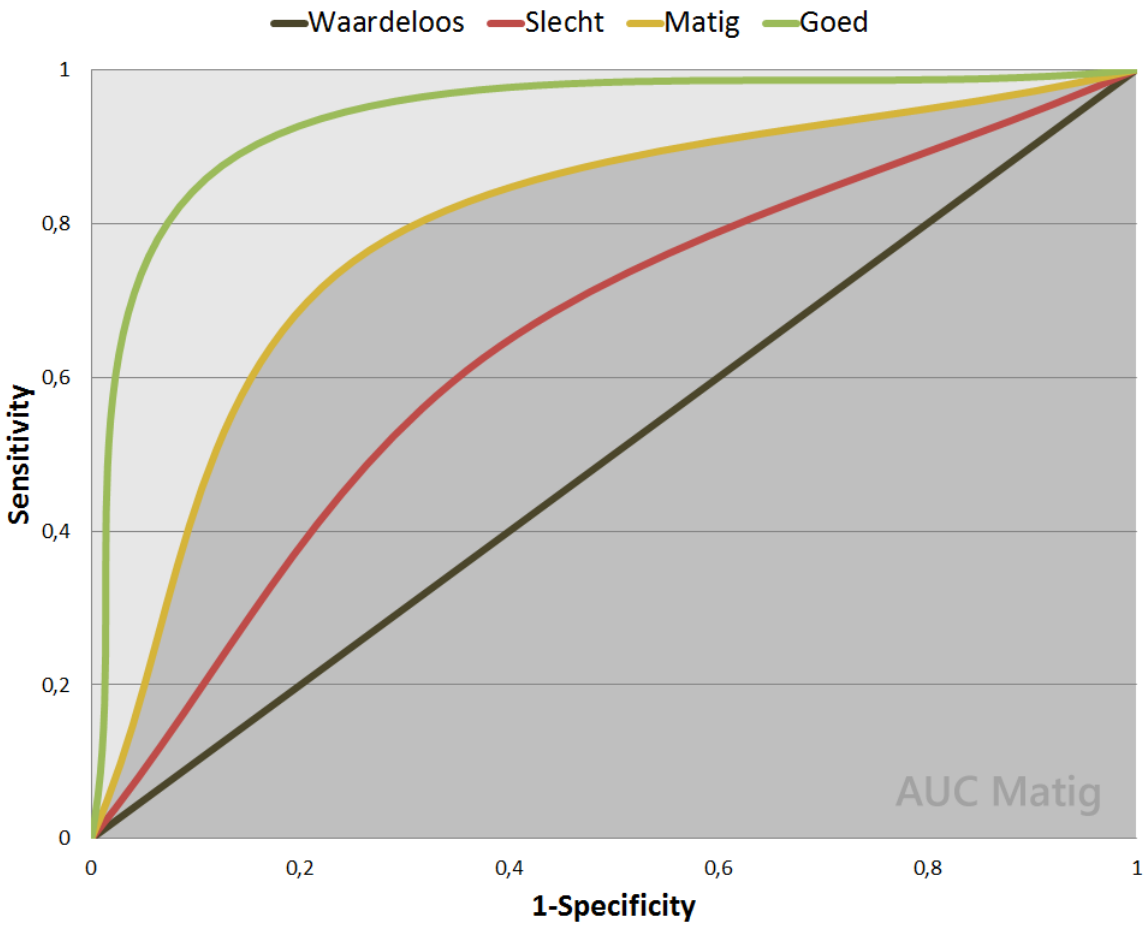
De gebouwde modellen moeten gevalideerd worden, zoals aangegeven als *Evaluation* in het CRISP-DM model. De modellen worden om meerdere redenen gevalideerd. Ten eerste om te controleren hoe goed (zie hoofdstuk 3.7.1 *Performance maat*) de verschillende modellen op zichzelf staand voorspellingen kunnen maken. Hiermee is te bepalen welk algoritme voor welke dataset de hoogste performance haalt. Daarnaast dient de validatie om de performance van de uiteindelijke modellen met elkaar te vergelijken. Als laatst wordt er gekeken naar de performance verandering van de toevoeging van het bedrag aan terugvorderingen ten opzichte van een model waarbij alleen kans op terugvordering een rol zou spelen.

Zoals in hoofstuk 3.3.1 *Dataset per kalenderjaar* bepaald is, wordt de data uit 2012 gebruikt als training- en validatie set en de data uit 2013 als test set. De train set dient om het model te bouwen, de validatie set om de parameters te tunen en de test set om het getunede model te beoordelen. Als methode om 2012 te gebruiken als zowel train- als validatie set wordt 10-cross-fold validatie gebruikt. Het voordeel van deze methode ten opzichte van een split volgens een vast percentage in train- en validatie set is dat alle gevallen gebruikt worden voor zowel training als validatie, en dat elk geval precies één keer gebruikt wordt voor validatie. Bij 10-cross-fold validatie wordt uit de tien ronden de gemiddelde performace maat berekend als uiteindelijke performance waarde. Deze 10-cross-fold validatie is echter niet voldoende om een uitspraak over de performance van het model te doen, aangezien de parameters van de algoritmen getuned worden tijdens dit process. Hierbij wordt geprobeerd de performance waarde die uit de 10-cross-fold validatie komt te verbeteren. Als het getunede model vervolgens op de test set toegepast wordt kan er pas volledig op basis van nieuwe data een uitspraak over de performance gedaan worden.

3.7.1 Performance maat

Om te bepalen hoe goed een model is moet er een performance maat gekozen worden waarmee de modellen vergeleken kunnen worden. De standaard maat welke Weka (zie hoofdstuk 3.1 *Gebruikte software*) biedt is de *accuracy* oftewel het percentage correct geclassificeerde gevallen. Dit is echter niet een goede maat voor ongebalanceerde data. Stel er is een dataset met 100 targets en 900 non-targets. Het te testen model is extreem simpel: het voorspelt altijd negatief. Dit model zal op deze dataset een accuracy van 90% geven, terwijl het onbruikbaar als voorspeller is aangezien het geen onderscheid kan maken tussen de targetsen non-targets. Een andere veelgebruikte maat welke Weka biedt en wél met ongebalanceerde data om kan gaan is de *AUC* oftewel de oppervlakte onder de ROC curve (**A**rea **U**nder the ROC **C**urve). De ROC is gebasseerd op de *sensitivity* en de *specificity*, oftewel de proportie targets welke correct voorspeld zijn en de proportie non-targets welke correct voorspeld zijn. Het gaat hier om de proportie en daarmee worden de sensitivity en de specificity dus uitgedrukt in waarden ten opzichte van het totaal aantal targetsrespectievelijk non-targets. Hierdoor speelt de scheve verdeling targetsen non-targets geen rol meer. De AUC wordt daarom in dit onderzoek gebruikt als performance maat bij het valideren en beoordelen van de modellen. Een uitgebreide vergelijking tussen de AUC en accuracy wordt beschreven in *Using AUC and Accuracy in Evaluating Learning Algorithms* [7].

Zie ter illustratie figuur 3.3 met vier mogelijke ROC curves. Een goede ROC curve heeft op elk punt een relatief hoge sensitivity en specificity, omdat dit betekent dat zowel de proportie correct voorspelde targets als correct voorspelde non-targets relatief groot is. Doordat op de y-as 1-specificity getoond wordt in plaats van specificity, loopt een goede ROC curve zo dicht mogelijk langs de linkerbovenhoek van de grafiek, net zoals de groene “Goed” curve dit doet. Dit betekent dus dat hoe beter de ROC curve is, hoe groter de oppervlakte onder de curve: de AUC. Dit maakt de AUC een performance maat welke in één getal uit te drukken is, toeneemt naarmate de performance van het model toeneemt en om kan gaan met ongebalanceerde data.



Figuur 3.3: Vier mogelijke ROC curves, waarbij het donkergrijze gebied de AUC aangeeft van de “Matig” ROC curve.

Hoofdstuk 4

Resultaten

Dit hoofdstuk toont de resultaten van het onderzoek naar zowel inzicht in de data als de performance van de verschillende modellen.

4.1 Inzicht in de data

In dit hoofdstuk worden de resultaten van het onderzoek getoond zoals omschreven in de aanpak in hoofdstuk 3.5 *Inzicht in de data*. Er wordt aandacht besteed aan een aantal correlaties tussen verklarende features, een aantal correlaties tussen verklarende features en de targets, en aan het percentage targets binnen elke waarde van de verklarende features.

4.1.1 Correlaties

Onderstaand worden de opvallendste correlaties besproken tussen de verklarende features onderling, tussen verklarende features en de target "terugvordering vanuit Materiële Controle" en tussen verklarende features en de target "bedrag teruggevorderd door Materiële Controle".

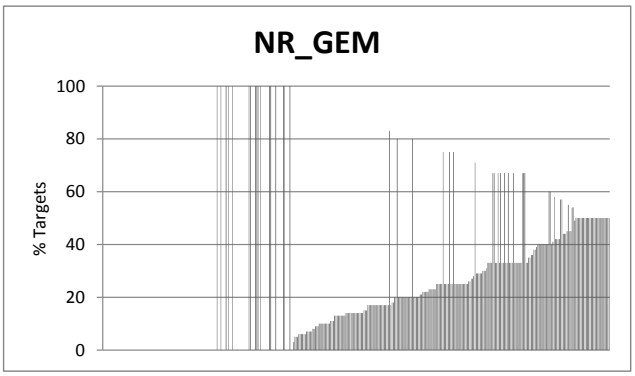
- Er is een perfect, stijgend lineair verband (1,00) tussen het bedrag aan terugboekingen op declaratieniveau (*BEDR_TERUGB_DECLNIV_NUM*) en het bedrag aan vrijwillige terugboekingen door de zorgverlener zelf (*BEDR_FOUT_ZELF_NUM*). Een vergelijkbaar, bijna perfect, stijgend lineair verband (0,96) is ook te zien tussen het percentage van het aantal terugboekingen op declaratieniveau ten opzichte van het totaal aantal declaraties (*AANTAL_TERUGB_DECLNIV_PERC*) en het percentage van het bedrag aan vrijwillige terugboekingen door de zorgverlener zelf ten opzichte van het totaal gedeclareerde bedrag

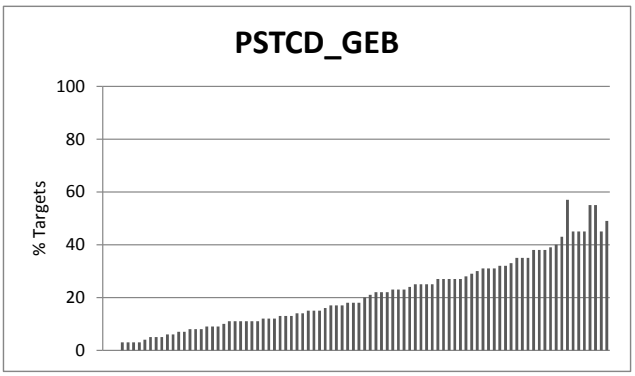
(*BEDR_FOUT_ZELF_PERC*). Gezien het feit dat de terugboekingen op declaratieniveau bestaan uit terugvorderingen door Formele Controle en vrijwillige terugboekingen door zorgverlener zelf, suggereert dit dat de terugboekingen op declaratieniveau bijna alleen maar bestaan uit vrijwillige terugboekingen door de zorgverleners.

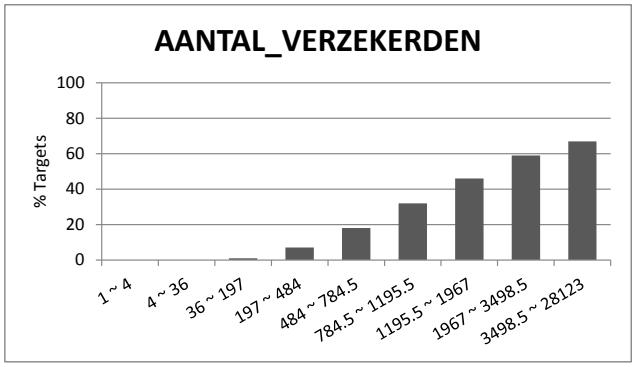
- Er is een perfect, stijgend lineair verband (1,00) tussen het aantal declaraties (*AANTAL_DECL*) en het aantal declaraties binnen de basisverzekering (*AANTAL_DECL_BV*). Gezien het feit dat het aantal declaraties bestaat uit het aantal declaraties binnen de basis- en aanvullendeverzekering, suggereert dit dat het aantal declaraties bijna alleen maar bestaat uit declaraties binnen de basisverzekering.
- Er is een bijna perfect, stijgend lineair verband (0,98) tussen het bedrag teruggevorderd door Formele Controle (*BEDR_FOUT_FC_NUM*) en het percentage van dat bedrag ten opzichte van het totaal gedeclareerde bedrag (*BEDR_FOUT_FC_PERC*).
- Er is een sterk, stijgend lineair verband (0,84) tussen het bedrag gedeclareerd in het voorgaande jaar (*BEDR_GEDECL_VJ*) en het huidige jaar (*BEDR_GEDECL_TOT*). Dit suggereert dat het ongebruikelijk is dat er een groot verschil bestaat tussen het totaal gedeclareerde bedrag van een zorgverlener tussen twee opvolgende jaren.
- Er is een matig, stijgende lineair verband van respectievelijk 0,51; 0,49; 0,46; 0,44 en 0,44 tussen de target "terugvordering vanuit Materiële Controle" en het grootste aantal declaraties per patiënt (*MAX_DECL_VZD*), het aantal declaraties (*AANTAL_DECL*), het aantal verzekerden (*AANTAL_VERZEKERDEN*), het aantal verschillende verwijzers (*AANTAL_VERWIJZERS*) en zorgverlener type 20 (*CODE_TYPE_20*), oftewel "standaard" apothekers (*in tegenstelling tot bijvoorbeeld apotheekhoudende huisartsen*). Deze laatste feature geeft aan of een zorgverlener type 20 is of niet. De feature is dus binair. Dit suggereert, in combinatie met de correlatie van 0,44, dat terugvorderingen vanuit Materiële Controle voornamelijk voorkomen bij "standaard" apothekers. Dit kan zo zijn omdat dit type zorgverlener relatief vaker onrechtmatige declaraties indient dan de andere types, maar dit kan ook komen doordat Materiële Controle ervoor gekozen heeft om in het verleden juist voornamelijk dit type zorgverlener te onderzoeken.
- Er is een bijna perfect, stijgend lineair verband (0,99) tussen de target "bedrag teruggevorderd door Materiële Controle" en het percentage van dit bedrag ten opzichte van het totaal gedeclareerde bedrag (*BEDR_FOUT_MC_PERC*). Daarnaast is er tussen deze target en het aantal declaraties waar door de zorgverlener een dummy code is gebruikt als verwijzer (*AANTAL_VERWIJZERS_DUMMY*) een matig tot sterk, lineair verband (0,64). Een dummy code betekent hier een code die door de zorgverlener ingevuld kan worden als het nummer van de verwijzer van de patiënt om wat voor een rede dan ook niet gegeven kan worden.

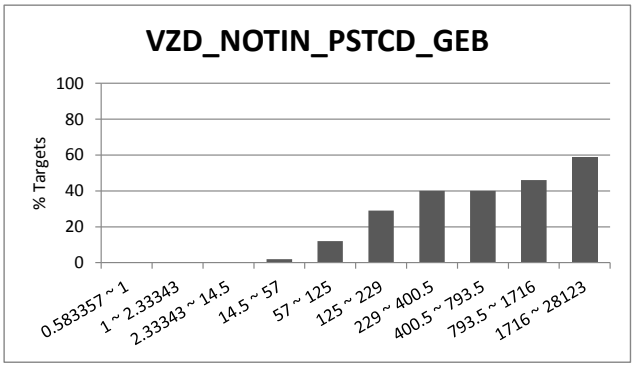
4.1.2 Eigenschappen van targets

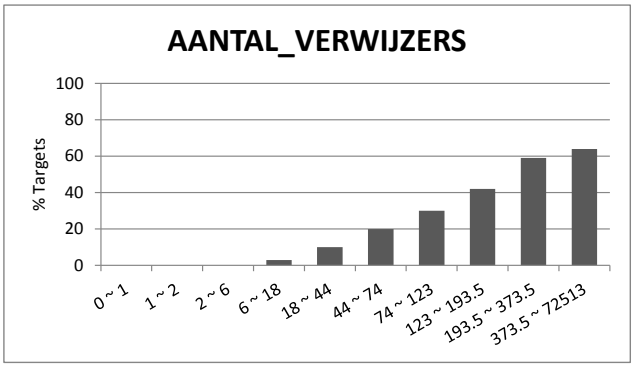
Naast de onderzochte correlaties is er ook per verklarende feature een histogram gemaakt welke voor elke mogelijke waarde het percentage targets visualiseerd, oftewel het percentage zorgverleners mét terugvorderingen. Deze histogrammen geven inzicht in wat de meest voorkomende eigenschappen zijn van zorgverleners met terugvorderingen. Hoe de grenzen van de bins in deze histogrammen bepaald is, is omschreven in hoofdstuk 3.5 *Inzicht in de data*. Alle histogrammen zijn te vinden in bijlage 6.3 *Percentage targets*. Bij de histogrammen van de nominale features *PSTCD_GEB* (postcodegebied) en *NR_GEM* (gemeentenummer) zijn de waarden welke de feature aan kan nemen (x-as) niet getoond, omdat dit er relatief veel zijn en ook niet deels getoond omdat de waarden niet-ordinaal zijn. De belangrijkste waarden van deze features worden wel besproken in dit hoofdstuk. Onderstaand worden de histogrammen getoond van de 15 features met waarden waarbij er meer dan 50% aan zorgverleners met terugvorderingen zijn.

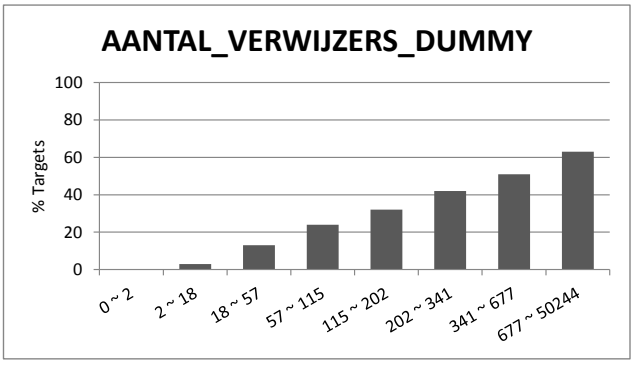


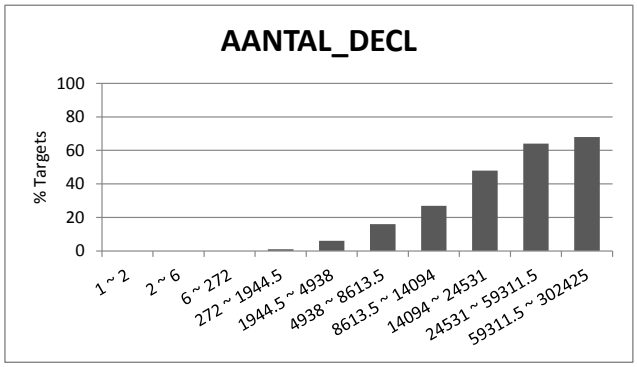


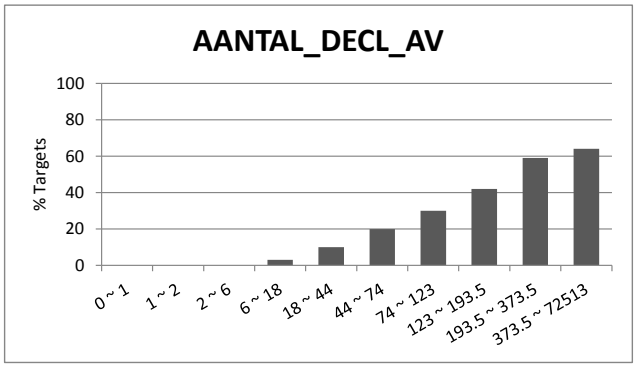


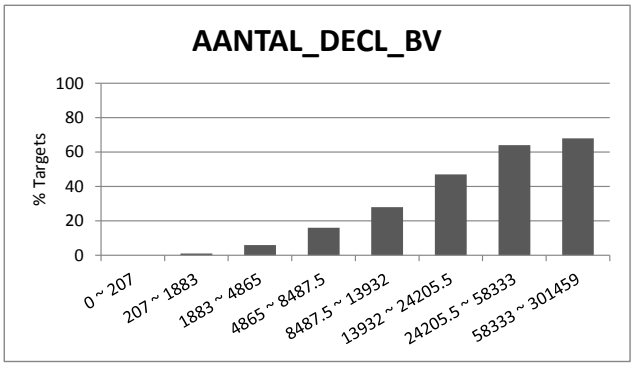


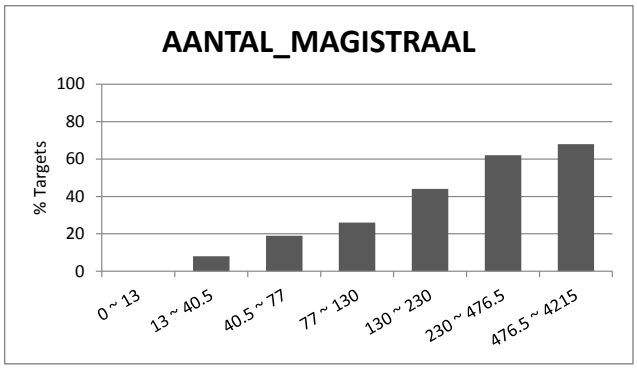


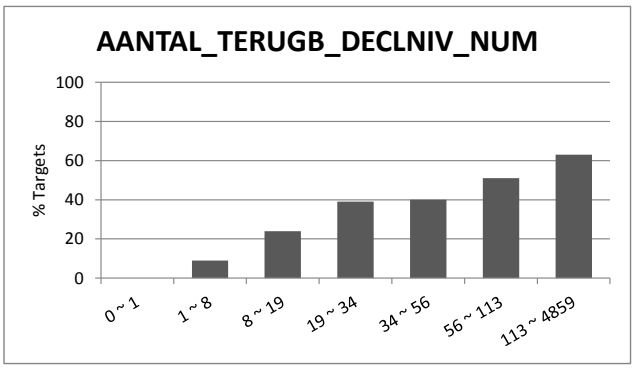


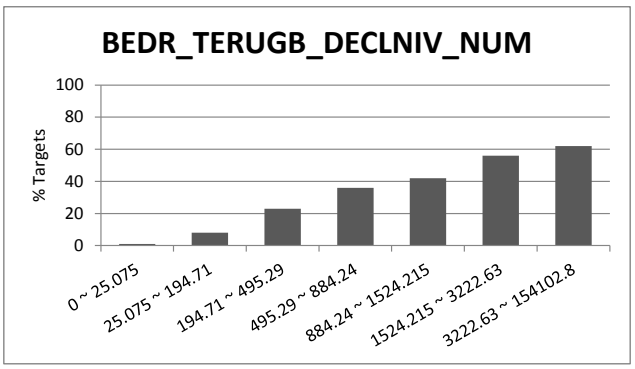

AANTAL_DECL
% Targets
100
80
60
40
20
0
1 ~ 2
2 ~ 6
6 ~ 272
272 ~ 1944.5
1944.5 ~ 4938
4938 ~ 8613.5
8613.5 ~ 14094
14094 ~ 24531
24531 ~ 59311.5
59311.5 ~ 302425

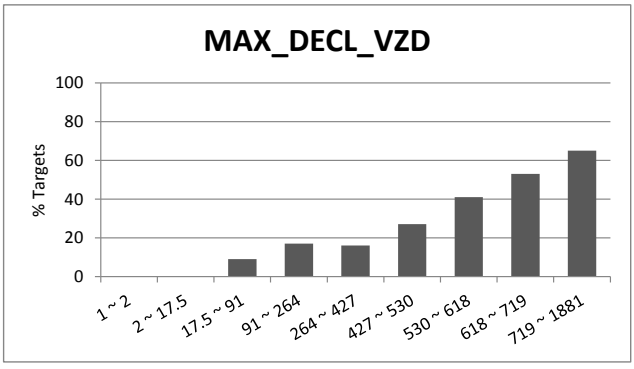

AANTAL_DECL_AV
% Targets
100
80
60
40
20
0
0 ~ 1
1 ~ 2
2 ~ 6
6 ~ 18
18 ~ 44
44 ~ 74
74 ~ 123
123 ~ 193.5
193.5 ~ 373.5
373.5 ~ 72513

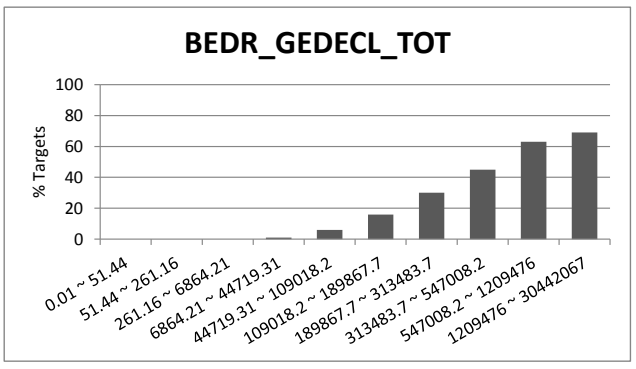

AANTAL_DECL_BV
% Targets
100
80
60
40
20
0
0 ~ 207
207 ~ 1883
1883 ~ 4865
4865 ~ 8487.5
8487.5 ~ 13932
13932 ~ 24205.5
24205.5 ~ 58333
58333 ~ 301459

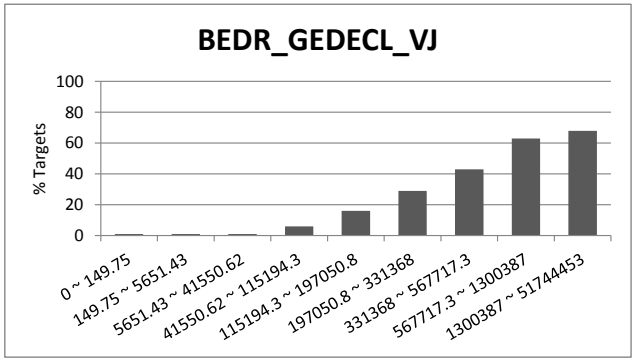

AANTAL_MAGISTRAAL
% Targets
100
80
60
40
20
0
0 ~ 13
13 ~ 40.5
40.5 ~ 77
77 ~ 130
130 ~ 230
230 ~ 476.5
476.5 ~ 4215


AANTAL_TERUGB_DECLNIV_NUM
% Targets
100
80
60
40
20
0
0 ~ 1
1 ~ 8
8 ~ 19
19 ~ 34
34 ~ 56
56 ~ 113
113 ~ 4859


BEDR_TERUGB_DECLNIV_NUM
% Targets
100
80
60
40
20
0
0 ~ 25.075
25.075 ~ 194.71
194.71 ~ 495.29
495.29 ~ 884.24
884.24 ~ 1524.215
1524.215 ~ 3222.63
3222.63 ~ 154102.8


MAX_DECL_VZD
% Targets
100
80
60
40
20
0
1 ~ 2
2 ~ 17.5
17.5 ~ 91
91 ~ 264
264 ~ 427
427 ~ 530
530 ~ 618
618 ~ 719
719 ~ 1881


BEDR_GEDECL_TOT
% Targets
100
80
60
40
20
0
0.01 ~ 51.44
51.44 ~ 261.16
261.16 ~ 6864.21
6864.21 ~ 44719.31
44719.31 ~ 109018.2
109018.2 ~ 189867.7
189867.7 ~ 313483.7
313483.7 ~ 547008.2
547008.2 ~ 1209476
1209476 ~ 30442067



Tabel 4.1 geeft een samenvatting van bovenstaande histogrammen. Het toont de features met de bijbehorende waarden, waarbinnen er meer dan 50% aan zorgverleners met terugvorderingen zijn. De feature *NR_GEM* is niet opgenomen in deze tabel en wordt later op zichzelf staand besproken.

Feature	**Omschrijving**	**Waarden**
PSTCD_GEB	Twee-cijferig postcodegebied	19; 20; 94
AANTAL_VERZEKERDEN	Aantal verzekerden	1967 ∼ 28123
VZD_NOTIN_PSTCD_GEB	Aantal verzekerden buiten postcodegebied van zorgverlener	1716 ∼ 28123
AANTAL_VERWIJZERS	Aantal verschillende verwijzers	193 ∼ 72513
AANTAL_VERWIJZERS_DUMMY	Aantal keer dummycode als verwijzer	677 ∼ 50244
AANTAL_DECL	Aantal declaraties	24531 ∼ 302425
AANTAL_DECL_AV	Aantal declaraties aanvullendeverzekering	193 ∼ 72513
AANTAL_DECL_BV	Aantal declaraties basisverzekering	24205 ∼ 301459
AANTAL_MAGISTRAAL	Aantal magistrale bereidingen	230 ∼ 4215
AANTAL_TERUGB_DECLNIV_NUM	Aantal terugboekingen op declaratieniveau	56 ∼ 4859
BEDR_TERUGB_DECLNIV_NUM	Bedrag terugboekingen op declaratieniveau	1524 ∼ 154103
MAX_DECL_VZD	Maximale aantal declaraties per verzekerde	618 ∼ 1881
BEDR_GEDECL_TOT	Gedeclareerd bedrag	547008 ∼ 30442067
BEDR_GEDECL_VJ	Bedrag gedeclareerd vorig jaar	567717 ∼ 51744453

Tabel 4.1: Features met waarden waarbinnen er meer dan 50% aan zorgverleners met terugvorderingen zijn.

Het is opvallend dat het bij alle continue features uit tabel 4.1 gaat om een aaneengesloten waarden reeks, namelijk de laatste of de laatste én een na laatste bin uit het bijbehrende histogram. Dit duidt erop, gezien features zoals aantal verzekerden, aantal declaraties en gedeclareerd bedrag, dat terugvorderingen vooral plaats vinden bij zorgverleners die voor deze features hoge waarden hebben. Dit kan komen doordat "grote" zorgverleners daadwerkelijk meer en duurdere onrechtmatige declaraties indienen, maar dit kan ook simpelweg komen doordat in het verleden slechts deze groep zorgverleners onderzocht zijn door Materiële Controle.

De nominale feature *NR_GEM* mist in tabel 4.1, aangezien deze feature veel waarden aanneemt en daarnaast op een eigen manier mooi gevisualiseerd kan worden. Het probleem dat ontstaat door dat de feature veel waarden aanneemt, is dat er relatief veel waarden zijn waar slechts een of enkele zorgverleners binnen vallen.

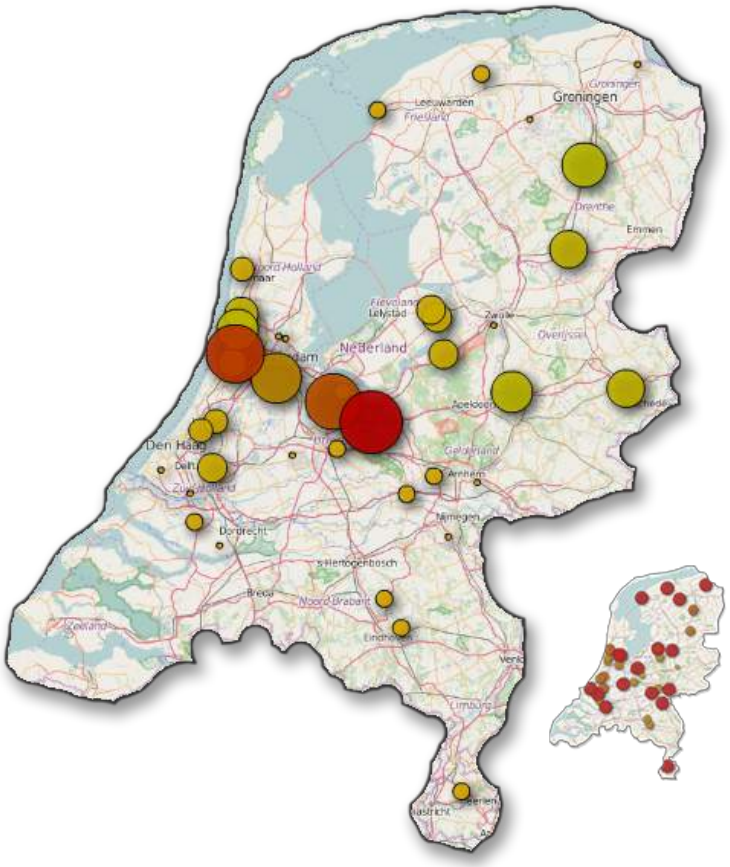
Als in dit geval nog steeds het percentage targets gebruikt zou worden om de feature te visualiseren, dan zouden de gemeenten niet eerlijk met elkaar vergeleken kunnen worden. Een gemeente met bijvoorbeeld een totaal van één zorgverlener, welke terugvorderingen heeft gehad, zou een percentage targets van 100% krijgen. Hier tegenover staat dat een gemeente met een totaal van 100 zorgverleners, waarvan 90 terugvorderingen gehad hebben, een percentage targets van 90% zou krijgen. Het is begrijpelijk dat in dit geval de gemeente met 100 zorgverleners als risicovoller wordt gezien dan de gemeente met slechts één zorgverlener, terwijl het percentage targets dit juist andersom aangeeft. Om dit probleem te verhelpen bestaan er naast het percentage targets meer *quality measures*, waaronder een aantal maten welke wel rekening houden met de absolute grootte van het aantal gevallen. Deze zijn uitgebreid beschreven in *Quality Measures in Data Mining* [6]. Een simpele variant is het vermenigvuldigen van het percentage targets met het totaal aantal zorgverleners. Dit zorgt ervoor dat de ranking van een gemeente strijgt naarmate het totaal aantal zorgverleners groter wordt, maar ook als het percentage targets groter wordt. Zo ook daalt de ranking, naarmate een van beide variabelen daalt. Tabel 4.2 geeft voor alle gemeenten waarin er meer dan 50% zorgverleners met terugvorderingen zijn deze ranking. Om de ranking mooier te maken worden de uitkomsten van de vermenigvuldiging door 100 gedeeld en afgerond op een geheel getal.

Gemeente	**Tot.**	**%T**	**Rank**	**Gemeente**	**Tot.**	**%T**	**Rank**	**Gemeente**	**Tot.**	**%T**	**Rank**
Amersfoort	26	58	15	Lansingerland	5	80	4	Neder-Betuwe	2	100	2
Haarlem	23	57	13	Leiderdorp	4	75	3	Dantumadiel	2	100	2
Hilversum	21	57	12	Voorschoten	4	75	3	Marum	1	100	1
Amstelveen	14	71	10	Bergen (NH.)	5	60	3	Hattem	1	100	1
Assen	10	80	8	Beverwijk	5	60	3	Heumen	1	100	1
Deventer	13	54	7	Renkum	3	67	2	Westervoort	1	100	1
Velsen	13	54	7	Bunnik	3	67	2	Eemnes	1	100	1
Hengelo (O.)	11	55	6	Bloemendaal	3	67	2	Montfoort	1	100	1
Hoogeveen	8	75	6	Oud-Beijerland	3	67	2	Landsmeer	1	100	1
Heemskerk	6	83	5	Nuenen, G. en N.	3	67	2	Oostzaan	1	100	1
Nunspeet	6	67	4	Sint-Oedenrode	3	67	2	Schiedam	1	100	1
Dronten	6	67	4	Harlingen	2	100	2	Strijen	1	100	1
Heemstede	6	67	4	Laren (NH.)	2	100	2	Westland	1	100	1
Overbetuwe	6	67	4	Nuth	2	100	2	Appingedam	1	100	1

Tabel 4.2: Lijst van gemeenten waarin er meer dan 50% zorgverleners met terugvorderingen zijn, waarbij het totaal aantal zorgverleners, het percentage zorgverleners met terugvorderingen (percentage targets, %T) en hun ranking gegeven wordt. Tevens verklaring van figuur 4.1.

Figuur 4.1 geeft vervolgens de visualisatie van de ranking zoals toegekend in tabel 4.2. Om het principe van de ranking duidelijk te maken is ter vergelijking in figuur 4.1 in de rechter onderhoek een kleinere variant gegeven van hoe de visualisatie eruit zou zien als er alleen met het percentage targets rekening gehouden wordt. Het lijkt hierbij alsof er door het hele land risicovolle gemeenten zijn, maar zoals tabel 4.2 verklaart hebben de meeste hiervan slechts een of twee zorgverleners. De grote visualisatie in figuur 4.1 geeft een

realistischer beeld van de geografische verdeling van risicovolle gemeenten.



Figuur 4.1: Visualisatie van de ranking van gemeenten waarin er meer dan 50% zorgverleners met terugvorderingen zijn. De kleine kaart in de rechter onderhoek geeft de visualisatie op basis van het percentage targets.

4.2 Performance van modellen

Dit hoofdstuk toont de resultaten van de modellen zoals gebouwd volgens hoofdstuk 3.6 *Modellen bouwen* en gevalideerd volgens hoofdstuk 3.7 *Modellen valideren*, met als performance maat de AUC, zoals omschreven in hoofdstuk 3.7.1 *Performance maat*.

Tabel 4.3 geeft voor elk gebruikt algoritme, per dataset, de AUC van het gebouwde model met standaard parameters. Zoals in hoofdstuk 3.7 *Modellen valideren* besproken is, is dit de AUC op basis van 10-cross-fold validatie. Ter herinnering: er zijn zes datasets. Ten eerste dataset 1, met als target zorgverleners met een bedrag aan terugvorderingen vanaf €500 tot €1500. Ten tweede dataset 2, met als target zorgverleners met een bedrag aan terugvorderingen vanaf €1500. Ten derde dataset 3, met als target alle zorgverleners met terugvorderingen. Elke dataset wordt in tabel 4.3 aangegeven als *Dx*, met als *x* het nummer van de dataset

zoals hierboven beschreven. Voor elke dataset is met behulp van de Classbalancer in Weka (zie hoofdstuk 3.6.2 *Gewogen datasets*) een gewogen dataset gemaakt. Deze wordt aangegeven als *DxG* met als *x* het nummer van de dataset zoals hierboven beschreven. De meest interresante AUC's worden in het zwart aangegeven, ten opzichte van de minder interresante grijze AUC's.

Type	**Naam (Weka)**	**D1**	**D1G**	**D2**	**D2G**	**D3**	**D3G**
Naive Bayes	NaiveBayes	0,83	0,83	0,91	0,91	0,90	0,90
Logistic Regression	SimpleLogistic	0,54	0,79	0,91	0,91	0,90	0,90
Support Vector Machine	LibSVM	0,50	0,50	0,54	0,54	0,75	0,75
K-nearest neighbours	IBk	0,57	0,57	0,68	0,68	0,71	0,71
Decision Table	DecisionTable	0,50	0,79	0,89	0,89	0,88	0,89
Decision Tree	J48 (C4.5)	0,50	0,56	0,54	0,85	0,88	0,87
Decision Forest	RandomForest	0,81	0,81	0,91	0,91	0,90	0,90

Tabel 4.3: De AUC's van de verschillende modellen, waarbij de standaard parameters van de algoritmen zijn gebruikt tijdens het bouwen van de modellen.

Tabel 4.3 laat ten eerste zien dat een gewogen dataset slechts een significante verbetering oplevert bij dataset 1 in combinatie met logistic regression, bij dataset 1 in combinatie met een decision table en bij dataset 2 in combinatie met een decision tree. Ook is te zien dat er voor elke dataset *Dx* een algoritme is, waarmee de AUC minimaal even groot is als de grootste AUC van de gewogen variant *DxG*. Om de modellen zo simpel mogelijk te houden, volgt hieruit het besluit dat de gewogen datasets niet verder gebruikt zullen worden. Daarnaast maken de support vector machine en k-nearest neighbours relatief slechte modellen en deze worden vanaf nu ook buiten beschouwing gelaten. Tabel 4.3 veranderd hiermee in tabel 4.4.

Type	**Naam (Weka)**	**D1**	**D2**	**D3**
Naive Bayes	NaiveBayes	0,83	0,91	0,90
Logistic Regression	SimpleLogistic	0,54	0,91	0,90
Decision Table	DecisionTable	0,50	0,89	0,88
Decision Tree	J48 (C4.5)	0,50	0,54	0,88
Decision Forest	RandomForest	0,81	0,91	0,90

Tabel 4.4: De AUC's van de beste modellen op basis van de ongewogen datasets, waarbij de standaard parameters van de algoritmen zijn gebruikt tijdens het bouwen van de modellen.

Aan de hand van tabel 4.4 is te zien dat naive bayes en een random forest bij alle drie de datasets behoren tot de beste algoritmen. De andere drie algoritmen werken slechts significant goed op dataset 2 en 3, of zelfs alleen op dataset 3. Sterker nog, de AUC's van deze algoritmen zijn voor elke dataset gelijk aan of kleiner dan die van naive bayes en een random forest. Dit komt omdat in de het eerste model de targets

waarschijnlijk lastiger te onderscheiden zijn van de non-targets, waardoor er minder algoritmen een goed resultaat opleveren. Naarmate het bedrag stijgt, wordt deze scheiding blijkbaar duidelijker en door meer algoritmen vindbaar. Om nogmaals de modellen zo simpel mogelijk te houden en om de best presterende modellen te behouden, volgt hieruit het besluit dat er een enkel algortime gebruikt zal worden voor alle modellen. Hiermee vallen logistische regressie, decision table en decision tree af. Om het maximale uit de twee overgebleven algoritmen te halen worden de parameters getuned, zoals omschreven in hoofdstuk 3.6.4 *Parameter tuning*. Hierbij worden de parameters van de modellen zo aangepast dat de AUC van de modellen toeneemt. Bij naive bayes blijkt er geen verbetering mogelijk te zijn. Voor een random forest blijkt voor elke dataset te gelden dat in de situatie welke de grootste AUC opleverd, de parameter "numFeatures" gelijk is aan *1*. Deze parameter bepaaldt het aantal willekeurig gekozen features waarop gesplitst wordt binnen elke boom in het random forest. De parameter "maxDepth" blijkt ook van belang te zijn. Deze parameter bepaaldt de maximale diepte per boom in het random forest. In de situatie welke de grootste AUC opleverd, geldt voor respectievelijk dataset 1, 2 en 3 dat deze parameter gelijk is aan *4*, *3* en *8*. De parameter tuning uit zich in de volgende AUC's:

Type	**Naam (Weka)**	**D1**	**D2**	**D3**
Naive Bayes	NaiveBayes	0,83	0,91	0,90
Decision Forest	RandomForest	0,83	0,92	0,91

Tabel 4.5: De AUC's van de beste getunede modellen, op basis van de ongewogen datasets.

Volgens tabel 4.5 lijkt een random forest het beste te werken, maar om een definitieve uitspraak te doen worden de modellen getest op de test set, zoals omschreven in hoofdstuk 3.7 *Modellen valideren*. De test set is niet gebruikt bij het trainen van de modellen, noch bij het tunen van de parameters, en daarom geschikt als laatste test. Tabel 4.6 geeft hiervan de resultaten.

Type	**Naam (Weka)**	**D1**	**D2**	**D3**
Naive Bayes	NaiveBayes	0,77	0,89	0,86
Decision Forest	RandomForest	0,82	0,91	0,90

Tabel 4.6: De AUC's van de beste getunede modellen, getest op de test set.

Zoals te zien in tabel 4.6 werkt een random forest inderdaad het beste. Opvallend is dat de AUC's van de random forest modellen bij de test set, in absolute zin, slechts een honderdste omlaag gaan ten opzichte van de 10-cross-fold validatie, terwijl dit bij de naive bayes modellen tot meer dan een zeventiente omlaag gaat. Dit suggereert dat in dit geval de random forest modellen minder *overfit* zijn dan de naive bayes modellen. Met *overfit* wordt bedoeld dat een model zo specifiek op zijn training set is afgesteld, dat het moeite heeft met correcte voorspellingen maken van nieuwe data.

Aan de hand van het model op dataset 3, waarbij de target alle zorgverleners met terugvorderingen zijn, kan de performance verandering onderzocht worden van de toevoeging van de tweede target "teruggevorderd bedrag", oftewel de splitsing in dataset 1 en 2. Om dit te bereiken, wordt de performance van de beste modellen op basis van dataset 1 en 2 uitgedrukt in de gemiddelde AUC hiervan, zoals gegeven in tabel 4.6

$$\frac{0,82 + 0,91}{2} = 0,865$$

Dit gemiddelde wordt vervolgens vergeleken met de AUC van het beste model op basis van dataset 3. Aan de hand hiervan is te berekenen dat de splitsing in dataset 1 en 2, oftewel de implementatie van de toevoeging van de tweede target "teruggevorderd bedrag", een verandering in performance op basis van de AUC tewege brengt van

$$\frac{0,900 - 0,865}{0,900} * 100 \approx -4\%$$

4.3 Gebruik van de modellen

In dit hoofdstuk wordt de *Deployment* omschreven zoals beschreven in het CRISP-DM model. Het legt uit hoe de modellen toegepast kunnen worden in de praktijk. In dit hoofdstuk wordt er met "modellen" gerefereerd naar de twee modellen uit tabel 5.1. Hierbij geldt weer dat de target voor model 1 zorgverleners met een bedrag aan terugvorderingen vanaf €500 tot €1500 is, en voor model 2 zorgverleners met een bedrag aan terugvorderingen vanaf €1500.

Doordat de modellen beide alle zorgverleners bevatten, maar beide een andere groep als target hebben, zal voor elke zorgverlener de sommatie van de verwachtte kans op terugvordering uit beide modellen, plus de kans op een terugvordering met een bedrag van minder dan €500, plus de kans op geen terugvordering ongeveer op 1 uitkomen. Ter herinnering tabel 4.7

Tabel 4.7: Verdeling van zorgverleners per dataset. Per dataset geeft een groen vakje de **target** aan en een rood vakje de **non-target**, waarbij elk vakje een groep zorgverleners representeert met een bedrag aan terugvorderingen binnen bepaalde grenzen.

Om tot een ranking te komen wordt in acht genomen dat zorgverleners met terugvorderingsbedragen onder de €500 en zorgverleners zonder terugvorderingen niet interresant zijn. De kans dat een zorgverlener in

een van deze groepen zit wordt dus niet expliciet meegenomen in de berekening van de ranking. Impliciet gebeurdt dit natuurlijk wel, want de kans dat een zorgverlener terugvorderingen vanaf €500 heeft wordt wél als interresant geacht en elke kans uit dit stelsel is afhankelijk van elkander. De ranking is dus gebasseerd op de expliciete kans van model 1 en model 2. Om de ranking te berekenen worden deze kansen vermendigvuldigd met waarden welke representatief zijn voor het model waar ze iets over zeggen. Vervolgens worden deze vermenigvuldigingen bij elkaar opgeteld. Een intuïtieve waarde is het gemiddelde terugvorderingsbedrag van alle zorgverleners binnen de dataset waar het model op gebasseerd is. Het probleem hiermee is dat deze gevoellig is voor uitschieters in de data. De mediaan biedt hier een oplossing voor. Voor dataset 1 is de mediaan €861,67. Rond dit af naar 860. Voor dataset 2 is de mediaan 3283,67. Rond dit af naar 3280. Stel model 1 schat een kans op terugvordering van 30% en model 2 schat een kans op terugvordering van 60%. Dan is de kans op terugvordering onder de €500 plus de kans op geen terugvordering dus 10%. De ranking van deze zorgverlener wordt dan

$$0,3 * 860 + 0,6 * 3280 = 2226$$

Kortom, deze ranking houdt rekening met de kans op terugvordering en de hoogte van het verwachtte terug te vorderen bedrag, waarbij bedragen onder de €500 genegeerd worden.

Hoofdstuk 5

Conclusie

Dit hoofdstuk bespreekt de conclusies welke getrokken kunnen worden uit het gehele onderzoek en komt hiermee terug op de onderzoeksvraag: hoe (goed) kan een voorspellingsmodel, met features gebasseerd op domeinkennis, een ranglijst maken van zorgverleners, waarbij de kans op terugvordering van declaraties en het verwachtte bedrag daarvan van belang zijn?

5.1 Modellen

Uit tabel 4.6 wordt duidelijk dat een random forest de beste modellen genereert. Tabel 5.1 toont de random forest modellen op basis van de twee datasets welke rekening houden met het terugvorderingsbedrag. Hierbij worden hun de getunede parameters gegeven, met bijbehorende waarden. Daarnaast is ook de performance van de modellen gegeven. Hierbij geldt weer dat de target voor dataset 1 zorgverleners met een bedrag aan terugvorderingen vanaf €500 tot €1500 is, en voor dataset 2 zorgverleners met een bedrag aan terugvorderingen vanaf €1500.

Dataset	**Model**	**Parameters**	**Param. waarden**	**AUC**
1	Random Forest	numFeatures; maxDepth	1;4	0,82
2	Random Forest	numFeatures; maxDepth	1;3	0,91

Tabel 5.1: Algoritme, getunede parameters en performance van de beste modellen op basis van de twee datasets welke rekening houden met het terugvorderingsbedrag.

Aan de hand van deze twee modellen kan een ranglijst op basis van kans en bedrag opgesteld worden, zoals gevraagd in de onderzoeksvraag. De ranglijst kan opgesteld worden door elke nieuwe zorgverlener te laten

classificeren door beide modellen en een ranking aan elke zorgverlener toe te kennen volgens onderstaande formule:

$$kans_1 * 860 + kans_2 * 3280$$

Er is een daling van 4% in de performance van de modellen die rekening houden met het terugvorderingsbedrag, ten opzichte van het model waarin alle zorgverleners met terugvorderingen de target zijn. Dit is acceptabel, omdat de splitsing in twee modellen er voor zorgt dat er naast de kans op terugvordering, ook een inschatting van het terug te vorderen bedrag gemaakt kan worden.

5.2 Aanbevelingen voor vervolgonderzoek

Tijdens de uitvoer van het onderzoek zijn er een aantal zaken opgevallen welke door de onderzoeker verbeterd kunnen worden om vergelijkbaar- en vervolgonderzoek te verbeteren:

- Voeg aan de dataset daadwerkelijk features uit het AGB-register toe die niet opgenomen zijn in de MIAZ Dimensie Zorgrelatie, zoals het aantal medewerkers in de apotheek van een zorgverlener.
- Onderzoek en voeg eventueel toe aan de dataset alle bedachte features aan de hand van domein kennis, in plaats van een selectie hiervan.
- Bouw de voorspellingsmodellen in SAS om zo alles binnen het standaard pakket van Zilveren Kruis uit te kunnen voeren.

Daarnaast zijn er een aantal zaken opgevallen welke door Zilveren Kruis verbeterd kunnen worden:

- Houd vanuit Materiële Controle bij welke zorgverleners in het verleden gecontroleerd zijn op onrechtmatige declaraties, in plaats van alleen de resultaten van de terugvorderingen. Dit helpt om onzuiverheden uit de modellen te halen. Deze onzuiverheden ontstaan doordat de terugvorderingsresultaten alleen data bevatten over zorgverleners welke in het verleden door Materiële Controle onderzocht zijn. De modellen hebben dus geen kennis van alle andere zorgverleners.
- Materiële Controle maakt veel terugvorderingen op bulkniveau waarbij alleen het bulk bedrag berekend wordt. Als naast het bedrag ook het aantal declaraties berekend zou worden, zou dit, zoals Materiële Controle zelf aangeeft te willen, als target meegenomen kunnen worden in de voorspellingsmodellen.
- Betere verslaglegging van Formele Controle: veel ROC dataregels missen een identificatie van zorgverlener en daarnaast is het teruggevorderde bedrag volgens de ROC database velen malen lager dan wat Formele Controle zelf rapporteert. Mocht de informatievoorziening in ROC verbeterd worden dan kan het mogelijk een belangrijke rol gaan spelen in de voorspellingsmodellen.

Hoofdstuk 6

Bijlagen

6.1 Features uit gesprekken met stakeholders

- Gedeclareerd bedrag
- Verschil in gedeclareerde bedrag en daadwerkelijke vergoeding
- Aantal declaraties
- Aantal verzekerden
- Splitsing in aanvullende- en basisverzekering van declaraties
- Verandering in gedeclareerd bedrag ten opzichte van vorig jaar
- Aantal passanten
- Aantal werknemers
- Aantal jaar zorgverlener reeds actief
- Type zorgverlener
- Type apotheek van zorgverlener
- Locatie van apotheek van zorgverlener
- Landgrens waaraan apotheek van zorgverlener grenst
- Aantal inwoners van de gemeente waarin zorgverlener actief is
- Zorgverlener doet mee aan ketenzorg ja/nee
- Zorgverlener heeft meerdere apotheken ja/nee
- Apotheek van zorgverlener valt onder koepel organisatie ja/nee
- Zorgverlener verricht magistrale bereidingen ja/nee
- Aantal magistrale bereidingen
- Aantal verschillende verwijzers

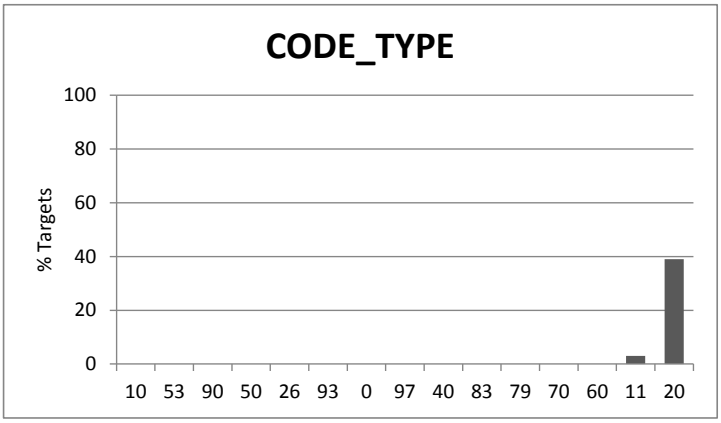
- Aantal keer dummy code gebruik als verwijzer
- Gemiddelde leeftijd verzekerden
- Aantal verzekerden buiten postcode gebied van zorgverlener
- Gemiddelde tijd dat verzekerde bij zorgverlener zit
- Afwijking van gemiddelde frequentie van uitgave per geneesmiddel
- Eerdere terugvorderingen bij zorgverlener
- Meest voorkomende reden van eerdere terugboeking
- Aantal declaraties op initiatief van zorgverlener teruggestort
- Bedrag op initiatief van zorgverlener teruggestort

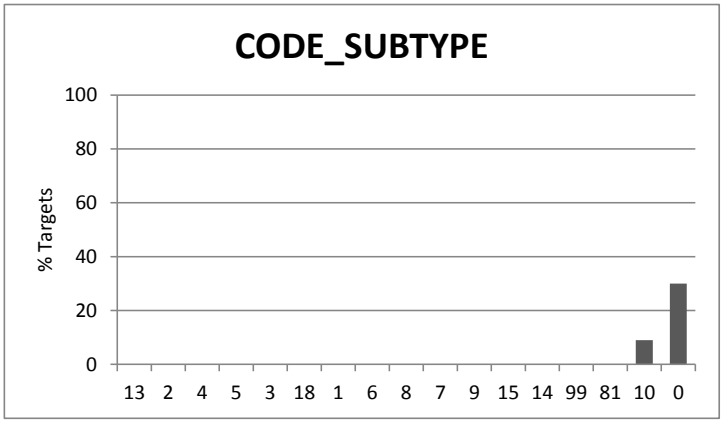
6.2 Geïmplementeerde features

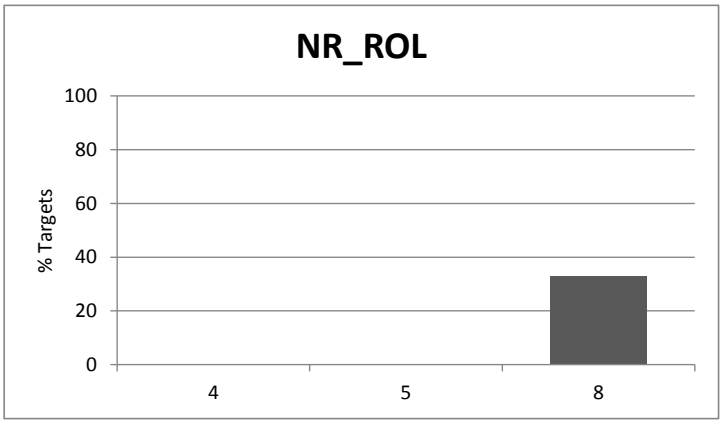
Feature	Type	Omschrijving
BEDR_GEDECL_TOT	Continu	Gedeclareerd bedrag huidig jaar
BEDR_GEDECL_VJ	Continu	Gedeclareerd bedrag vorig jaar
TOENAME_KOSTEN_PERC	Continu	Verandering in gedeclareerd bedrag
AANTAL_DECL	Continu	Aantal declaraties
AANTAL_DECL_BV	Continu	Aantal declaraties binnen de basisverzekering
AANTAL_DECL_AV	Continu	Aantal declaraties binnen de aanvullendeverzekering
AANTAL_VERZEKERDEN	Continu	Aantal verzekerden
NIEUWKOMER	Binair	Zorgverlener heeft sinds huidig jaar een contract bij Zilveren Kruis
MAX_DECL_VZD	Continu	Maximaal aantal declaraties per verzekerde
GEM_DECL_VZD	Continu	Gemiddeld aantal declaraties per verzekerde
GEM_LEEFTIJD_VZD	Continu	Gemiddelde leeftijd verzekerde
NR_ROL	Nominaal	Type zorgverlener
CODE_TYPE	Nominaal	Type apotheek van zorgverlener
CODE_SUBTYPE	Nominaal	Subtype apotheek van zorgverlener
PROV_NAAM	Nominaal	Locatie van grootste apotheek van zorgverlener: provincie
PSTCD_GEB	Nominaal	Locatie van grootste apotheek van zorgverlener: postcodegebied *(12 in 1234AB)*
VZD_NOTIN_PSTCD_GEB	Continu	Aantal verzekerden buiten postcode gebied van zorgverlener
GRENS	Nominaal	Landgrens waaraan apotheek van zorgverlener grenst
NR_GEM	Nominaal	Locatie van grootste apotheek van zorgverlener: gemeente
INWONERS	Continu	Aantal inwoners van de gemeente waarin zorgverlener actief is
MAGISTRAAL	Binair	Zorgverlener verricht magistrale bereidingen ja/nee
AANTAL_MAGISTRAAL	Continu	Aantal magistrale bereidingen
AANTAL_VERWIJZERS	Continu	Aantal verschillende verwijzers
VERWIJZERS_DUMMY	Binair	Zorgverlener gebruikt dummycode(s) als verwijzer
AANTAL_VERWIJZERS_DUMMY	Continu	Aantal keer dummy code gebruik als verwijzer
TERUGVRD_FC	Binair	Terugvordering vanuit Formele Controle in huidig jaar
TERUGVRD_FR	Binair	Terugvordering wegens bewezen fraude zaak in huidig jaar
BEDR_FOUT_FC_NUM	Continu	Teruggevorderd bedrag vanuit Formele Controle in huidig jaar
BEDR_FOUT_FR_NUM	Continu	Teruggevorderd bedrag wegens bewezen fraude zaak in huidig jaar
BEDR_FOUT_FC_PERC	Continu	Percentage teruggevorderd bedrag vanuit Formele Controle in huidig jaar t.o.v. gedeclareerd bedrag
BEDR_FOUT_FR_PERC	Continu	Percentage teruggevorderd bedrag wegens bewezen fraude zaak in huidig jaar t.o.v. gedeclareerd bedrag
AANTAL_TERUGB_DECLNIV_NUM	Continu	Aantal terugboekingen op declaratieniveau
AANTAL_TERUGB_DECLNIV_PERC	Continu	Percentage terugboekingen op declaratieniveau t.o.v. gedeclareerd bedrag
BEDR_TERUGB_DECLNIV_NUM	Continu	Bedrag teruggeboekt op declaratieniveau
BEDR_TERUGB_DECLNIV_PERC	Continu	Percentage teruggeboekt bedrag op declaratieniveau t.o.v. gedeclareerd bedrag
BEDR_FOUT_ZELF_NUM	Continu	Bedrag op initiatief van zorgverlener teruggestort
BEDR_FOUT_ZELF_PERC	Continu	Percentage teruggestort bedrag op initiatief van zorgverlener t.o.v. gedeclareerd bedrag

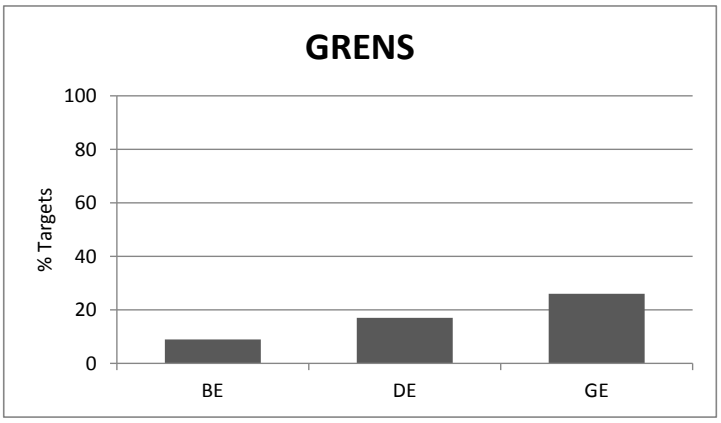
6.3 Percentage targets

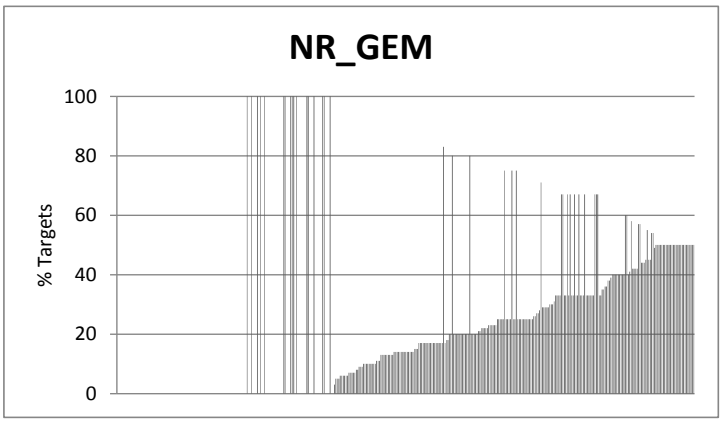
De betekenis van de afkortingen van features, zoals *CODE_TYPE*, zijn te vinden in bijlage 6.2 *Geïmplementeerde features*

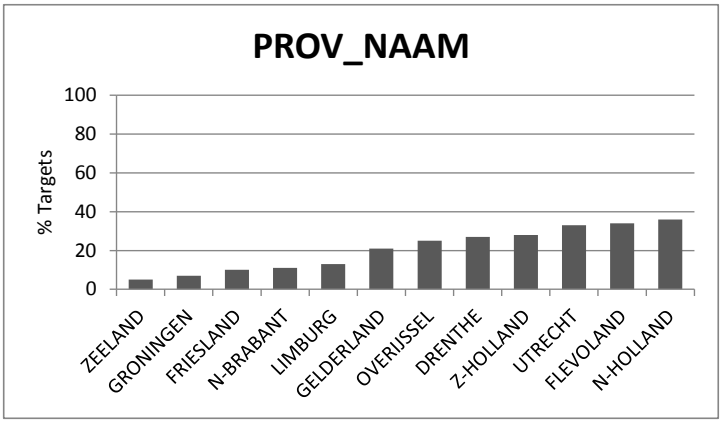


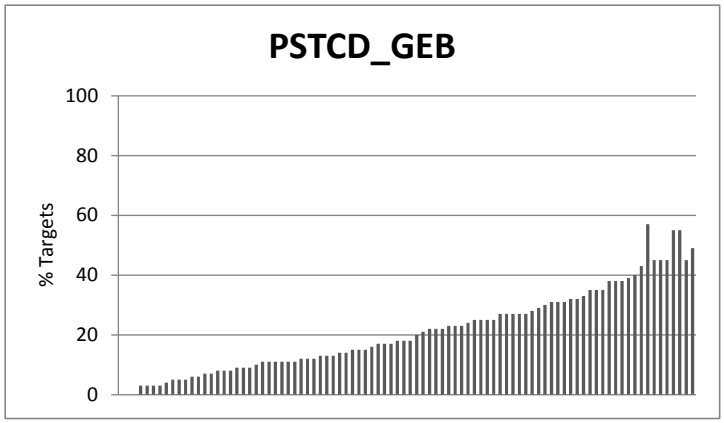


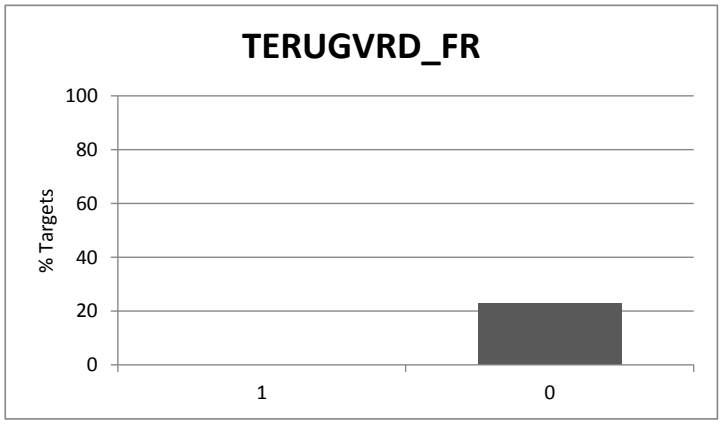


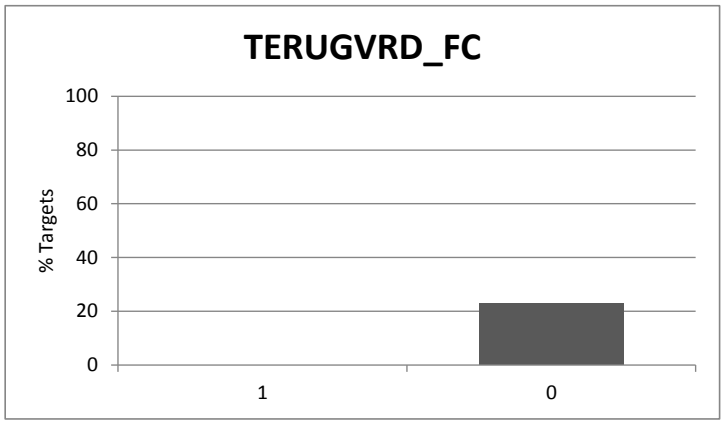


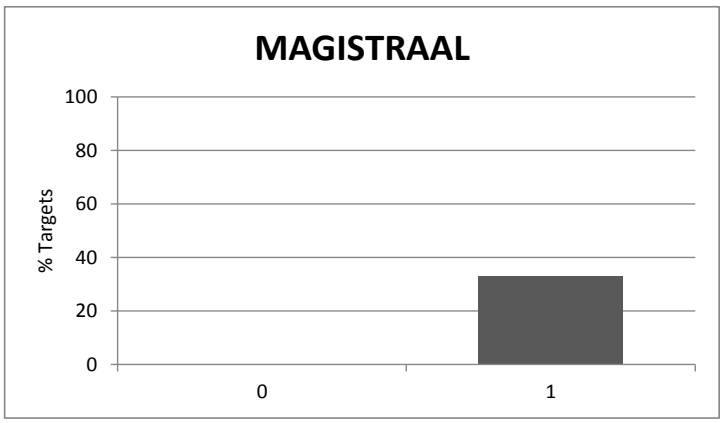


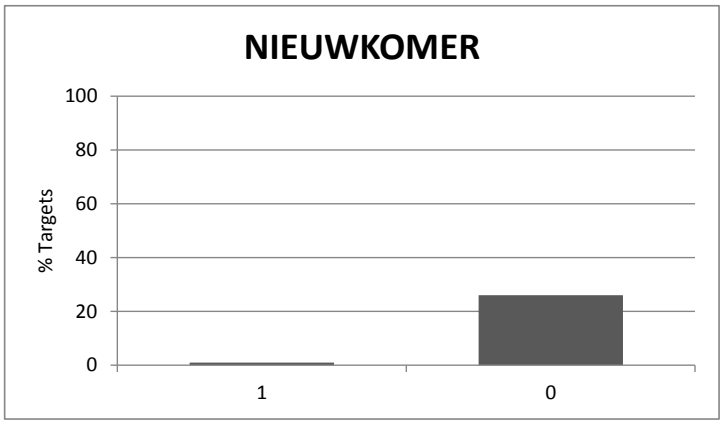


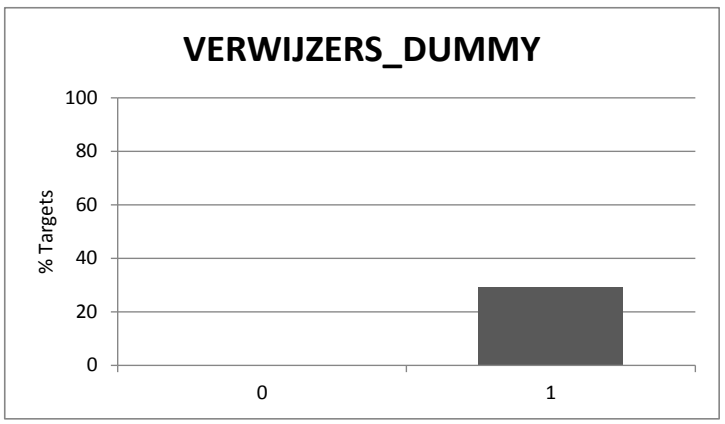

PSTCD_GEB
% Targets
100
80
60
40
20
0

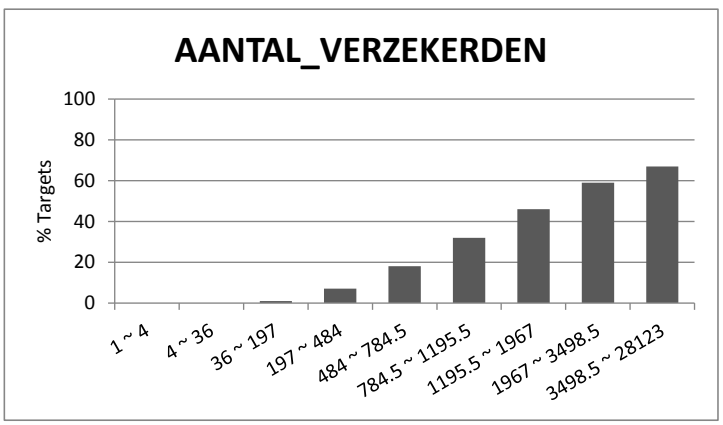

TERUGVRD_FR
% Targets
100
80
60
40
20
0
1
0

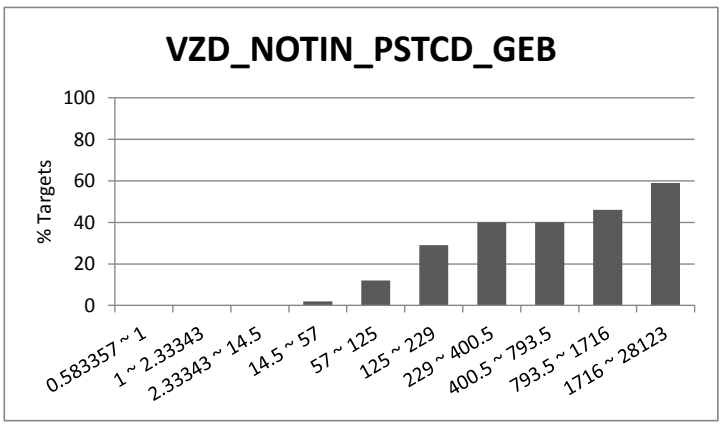

TERUGVRD_FC
% Targets
100
80
60
40
20
0
1
0

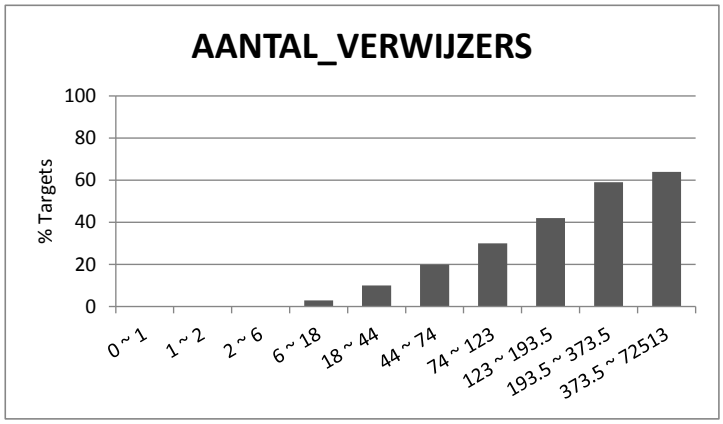

MAGISTRAAL
% Targets
100
80
60
40
20
0
0
1

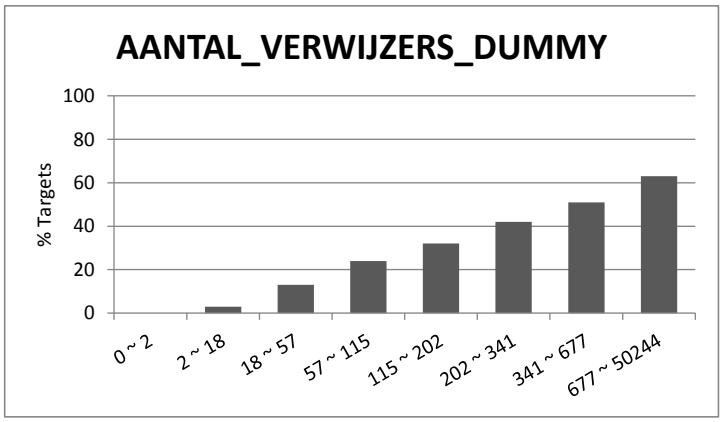

NIEUWKOMER
% Targets
100
80
60
40
20
0
1
0

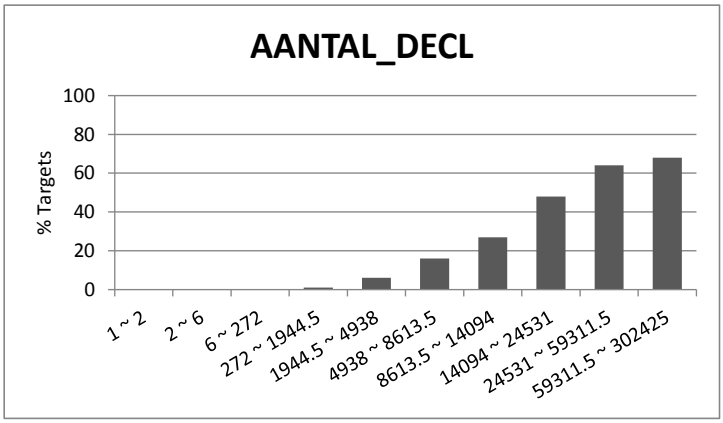

VERWIJZERS_DUMMY
% Targets
100
80
60
40
20
0
0
1

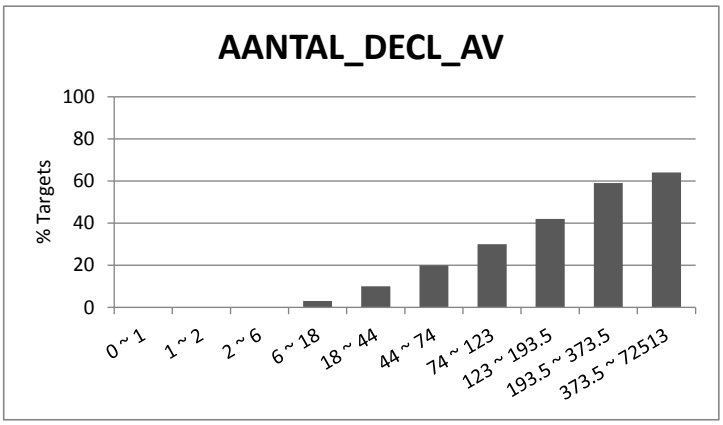

AANTAL_VERZEKERDEN
% Targets
100
80
60
40
20
0
1 ~ 4
4 ~ 36
36 ~ 197
197 ~ 484
484 ~ 784.5
784.5 ~ 1195.5
1195.5 ~ 1967
1967 ~ 3498.5
3498.5 ~ 28123

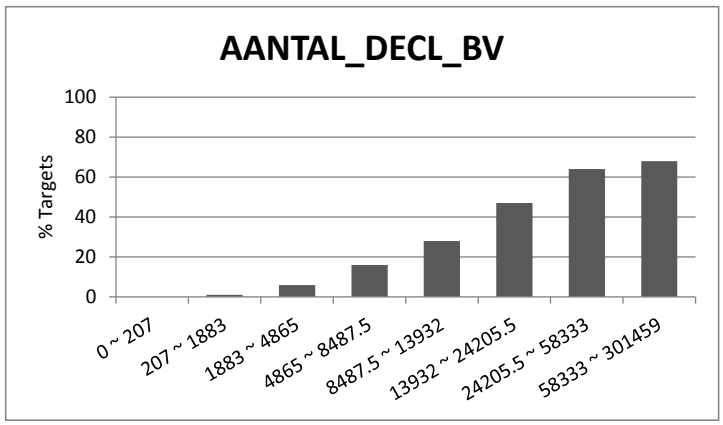

VZD_NOTIN_PSTCD_GEB
% Targets
100
80
60
40
20
0
0.583357 ~ 1
1 ~ 2.33343
2.33343 ~ 14.5
14.5 ~ 57
57 ~ 125
125 ~ 229
229 ~ 400.5
400.5 ~ 793.5
793.5 ~ 1716
1716 ~ 28123

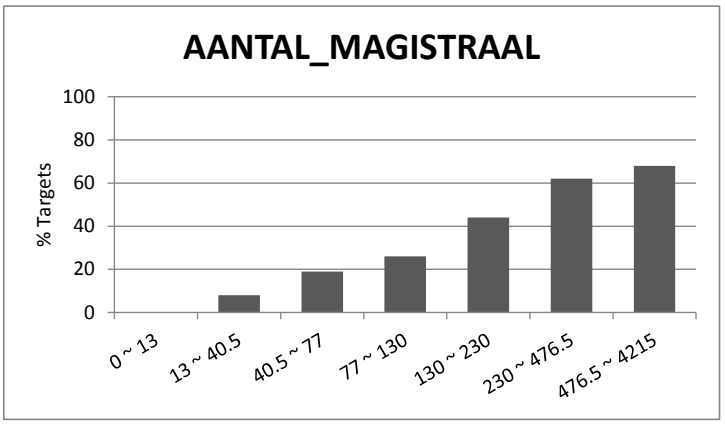

AANTAL_VERWIJZERS
% Targets
100
80
60
40
20
0
0 ~ 1
1 ~ 2
2 ~ 6
6 ~ 18
18 ~ 44
44 ~ 74
74 ~ 123
123 ~ 193.5
193.5 ~ 373.5
373.5 ~ 72513

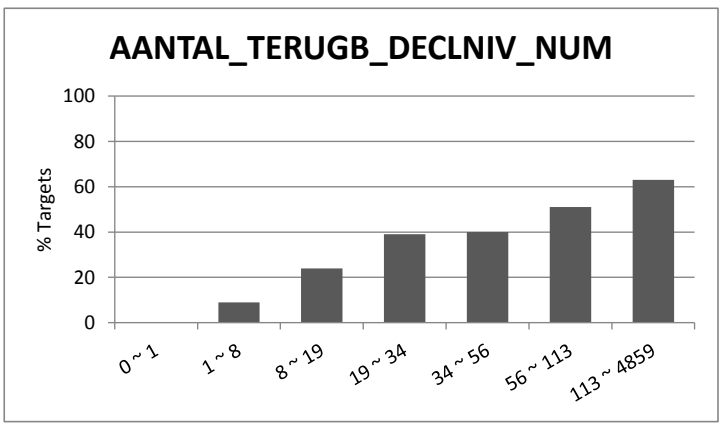

AANTAL_VERWIJZERS_DUMMY
% Targets
100
80
60
40
20
0
0 ~ 2
2 ~ 18
18 ~ 57
57 ~ 115
115 ~ 202
202 ~ 341
341 ~ 677
677 ~ 50244

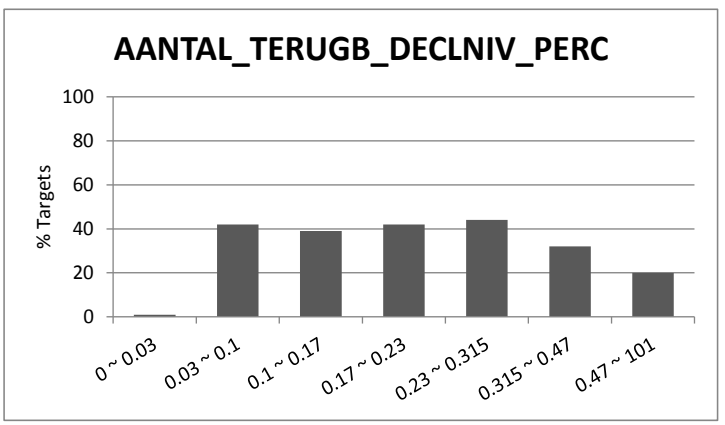

AANTAL_DECL
% Targets
100
80
60
40
20
0
1 ~ 2
2 ~ 6
6 ~ 272
272 ~ 1944.5
1944.5 ~ 4938
4938 ~ 8613.5
8613.5 ~ 14094
14094 ~ 24531
24531 ~ 59311.5
59311.5 ~ 302425

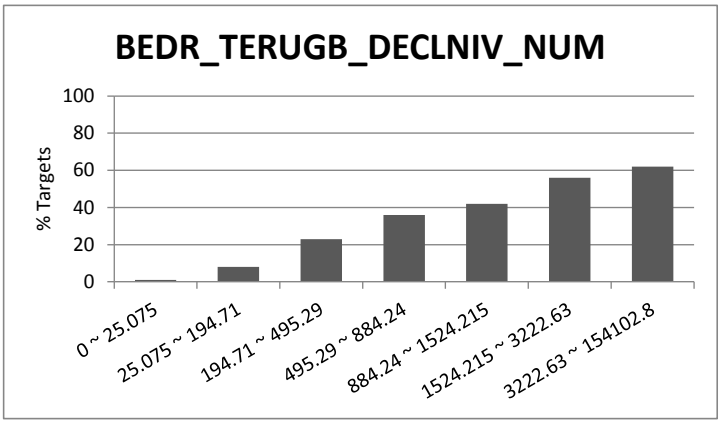

AANTAL_DECL_AV
% Targets
100
80
60
40
20
0
0 ~ 1
1 ~ 2
2 ~ 6
6 ~ 18
18 ~ 44
44 ~ 74
74 ~ 123
123 ~ 193.5
193.5 ~ 373.5
373.5 ~ 72513

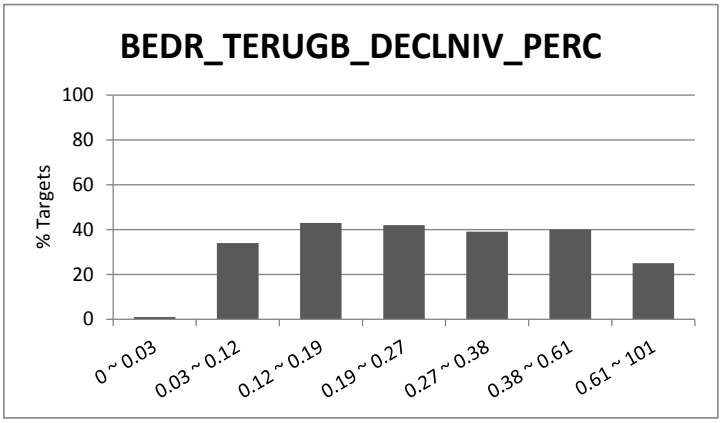

AANTAL_DECL_BV
% Targets
100
80
60
40
20
0
0 ~ 207
207 ~ 1883
1883 ~ 4865
4865 ~ 8487.5
8487.5 ~ 13932
13932 ~ 24205.5
24205.5 ~ 58333
58333 ~ 301459

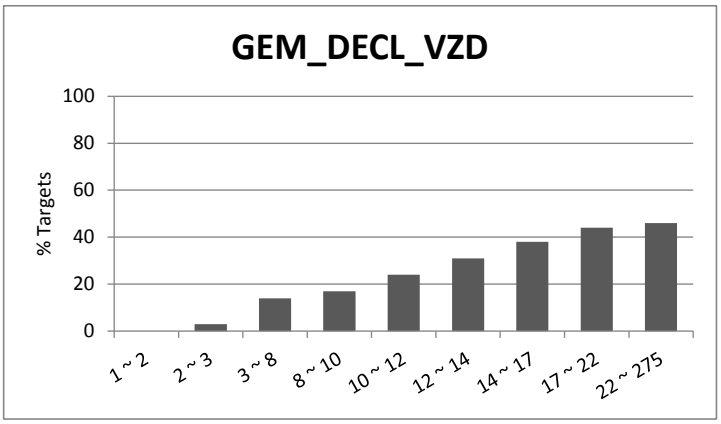

AANTAL_MAGISTRAAL
% Targets
100
80
60
40
20
0
0 ~ 13
13 ~ 40.5
40.5 ~ 77
77 ~ 130
130 ~ 230
230 ~ 476.5
476.5 ~ 4215

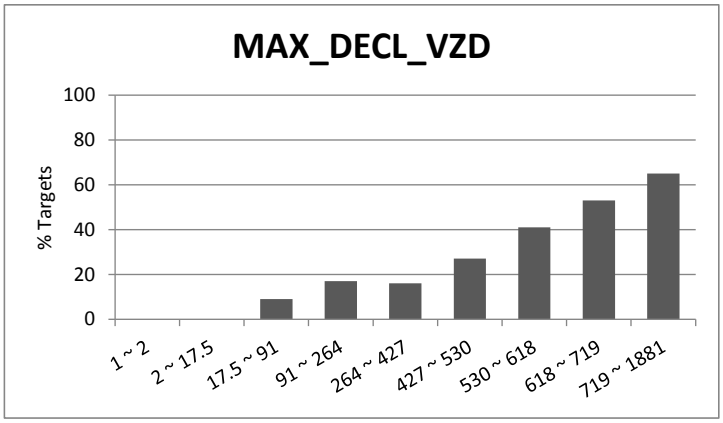

AANTAL_TERUGB_DECLNIV_NUM
% Targets
100
80
60
40
20
0
0 ~ 1
1 ~ 8
8 ~ 19
19 ~ 34
34 ~ 56
56 ~ 113
113 ~ 4859

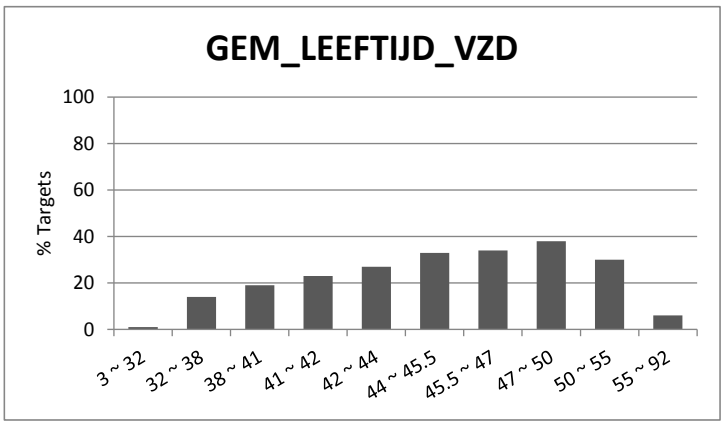

AANTAL_TERUGB_DECLNIV_PERC
% Targets
100
80
60
40
20
0
0 ~ 0.03
0.03 ~ 0.1
0.1 ~ 0.17
0.17 ~ 0.23
0.23 ~ 0.315
0.315 ~ 0.47
0.47 ~ 101

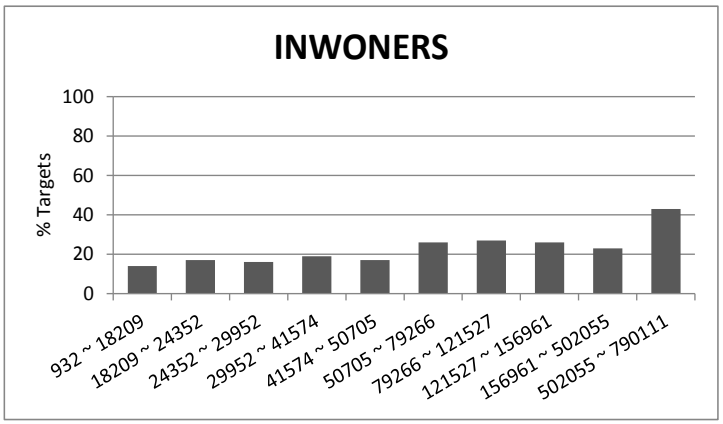

BEDR_TERUGB_DECLNIV_NUM
% Targets
100
80
60
40
20
0
0 ~ 25.075
25.075 ~ 194.71
194.71 ~ 495.29
495.29 ~ 884.24
884.24 ~ 1524.215
1524.215 ~ 3222.63
3222.63 ~ 154102.8

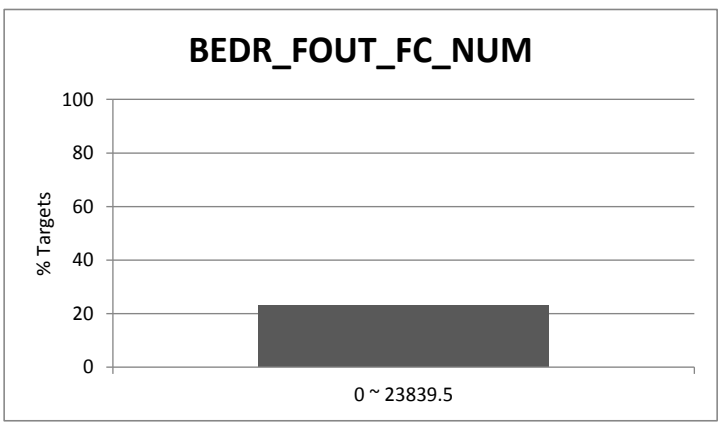

BEDR_TERUGB_DECLNIV_PERC
% Targets
100
80
60
40
20
0
0 ~ 0.03
0.03 ~ 0.12
0.12 ~ 0.19
0.19 ~ 0.27
0.27 ~ 0.38
0.38 ~ 0.61
0.61 ~ 101

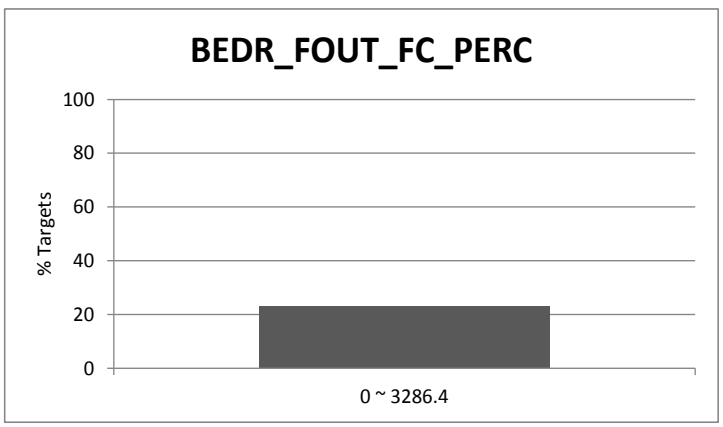

GEM_DECL_VZD
% Targets
100
80
60
40
20
0
1 ~ 2
2 ~ 3
3 ~ 8
8 ~ 10
10 ~ 12
12 ~ 14
14 ~ 17
17 ~ 22
22 ~ 275

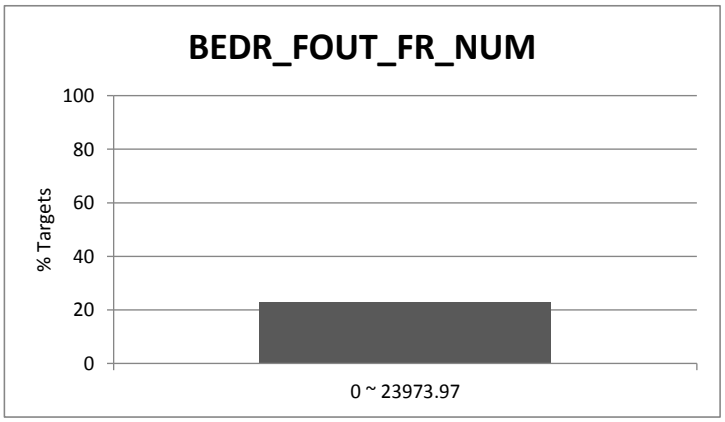

MAX_DECL_VZD
% Targets
100
80
60
40
20
0
1 ~ 2
2 ~ 17.5
17.5 ~ 91
91 ~ 264
264 ~ 427
427 ~ 530
530 ~ 618
618 ~ 719
719 ~ 1881

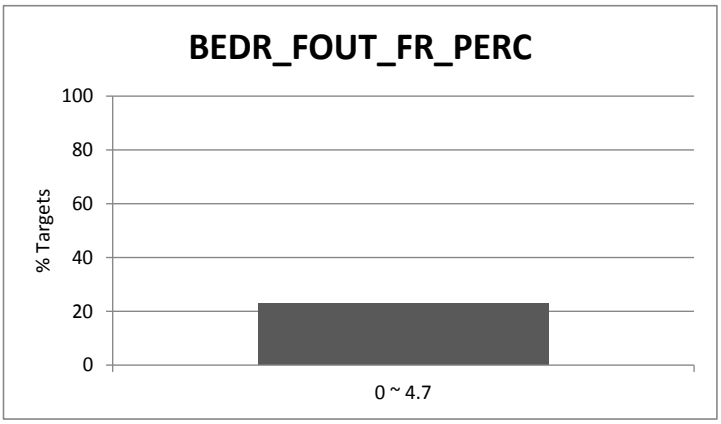

GEM_LEEFTIJD_VZD
% Targets
100
80
60
40
20
0
3 ~ 32
32 ~ 38
38 ~ 41
41 ~ 42
42 ~ 44
44 ~ 45.5
45.5 ~ 47
47 ~ 50
50 ~ 55
55 ~ 92

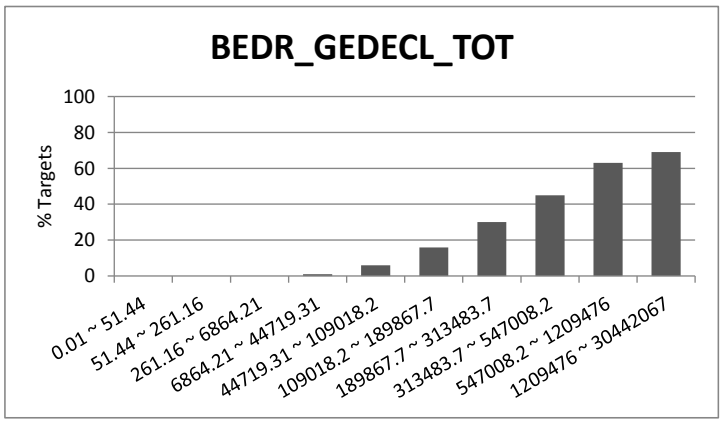

INWONERS
% Targets
100
80
60
40
20
0
932 ~ 18209
18209 ~ 24352
24352 ~ 29952
29952 ~ 41574
41574 ~ 50705
50705 ~ 79266
79266 ~ 121527
121527 ~ 156961
156961 ~ 502055
502055 ~ 790111

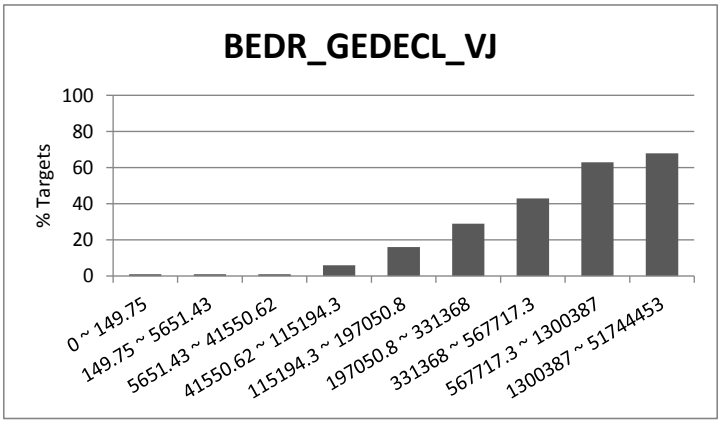

BEDR_FOUT_FC_NUM
% Targets
100
80
60
40
20
0
0 ~ 23839.5

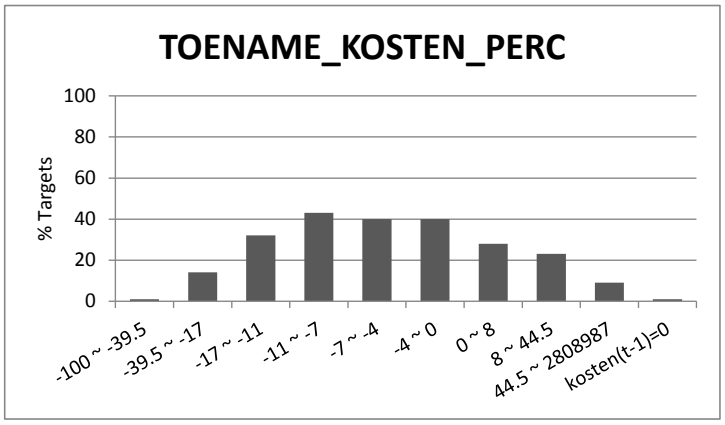

BEDR_FOUT_FC_PERC
% Targets
100
80
60
40
20
0
0 ~ 3286.4

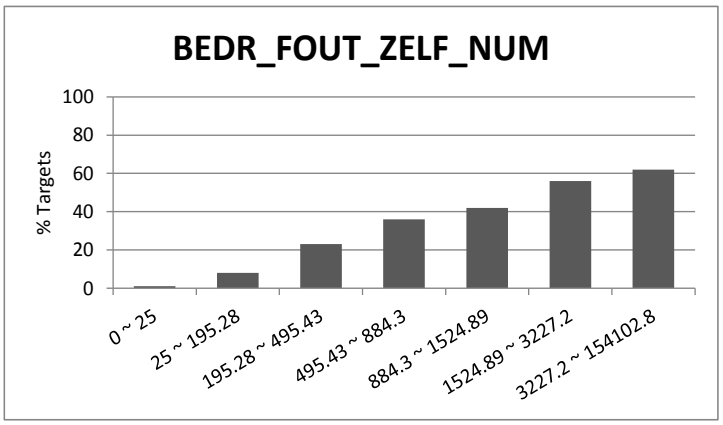

BEDR_FOUT_FR_NUM
% Targets
100
80
60
40
20
0
0 ~ 23973.97

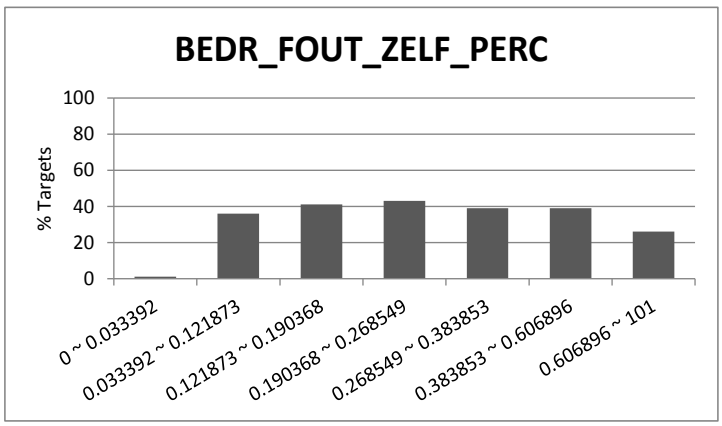

BEDR_FOUT_FR_PERC
% Targets
100
80
60
40
20
0
0 ~ 4.7


BEDR_GEDECL_TOT
% Targets
100
80
60
40
20
0
0.01 ~ 51.44
51.44 ~ 261.16
261.16 ~ 6864.21
6864.21 ~ 44719.31
44719.31 ~ 109018.2
109018.2 ~ 189867.7
189867.7 ~ 313483.7
313483.7 ~ 547008.2
547008.2 ~ 1209476
1209476 ~ 30442067


BEDR_GEDECL_VJ
% Targets
100
80
60
40
20
0
0 ~ 149.75
149.75 ~ 5651.43
5651.43 ~ 41550.62
41550.62 ~ 115194.3
115194.3 ~ 197050.8
197050.8 ~ 331368
331368 ~ 567717.3
567717.3 ~ 1300387
1300387 ~ 51744453


TOENAME_KOSTEN_PERC
% Targets
100
80
60
40
20
0
-100 ~ -39.5
-39.5 ~ -17
-17 ~ -11
-11 ~ -7
-7 ~ -4
-4 ~ 0
0 ~ 8
8 ~ 44.5
44.5 ~ 2808987
kosten(t-1)=0


BEDR_FOUT_ZELF_NUM
% Targets
100
80
60
40
20
0
0 ~ 25
25 ~ 195.28
195.28 ~ 495.43
495.43 ~ 884.3
884.3 ~ 1524.89
1524.89 ~ 3227.2
3227.2 ~ 154102.8


BEDR_FOUT_ZELF_PERC
% Targets
100
80
60
40
20
0
0 ~ 0.033392
0.033392 ~ 0.121873
0.121873 ~ 0.190368
0.190368 ~ 0.268549
0.268549 ~ 0.383853
0.383853 ~ 0.606896
0.606896 ~ 101

Bibliografie

[1] Donalek, C. *Supervised and Unsupervised Learning*, april 2011. http://www.astro.caltech.edu/~george/aybi199/Donalek_Classif.pdf.

[2] Dr. Langejan, T.W., voorzitter Raad van Bestuur, namens De Nederlandse Zorgautoriteit. *Rapport Onderzoek Zorgfraude*, december 2013. https://www.nza.nl/1048076/1048181/Rapport_Onderzoek_Zorgfraude.pdf.

[3] Drs. Kursten, J.C.E., Unitmanager Eerstelijns Zorg en Ketens, namens De Nederlandse Zorgautoriteit. *Beleidsregel Prestatiebeschrijvingen voor Farmaceutische zorg 2015*, 2015. https://www.nza.nl/98174/139255/1036985/TB-CU-5075-03_Farmaceutische_zorg.pdf.

[4] Fawcett, T. *ROC Graphs: Notes and Practical Considerations for Data Mining Researchers*, januari 2003. http://www.hpl.hp.com/techreports/2003/HPL-2003-4.pdf.

[5] Gordon, L. Using Classification and Regression Trees (CART) in SAS® Enterprise Miner™ For Applications in Public Health. *SAS Global Forum 2013*, (089), 2013. http://support.sas.com/resources/papers/proceedings13/089-2013.pdf.

[6] Guillet, F.; Hamilton, J. *Quality Measures in Data Mining*. Springer, 2007. ISBN: 978-3-540-44918-8.

[7] Huang, J.; Ling, C.X. *Using AUC and Accuracy in Evaluating Learning Algorithms*, december 2003. http://home.cse.ust.hk/~qyang/Teaching/537/Papers/AUC-evaluation.pdf.

[8] Kantardzic, M. *Data Mining Concepts, Models, Methods and Algorithms*. Wiley-IEEE Computer Society Press, september 2011. ISBN: 978-0-470-89045-5.

[9] Maimon, O.; Rokach, L. *Data Mining and Knowledge Discovery Handbook*. Springer, 2005. ISBN: 978-0-387-24435-8.

[10] Maimon, O.; Rokach, L. *Data Mining with Decision Trees: Theory and Applications*. World Scientific, december 2007. ISBN: 978-9-812-77171-1.

[11] Molina, M.M.; Luna, J.M.; Romero, C.;Ventura, S. *Meta-learning approach for automatic parameter tuning: A case study with educational datasets*, 2012. http://educationaldatamining.org/EDM2012/uploads/procs/Short_Papers/edm2012_short_5.pdf.

[12] Shearer C. The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing*, 5(4), 2000. https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf.

[13] Srinivas, K.; Kavihta Rani, B.; Dr. Govrdhan, A. Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. *International Journal on Computer Science and Engineering*, (02). http://www.enggjournals.com/ijcse/doc/IJCSE10-02-02-25.pdf.

[14] Witten, I.H.; Frank, E.; Hall, M.A. *Data Mining: Practical Machine Learning Tools and Techniques (Third Edition)*. Morgan Kaufmann, januari 2011. ISBN: 978-0-12-374856-0.