Onderzoek Machine Learning in de Risicoverevening

Eindrapportage, 29 mei 2020

i2i | Gupta Strategists |

Ministerie van Volksgezondheid, Welzijn en Sport

Inhoudsopgave

Over dit rapport

Dit rapport beschrijft de resultaten van een brede verkenning van de mogelijkheden die machine learning algoritmes kunnen bieden voor de risicoverevening. Deze verkenning is uitgevoerd door Gupta Strategists en i2i (intelligence to integrity) in opdracht van het Ministerie van Volksgezondheid, Welzijn en Sport.

Broncode

De broncode van de analyses uitgevoerd in het kader van dit onderzoek zijn, inclusief toelichting, te downloaden op:

https://github.com/intelligence2integrity/ml-risicoverevening

Samenstelling projectteam

Het projectteam voor dit onderzoek bestond uit:

- **Gupta Strategists**: Roxanne van Donselaar-Busschers, Daan Livestro, Sjors Oudshoorn
- **i2i**: Birgitta de Gruijter, Jules van Ligtenberg, Diederik Perdok, Michel Taal

Het projectteam werd bijgestaan door prof. dr. Tom Heskes, hoogleraar Data Science aan de Radboud Universiteit.

Voor meer informatie over dit onderzoek kunt u contact opnemen met:

- Daan Livestro (Daan.Livestro@gupta-strategists.nl), of
- Birgitta de Gruijter (Birgitta.de.Gruijter@i2i.eu)

Samenvatting

Het Nederlandse risicovereveningssysteem werkt goed in haar belangrijkste doelstellingen: het creëren van een gelijk speelveld voor zorgverzekeraars en het daarbij zoveel mogelijk wegnemen van prikkels tot risicoselectie. Het systeem is dan ook wereldwijd koploper en wordt nauwkeurig gevolgd door andere landen. Evenwel lijkt de piek aan de mogelijkheden van het huidige model bereikt, getuige ook het feit dat partijen overeengekomen zijn dat het systeem binnenkort naar een onderhoudssysteem over kan gaan. Daarom is nu een goed moment om niet langer alleen te kijken naar incrementele verbeteringen van de huidige methode, maar de methode *an sich* te evalueren. Tegen deze achtergrond verkennen we in dit onderzoek de potentiële meerwaarde van machine learning technieken voor de risicoverevening.

Machine learning wordt overal om ons heen toegepast, bijvoorbeeld bij het segmenteren van klanten door supermarktketens, het plaatsen van gerichte online-advertenties door Facebook en het bepalen van energiebehoeftes door leveranciers. Ook in relatie tot risicoverevening is in (internationale) studies en eerdere WOR-onderzoeken voorzichtige ervaring opgedaan. Meerdere studies suggereren dat machine learning modellen zorgkosten mogelijk beter kunnen voorspellen dan het huidige risicovereveningsmodel.

Dit onderzoek is een brede verkenning van de mogelijkheden die machine learning algoritmes kunnen bieden voor de risicoverevening. Daarbij richt het onderzoek zich met name op het verbeteren van de verevenende werking middels het vervangen van het huidige OLS-model door een machine learning algoritme. Bijvangst zijn aanknopingspunten voor verbeteringen aan de huidige systematiek.

Onderzoeksopzet

Met behulp van literatuuronderzoek en expert opinion zijn mogelijke algoritmes geïdentificeerd en beoordeeld op vijf criteria: praktische toepasbaarheid, bewijskracht, verwachte verevenende werking, verwachte robuustheid en interpreteerbaarheid van resultaten. Dit heeft geleid tot de keuze voor de volgende vijf algoritmes:

- **M1 – Decision Tree**. De eenvoudigste klasse van alle tree-based modellen wordt toegepast op de OT-dataset. Net als OLS produceert het goed interpreteerbare resultaten doordat het algoritme middels een duidelijke beslisboom komt tot gemodelleerde zorgkosten per verzekerde.
- **M2 – Piecewise Regressie.** Deze vorm van regressie passen we toe op de OT-dataset verrijkt met leeftijd in jaren. Het kan gezien worden als een logische uitbreiding van OLS. In dit model wordt leeftijd gebruikt om de data te segmenteren voor separate lineaire regressies.
- **M3 – Random Forest.** Dit complexere tree-based algoritme passen we toe op de verrijkte OT-dataset. Het algoritme kan goed omgaan met niet-lineaire interacties in de data waardoor het een hoog voorspellend vermogen heeft, maar produceert lastig te interpreteren resultaten omdat het een gemiddelde is van meerdere losse beslisbomen.
- **M4 – Gradient Boosting Machine.** Dit algoritme passen we toe op de brondataset, d.w.z. de OT-dataset aangevuld met leeftijd in jaren en onderliggende gegevens voor morbiditeitscriteria. Gradient Boosting Machines zijn de meest krachtige van alle tree-based algoritmes en scoren vaak hoog bij internationale machine learning competities. De

verevenende werking is naar verwachting gunstig, hoewel ook dit algoritme daarvoor inlevert op interpreteerbaarheid.

- **M5 – Artificial Neural Network.** Het Artificial Neural Network passen we toe op de brondataset. Dit algoritme is in meerdere studies met succes toegepast op risicoverevening, maar is door grote mate waarin het interacties toelaat een van de meest lastig te interpreteren modellen.

In dit onderzoek zijn de modellen geoptimaliseerd naar R^2. Dit is binnen de risicoverevening de meest gangbare maat, en zorgt dat de voorspelling van een model een benadering is van de gemiddelde werkelijke kosten.

De geselecteerde algoritmes variëren in de mate waarin ze afwijken van de huidige methodiek en bieden daarmee een brede verkenning van de mogelijkheden van machine learning voor de risicoverevening.

Om bij het ontwikkelen en valideren van de modellen overfitting te voorkomen is voor dit onderzoek 30% van de unieke verzekerden willekeurig toegewezen aan een testset. De resterende 70% van de verzekerden vormt de trainingsset. Alle algoritmes zijn uitsluitend getraind met behulp van deze trainingsset. Bij het ontwikkelen van ieder model binnen de trainingsset is gebruik gemaakt van 10-fold cross validation. Na definitieve vaststelling van de modelparameters zijn de maatstaven van ieder model berekend voor de testset.

Tenslotte gaven interviews met belanghebbenden en betrokkenen belangrijke inzichten met betrekking tot de consequenties van het toepassen van machine learning op de huidige werkwijze binnen risicovereveningscyclus.

Resultaten

In dit onderzoek is in een beperkte bouwtijd van 3 maanden met machine learning technieken een eerste verbetering in verevenende werking (R^2) gerealiseerd van 35.1% naar 38.5%. Hiermee is aangetoond dat machine learning technieken in potentie meerwaarde hebben voor de Nederlandse risicoverevening.

Naast een verbetering in R^2, waarop is geoptimaliseerd, zijn gedetailleerde maatstaven berekend, waaruit een aantal aanvullende conclusies te destilleren valt:

- **Model M1 - Decision Tree** scoort iets minder goed dan het huidige model op nagenoeg alle maatstaven. Op zowel individu-, subgroep-, als verzekeraarsniveau worden de maatstaven negatief beïnvloed.
- **Model M2 – Piecewise Regressie** behaalt een verbetering op R^2 ten opzichte van het huidige model. Opmerkelijk is dat een vergelijking met OLS op alle individuele segmenten behalve het segment 0-8 jarigen een verbetering van R^2 laat zien. Hier staat echter tegenover dat er substantieel meer verzekerden zijn met een negatief normbedrag. Dit komt vooral doordat het model gebruik maakt van een leeftijdssegment van 0-8 jaar, en dus het afwijkende kostenpatroon van 0-jarigen niet apart als segment meeneemt. De gemodelleerde zorgkosten van een deel van de 8-jarigen worden hierdoor negatief. Op subgroep- en verzekeraarsniveau is een kleine verslechtering waarneembaar op het merendeel van de maatstaven. Overigens zijn de uitkomsten op de maatstaven beter als dit model in beperkt aangepaste variant wordt toegepast, door een regulier OLS uit te voeren op de gevonden leeftijdssegmenten.
- **Model M3 – Random Forest** scoort op alle individuele maatstaven beter dan het huidige model. Op dezelfde data laat het model ten opzichte van OLS een verbetering in R^2 zien

van 35,1% naar 36,3%. Op subgroep- en verzekeraarsniveau laten verschillende maatstaven zowel een verbetering als een verslechtering zien. Het toevoegen van continue leeftijd heeft een beperkt positieve impact op de maatstaven voor dit model.

- **Model M4 – Gradient Boosting Machine** behaalt gunstige resultaten op bijna alle maatstaven. Op dezelfde data laat het model ten opzichte van OLS een verbetering in R^2 zien van 35,1% naar 36,3%. Bij het toevoegen van extra brondata blijkt de meerwaarde ten opzichte van OLS pas echt goed: de R^2 verbetert dan naar 38,5%, terwijl een OLS model op dezelfde data een R^2 van 36,1% realiseert. Met name de verbeteringen op individuele maatstaven vallen op en zijn de beste van alle modellen. Enkel de bandbreedte van het resultaat van middel en grote verzekeraars verslechtert licht.
- **Model M5 – Artificial Neural Network** behaalt gunstige resultaten op alle maatstaven. Op dezelfde data laat het model ten opzichte van OLS een verbetering in R^2 zien van 35,1% naar 36,3%. Net als bij de Gradient Boosting Machine blijkt de meerwaarde ten opzichte van OLS pas echt goed wanneer we extra brondata toevoegen: de R^2 verbetert dan naar 38,2%, terwijl een OLS model op dezelfde data een R^2 van 36,2% realiseert. Het complete getrainde model, dus inclusief aanvullende data, laat op zowel individu-, subgroep- als verzekeraarsniveau verbeteringen zien ten opzichte van een OLS model op dezelfde aanvullende dataset. Vooral de uitkomsten op verzekeraarsniveau zijn zeer gunstig en de beste van alle modellen. Alle bandbreedtes worden smaller. Dit vertaalt zich ook in de hoogste gewogen gemiddelde absolute resultaatverschuiving (GGARV).

Over de verschillende modellen blijven de belangrijkste risicokenmerken, dat wil zeggen de kenmerken die het meest bijdragen aan de verevenende werking, min of meer gelijk. De kenmerken MHK en MVV behoren in ieder model tot de belangrijkste kenmerken – zij blijven net als bij de OLS onverminderd belangrijk om tot een goede voorspelling van zorgkosten te komen. Wel neemt het relatieve belang van MVV af in modellen M1 t/m M5 ten opzichte van het uitgangsmodel M0, en neemt het belang van MHK toe. Dit effect is het grootst bij de *tree-based* modellen M1, M3 en M4. Opvallend is verder dat kenmerken als HKG, FDG, AVI en SES in modellen M4 en M5 niet tot de top 10 belangrijkste kenmerken behoren.

Naast conclusies omtrent verevenende werking leidde dit onderzoek ook tot enkele verdiepende inzichten die mogelijk van belang kunnen zijn voor de reguliere risicoverevening:

- De Decision Tree analyse karakteriseert vijf zeer grote groepen verzekerden met slechts een beperkt aantal kenmerken. In totaal bevatten deze vijf groepen 44% van de verzekerden. Uit de gebruikte kenmerken van deze vijf groepen blijkt dat met name relatief gezonde verzekerden goed gevangen kunnen worden met minder kenmerken dan het huidige OLS hanteert. Het is aannemelijk dat OLS op sommige van deze groepen overfit – het gebruik van méér vereveningscriteria voor deze verzekerden verslechtert mogelijk de voorspelling. Het OLS model kan wellicht van dit inzicht profiteren door rekening te houden met de interactie door parameters die de decision tree niet gebruikt (zoals, voor een van de groepen, FKG) op 0 te zetten voor de betreffende groep in de OLS.
- Uit de Piecewise Regressie analyse volgt nog een andere interessante observatie. De relatieve meerwaarde van een hoge risicoklasse, bijvoorbeeld MHK8, ten opzichte van de referentieklasse, MHK0 neemt af bij oudere verzekerden. Het ontvangen van het MHK8-normbedrag conform de reguliere OLS zou hebben geleid tot een overschatting van de zorgkosten. Of dit echt zo is valt op basis van dit onderzoek niet met zekerheid te zeggen. Wel zien we dat op het merendeel van verzekerdengroepen op basis van MHK en leeftijdssegment model M2 de resultaten beter voorspelt dan het uitgangsmodel.

Vervolgonderzoek zou nader op deze observatie in kunnen gaan en bepalen of segmentatie van kenmerken als MHK naar leeftijd het vereveningsresultaat kan verbeteren.

Consequenties van het toepassen van machine learning in de risicoverevening

Naast de positieve resultaten op R^2 van de inzet van machine learning modellen (resultaat), zijn de consequenties van het toepassen van machine learning (haalbaarheid) ook een belangrijk criterium in de aantrekkelijkheid van machine learning voor de risicoverevening.

Wanneer een van de machine learning technieken het huidige OLS-model zou vervangen heeft dit consequenties op alle aspecten van de risicovereveningscyclus. In dit onderzoek is gekeken naar:

- **Randvoorwaarden**: wet- en regelgeving, betrokken partijen en infrastructuur
- **Verbetercyclus**: uitvoering en interpretatie van verbeteronderzoeken
- **Onderhoudscyclus**: consequenties voor de diverse processtappen – Gegevensfase, OT en Normbedragenfase
- **Uitvoering**: consequenties voor het afrekeningsproces en voor ex-post mechanismen
- **Interpretatie en prikkelwerking**: validiteit, stabiliteit en homogeniteit, transparantie en eenvoud en prikkelwerking

Voor de decision tree (M1) bestaan vooral zorgen rondom stabiliteit van de resultaten. Piecewise lineaire regressie (M2) lijkt in veel opzichten op het huidige OLS-model, implementatie daarvan zou in dat opzicht dus eenvoudiger zijn.

Vooral modellen M3, M4 en M5 kennen behoorlijke implementatiedrempels. Voor toepassing ter vervanging van de huidige OLS is mogelijk aanpassing van wet- en regelgeving nodig, omdat deze modellen niet leiden tot “gewichten per criteria” zoals de Zorgverzekeringswet voorschrijft. Maar ook het uitvoeren van onderzoeken in de verbeter- en onderhoudscycli wordt onder deze modellen complexer. Naast het maken en testen van *features* (risicokenmerken) is enige extra tijd nodig om hyperparameters te optimaliseren, en wanneer gedetailleerde onderliggende data wordt gebruikt (zoals in M4 en M5) is het denkbaar dat een zeer gedetailleerde verzekerdenraming op alle kenmerken niet langer mogelijk is.

Conclusie: hoe verder met machine learning in de risicoverevening?

Een aantal modellen getest in dit onderzoek zijn laagdrempelig toepasbaar en zouden al snel meerwaarde kunnen bieden:

- Model M2 (Piecewise Lineaire Regressie) is een logische extensie van de huidige OLS en zou, met beperkte doorontwikkeling, potentieel snel geschikt zijn om de OLS te vervangen, met name om beter om te kunnen gaan met interacties tussen leeftijd en reguliere vereveningskenmerken zoals FKG of DKG. Het model is voor dit doel ook goed inzetbaar in de verbetercyclus.
- Model M1 (Decision Tree) is gemakkelijk uit te voeren, en heeft het ook in dit onderzoek geleid tot identificatie van grote, kostenhomogene groepen. Het periodiek uitvoeren van dergelijk onderzoek kan helpen om interacties tussen kenmerken op eenvoudige wijze inzichtelijk te maken, en vast te stellen of bijvoorbeeld interactietermen van toegevoegde waarde kunnen zijn.

Modellen M4 (Gradient Boosting Machine) en M5 (Artificial Neural Network) zijn zeer veelbelovende modellen. De resultaten zijn overtuigend wanneer deze modellen van meer data gebruik kunnen maken, maar zelfs op precies dezelfde data laten deze modellen (enigszins) betere resultaten zien dan de OLS. Daarbij is van belang dat de machine learning modellen op de beperkte

datasets niet uitgebreid geoptimaliseerd zijn, en verdere verbetering wellicht nog mogelijk is. Ook wanneer deze modellen voorlopig de OLS nog niet zullen *vervangen* kunnen ze als aanvullende onderzoekstechniek al snel van meerwaarde blijken als benchmark voor het 'best haalbare' vereveningsresultaat (bijvoorbeeld als onderdeel van de OT), en als onderdeel van partiële onderzoeken in de verbetercyclus.

Vooral wanneer meer informatie wordt toegevoegd laten deze modellen duidelijk meerwaarde zien. Het lijkt dat het vermogen van deze modellen om voorspelkracht te halen uit interacties daarmee nog meer uit de verf komt. Daarbij was dit slechts een verkennend onderzoek, en het is mogelijk dat in de toekomst met verdere optimalisatie nog betere resultaten behaald kunnen worden. Daarentegen zijn dit ook de modellen waarvan implementatie grotere consequenties zal hebben voor de huidige manier van werken, wat maakt dat de implementatiedrempel hoger is.

Aanbevelingen m.b.t. vervolgonderzoek op dit gebied

Naast de specifieke aanbevelingen voor implementatie van de verschillende modellen leidt dit onderzoek ook tot een aantal algemenere aanbevelingen voor vervolgonderzoek naar machine learning in de risicoverevening:

- Onze belangrijkste aanbeveling is om deze machine learning modellen enkele jaren parallel te draaien *naast* de reguliere OLS om geleidelijk ervaring op te bouwen met robuustheid en implementatie-aspecten van deze modellen, en om de belangrijkste onzekerheden te adresseren. Ook zouden de best presterende modellen retrospectief getest kunnen worden over meerdere jaren.
- Daarnaast zou het interessant zijn om te onderzoeken hoe machine learning modellen presteren wanneer inhoudelijk minder wenselijke aspecten van het huidige model worden weggelaten. In dit onderzoek zijn modellen getraind op *minimaal dezelfde features* als in het huidige model beschikbaar zijn, maar niet op een *beperktere* set aan features. Hoe waardevol zijn deze modellen wanneer bijvoorbeeld MHK of MVV niet meegenomen wordt als feature?
- In dit onderzoek is alleen het somatisch model verkend. Bij toekomstig onderzoek zou het ook nuttig zijn om de andere modellen, met name het GGZ-model, te verkennen. Ook is het zinvol om te verkennen in hoeverre machine learning kan helpen om te komen tot minder modellen in het algemeen.
- In dit onderzoek was het hoofddoel om modellen te vinden die leiden tot een zo hoog mogelijke R^2. Het zou nuttig zijn om in vervolgonderzoek ook te kijken naar andere doelfuncties, zoals het resultaat op subgroepen.

Aanbevelingen voor de reguliere risicovereveningscyclus

Tenslotte heeft dit onderzoek geleid tot inzichten die mogelijk relevant kunnen zijn bij vervolgonderzoek binnen de reguliere risicovereveningscyclus.

- Het is van belang om bij evaluatie van de huidige risicoverevening ook andere stappen dan de modelontwikkeling zelf te toetsen. In dit onderzoek kwam bijvoorbeeld de vraag voorbij in hoeverre het schalen van de OT-dataset naar de verzekerdenraming invloed heeft op de kwaliteit van de verevening. Hoe erg is het als we dit niet zo precies kunnen doen, bijvoorbeeld omdat de kenmerken te gedetailleerd zijn? Daarnaast kan bijvoorbeeld het samenvoegen van verschillende modellen (somatisch, GGZ en eigen risico) voor een totaalresultaat per verzekerde nuttig zijn om apart te testen. De belangrijkste aanbeveling op dit gebied is om het *geheel* aan stappen van het creëren van de gegevensset tot en met het uitvoeren van de definitieve vaststelling geprotocolleerd en integraal te testen.

- De Decision Tree analyse identificeerde enkele grote groepen gezonde verzekerden waarvoor het bieden van meer informatie aan het model ongunstig lijkt. Het is het onderzoeken waard of het aanpassen van de OLS zodat deze rekening houdt met de groepen geïdentificeerd in de decision tree analyse kan leiden tot een beter vereveningsresultaat in het huidige vereveningsmodel.
- Piecewise Regressie laat zien dat vooral MHK mogelijk interactie vertoont met leeftijd. Of dit echt zo is valt op basis van dit onderzoek niet met zekerheid te zeggen. Wel zien we dat op het merendeel van verzekerdengroepen op basis van MHK en leeftijdssegment model M2 de resultaten beter voorspelt dan het uitgangsmodel. Vervolgonderzoek zou nader op deze observatie in kunnen gaan en bepalen of segmentatie van kenmerken als MHK naar leeftijd het vereveningsresultaat kan verbeteren.
- Het uitsplitsen van FKG o.b.v. aantal ddd's lijkt van toegevoegde waarde voor de voorspelkracht, ook bij de een OLS model. Het is, bijvoorbeeld bij groot onderhoud, zinvol te verkennen of een fijnmaziger model, of in ieder geval nog specifiekere drempelwaardes per FKG, kan leiden tot een beter vereveningsmodel. Daarbij moet een verbetering van verevenende werking nadrukkelijk wel afgewogen worden tegen potentieel onwenselijke prikkels op het gebied van doelmatigheid.
- Leeftijd in jaren, aantal verpleegdagen, aantal zorgproducten, aantal diagnoses en aantal operaties zijn allen mogelijke *feature*s die de voorspelkracht van het OLS-model kunnen vergroten, en zijn daarmee het onderzoeken waard. Daarbij is vooral ook belangrijk om prikkelwerking mee te nemen in de evaluatie.
- Het uitsplitsen van DKG naar dx-groepen lijkt van toegevoegde waarde voor de voorspelkracht. Bij het recente groot onderhoud DKG (WOR 988) en in de pre-OT 2020 (WOR 990) is deze mogelijkheid reeds verkend, en is een fijnmaziger diagnosecategorisering opgenomen. Wij doen daarom nu geen aanvullende aanbeveling op dit gebied.

1 Inleiding

1.1 Aanleiding

Het Nederlandse risicovereveningssysteem werkt goed in haar belangrijkste doelstellingen: het creëren van een gelijk speelveld voor zorgverzekeraars en het daarbij zoveel mogelijk wegnemen van prikkels tot risicoselectie. Het systeem is dan ook wereldwijd koploper en wordt nauwkeurig gevolgd door andere landen. Evenwel lijkt de piek aan de mogelijkheden van het huidige model bereikt, getuige ook het feit dat partijen overeengekomen zijn dat het systeem binnenkort naar een onderhoudssysteem over kan gaan. Daarom is nu een goed moment om niet langer alleen te kijken naar incrementele verbeteringen van de huidige methode, maar de methode *an sich* te evalueren. Dat kan middels machine learning algoritmes die een volledig nieuwe wijze van risicoverevening mogelijk maken.

Machine learning wordt overal om ons heen toegepast, bijvoorbeeld bij het segmenteren van klanten door supermarktketens, het plaatsen van gerichte advertenties online door Facebook en het bepalen van energiebehoeftes door leveranciers. Ook in relatie tot risicoverevening is in (internationale) studies en eerdere WOR-onderzoeken voorzichtige ervaring opgedaan. Het onderzoek van Ismail (2018) [1] dat random forest en een gradient boosting machine heeft toegepast op de OT-data, suggereert dat machine learning modellen zorgkosten mogelijk beter kunnen voorspellen dan het huidige risicovereveningsmodel[1]. Internationaal heeft bijvoorbeeld onderzoek door Rose (2016) aangetoond dat het mogelijk is met machine learning technieken de voorspellende kracht te verbeteren én de formule te vereenvoudigen [2]. In enkele WOR-onderzoeken is daarnaast ervaring opgedaan met toepassingen van machine learning bij de constructie van risicoklassen [3], [4].

In dit onderzoek verkennen we welke mogelijkheden machine learning algoritmes bieden voor de risicoverevening. Daarbij richt het onderzoek zich met name op het verbeteren van de verevenende werking middels het vervangen van het huidige OLS-model door een machine learning algoritme. Bijvangst zijn aanknopingspunten voor verbeteringen aan de huidige systematiek. De potentie van machine learning algoritmes om de verevenende werking te verbeteren is drieledig: 1) het ontdekken van verbanden in de data die nu nog verborgen zijn; 2) het verruimen van beperkende modelaannames; 3) de mogelijkheid om meer en gedetailleerdere data aan het model toe te voegen. Dat laatste biedt tevens het potentiële voordeel dat het risicovereveningsmodel minder afhankelijk kan worden van de bestaande datastructuur met zorgvuldig opgebouwde risicoklassen.

Gegeven de limitaties van het huidige OLS-regressiemodel en de potentiële winst die te behalen is met machine learning technieken, voeren we een verkenning uit naar de mogelijkheden die machine learning wel én niet kan bieden voor de Nederlandse risicoverevening. Daarbij brengen we in kaart welke veranderingen dit meebrengt voor de verschillende fasen van de risicovereveningscycylus.

1.2 Doelstelling

Het doel van dit onderzoek is een brede verkenning van de waarde van machine learning toepassingen voor de risicoverevening, inclusief een beschrijving van toepasbare methodes, resultaten van toepassing, consequenties voor uitvoer van de risicoverevening en eventuele

[1] https://equalis.nl/verslaat-machine-learning-het-huidige-risicovereveningsmodel; onderzoek is niet publiek beschikbaar

aanknopingspunten voor verbeteringen binnen de huidige systematiek. Daarmee beantwoordt deze rapportage de onderstaande onderzoeksvraag en bijbehorende deelvragen.

Onderzoeksvraag: *'Wat is de toegevoegde waarde van machine learning voor de risicoverevening?'*

Deelvragen:

1. Van welke technieken is het zinvol om de toepasbaarheid voor de verevening verder te onderzoeken?
2. Welke resultaten geven deze technieken in termen van verevende werking?
3. Welke overige resultaten geven deze technieken die bruikbaar zijn in het huidige vereveningsmodel?
4. Welke andere consequenties (voor- en nadelen) van de toepassing van deze technieken voor de risicoverevening zijn er?
5. Formuleer op basis van de uitwerking van bovenstaande vragen een advies over toepassing van machine learning in de risicoverevening.

1.3 Onderzoeksopzet

Voor dit onderzoek gebruiken we de Overall Toets (OT) data van 2020. Daarnaast heeft Zorginstituut Nederland aanvullende bronbestanden aangeleverd met betrekking tot de morbiditeitscriteria in het OT-bestand. Zowel het OT-bestand als de aanvullende bronbestanden zijn gevalideerd aan de hand van WOR 973 en referentiebestanden [5].

Met behulp van literatuuronderzoek en expert opinion zijn mogelijke algoritmes geïdentificeerd en beoordeeld op vijf criteria: praktische toepasbaarheid, bewijskracht, verwachte verevenende werking, verwachte robuustheid en interpreteerbaarheid van resultaten. Dit heeft geleid tot de keuze voor de volgende vijf algoritmes: Decision Tree, Piecewise Regressie, Random Forest, Gradient Boosting Machine en Artificial Neural Network. De geselecteerde algoritmes variëren in de mate waarin ze afwijken van de huidige methodiek en bieden daarmee een brede verkenning van de mogelijkheden van machine learning voor de risicoverevening.

Om bij het ontwikkelen en valideren van de modellen overfitting te voorkomen is voor dit onderzoek 30% van de unieke verzekerden willekeurig toegewezen aan een testset. De resterende 70% van de verzekerden vormt de trainingsset. Alle algoritmes zijn uitsluitend getraind met behulp van deze trainingsset. Na definitieve vaststelling van de modelparameters zijn de maatstaven van ieder model berekend voor de testset.

Tenslotte gaven interviews met belanghebbenden en betrokkenen belangrijke inzichten met betrekking tot de implicaties van machine learning toepassingen de risicovereveningscyclus.

1.4 Opbouw rapportage

Het volgende hoofdstuk beschrijft het literatuuronderzoek naar machine learning toepassingen in de risicoverevening en de selectie van algoritmes voor dit onderzoek. Daarmee geeft dit hoofdstuk antwoord op deelvraag 1. Hoofdstuk 3 beschrijft de gebruikte databronnen en de toepassing van deze data in het onderzoek. Hoofdstuk 4 beschrijft de werking van de modellen en geeft de kwalitatieve resultaten van de verschillende modellen (deelvraag 2). Hoofdstuk 5 duidt de resultaten van de verschillende modellen en beschrijft lessen die we uit dit onderzoek kunnen trekken voor het huidige vereveningsmodel (deelvraag 3). Hoofdstuk 6 benoemt de implicaties voor de vereveningssystematiek van het toepassen van machine learning algoritmes (deelvraag 4). Tenslotte geeft hoofdstuk 7 een advies over de toepasbaarheid van machine learning voor de risicoverevening en aanbevelingen voor vervolgonderzoek (deelvraag 5).

2 Onderzoekskader voor Machine Learning toepassingen in de risicoverevening

Dit hoofdstuk beschrijft de selectie van machine learning algoritmes die we in dit onderzoek toetsen. Daartoe schetst paragraaf 2.1 eerst de relevante afbakening van machine learning toepassingen binnen de risicoverevening. Paragraaf 2.2 licht de selectiecriteria toe die we hanteren voor dit onderzoek. Paragraaf 2.3 beschrijft de verschillende klassen van machine learning die toegepast kunnen worden. Tenslotte beargumenteert paragraaf 2.4 de keuze voor vijf algoritmes op basis van de gehanteerde criteria.

2.1 Afbakening machine learning toepassing binnen risicoverevening

2.1.1 Het somatische vereveningsmodel kent 12 vereveningscriteria en meer dan 200 binaire risicoklassen

In dit onderzoek beperken wij ons tot het somatische risicovereveningsmodel – dit model heeft tot doel het zo goed mogelijk voorspellen van somatische zorgkosten op het niveau van individuele verzekerden. Het huidige somatische risicovereveningsmodel maakt gebruik van Ordinary Least Squares (OLS)-regressie op 12 vereveningscriteria:

- Leeftijd en geslacht
- Aard van het inkomen (AVI)
- Sociaaleconomische status (SES)
- Personen per adres (PPA)
- Regio
- Primaire diagnosekostengroep (pDKG)
- Secundaire diagnosekostengroep (sDKG)
- Farmaciekostengroep (FKG)
- Hulpmiddelenkostengroep (HKG)
- Fysiotherapiediagnosegroep (FDG)
- Meerjarige kosten van V&V (MVV), en
- Meerjarig hoge kosten (MHK)

Deze vereveningscriteria zijn verdeeld in ongeveer 200 binaire risicoklassen waarvoor normbedragen worden berekend (zie bron [5] voor een beschrijving van deze risicoklassen).

Voor de toepassing van machine learning op risicoverevening veralgemeniseren we de terminologie van criteria en risicoklassen naar ‘variabelen’, omdat veelal geen sprake meer is van risicoklassen. Daarnaast spreken we over ‘gemodelleerde zorgkosten’ (ook wel: ‘gemodelleerde waarde’ of ‘modelwaarde’) als het gaat over de ‘voorspelling’ die een machine learning-algoritme (afgekort tot ML-algoritme) maakt voor de zorgkosten van een individu op basis van de variabelen.

2.1.2 Machine learning is relevant op verschillende toepassingsniveaus van data

In dit onderzoek verkennen we de toepassing van machine learning op de Overall Toets (OT)-dataset, maar ook op datasets met een rijker informatieniveau. Machine learning is namelijk bij uitstek geschikt om verbanden te vinden in datasets met meerdere continue en discrete variabelen. Voor dit onderzoek zien we vier relevante toepassingsniveaus (zie Figuur 1):

a) **Modelleren zorgkosten op de OT-dataset.** De OT-dataset bevat de 12 criteria verdeeld over ongeveer 200 binaire risicoklassen zoals in paragraaf 2.1.1 beschreven.

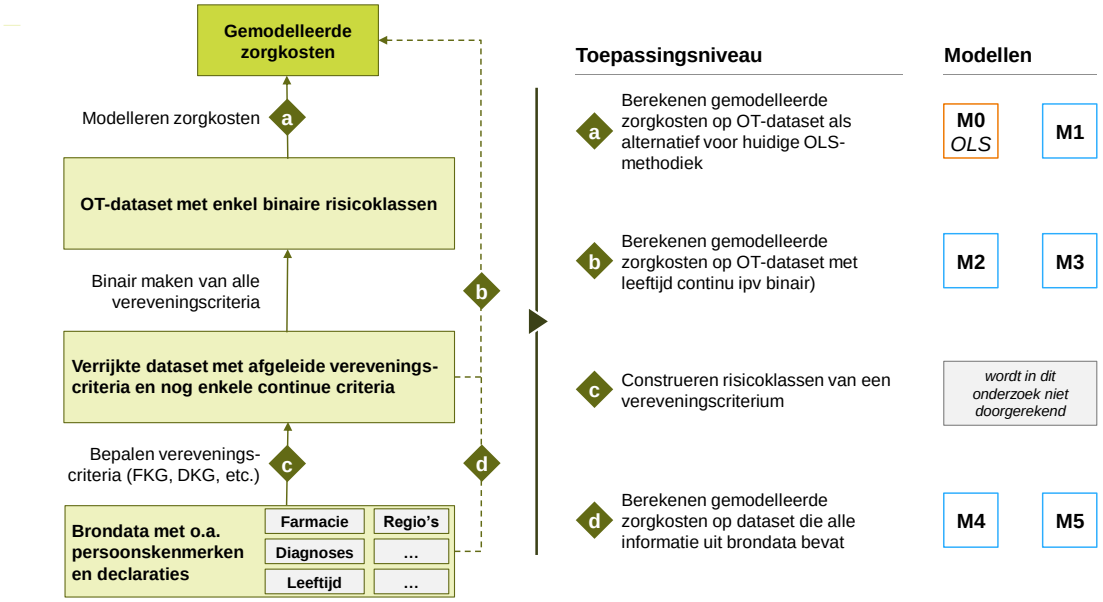
b) **Modelleren zorgkosten op een verrijkte OT-dataset.** Deze verrijkte OT-dataset lijkt sterk op de huidige OT-dataset, maar bevat leeftijd als continue variabele in plaats van in de vorm van risicoklassen van vijf jaar.

c) **De constructie van risicoklassen.** Voor ieder vereveningscriterium worden risicoklassen in een separate stap geconstrueerd uit combinaties van variabelen in een bronbestand. Dit toepassingsniveau is geen opzichzelfstaande vereveningsmethode – op de uiteindelijke risicoklassen volgt bijvoorbeeld OLS-regressie om tot gemodelleerde zorgkosten te komen.

d) **Modelleren zorgkosten op de brondataset.** Op deze dataset is nog geen enkele bewerking van variabelen toegepast – er zijn geen risicoklassen gedefinieerd. Voor dit onderzoek zullen we, vanuit zowel praktische als theoretische overwegingen,[2] alleen voor de morbiditeitscriteria de brondata gebruiken en voor de overige criteria bestaande risicoklassen.

Voor dit onderzoek verkennen we alle toepassingsniveaus waarop direct zorgkosten gemodelleerd worden, oftewel toepassingsniveaus a, b en d. Voor ieder van deze niveaus rekenen we twee modellen door (inclusief OLS als M0-model). Voor toepassingsniveau c geldt dat deze beter past binnen regulier groot onderhoud van een specifiek vereveningscriterium dan binnen dit onderzoek, en daarnaast potentieel zeer onderzoeksintensief is. Voor toepassingsniveau c rekenen we dan ook geen modellen door.

Figuur 1: op vier verschillende dataniveaus worden zes machine learning modellen verkend.



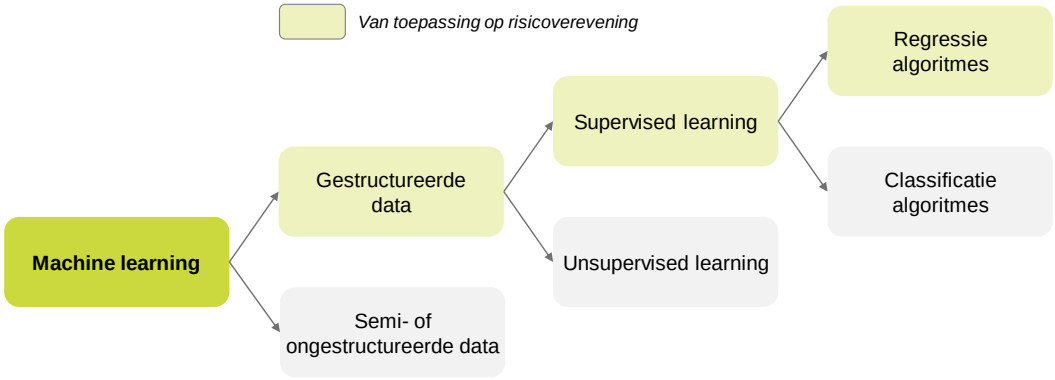
2.1.3 Supervised regression algoritmes zijn toepasbaar voor de risicoverevening

Er bestaan vele varianten van machine learning, ieder met een eigen toepassing. Logischerwijs is niet ieder algoritme even geschikt voor risicoverevening. Risicoverevening modelleert zorgkosten op basis van gestructureerde data en vraagt om *supervised regression* algoritmes (zie Figuur 2):

[2] Door alleen voor de morbiditeitscriteria (leeftijd en geslacht, FKG, DKG, HKG, FDG) de brondata te gebruiken voorkomen we dat de data onnodige persoonsgegevens bevat. Bovendien zijn de morbiditeitscriteria het meest voorspellend in het huidige model en passen zij het best bij de achterliggende gedachte van risicoverevening, en zijn ze daarmee de beste kandidaten voor verder onderzoek.

- **Gestructureerde data** – de risicovereveningsdata is getabelleerd en kent daarmee een duidelijke structuur. Iedere rij van de tabel komt overeen met één record (verzekerde). Iedere kolom geeft een variabele van de verzekerde weer, zoals leeftijd of AVI. De andere datavorm is (semi-)ongestructureerde data, zoals tekstbestanden en afbeeldingen. We beschouwen alleen algoritmes die toepasbaar zijn op gestructureerde data.
- **Supervised** – als de gewenste uitkomst bekend is spreekt men van *supervised* machine learning. Op toepassingsniveaus a, b en d is de uitkomst – de gemodelleerde zorgkosten – bekend. Bij de toepassing van *unsupervised* machine learning is de gewenste uitkomst niet bekend. Het derde toepassingsniveau, het construeren van risicoklassen, is hier een voorbeeld van. Omdat we het derde toepassingsniveau niet verder verkennen worden unsupervised algoritmes niet onderzocht in dit rapport.
- **Regressie** – tot slot is risicoverevening een vorm van regressie: het voorspellen van een continue, numerieke waarde. Er zijn ook algoritmes die sterker zijn in classificatie: het toewijzen van datapunten aan vooraf bekende klassen. In dit onderzoek worden alleen algoritmes geschikt voor regressie verkend.

Figuur 2: risicoverevening vraagt om algoritmes die behoren tot de categorie supervised regression op gestructureerde data.



2.2 Selectiecriteria algoritmes

Zoals reeds beschreven in paragraaf 2.1 is randvoorwaardelijk bij de selectie van machine learning algoritmes voor dit onderzoek dat het algoritme zorgkosten kan modelleren op basis van tabelvormige data (oftewel *supervised regression* algoritmes). Daarnaast moeten alle geselecteerde algoritmes een bestaand algoritme betreffen waarmee eerder ervaring opgedaan is in (semi-)wetenschappelijke studies.

Voor de definitieve selectie van algoritmes hanteren we daarnaast vijf selectiecriteria. Deze criteria zijn:

1. **Praktische toepasbaarheid:** in hoeverre is de methode toe te passen in de huidige risicovereveningscyclus? Het huidige model wordt al jaren toegepast en geeft betrokkenen duidelijkheid op basis van jaarlijks gepubliceerde normbedragen. Veel ML-algoritmes leveren echter geen normbedragen op maar een complexe combinatie van variabelen, of ze gebruiken een andere methodiek zoals vergelijking van verzekerden met een grote dataset. Deze algoritmes scoren *lager* op dit criterium. Bij andere ML-algoritmes schaalt de rekentijd niet-lineair met het aantal records, waardoor zij niet te onderzoeken zijn in de looptijd van dit onderzoek. Dergelijke algoritmes *voldoen niet* aan dit criterium en vallen daarom af.

2. **Bewijskracht:** in hoeverre is de techniek eerder toegepast, in literatuur of in praktische setting, op risicoverevening of vergelijkbare problemen? OLS-regressie is de standaardmethode voor risicoverevening in veel landen [6] en wordt daarom in veel vergelijkende onderzoeken meegenomen. Bij andere algoritmes is het aantal wetenschappelijke en praktische toepassingen lager. Dergelijke algoritmes scoren *lager* op dit criterium.

3. **Verevenende werking:** in hoeverre sluiten de gemodelleerde zorgkosten aan bij de werkelijke zorgkosten? Het huidige model is de afgelopen jaren consequent in staat geweest meer dan 30% van de variatie in werkelijke zorgkosten te modelleren [5], [7]. ML-algoritmes die de zorgkosten van verzekerden naar verwachting beter modelleren scoren *hoger* op dit criterium.

4. **Robuustheid:** in hoeverre produceert de methode stabiele uitkomsten en is het robuust voor andere instellingen van hyperparameters? Ieder jaar wordt het vereveningsmodel van het voorgaande jaar toegepast op de nieuwe dataset – de verschuivingen van normbedragen per risicoklasse zijn doorgaans zeer beperkt [5] en grotendeels te verklaren door toegenomen zorgkosten. Bij complexere algoritmes is het echter goed mogelijk dat deze gevoeliger zijn voor verschuivingen in de onderliggende data. Dergelijke algoritmes scoren *lager* op dit criterium.

5. **Interpreteerbaarheid:** in hoeverre is de methode begrijpelijk en zijn de uitkomsten goed uitlegbaar? De resultaten van het huidige vereveningsmodel zijn gemakkelijk interpreteerbaar – de normbedragen die gekoppeld zijn aan de verschillende risicoklassen zijn onafhankelijk van elkaar toe te passen op de volledige verzekerdenpopulatie. Bovendien zijn de relatieve hoogtes van de normbedragen veelal goed uitlegbaar op basis van bekende relaties tussen gezondheidsstatus en vervolgkosten. Er zijn echter ook ML-algoritmes die minder gemakkelijk interpreteerbaar zijn, doordat zij bijvoorbeeld verzekerden indelen op basis van combinaties van vele variabelen. Dergelijke algoritmes scoren *lager* op dit criterium.

Tot slot streven we er in dit verkennende onderzoek naar een zo breed mogelijk spectrum aan ML-algoritmes te onderzoeken. Dat wil zeggen dat we zoeken naar algoritmes die gestoeld zijn op verschillende principes en uitgangspunten en die variëren in de mate waarin ze ingrijpen op de huidige methodiek. Bij gelijke score op de andere criteria geven wij de voorkeur aan het algoritme dat het *minst* lijkt op de andere geselecteerde algoritmes.

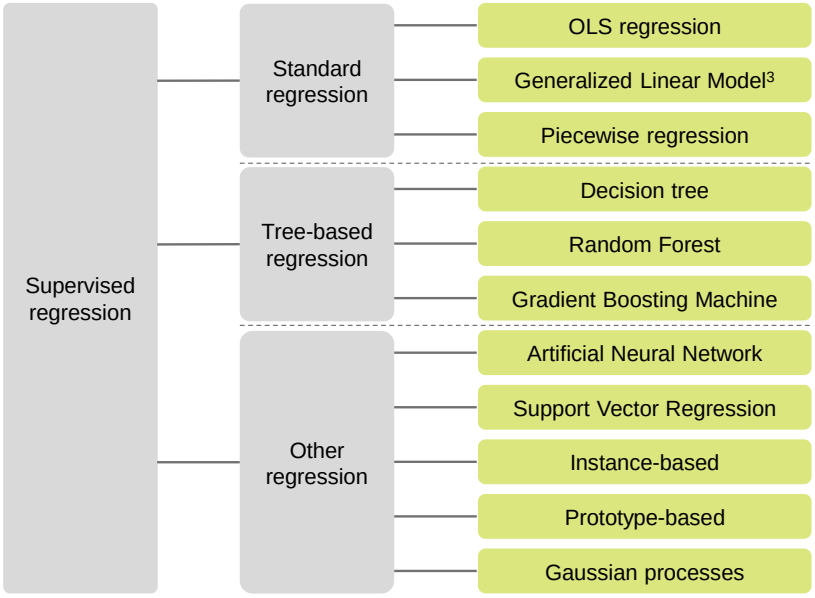
2.3 Uitkomsten literatuuronderzoek

Machine learning is een zeer breed begrip. Er bestaat een grote diversiteit aan algoritmes, met verschillende eigenschappen en wisselende toepasbaarheid voor risicoverevening. Zoals beschreven in paragraaf 2.1 en paragraaf 2.2 verkennen wij in dit onderzoek alleen ML-algoritmes die supervised regression kunnen uitvoeren op getabelleerde data en waarmee in (semi-) wetenschappelijk onderzoek ervaring opgedaan is bij vergelijkbare problemen. Alle verschillende vormen van machine learning die we in deze paragraaf beschrijven voldoen aan deze randvoorwaarden.

Uit literatuuronderzoek blijkt dat er voor supervised regression 11 gangbare vormen zijn[3] (zie Figuur 3). Ieder van deze vormen kent weer vele specifieke uitvoeringen; zo kan een Generalized Linear Model gebruik maken van onder andere lineaire regressie, logistische regressie of Poisson regressie. In een later stadium van dit onderzoek beschrijven we de exacte toepassing van de geselecteerde algoritmes.

Binnen de supervised regressie algoritmes maken we onderscheid tussen standaard regressie-algoritmes, tree-based algoritmes en overige regressie-algoritmes. Hieronder beschrijven we deze drie categorieën en de onderliggende algoritmes.

Figuur 3: elf algoritmes zijn in dit onderzoek overwogen, op te delen in drie categorieën: standard regression, tree-based regression en other regression.



2.3.1 Standard regression algoritmes

De standaard regressieanalyses behoren tot een aantal van de meeste gebruikte algoritmes van machine learning en leunen sterk op statistiek. Dit type algoritme zoekt de relatie volgens een bepaalde vergelijking tussen de verklarende variabelen en de afhankelijke variabele (of gemodelleerde waarde, voor risicoverevening de zorgkosten), waarbij het kwadraat van de afstand van alle datapunten tot de gemodelleerde waarden zo klein mogelijk is.

Standaard regressiemethodes zijn praktisch zeer goed toepasbaar: ze zijn goed uit te rekenen met conventionele rekenkracht en resulteren in normbedragen die eenieder kan toepassen op individuele verzekerden. Deze algoritmes worden veel toegepast in literatuur, al verschilt de bewijskracht per toepassingsvorm. De verwachte verevenende werking verschilt per algoritme, maar is over het algemeen lager dan van andere klassen van machine learning. Wel zijn de algoritmes relatief robuust voor veranderingen in de onderliggende data[5] en zijn uitkomsten gemakkelijk te interpreteren doordat ze normbedragen opleveren. De interpretatie wordt wel lastiger naarmate het model complexer wordt.

[3] Uiteraard zijn er meer vormen van supervised regression denkbaar; deze vallen of in één van de beschreven categorieën of worden niet succesvol toegepast op vergelijkbare vraagstukken
[4] Generalized Linear Model exclusief OLS-regressie, omdat het voor risicoverevening relevant is dit als separate categorie te beschouwen
[5] Dit geldt zolang model niet te veel variabelen krijgt waardoor overfitting van de data plaatsvindt

We kunnen drie vormen onderscheiden, die in toenemende mate complexiteit toevoegen:

1. **Ordinary Least Squares-regressie** is een van de meest gebruikte regressievormen. Het algoritme zoekt naar de combinatie van weegfactoren van alle variabelen welke leidt tot de lijn met gemiddeld de laagste kwadratische afstand tussen de gemodelleerde waarden en de werkelijke waarden. Het huidige risicovereveningsmodel gebruikt deze methode. De resultaten van lineaire regressie zijn gemakkelijk interpreteerbaar en zeer stabiel over verschillende jaren, maar blijven achter in verevenende werking ten opzichte van geavanceerdere algoritmes [8].
2. **Het Generalized Linear Model** is een generalisatie van OLS-regressie, waarbij de linkfunctie, die de relatie tussen de variabelen en de modelwaarden beschrijft, ook niet-lineaire interacties toelaat. De gekozen vorm van de linkfunctie is in sterke mate bepalend voor het voorspellend vermogen van dit model. Het gebruik van niet-lineaire linkfuncties leidt mogelijk tot betere verevenende werking dan OLS-regressie indien passende interacties tussen variabelen gemodelleerd worden; dit gaat wel in beperkte mate ten koste van interpreteerbaarheid.
3. **Piecewise Regressie** (ook wel gesegmenteerde regressie genoemd) past regressie toe op subsets (segmenten) van de data. Er wordt gesegmenteerd op één of meerdere variabelen. Vervolgens voert het algoritme een regressie uit op elk van de segmenten. Hoewel per segment ieder type regressie mogelijk is, overweegt dit onderzoek alleen het herhaaldelijk toepassen van lineaire regressie per segment, om een goede vergelijking met OLS-regressie te faciliteren. In eerder onderzoek bood piecewise regressie meer voorspellend vermogen dan OLS-regressie, wanneer toegepast op risicoverevening met extreme uitschieters [8], [9]. Net als voor het generalized linear model gaat deze extra verevenende werking wel ten koste van interpreteerbaarheid doordat het per segment afzonderlijke sets van normbedragen oplevert.

2.3.2 Tree-based algoritmes

Tree-based algoritmes maken gebruik van een keuzeboom, waarbij de dataset bij iedere stap wordt gesplitst in (doorgaans twee) subsets. Per stap berekent het algoritme welke variabele en welke splitsingswaarde de dataset het best opdelen. Iedere subset wordt opnieuw gesplitst, vaak (maar niet noodzakelijk) op basis van een andere variabele. Een keuzeboom kan op zichzelf worden gebruikt in een Decision Tree algoritme, of worden gecombineerd in de Random Forest en Gradient Boosting Machine algoritmes.

Tree-based algoritmes hebben zich eerder bewezen voor de toepassing op risicoverevening [8]–[11]. Deze algoritmes hebben vaak een goede verevenende werking en leveren robuuste resultaten. Naarmate de algoritmes complexer worden neemt ook de verevenende werking toe, hoewel dit vaak ten koste gaat van de interpreteerbaarheid.

4. Een **Decision Tree** algoritme maakt gebruik van één keuzeboom, die bij grote datasets met veel variabelen honderden splitsingen kan bevatten. Het algoritme clustert de volledige data in subsets met eigen specifieke kenmerken en bijbehorende gemodelleerde waarde. Doorgaans leidt de Decision Tree tot goed interpreteerbare resultaten, met een goede verevenende werking als de data duidelijke clusters bevat [8], [10].
5. Het **Random Forest** algoritme gebruikt niet één keuzeboom maar verevent op basis van de gemiddelde uitkomst van meerdere kleine en onafhankelijk gemodelleerde keuzebomen. Waar *Decision Trees* de neiging hebben tot overfitting middelt het RF-algoritme de uitkomst van verschillende bomen om zo tot een beter generaliseerbaar model van de zorgkosten te komen. Daarvoor is dan wel vereist dat de individuele bomen zo weinig mogelijk met elkaar gecorreleerd zijn. Daarom wordt elke boom gemaakt op een willekeurige subset van de data en van de variabelen. Het algoritme herhaalt dit proces meerdere keren, waardoor het

uiteindelijk vele honderden simpele keuzebomen kan bevatten. Ook is het algoritme robuust voor veranderingen in de dataset doordat het meer dan één voorspelling gebruikt. Hier staat tegenover dat het model minder interpreteerbaar is dan het gebruik van een enkele keuzeboom [1], [2], [8], [9], [12], [13].

6. **Gradient Boosting Machines** maken ook gebruik van meerdere simpele keuzebomen. Het algoritme leert deze keuzebomen sequentieel; iedere nieuwe iteratie van de boom maakt gebruik van de restfout van de vorige keuzeboom om het model verder te trainen. Hierdoor benadert de gemodelleerde uitkomst iedere stap iets dichter de werkelijke waarde. Tijdens het verevenen telt het algoritme de uitkomsten van elke iteratie op om tot een uiteindelijke gemodelleerde waarde te komen voor de zorgkosten van een verzekerde. Zowel uit literatuuronderzoek als bij internationale machine learning competities[6] blijkt dat dit een van de meest krachtige vormen van regressie is met een sterk voorspellend vermogen en robuuste uitkomsten. Net als bij het Random Forest algoritme gaat dit ten koste van de interpreteerbaarheid van de uitkomsten [1], [9], [13], [14].

2.3.3 Overige regressie algoritmes

Vijf andere regressie-algoritmes zijn niet in een van bovenstaande categorieën te vatten: Artificial Neural Networks (ANN), Support Vector Regression (SVR), Instance-based Regression, Prototype-based Regression en Gaussian Processes. Deze machine learning vormen zijn diffuus in werkingsprincipes en verwachte uitkomst op de selectiecriteria. Wel is duidelijk dat zij minder vaak zijn toegepast op risicoverevening dan de eerdergenoemde algoritmes, en dat zij doorgaans minder goed toepasbaar zijn op grote datasets met veel variabelen.

7. **Artificial Neural Networks** (ook wel *deep learning*) modeleren complexe niet-lineaire verbanden tussen variabelen en modelwaarden in een netwerk dat bestaat uit ‘neuronen’ die georganiseerd zijn in verschillende lagen. De inputdata vormt de eerste laag en de gemodelleerde waarden vormen de laatste laag. Daartussen kunnen verschillende verborgen lagen zitten. In iedere laag verwerken alle neuronen zelfstandig de ‘inputsignalen’ en geven zij een ‘outputsignaal’ door naar de volgende laag neuronen. Ieder neuron ontvangt een signaal van de neuronen uit de hogere laag, en stuurt een signaal naar de neuronen in de lagere laag – er zijn dus potentieel zeer veel en zeer gelaagde interacties tussen de inputvariabelen mogelijk. Ieder signaal heeft een numeriek gewicht. Het algoritme zoek naar de optimale combinatie van gewichten van alle signalen zodanig dat de modelwaarden de werkelijke waarden zo goed mogelijk benaderen. Artificial Neural networks zijn met succes getoetst op het modelleren van zorggebruik van verzekerden [9]. Uit het literatuuronderzoek blijkt dat deze algoritmes nauwkeuriger kunnen zijn dan lineaire regressie, maar vanwege de complexe interactiemogelijkheden tussen variabelen wel tot moeilijk uitlegbare resultaten leiden.
8. **Support Vector Regression** is een specifieke vorm van regressie om lineaire en niet-lineaire verbanden te vangen. Het algoritme combineert de beschikbare variabelen dusdanig dat zoveel mogelijk punten in de dataset zich binnen een gedefinieerde afstand (de foutmarge) van deze lijn bevinden. Soms lukt dit onvoldoende met alleen lineaire combinaties van de variabelen: in dat geval kan een specifieke *kernel* gekozen worden om de datapunten beter te scheiden. De kernel transformeert de bestaande variabelen, bijvoorbeeld middels een vermenigvuldiging of een logaritmische functie. Dankzij deze transformatie is het vaak beter mogelijk relatie in de data te vatten. Helaas neemt het aantal berekeningen van een support vector regression algoritme niet lineair toe met het aantal datapunten, waardoor het moeilijk toepasbaar is op grote datasets [11].

[6] Bijvoorbeeld het identificeren van fraude op: kaggle.com/c/ieee-fraud-detection

9. **Instance-based** algoritmes slaan alle vooraf bekende datapunten op als trainingsset. Zij vergelijken iedere nieuwe record met alle records in de trainingsset en bepalen op basis van vergelijkbaarheid de modelwaarde. Instance-based algoritmes zijn vatbaar voor overbodige variabelen die de vergelijkbaarheid met andere datapunten beïnvloeden: met behulp van *dimensionality reduction* algoritmes (zie kader) wordt voorkomen dat variabelen met laag voorspellend vermogen de regressie beïnvloeden. Ook deze vorm van machine learning schaalt niet lineair met het aantal records in de dataset, waardoor het moeilijk toepasbaar is op het modelleren van zorgkosten van een grote populatie [8], [9], [14], [15].
10. **Prototype-based algoritmes** zijn vergelijkbaar met de instance-based algoritmes, maar maken gebruik van prototypes als samenvatting van een (groot) aantal records. Het model vindt deze prototypes door bijvoorbeeld eerst de data te clusteren of door het aantal records dat gebruikt wordt bij de instance-based algoritmes bij voorbaat te reduceren. Hoe meer een prototype lijkt op een te voorspellen record, hoe zwaarder dit prototype mee zal wegen in de schatting van de modelwaarde. Net als bij de instance-based methode zijn deze algoritmes vanwege de grote hoeveelheid benodigde rekenkracht niet goed toepasbaar op grote datasets.
11. **Gaussian Processes** modelleren functies met onzekerheid – in de buurt van bekende datapunten is de onzekerheid laag, maar tussen datapunten in neemt de onzekerheid toe. Hiervoor maakt het algoritme de aanname dat de geobserveerde uitkomsten gezamenlijk een Gaussiaanse verdeling volgen. De modelwaarde van records tussen bestaande datapunten in is afhankelijk van de gekozen instellingen van het algoritme. Ook deze vorm van machine learning is moeilijk toepasbaar op grote datasets omdat het niet-lineair schaalt met het aantal records in de set.

Bovenstaande algoritmes zijn allen in staat tot een volledige regressie op de dataset. Daarnaast zijn er enkele aanvullende technieken die regelmatig gebruikt worden: *regularization* en *dimensionality reduction.*

Regularization en dimensionality reduction worden in machine learning gebruikt om overfitting op de data te voorkomen. Dit kan bijvoorbeeld interessant zijn bij piecewise regression modellen, waarbij de dataset eerst gesegmenteerd wordt alvorens een regressie uit te voeren, waardoor er opeens veel meer variabelen zijn in verhouding tot de omvang van de datasets waarop regressie wordt toegepast. Middels een *penalty* op het aantal mee te nemen variabelen of een *principal component analysis* worden variabelen met veel ruis minder zwaar gewogen in het model – of zelfs verwijderd – dan variabelen met een hoog voorspellend vermogen. Uit literatuur volgt dat dit het voorspellend vermogen van algoritmes op nieuwe records kan doen toenemen [15]. In de modellen in dit onderzoek is deze techniek niet toegepast, dit kan derhalve nog een aanknopingspunt zijn voor toekomstig onderzoek zijn.

2.4 Selectie van algoritmes voor dit onderzoek

De 11 vormen van machine learning zijn gescoord op de vijf criteria uit paragraaf 2.2 met behulp van literatuuronderzoek en expertinterviews (zie Tabel 1). De toelichting op de scores volgt uit de omschrijving per algoritme in paragraaf 2.3.

Vijf algoritmes zijn niet geselecteerd voor verder onderzoek om twee verschillende redenen:

1. Support Vector Regression, Instance-Based Regression, Prototype-Based regression en Gaussian Processes schalen allen niet-lineair met het aantal records[7]. Dat maakt deze algoritmes praktisch niet toepasbaar voor dit onderzoek. Bovendien geldt voor deze algoritmes dat er relatief weinig bewijskracht is; zowel in internationale machine learning competities als wetenschappelijke literatuur worden andere algoritmes vaker toegepast.
2. Generalized linear model regressie is niet gekozen omdat we voor de diversiteit slechts één standaard regressietechniek naast OLS willen toetsen. Piecewise regressie en generalized linear model algoritmes scoren op alle criteria gelijk. De voorkeur gaat uiteindelijk uit naar piecewise regressie, omdat de huidige OLS-methodiek al een specifieke variant is van een generalized linear model.

Tabel 1: de beoordeling van 11 vormen van machine learning op vijf criteria: praktische toepasbaarheid, bewijskracht, verevenende werking, robuustheid en interpreteerbaarheid. Een lage score is ongunstig op dat criterium, een hoge score is gunstig. Wanneer een algoritme niet toepasbaar is voor risicoverevening vanwege een criterium is dit in oranje aangegeven.

	Praktisch toepasbaar	Bewijskracht	Verevenende werking	Robuustheid	Interpreteerbaarheid	Selectie
Standard regression						
OLS-Regression	Hoog	Hoog	Laag	Hoog	Hoog	**M0**
Generalized Linear Model	Hoog	Midden	Midden	Hoog	Midden	-
Piecewise regression	Hoog	Midden	Midden	Hoog	Hoog	**M2**
Tree-based regression						
Decision Tree	Midden	Midden	Midden	Midden	Hoog	**M1**
Random Forest	Midden	Hoog	Hoog	Hoog	Laag	**M3**
Gradient Boosting Machine	Midden	Hoog	Hoog	Hoog	Laag	**M4**
Other regression						
Artificial Neural Network	Midden	Midden	Midden	Midden	Laag	**M5**
Support Vector Regression	Niet	Midden	Midden	Midden	Laag	-
Instance-based	Niet	Laag	Midden	Hoog	Hoog	-
Prototype-based	Niet	Laag	Midden	Hoog	Hoog	-
Gaussian processes	Niet	Laag	Midden	Hoog	Laag	-

Op basis van de scoring van de algoritmes komen we tot de volgende selectie van modellen (zie ook Figuur 4)[8]:

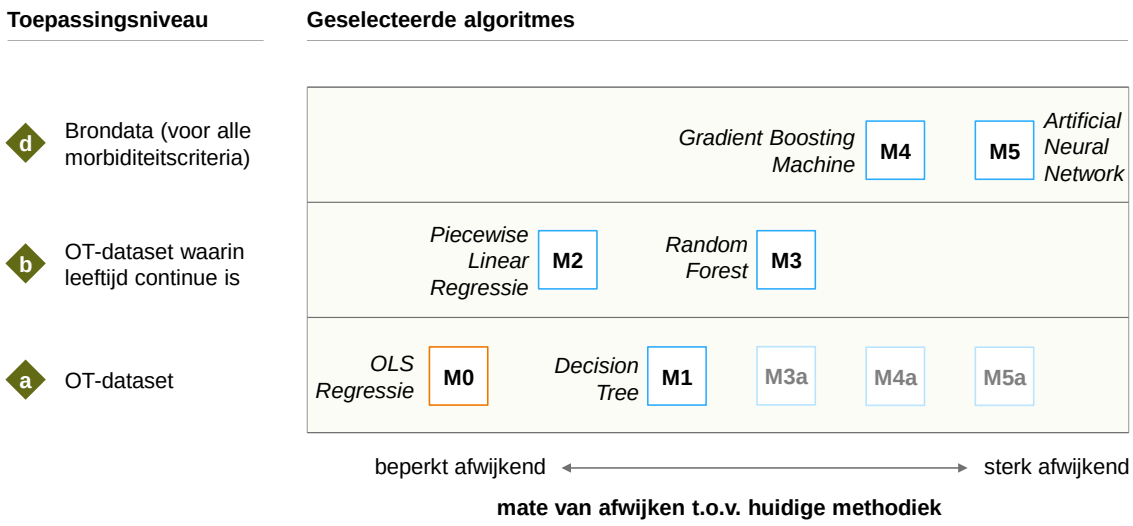
- **M0 – OLS**. Dit is het huidige model dat toegepast wordt op de OT-dataset. Het dient als baseline om de andere vijf modellen mee te vergelijken.
- **M1 – Decision Tree**. De eenvoudigste klasse van alle tree-based modellen wordt ook toegepast op de OT-dataset. Net als OLS produceert het goed interpreteerbare resultaten doordat het

[7] Er bestaan methodes om het aantal berekeningen van deze algoritmes te verkleinen, maar deze verlagen daarmee over het algemeen ook het voorspellend vermogen van het algoritme.
[8] Per model zijn vaak meerdere onderliggende algoritmes geschikt. De definitieve selectie van het algoritme wordt in een later stadium van dit project gemaakt.

algoritme middels een duidelijke beslisboom komt tot gemodelleerde zorgkosten per verzekerde.

- **M2 – Piecewise Regressie.** Deze vorm van regressie passen we toe op de verrijkte OT-dataset. Het kan gezien worden als een logische uitbreiding van OLS. In dit model zal leeftijd gebruikt worden om de data te segmenteren voor separate lineaire regressies.
- **M3 – Random Forest.** Dit complexere tree-based algoritme passen we toe op de verrijkte OT-dataset. Het algoritme kan goed omgaan met niet-lineaire interacties in de data waardoor het een hoog voorspellend vermogen heeft, maar produceert lastig te interpreteren resultaten omdat het een gemiddelde is van meerdere losse keuzebomen.
- **M4 – Gradient Boosting Machine.** Dit algoritme passen we toe op de brondataset. Gradient Boosting Machines zijn de meest krachtige van alle tree-based algoritmes en scoren vaak hoog bij internationale machine learning competities. De verevenende werking is naar verwachting gunstig, hoewel ook dit algoritme daarvoor inlevert op interpreteerbaarheid.
- **M5 – Artificial Neural Network.** Het Artificial Neural Network passen we toe op de brondataset. Dit algoritme is in meerdere studies met succes toegepast op risicoverevening, maar is door grote mate waarin het interacties toelaat een van de meest lastig te interpreteren modellen.

Figuur 4: vijf machine learning modellen zijn geselecteerd voor nadere verkenning in dit onderzoek op drie toepassingsniveaus[9].



Gedurende het onderzoek is bij de begeleidingscommissie veel interesse gebleken om de nieuwe modellen ook door te rekenen zonder het toepassingsniveau aan te passen. Zo is het effect van het toepassen van een ander predictiemodel los te trekken van het effect van het verrijken van de data die modellen tot hun beschikking hebben om de predictie te maken. In Figuur 4 staan deze extra modelvarianten (M3a, M4a en M5a) aangeduid in grijs. We merken op dat het voor model M2 (Piecewise Regressie) niet mogelijk is om alleen de OT-dataset te gebruiken, omdat de continue leeftijd nodig is om de segmenten te definiëren.

Voor deze extra modelvarianten rapporteren we in dit onderzoek slechts beperkt uitkomsten, namelijk daar waar het relevant is om het effect van het model te onderscheiden van het effect van het toepassingsniveau.

[9] Zoals beschreven in paragraaf 2.1 verkent dit onderzoek geen algoritmes voor toepassingsniveau 3 – het bepalen van vereveningscriteria (zoals FKG's en DKG's) uit de brondata.

3 Databronnen en toepassing in dit onderzoek

Dit hoofdstuk beschrijft welke data beschikbaar was en hoe deze data toegepast is voor dit onderzoek. Paragraaf 3.1 beschrijft de gebruikte databronnen en de wijze waarop deze gevalideerd is. Paragraaf 3.2 beschrijft vervolgens welke inputparameters op basis van deze databronnen gebruikt zijn voor de verschillende modellen. Paragraaf 3.3 gaat in op de wijze waarop overfitting in dit onderzoek voorkomen wordt door de dataset op te delen in een trainingsset en een testset en *10-fold cross validation* te gebruiken binnen de trainingsset. Tenslotte toont paragraaf 3.4 aan dat de selectie van de testset dusdanig uitgevoerd is dat hiermee een voldoende representatieve subset van de volledige dataset is gecreëerd.

3.1 Beschikbare databronnen

Voor dit onderzoek verrijken we het OT-bestand 2020 met onderliggende bronbestanden die betrekking hebben op leeftijd en morbiditeitscriteria. Zorginstituut Nederland heeft hiertoe vier aanvullende bestanden aangeleverd:

1. **Persoonskenmerken 2016 en 2017**. Deze bestanden bevatten de leeftijd op 30 juni 2016 en 2017 van alle verzekerden in jaren. In totaal bevatten deze bestanden de leeftijd van 16.912.322 en 17.014.864 unieke verzekerden, respectievelijk. Voor 241.559 unieke verzekerden (89.925 verzekerdenjaren) uit het OT-bestand is in beide bestanden géén leeftijdsinformatie beschikbaar. Zorginstituut Nederland heeft daarom nog twee aanvullende bestanden aangeleverd met het geboortejaar van alle mensen die op enig moment verzekerd waren in 2016 en 2017. Met deze aanvullende data is voor alle verzekerden waarvoor leeftijdsinformatie nog ontbrak de leeftijd vast te stellen[10].
2. **Declaratiebestand farmaceutische zorg 2016**. Dit bestand bevat alle declaraties van extramuraal medicijngebruik binnen de somatische zorg (dus exclusief gebruik van intramurale geneesmiddelen en add-on medicatie). Van deze declaraties is gegeven: ATC-code, productcode, ZI-artikelnummer en DDD-factor[11]. In totaal bevat dit bestand 242.383.160 declaratieregels van 11.556.060 unieke verzekerden. Deze data is gevalideerd door vast te stellen dat voor iedere verzekerde in het OT-bestand met een FKG-classificatie tenminste één bijbehorend middel (inclusief de informatie uit het add-on-bestand) is gedeclareerd[12].
3. **Declaratiebestand add-on geneesmiddelen 2016.** Dit bestand bevat alle declaraties van add-ongeneesmiddelen. Dit bestand bevat 1.417.953 declaratieregels van 197.445 unieke verzekerden. Separaat is een uitvoeringstabel aangeleverd met een vertaling van declaratiecodes naar ATC-codes en DDD-waarde. Deze dataset is gevalideerd in combinatie met het declaratiebestand farmaceutische zorg, zoals hierboven beschreven.
4. **Declaratiebestand DBC-somatisch 2016**. Dit bestand bevat declaraties van alle DBC-trajecten: DBC-code, diagnosecode, specialismecode en zorgproductcode. Dit bestand bevat 22.262.516 declaratieregels van 6.980.142 unieke verzekerden. Deze data is gevalideerd door vast te stellen dat voor iedere verzekerde in het OT-bestand met een DKG-classificatie tenminste één bijbehorende diagnose of zorgproductcode is gedeclareerd.

Bijlage A vat alle gegevens uit de bronbestanden samen en beschrijft welke gegevens beschikbaar zijn voor de verschillende modellen.

[10] Voor verzekerden met discrepantie tussen de datasets nemen we de gemiddelde leeftijd over de bestanden
[11] Deze gegevens zijn door Zorginstituut Nederland gekoppeld op basis van de G-Standaard
[12] Voor slechts 17 verzekerden is deze toewijzing niet reproduceerbaar

3.2 Inputparameters van de modellen

De in paragraaf 3.1 beschreven data wordt in verschillende mate beschikbaar gesteld voor de modellen M1 t/m M5. Model M1 (Decision Tree) gebruikt enkel het OT-bestand. M2 en M3 gebruiken het OT-bestand aangevuld met de leeftijd in jaren. Hiervoor koppelden we data uit de aangeleverde persoonskenmerken datasets aan het OT-bestand.

M4 en M5 maken gebruik van aanvullende data met betrekking tot farmaciegebruik en MSZ-diagnoses. Dergelijke declaratiebestanden kunnen vele zeer gedetailleerde data-items per verzekerde bevatten. Alleen al het farmaciebestand bevat gemiddeld meer dan 20 declaratieregels per verzekerde met farmaciegebruik. Voor elk van deze regels is dan weer informatie zoals dosis, ATC-code, etc. beschikbaar. Op deze informatie zijn tevens bewerkingen mogelijk welke zeer relevante input voor het uiteindelijke model kunnen zijn. Als we bijvoorbeeld willen dat een model rekening houdt met factoren zoals 'totale jaardosis per ATC-code', 'totaal aantal declaraties', 'aantal ddd per fkg' etc. dan is voorbewerking nodig – machine learning modellen doen dit niet uit zichzelf.

Daarom is het noodzakelijk om zogenaamde *features* vast te stellen die een machine learning algoritme daadwerkelijk kan gebruiken om een model te bepalen. Het construeren van features is niet nieuw: bijvoorbeeld FKG's en DKG's zijn (complexe) features die op basis van brondata geconstrueerd zijn om per verzekerde gerichte input aan het model te geven. Maar er zijn ook ontelbaar veel andere features denkbaar per verzekerde, zoals: aantal farmacierecepten, aantal pakketjes per ATC-code, of een binaire variabele per diagnose of per diagnosecluster.

Het is geenszins mogelijk om in de beperkte scope van dit onderzoek een uitputtende zoektocht naar de beste features uit te voeren. Voor dit onderzoek is daarom, na afstemming met de begeleidingscommissie, alleen een aantal veelbelovende features getoetst in modellen M4 en M5, namelijk:

- **Dx-groepen**. Een binaire variabele per Dx-groep die aangeeft of er voor een verzekerde een diagnose binnen de betreffende Dx-groep geregistreerd is in het voorgaande jaar. Dat zijn 198 extra variabelen.
- **FKG_DDD**. Een continue variabele per FKG die aantal DDD's (gesommeerd over alle geneesmiddelen binnen de betreffende FKG) van een verzekerde binnen de betreffende FKG geeft. Dat zijn 36 extra variabelen[13].
- **Ligdagen_aantal.** Een continue variabele met aantal ligdagen van een verzekerde in het voorgaande jaar. Dit is bepaald op basis van gemiddelden in NZa zorgproductprofielen[14] van de aangeleverde zorgproducten, omdat zorgactiviteiten niet beschikbaar zijn.
- **Operaties_aantal.** Een continue variabele met aantal operaties van een verzekerde in het voorgaande jaar. Dit is bepaald op basis van gemiddelden in NZa normprofielen van de aangeleverde zorgproducten, omdat zorgactiviteiten niet voor ons beschikbaar waren.
- **Diagnoses_aantal.** Aantal unieke diagnoses van een verzekerde in het voorgaande jaar.
- **Specialismen_aantal.** Aantal unieke specialismen dat een verzekerde bezocht heeft in het voorgaande jaar.

[13] Hypertensie behandelen we los van diabetes, dus in plaats van een FKG diabetes type I met hypertensie en FKG diabetes type II met hypertensie hanteren we diabetes type I, diabetes type II en hypertensie als drie categorieën. Daardoor kan ook effect van hypertensie los meegenomen worden, aangezien hypertensie een groter risico op hartfalen, nierfunctiestoornissen en CVA met zich meebrengt.
[14] Op basis van de versie gepubliceerd op 17 november 2016

- **Zorgproducten_aantal.** Aantal geopende zorgproducten van een verzekerde in het voorgaande jaar.

Uiteindelijk zijn model M4 en M5 vormgegeven met de features zoals weergegeven in Tabel 2. Door de lange rekentijd is voor M4 geen variant meer doorgerekend inclusief diagnoses_aantal, specialismen_aantal en zorgproducten_aantal. Leeftijd in jaren was wel beschikbaar voor model M5 maar bleek tijdens het trainen niet te leiden tot een beter model.

Tabel 2: Overzicht van features in modellen M4 en M5.

	M4 – Gradient Boosting Machine	M5 – Artificial Neural Network
OT-data	X	X
Leeftijd in jaren	X	
Dx-groepen	X	X
FKG_DDD	X	X
Ligdagen_aantal	X	X
Operaties_aantal	X	X
Diagnoses_aantal		X
Specialismen_aantal		X
Zorgproducten_aantal		X

3.3 Voorkomen van overfitting

3.3.1 Opdeling van de data in trainingsset en testset

Binnen machine learning is het gebruikelijk een algoritme *out-of-sample* te beoordelen, dat wil zeggen op een andere dataset dan waarop het algoritme is getraind. Deze manier van werken zorgt ervoor dat we niet een model dat 'overfitted' is onterecht als goed classificeren. Bij 'overfitting' modelleert het algoritme algoritme in feite random variatie in plaats van variatie die gedreven wordt door de inputvariabelen, waardoor het model een hoog voorspellend vermogen heeft op reeds bekende data, maar niet goed generaliseert naar onbekende data. Om overfitting te voorkomen is voor dit onderzoek 30% van de data (5.158.453 verzekerden[15]) willekeurig toegewezen aan een testset, de resterende 70% is onderdeel van een trainingsset.[16,17]

Alle algoritmes worden uitsluitend getraind met behulp van de trainingsset: met de hierin aanwezige data wordt het model vormgegeven, bijvoorbeeld door de normbedragen uit te rekenen of relevante splitsingscriteria in een beslisboom te bepalen. De werking van de modellen is pas aan het eind van het onderzoek, na definitieve vaststelling van alle modelparameters, eenmalig getest op de testset.

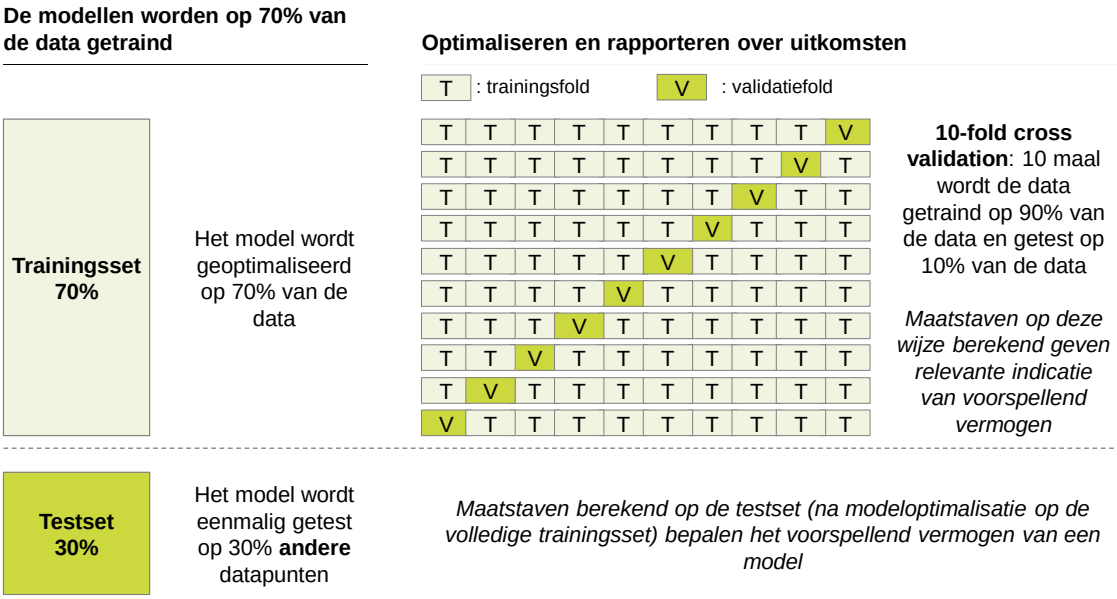
[15] Hierbij is rekening gehouden met verzekerden die door overstappen in het lopende jaren meerdere regels beslaan.
[16] Idealiter zouden we de scheiding tussen trainingsset en testset maken over de tijd, dus een dataset van het volgende jaar gebruiken als testset om daadwerkelijk het voorspellend vermogen van het model te evalueren.
[17] Voor een volgend onderzoek raden wij aan om 40% van de data apart te zetten als testset, omdat gebleken is dat het voorspellend vermogen van alle modellen op de testset iets beter is dan op de trainingsset

3.3.2 *Cross-validation*

Om ook binnen de trainingsset overfitting te voorkomen, maken we gebruik van 10-fold cross-validation. In deze methode wordt de trainingsset gesplitst in 10 willekeurige subsets of *folds*[18]. Steeds wordt het model getraind op 9 folds en getest op 1 fold. Dat wil zeggen dat 9 folds input zijn voor het vormgeven van het model, bijvoorbeeld het bepalen van normbedragen per risicoklasse of het opstellen van de keuzeboom. Voor de verzekerden in de resterende fold wordt het model vervolgens toegepast om de zorgkosten te modelleren. Doordat het model dit 10 keer uitvoert, waarbij steeds een andere fold als validatie wordt gebruikt, wordt de volledige trainingsset *out-of-sample* voorspeld[19].

De splitsing tussen trainingsset en testset en de werking van de 10-fold cross validation binnen de trainingsset is visueel weergegeven in Figuur 5.

Figuur 5: splitsing van data in een trainingsset en een testset en toepassing van 10 fold cross validation om out-of-sample maatstaven te bepalen.



3.4 Kenmerken van de trainingsset en testset

De testset bestaat uit 30% willekeurig geselecteerde verzekerden. De trainingsset bevat de overige 70% van de verzekerden. Tabel 3 geeft de beschrijvende statistieken van de volledige OT-dataset, trainingsset en testset weer[20].

De gemiddelde zorgkosten liggen 5 euro (0,2%) hoger in de testset dan in de volledige set. Ook de standaarddeviatie van de zorgkosten is met 51 euro (0,6%) iets hoger in de testset dan in de volledige set. Leeftijd, geslacht en het aantal verzekerden met een FKG, pDKG, sDKG, FDG, HKG, MHK of MVV zijn zeer vergelijkbaar verdeeld over de trainingsset en testset, met een maximale afwijking van 0,04-procentpunt (percentage man).

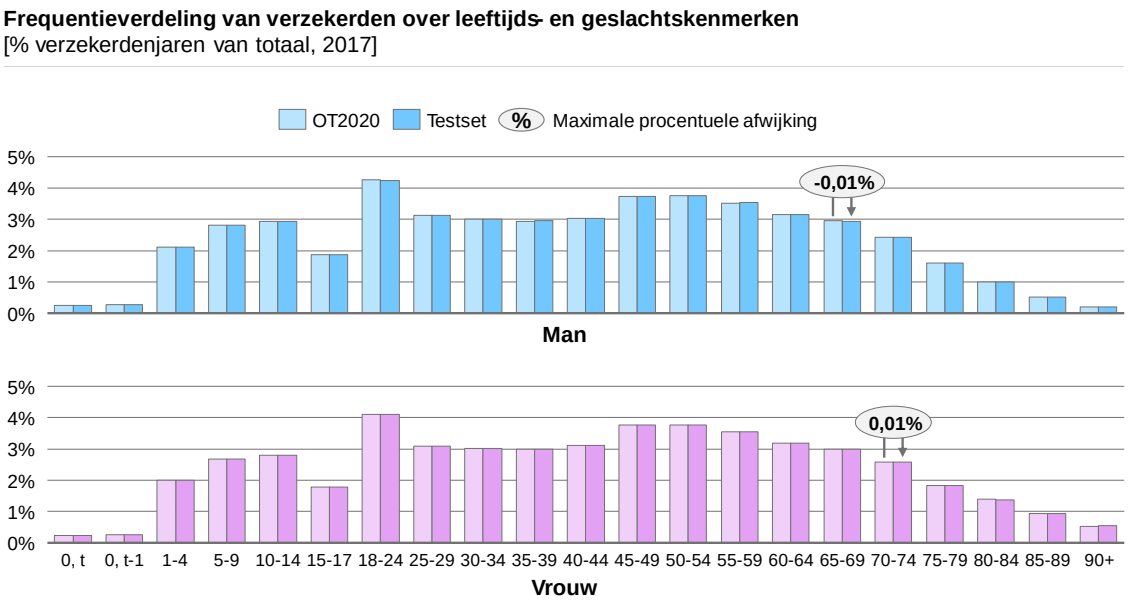
[18] Voor ieder model worden dezelfde folds gehanteerd.
[19] Vanwege beperkte rekencapaciteit worden modelparameters eerst op een kleine set getest middels hold-out validatie. Pas wanneer dit een kansrijke combinatie van modelparameters oplevert wordt de 10-fold cross-validation toegepast.
[20] Uiteraard zijn deze statistieken pas bepaald na definitieve vaststelling van de modellen

Tabel 3: Beschrijvende statistiek van de volledige set (OT2020), de trainingsset en de testset.

	OT2020	trainingsset	testset
Aantal verzekerden	17.194.536	12.036.083 (70%)	5.158.453 (30%)
Aantal verzekerdenjaren	16.839.948	11.787.901	5.052.047
Gemiddelde zorgkosten	€ 2.355	€ 2.352	€ 2.360
Standaarddeviatie zorgkosten	€ 8.525	€ 8.503	€ 8.576
% man	49,45%	49,47%	49,43%
Mediaan van leeftijdsgroep	40-44	40-44	40-44
% met FKG > 0	16,71%	16,70%	16,71%
% met meer dan 1 FKG	3,73%	3,73%	3,72%
% met pDKG > 0	8,88%	8,88%	8,88%
% met sDKG > 0	3,90%	3,90%	3,89%
% met FDG > 0	1,93%	1,93%	1,93%
% met HKG > 0	3,75%	3,75%	3,74%
% met MHK > 0	46,06%	46,06%	46,06%
% met MVV > 0	2,38%	2,38%	2,37%
% met FKG, pDKG, sDKG, FDG, HKG, MHK of MVV > 0	97,69%	97,69%	97,68%

De verdeling over de verschillende leeftijd-geslachtskenmerken in de volledige set en de testset is weergegeven in Figuur 6. De grootste absolute afwijking voor mannen bedraagt 0,01-procentpunt, binnen de leeftijdsgroep 65-69. Voor vrouwen is de leeftijdsgroep 70-74 het minst gelijk verdeeld, met eveneens een verschil van 0,01-procentpunt.

Figuur 6: frequentieverdeling van verzekerdenjaren over leeftijds- en geslachtskenmerken.



Voor ieder van de vereveningscriteria is onderzocht of deze significant anders verdeeld is binnen de volledige set en de testset. Voor alle vereveningscriteria, uitgezonderd FKG's, is hiervoor gebruik gemaakt van de Chi-Square significantietest. Deze test onderzoekt of twee datasets een significant andere distributie van risicoklassen hebben zonder aannames te doen over de onderliggende verdeling. De analyse van de distributie van de FKG's vraagt om een andere benadering. De onderliggende risicoklassen zijn, in tegenstelling tot de overige vereveningscriteria, onafhankelijk

van elkaar: een verzekerde kan meerdere risicoklassen hebben[21]. Voor iedere risicoklasse is een binomiale test uitgevoerd. Deze test vergelijkt de kans op een specifieke FKG in de volledige set met de werkelijk voorkomende frequentie in de testset.

Voor geen van de onderzochte vereveningscriteria zien we statistisch significante afwijkingen tussen beide datasets, behalve voor FKG 37 (extreem hoge kosten cluster 3), welke in de testset vaker voorkomt dan verwacht. FKG 37 heeft een prevalentie van <0,001% en is significant afwijkend gedistribueerd met een p-waarde van 0,005.

Samenvattend concluderen we dat de testset en trainingsset zeer goed overeenkomen.

[21] Op 12 uitzonderingen na zijn alle FKG's onafhankelijk. Deze uitzonderingen zijn beschreven in het FKG ATC Referentiebestand FKG's 2020, gepubliceerd door het Zorginstituut Nederland. Deze uitzonderingen zijn niet meegenomen in de statische vergelijking.

4 Resultaten

Dit hoofdstuk beschrijft de voorspellende waarde en verevenende werking van de vijf onderzochte modellen. Paragraaf 4.1 verklaart de werking van de algoritmes en de geselecteerde hyperparameters in meer detail. Paragraaf 4.2 beschrijft de uitkomst van de modellen op de binnen de risicoverevening gebruikelijke maatstaven.

4.1 Beschrijving modellen

In deze paragraaf beschrijven we de werking van de getoetste machine learning algoritmes. Ieder van deze machine learning algoritmes maakt gebruik van verschillende *hyperparameters* die de werking van het algoritme beïnvloeden. Een hyperparameter geeft een instelling van het algoritme weer, zoals het aantal *hidden layers* van een artificial neural network, de diepte van bomen in een random forest of het minimaal aantal datapunten per cluster in een decision tree.

De combinatie van hyperparameters met de hoogste gevonden 10-fold cross-validated R^2 bepaalt de uiteindelijke modelinstellingen. Met deze geselecteerde combinatie van hyperparameters wordt het algoritme vervolgens nog eenmaal getraind op de volledige trainingsset. Het getrainde model wordt tenslotte gebruikt om voor de verzekerden in de testset de zorgkosten te modelleren.

In dit onderzoek worden de modellen geoptimaliseerd naar R^2. Dit is binnen de risicoverevening de meest gangbare maat (zie bijvoorbeeld WOR 973 [5] en Kan (2019) [15]). Het minimaliseren van een kwadratische fout, zoals gebeurt bij R^2, betekent dat de voorspelling van een model een benadering is van de gemiddelde werkelijke kosten in plaats van de mediaan van de kosten (zoals bij optimalisatie naar GGAA het geval is). Dit is passend voor een risicovereveningsmodel waarin verzekeraars gemiddeld goed gecompenseerd moeten worden voor het risicoprofiel van hun verzekerdenpool en waarbij het resultaat relatief stabiel blijft bij overstapbewegingen.

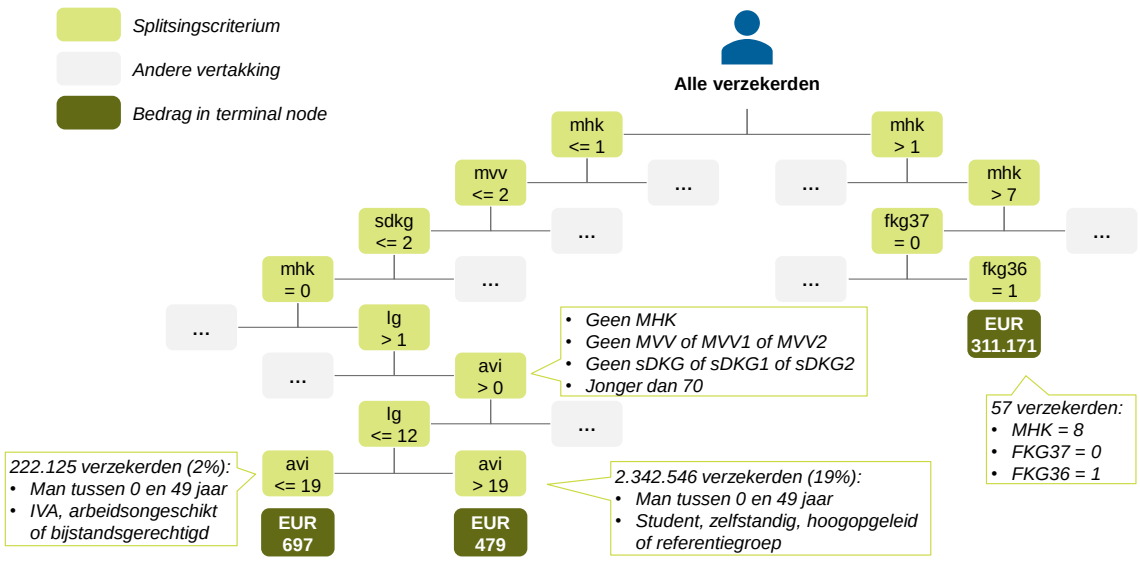
4.1.1 M1 – Decision Tree

Figuur 7 geeft een vereenvoudigde weergave van de decision tree op de trainingsset. Middels de risicokenmerken beschikbaar in de OT-dataset heeft het CART decision tree algoritme (Breiman et al., 1984 [16]) de volledige dataset opgedeeld in kleine subsets.

Het algoritme maakt gebruik van de trainingsset om het model vorm te geven. Bij iedere splitsing berekent het algoritme welk criterium en welke splitsingswaarde de variatie binnen de betreffende subset het meest verlaagt. Dit betekent dat de risicokenmerken die als eerst worden gebruikt de hoogste voorspellende waarde hebben. Wanneer de variatie onvoldoende afneemt, de subdataset te klein is of het maximale aantal sequentiële splitsingen is bereikt wordt een set niet verder gesplitst: op deze *terminal node* berekent het algoritme het gemiddelde van de zorgkosten van alle bijbehorende verzekerden. Met de uiteindelijk resulterende boom plaatst het algoritme iedere te verevenen verzekerde in één van de terminal nodes. De bijbehorende waarde geeft de gemodelleerde zorgkosten van deze verzekerde.

De in Figuur 7 weergegeven boom bevat 872 unieke combinaties van risicokenmerken; de langste vertakking van deze boom maakt gebruik van 23 risicokenmerken (dat zijn 9 unieke risicokenmerken, oftewel allen behalve AVI, regio en FDG), de kortste van slechts 4; de grootste eindgroep bevat 19% van de verzekerden, de kleinste groepen slechts 9 verzekerden.

Figuur 7: vereenvoudigde weergave van een Decision Tree; weergegeven zijn drie van de 872 terminal nodes[22]



Bij het trainen van het model is gebruik gemaakt van een gecomprimeerde dataset: alle verzekerden met dezelfde unieke combinatie van risicoklassen zijn samengevoegd tot één verzekerdenset met gewogen gemiddelde zorgkosten. Hierdoor neemt de rekentijd van het model sterk af. Bijlage B beschrijft alle instellingen van het algoritme waarmee deze boom tot stand is gekomen; de belangrijkste hiervan zijn:

- Minsplit = 60. Het algoritme onderzoekt alleen of een splitsing wenselijk is als de te splitsen subset minimaal 60 unieke verzekerdensets bevat.
- Minbucket = 9. Het algoritme plaatst minimaal 9 verzekerdensets in iedere *terminal node* om het vereveningsbedrag te berekenen.
- CP (complexity parameter) = 0.000015. Het algoritme splitst een subset wanneer de incrementele verbetering in R^2 groter is dan de complexity parameter. Potentiële splitsingen die onvoldoende verbetering opleveren worden niet uitgevoerd.
- MaxDepth = 30. Het algoritme onderzoekt maximaal 30 opeenvolgende splitsingen.

Tot slot negeert het model alle combinaties van risicoklassen die in totaal slechts één dag verzekerd waren in het voorspellingsjaar. Deze verzekerden zorgen veelal voor grote uitschieters, doordat de zorgkosten per verzekerde geschaald worden naar een volledig jaar. Uiteraard wordt deze bewerking *niet* uitgevoerd op de validatie- en testset. Deze procedure heeft als voordeel dat geen van de *terminal nodes* een uitzonderlijk hoog vereveningsbedrag krijgen door één uitschieter in de trainingsdata. Een decision tree is relatief gevoelig voor uitschieters, waardoor het voorspellend vermogen verbetert met de uitsluitprocedure. Een vergelijkbare procedure leek voor de Random Forest overigens niet te leiden tot een beter voorspellend vermogen, al is dit niet uitputtend getest en valt niet uit te sluiten dat het alsnog tot enige verbetering zou kunnen leiden[23].

[22] Nota bene: de weergegeven aantallen slechts binnen de trainingsset zijn, oftewel op 70% van de data
[23] Voor zowel de Decision Tree als Random Forest is eveneens een andere procedure getoetst: het plaatsen van een bovengrens op de mee te nemen zorgkosten uit de trainingsset. Deze bovengrens is voor iedere leeftijds-geslachtscategorie 90 standaarddeviaties boven de gemiddelde zorgkosten. Voor verzekerden met

4.1.2 M2 – Piecewise Regressie

Het Piecewise Regressie algoritme breekt de OT-dataset op in segmenten op basis van de leeftijd van verzekerden. Voor ieder van deze segmenten rekent het algoritme apart de normbedragen behorende bij de verschillende risicoklassen uit. Dit resulteert in 7 leeftijdsegmenten met ieder eigen normbedragen.

Het algoritme onderzoekt welke splitsingen op de variabele leeftijd de meeste interactie vertonen met zorgkosten en de andere risicokenmerken. Hiertoe gebruikt het een multivariate-analyse die sterk lijkt op het huidige OLS-model[24]. Iedere mogelijke leeftijdssplitsing wordt onderzocht; de splitsing die het meest gunstige effect laat zien o.b.v. het OLS-model wordt daadwerkelijk toegepast. We voorkomen overfitting door een minimaal aantal verzekerden per splitsing af te dwingen. Dit heeft echter wel invloed op bijvoorbeeld de 0-jarigen, die niet als aparte groep geselecteerd worden door het model. Dit model kan op verschillende manier verbeterd worden in verevenende werking en uitlegbaarheid. De verevenende werking zou verbeterd kunnen worden door een lasso-regressie uit te voeren per segment. Dit geeft een lineair model met normbedragen maar hanteert een penalty op de hoogte van de normbedragen om overfitting te voorkomen. De uitlegbaarheid van de resulterende normbedragen kan verbeterd worden door een regulier OLS te draaien op de gevonden leeftijdssegmenten. Dit laatste voeren we uit in hoofdstuk 5 om de uitkomsten te interpreteren voor toepassing in het huidige OLS.

Dit algoritme maakt gebruik van twee hyperparameters die afwijken van standaardwaarden. Zie Bijlage C voor de overige hyperparameters:

- % fractie = 1%. Het model berekent de best passende leeftijdssplitsing op basis van 1% van de trainingsset. Dit komt overeen met ongeveer 100.000 verzekerden tijdens de 10-fold cross-validation procedure. Grotere fracties doen de benodigde rekenkracht enorm toenemen, terwijl door ons uitgevoerde analyses met fracties van 2,5% en 5% laten zien dat de kwaliteit van het uiteindelijke model niet verbetert.
- Minsize = 10.000. Iedere subset moet minimaal 20.000 (2 x minsize) verzekerdenjaren bevatten. Kleinere sets worden niet verder gesplitst.

4.1.3 M3 – Random Forest

Het Random Forest algoritme gebruikt de trainingsdata om een groot aantal onafhankelijke keuzebomen te modelleren. Iedere keuzeboom maakt een willekeurige selectie mét terugleg van de verzekerden uit de trainingsset – dusdanig dat het aantal geselecteerde verzekerden overeenkomt met de omvang van de trainingsset. Het algoritme kan daarbij een verzekerde meerdere keren selecteren voor dezelfde boom, waardoor de bomen op ongelijke en onafhankelijke datasets worden getraind. Ook maakt het algoritme een willekeurige selectie van 20 risicokenmerken per split. Gegeven deze selecties stelt het algoritme een zo goed mogelijke keuzeboom op (zie paragraaf 4.1.1 voor de werking van een Decision Tree). Figuur 8 visualiseert de werking van dit algoritme.

---

hogere zorgkosten worden de kosten gelijkgesteld aan de bovengrens. Het totaal afgekapte bedrag wordt vervolgens uniform verdeeld over alle verzekerden binnen dezelfde leeftijdsgroep in de trainingsset. Voor de decision tree werkte deze methodiek minder goed en voor de Random Forest bleek de methode niet te leiden tot een beter voorspellend vermogen.

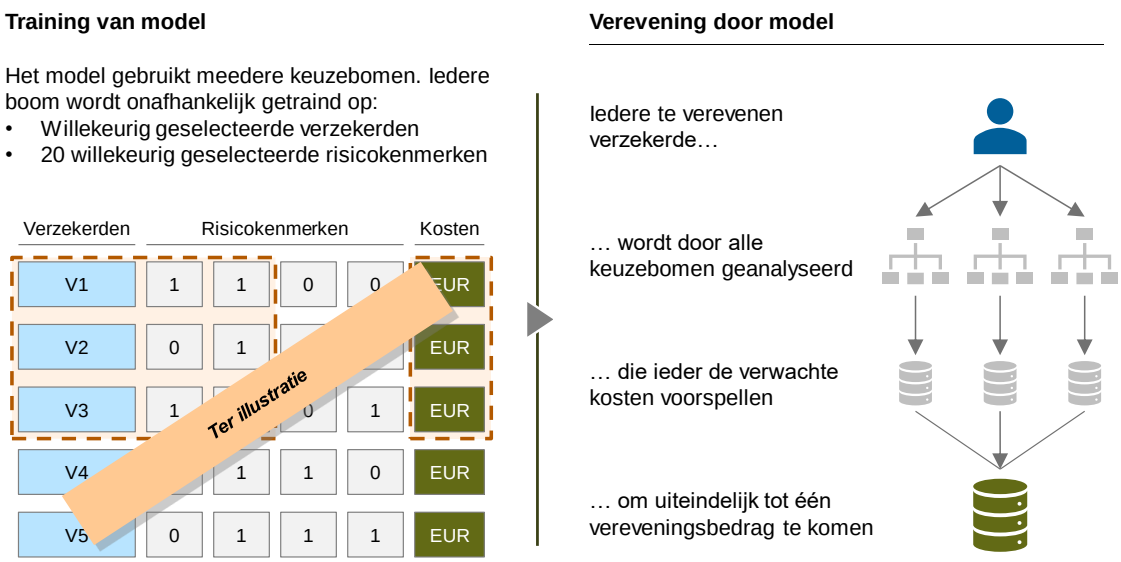
[24] Met drie verschillen: 1) leeftijd is een continue variabele met pér leeftijdssegment een lineair verband tussen leeftijd en zorgkosten; 2) geslacht is een aparte binaire variabele; 3) leeftijd is niet langer onderdeel van avi, ses en ppa: zij hebben nu respectievelijk 7, 4 en 4 risicoklassen.

Het algoritme bepaalt voor alle verzekerden op basis van elk van de 200 keuzebomen de gemiddelde gemodelleerde zorgkosten. Het gemiddelde van deze 200 voorspellingen vormt vervolgens de definitieve voorspelling voor de betreffende verzekerde.

Alle hyperparameters van dit algoritme zijn beschreven in Bijlage D. De belangrijkste instellingen hiervan zijn:

- Sample.fraction = 1. Het totaal aantal willekeurige verzekerden per boom is altijd gelijk aan het aantal verzekerden in de trainingsset.
- Replace = True. Elke keuzeboom gebruikt de gegevens van willekeurig geselecteerde verzekerden uit de trainingsset mét terugleg – hierdoor kan een verzekerde meerdere keren worden geselecteerd voor dezelfde boom.
- Mtry = 20. Iedere split gebruikt 20 willekeurig geselecteerde risicokenmerken van de betreffende verzekerden om het algoritme vorm te geven.
- Numtrees = 200. Het algoritme traint 200 onafhankelijke keuzebomen die ieder de zorgkosten van alle verzekerden modelleren.
- Min.node.size = 30. Iedere splitsing bevat minimaal 30 verzekerden.
- Max.depth = NULL. Er staat geen limiet op de diepte die iedere boom kan behalen.

Figuur 8: het Random Forest algoritme maakt 200 onafhankelijke keuzebomen – elk van deze bomen overweegt 20 risicokenmerken per splitsing en een willekeurige combinatie van de verzekerden uit de trainingsset



4.1.4 M4 – Gradient Boosting Machine

Ook de Gradient Boosting Machine maakt gebruik van eenvoudige keuzebomen om de zorgkosten van verzekerden te modelleren. In plaats van deze bomen parallel toe te passen, zoals het Random Forest algoritme, worden de bomen in serie geplaatst. Dit is weergegeven in Figuur 9.

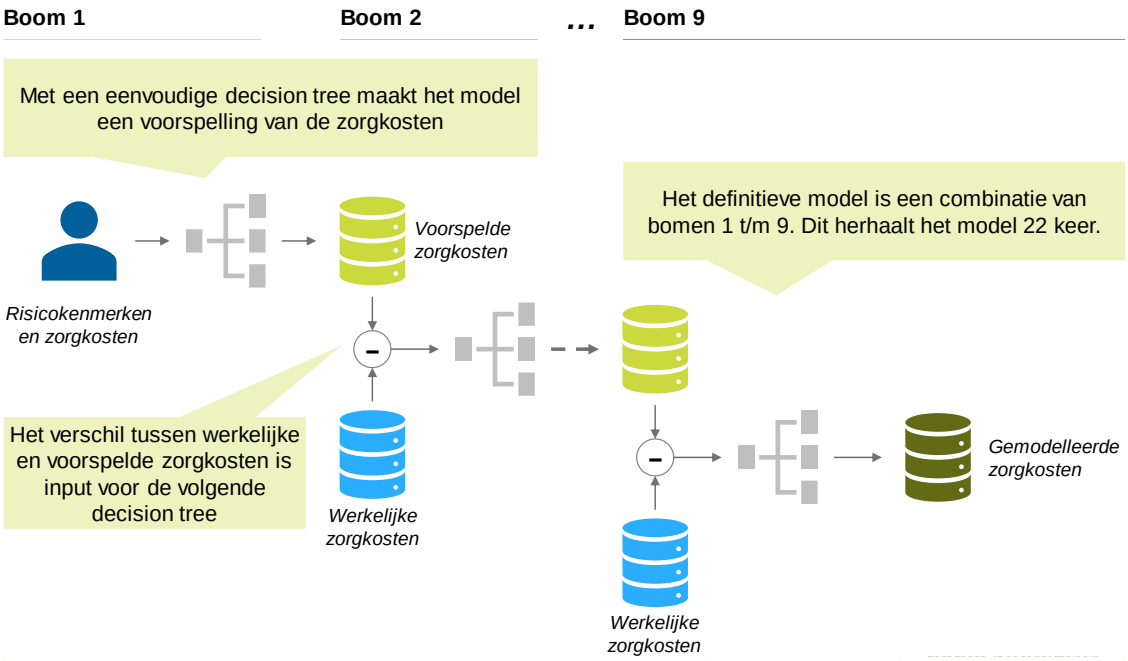
De eerste boom maakt een modellering op basis van de werkelijke zorgkosten in de trainingsset. De tweede boom vergelijkt de gemodelleerde kosten met de werkelijke kosten – het verschil hiertussen wordt door deze boom gebruikt als input om de splitsingen te bepalen. Dit herhaalt zich

in totaal 9 keer; de combinatie van deze bomen achter elkaar vormt een voorspellend model van de zorgkosten.

Het model herhaalt deze procedure in totaal 22 keer. Deze 22 combinaties van 9 bomen vormen gezamenlijk het definitieve model. Ieder van deze combinatie van bomen is onafhankelijk van de anderen – hiermee verkleint het model de kans op overfitting.

Om het risico op overfitting verder te verkleinen worden alle unieke kenmerk-combinaties die exact één dag in het bestand voorkomen verwijderd. Deze procedure is ook toegepast op de losse Decision Tree (zie paragraaf 4.1.1).

Figuur 9: conceptweergave van de geïmplementeerde Gradient Boosting Machine.



Een Gradient Boosting algoritme heeft veel hyperparameters om te optimaliseren. Hieronder volgen de belangrijkste; in Bijlage E is een gedetailleerde lijst opgenomen.

- Nrounds = 9. Het algoritme maakt gebruik van 9 sequentiële keuzebomen, die gezamenlijk leiden tot een modellering van de zorgkosten.
- Number_of_tracks = 22. Het model herhaalt bovenstaande procedure 22 keer – iedere procedure is volledig onafhankelijk van de overige procedures. De gemiddelde voorspelling van alle procedures vormt de definitieve modellering van de zorgkosten.
- Subsample = 0.632. Iedere keuzeboom maakt gebruik van 63,2% van de beschikbare datapunten om de splitsingen op te baseren.
- Min_child_weight = 32. Een subset wordt alleen verder gesplitst wanneer deze gegevens van minimaal 32 verzekerden bevat.[25]

[25] De hyperparameter min_child_weight is afhankelijk van het nummer van de procedure (1 t/m 22). De daadwerkelijk gebruikte waarde is gelijk aan 32 * procedure nummer / 22

4.1.5 M5 – Artificial Neural Network

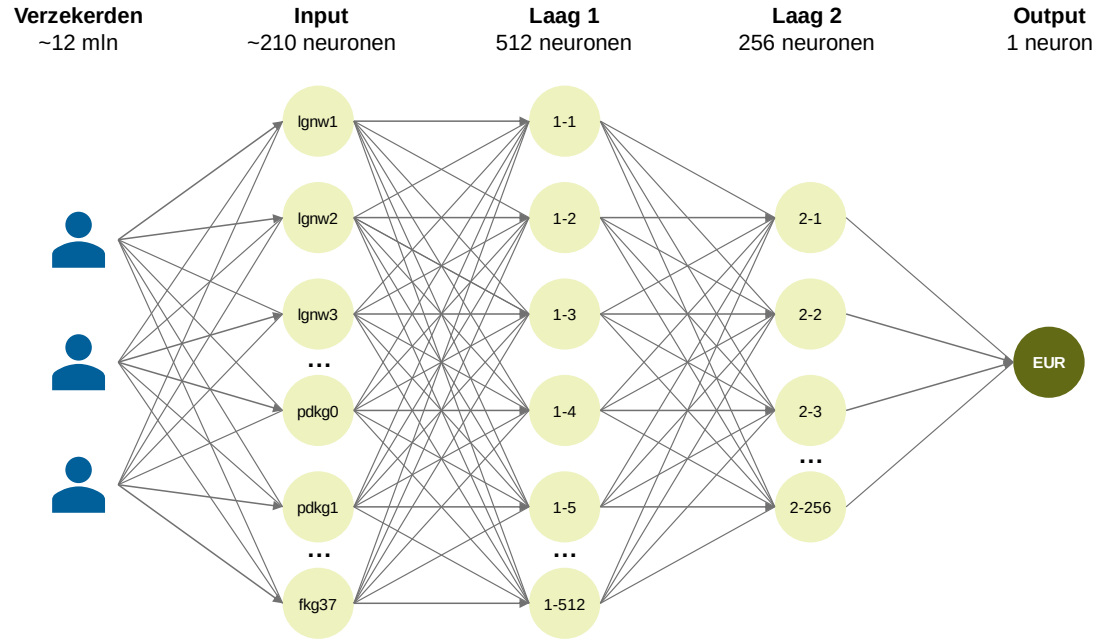
Neurale netwerken kunnen in hun uitwerking zeer complex lijken, maar zijn net als bijvoorbeeld OLS in feite terug te voeren tot één (zeer lange en niet-lineaire) formule. Ter illustratie is in Bijlage F een eenvoudig voorbeeld van de werking van een Artificial Neural Network nader uitgewerkt.

Het voor dit onderzoek geïmplementeerde neurale netwerk om de zorgkosten te modeleren is schematisch weergegeven in Figuur 10.

De *input layer* bevat van alle verzekerden de gegevens per risicokenmerk uit de trainingsset. Deze gegevens worden doorgestuurd naar elk van de 512 neuronen uit de eerste laag. Elk neuron maakt een gewogen optelsom van de activaties van de neuronen in een laag eerder. Vervolgens wordt hierop een niet-lineaire activatiefunctie toegepast (de 'ReLU') om zo de activatie van dat enkele neuron te berekenen. Uitzonderingen hierop zijn de inputneuronen en het outputneuron. De activatie van de inputneuronen is gewoon gelijk aan de waarde van de betreffende feature, en het outputneuron combineert alle signalen uit deze tweede hidden layer tot één voorspelling van zorgkosten en heeft dus geen activatiefunctie.

Het algoritme zoekt uiteindelijk naar de beste combinaties van factoren waarmee de neuronen de betreffende inputsignalen wegen. Vertaald naar onderstaande figuur betekent dit dat elke 'pijl' een weegfactor krijgt. Ieder neuron heeft zelf ook een 'bias', een getal dat de gemiddelde activatie van het neuron over de gehele trainingsset beïnvloedt. De 'bias' is daarmee vergelijkbaar met *intercept* in een lineair model.

Figuur 10: het geïmplementeerde Artificial Neural Network maakt gebruik van twee hidden layers met 512 en 256 neuronen.



De *weegfactoren* en *biases* worden geoptimaliseerd middels *batches* – combinaties van datapunten – van 2,048 verzekerden.[26, 27] Uiteindelijk wordt ieder datapunt uit de trainingsset 45 keer als input aan het model gegeven (dit is het aantal *epochs*). Bij iedere herhaling van batches

[26] De versie van dit model op OT-data (M5a) hanteert 1.024 batches
[27] Voor een volledig overzicht van verschillen tussen modellen M5 en M5a, zie bijlage G

en epochs optimaliseert het model de weegfactoren verder. In deze optimalisatieprocedure maakt het algoritme gebruik van de *drop-out* functie: bij iedere iteratie wordt willekeurig 35% van de neuronen uit de hidden layers uitgeschakeld. Dit dwingt het algoritme om de weegfactoren zo veel mogelijk onafhankelijk te trainen en beperkt de kans op overfitting.

De volledige lijst met hyperparameters van het Artificial Neural Network is in Bijlage G opgenomen. De belangrijkste van deze parameters zijn:

- Aantal *hidden layers* = 2; aantal neuronen per layer = 512 en 256, respectievelijk. Het aantal lagen en aantal neuronen is gerelateerd aan de complexiteit van de relatie tussen de risicokenmerken en de te modelleren zorgkosten.
- *Dropout rate* = 0.35. Iedere iteratie van de training wordt 35% van de neuronen uit de hidden layers op non-actief geplaatst. Tijdens het voorspellen zijn overigens altijd alle neuronen actief (maar de activaties van neuronen uit hidden layers worden dan met 0.65 vermenigvuldigd).
- *Batchsize* = 2.048, aantal *epochs* = 45 en *shuffle* = true. Het algoritme past de weegfactoren aan nadat het de voorspelling op 2.048 verzekerden heeft geëvalueerd. Het algoritme gebruikt elk datapunt 45 keer om te trainen. Na iedere epoch wordt de data willekeurig verdeeld over nieuwe batches.

4.2 Voorspellend vermogen en verevenende werking

Deze paragraaf beschrijft het voorspellend vermogen van de modellen en de verevenende werking van de modellen.

4.2.1 Voorspellend vermogen van modellen

Tabel 4 bevat het voorspellend vermogen van de verschillende modellen: de R^2 op de testset. Voor de vergelijkbaarheid wordt de R^2 ook gerapporteerd voor M0 op de testset van de exact zelfde datasets als gebruikt voor de modellen M1-M5. Modellen M4 en M5 zijn vooral geoptimaliseerd voor dataniveau d[28]. Het is niet uitgesloten dat de gerapporteerde uitkomsten op de OT-dataset (niveau a) nog zouden kunnen verbeteren door verdere optimalisatie van de hyperparameters.

Tabel 4: voorspellend vermogen van de zes modellen (R^2)

	M0 OLS	M1 Decision Tree	M2 Piecewise lineaire regressie	M3 Random Forest	M4 Gradient Boosting Machine	M5 Artificial Neural Network
OT-dataset (dataniveau a)	35,1%	35,0%		36,3%	36,3%	36,3%
OT met leeftijd continu (dataniveau b)	35,1%		35,3%	36,3%		
OT met brondata (dataniveau d1)	36,1%				38,5%	
OT met brondata (dataniveau d2)	36,2%					38,2%

Groene *waarden zijn beter dan M0;* Oranje *waarden zijn minder goed dan M0. NB: dataniveau d1 = data meegenomen in model M4, dataniveau d2 = data meegenomen in model M5)*

De Gradient Boosting Machine (M4) heeft het hoogste voorspellend vermogen gehaald met een R^2 van 38,5%. De OLS scoorde 2,4-procentpunt lager op dezelfde dataset. M3, M4 en M5 waren beter in staat de zorgkosten van verzekerden te voorspellen op basis van risicokenmerken in de OT-

[28] Voor volledig overzicht van de gebruikte data per model zie paragraaf 3.2

dataset dan OLS. Ze haalden allen een R^2 van 36,3% op de OT-dataset en deden het daarmee 1,2-procentpunt beter dan OLS. Ook Piecewise lineaire regressie (M2) heeft een hoger voorspellend vermogen dan OLS op dezelfde dataset met 0,2-procentpunt. Echter gaat deze relatief beperkte verbetering wel gepaard met bijna een verzesvoudiging van het aantal normbedragen doordat normbedragen per leeftijdssegment worden vastgesteld. De Decision Tree (M1) heeft een 0,1-procentpunt lager voorspellend vermogen dan OLS op de OT-dataset, maar hanteert hiervoor wel substantieel minder unieke combinaties van verzekerdenkenmerken.

Bijlage I geeft de R^2 van ieder van de modellen bepaald via 10-fold cross validation in de trainingsset en per fold in de trainingsset weer. Hoewel het eigenlijk niet wenselijk is om conclusies te trekken over individuele folds geeft het wel een beeld van de robuustheid van de resultaten van de modellen. Op 8 van de 10 folds behaalt de Gradient Boosting Machine de beste voorspelling. Het Artificial Neural Network algoritme voorspelt het best op de overige twee folds, maar ook op deze folds behaalt de Gradient Boosting Machine slechts 0,2-procentpunt lagere score.

4.2.2 Normalisatie van voorspelling

De som van de gemodelleerde zorgkosten van alle verzekerden dient gelijk te zijn aan de som van de werkelijke kosten[29], zodat er in een jaar niet meer uitgekeerd wordt dan het budget dat beschikbaar is. Het uitgangsmodel 2020, getraind op de volledige OT-dataset en gebruikmakend van OLS-regressie voldoet hier altijd aan. Voor de modellen M0 t/m M5 geldt dit niet. Dit heeft twee oorzaken:

1. Modellen M1, M3, M4 en M5 modelleren in opzet niet naar de gemiddelde zorgkosten. Hierdoor kan de som van de gemodelleerde kosten afwijken van de som van de werkelijke kosten.
2. Alle modellen M0 t/m M5 voorspellen de zorgkosten met behulp van de trainingsset – een ongelijke distributie in risicokenmerken en werkelijke zorgkosten tussen de trainingsset en de testset kan leiden tot een afwijking van de gemodelleerde uitgaven ten opzichte van het totale beschikbare budget.

Dit is een bekend probleem dat door onder andere Ismail (2018) is ervaren [1]. Om aansluiting te vinden bij het macrobudget zijn per model de gemodelleerde zorgkosten genormaliseerd naar de werkelijke zorgkosten middels een correctiefactor. Voor de modellen M0, M1, M2, M3, M4a, M5a, M4d en M5d zijn deze correctiefactoren respectievelijk 1,0016, 1,0013, 1,0034, 0,9965, 1,0118 en 0,9980. De grootste correctiefactor is nodig voor M4, de Gradient Boosting Machine, waar zonder normaliseren de som van de gemodelleerde zorgkosten ruim 1% lager uitvallen dan de som van de werkelijke zorgkosten.

Het voorspellend vermogen weergegeven in tabel 4 en besproken in paragraaf 4.2.1 is gebaseerd op onaangepaste zorgkosten, omdat dit de beste weergave is van de voorspelkracht van het model. De gerapporteerde maatstaven in de volgende paragraaf zijn wel berekend op basis van de genormaliseerde gemodelleerde zorgkosten per individu.

4.2.3 Verevenende werking van modellen

De verevenende werking van alle modellen wordt beoordeeld middels maatstaven op vier niveaus: normbedragen, individuen, subgroepen en verzekeraars. Tabel 5 geeft de maatstaven die berekend zijn op de testset. Het uitgangsmodel 2020 is getraind en getest op de volledige dataset en daarom niet representatief voor de vergelijking met de modellen M1-M5. De maatstaven in de

[29] Dat wil zeggen, werkelijke kosten in de OT dataset, dus werkelijke kosten 2017

kolom M0 – OLS zijn wel op dezelfde wijze in de testset bepaald: de normbedragen zijn bepaald op de trainingsset om vervolgens zorgkosten te modelleren in de testset.

Uit de berekende maatstaven van de vijf ontwikkelde modellen en het huidige model is een aantal conclusies te destilleren:

- **Model M1 - Decision Tree** scoort iets minder goed op nagenoeg alle maatstaven dan het huidige model. Op zowel individu-, subgroep-, als verzekeraarsniveau worden de maatstaven negatief beïnvloed.
- **Model M2 – Piecewise Regressie** behaalt een verbetering op R^2 ten opzichte van het huidige model. Hier staat echter tegenover dat er substantieel meer verzekerden zijn met een negatief normbedrag. Dit hoge aantal wordt veroorzaakt doordat het model een lineair verband veronderstelt tussen leeftijd in jaren en zorgkosten. Voor het leeftijdssegment 0-8 jaar geldt echter dat verzekerden geboren in het vereveningsjaar significant duurder zijn, waardoor dit verband niet bestaat. De gemodelleerde zorgkosten van een deel van de 8-jarigen worden hierdoor negatief. Ook op subgroep- en verzekeraarsniveau is een verslechtering waarneembaar op het merendeel van de maatstaven. Overigens zijn de uitkomsten op de maatstaven beter als dit model in beperkt aangepaste variant wordt toegepast, door een regulier OLS uit te voeren op de gevonden leeftijdssegmenten; zie voor meer toelichting paragraaf 5.3 en Bijlage L.
- **Model M3 – Random Forest** scoort op alle individuele maatstaven een beter resultaat dan het huidige model. Op dezelfde data laat het model ten opzichte van OLS een verbetering in R^2 zien van 35,1% naar 36,3%. Op subgroep- en verzekeraarsniveau laten verschillende maatstaven zowel een verbetering als een verslechtering zien. Het toevoegen van continue leeftijd heeft een beperkt positieve impact op de maatstaven voor dit model.
- **Model M4 – Gradient Boosting Machine** behaalt gunstige resultaten op bijna alle maatstaven. Op dezelfde data laat het model ten opzichte van OLS een verbetering in R^2 zien van 35,1% naar 36,3%. Bij het toevoegen van extra brondata blijkt de meerwaarde ten opzichte van OLS pas echt goed: de R^2 verbetert dan naar 38,5%, terwijl een OLS model op dezelfde data een R^2 van 36,1% realiseert. Met name de verbeteringen op individuele maatstaven vallen op en zijn de beste van alle modellen. Enkel de bandbreedte van het resultaat van middel en grote verzekeraars verslechtert licht.
- **Model M5 – Artificial Neural Network** behaalt gunstige resultaten op alle maatstaven. Op dezelfde data laat het model ten opzichte van OLS een verbetering in R^2 zien van 35,1% naar 36,3%. Net als bij de Gradient Boosting Machine blijkt de meerwaarde ten opzichte van OLS pas echt goed wanneer we extra brondata toevoegen: de R^2 verbetert dan naar 38,2%, terwijl een OLS model op dezelfde data een R^2 van 36,2% realiseert. Het complete getrainde model, dus inclusief aanvullende data, laat op zowel individu-, subgroep- als verzekeraarsniveau verbeteringen zien ten opzichte van een OLS model op dezelfde aanvullende dataset. Vooral de uitkomsten op verzekeraarsniveau zijn zeer gunstig en de beste van alle modellen. Alle bandbreedtes worden smaller. Dit vertaalt zich ook in de hoogste gewogen gemiddelde absolute resultaatverschuiving (GGARV).

Bijlage H vergelijkt alle modellen met een OLS op dezelfde dataset. Voor de modellen M3-M5, waarin extra brongegevens worden gebruik, geeft dit aanvullende informatie. Het blijkt dat M3-M5 op alle individuele maatstaven beter presteren dan een OLS model op dezelfde data. Bijvoorbeeld de R^2 van OLS op dezelfde data als M3 is 35,1% en voor dezelfde data als M4 en M5 is de R^2 36,1%. Samengevat zien we dat M3-M5 allen een verbetering zijn ten opzichte van OLS op dezelfde data.

Figuur 11 geeft het financieel resultaat per leeftijd (in jaren) van de modellen binnen de testset. Alle modellen volgen grotendeels dezelfde lijn – volatiele onder- of overcompensatie bij verzekerden jonger dan 5 of ouder dan 70, en een relatief stabiele en nauwkeurige voorspelling daartussen. De onder- en overcompensatie bij verzekerden jonger dan 40 is gemiddeld het laagst bij model M4, waar model M0 en M2 juist het nauwkeurigst voorspellen bij de verzekerden ouder dan 70 en 80, respectievelijk.

Figuur 12 geeft het financieel resultaat per risicodrager van de modellen binnen de testset. Voor M5 komt het financieel resultaat voor 16 van de 24 risicodragers dichter bij nul. Dat is consistent met de gerapporteerde ontwikkelingen op bandbreedte van het financieel resultaat. M4 doet het vergelijkbaar voor veel risicodragers, maar verkleint de bandbreedte minder omdat het financiële resultaat van de uitschieters minder sterk verbetert. M1, M2 en M3 vergroten juist de bandbreedte door een sterker negatief en positief financieel resultaat op de twee uiterste risicodragers.

Tabel 5: maatstaven van modellen op testset

Niveau	Maatstaf		Uitgangs-model 2020	M0 OLS	M1 Decision Tree	M2 Piecewise lineaire regressie	M3 Random Forest	M4 Gradient Boosting Machine	M5 Artificial Neural Network	M3a[b] Random Forest	M4a[b] Gradient Boosting Machine	M5a[b] Artificial Neural Network	M0d1[c] OLS op data van model M4
Individu	R^2 x 100%		34,4%	35,1%	35,0%	35,3%	36,3%	38,5%	38,2%	36,3%	36,3%	36,3%	36,1%
	CPM x 100%		33,5%	33,6%	33,6%	32,9%	34,2%	35,7%	35,7%	34,1%	34,0%	34,1%	34,1%
	GGAA		1.980	1.984	1.983	2.004	1.965	1.920	1.920	1.969	1.971	1.968	1.969
	Standaarddeviatie resultaten		6.906	6.909	6.915	6.898	6.843	6.726	6.741	6.845	6.843	6.843	6.854
	# met negatieve normkosten		15.054	4.412	-	62.137	-	58	544	-	19	-	4.390
Subgroepen	GGAA op alle subgroepen		1.011	1.132	1.124	1.176	1.091	1.047	1.058	1.080	1.141	1.099	1.114
	Res. 15% laagste kosten in t-3		107	112	171	140	144	107	111	159	159	150	108
	Res. 15% hoogste kosten in t-3		-124	-129	-147	-105	-131	-118	-90	-191	-124	-158	-149
Verzekeraar	R^2 x 100%		99,1%	98,9%	98,3%	98,6%	98,8%	99,0%	99,0%	98,8%	98,3%	98,9%	98,9%
	GGAA		26	29	38	31	30	26	21	31	32	30	28
		Allen	302	310	347	394	347	302	235	315	434	294	323
		Excl. 2	115	180	191	176	161	138	96	167	165	185	148
	Bandbreedte van resultaten	Klein	284	279	319	335	292	258	230	265	409	274	259
		Middel	154	157	192	179	167	160	138	172	168	160	178
		Groot	64	63	99	64	71	77	57	79	60	84	67
		Niet-concern	167	186	209	176	161	138	146	167	183	185	148
		Concern	302	310	347	394	347	302	235	315	434	294	323
	GGARV		8	-	14	3	4	12	19	6	6	6	4

Groene *waarden zijn beter dan M0;* Oranje *waarden zijn minder goed dan M0.*

[a] Als gerapporteerd in WOR 973 en gereproduceerd voor dit onderzoek; het kleine verschil op resultaat 15% laagste kosten t-3 mogelijk door afronding of zeer beperkte verschillen in data

[b] model doorgerekend op enkel OT-dataset

[c] Ter illustratie, OLS uitgevoerd op dataset van model M4. Voor vergelijking met andere OLS modellen op aanvullende datasets, zie Bijlage H

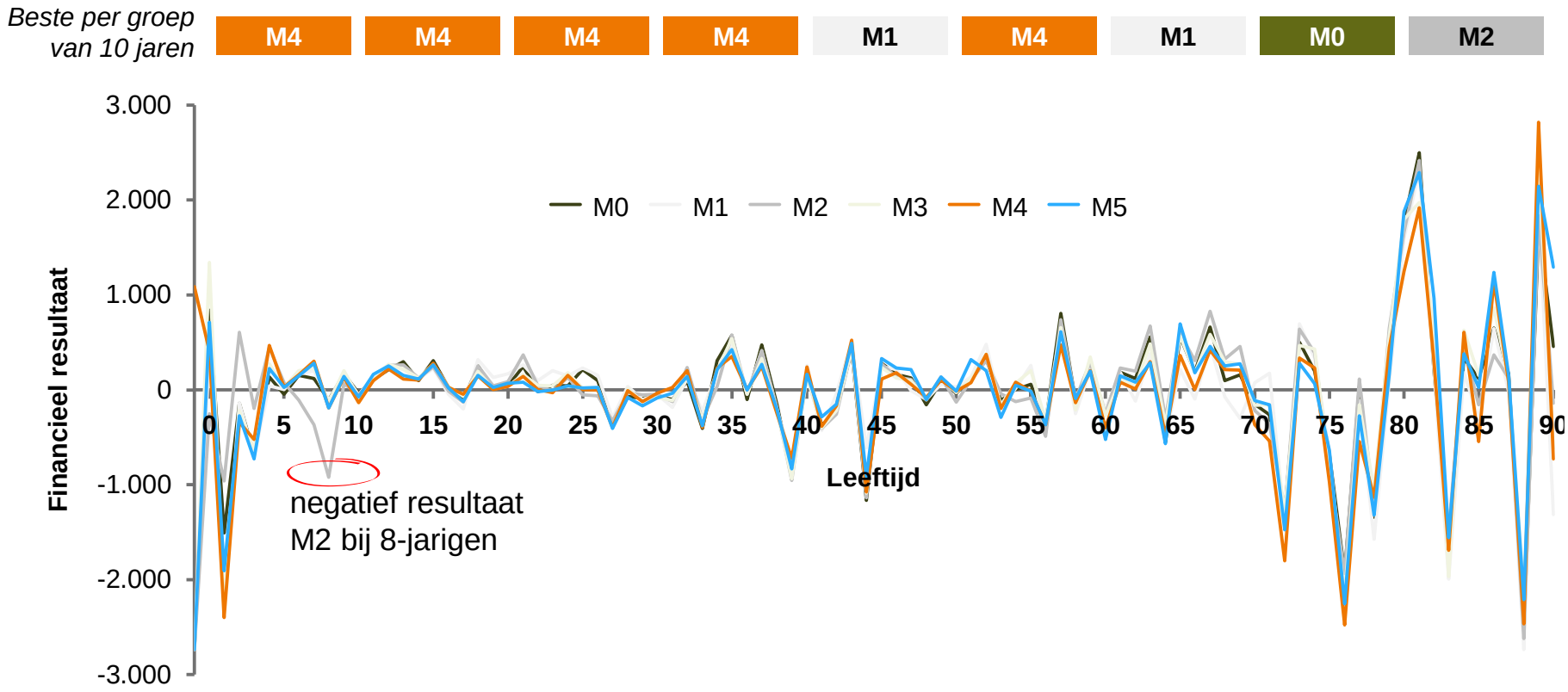
[d] groot verschil tussen uitgangsmodel en modellen M0-M5 wordt veroorzaakt doordat uitgangsmodel op volledige data (~17 miljoen verzekerdenjaren) bepaald is en overige modellen slechts op 30% van de set (testset)

[e] Per model zijn de subgroepen gedefinieerd op basis van de subgroepen uit het uitgangsmodel (1,85 miljoen subgroepen)

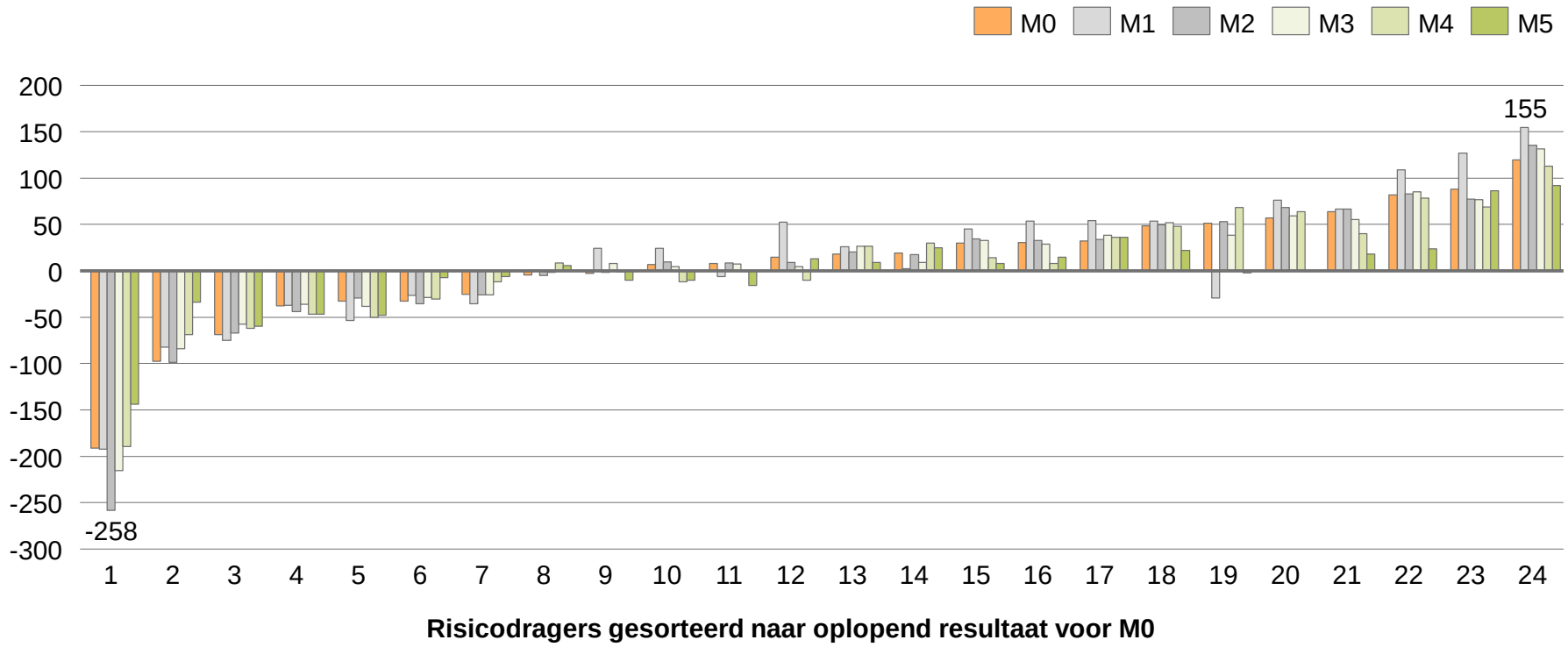
[f] Op deze regel staat de bandbreedte van de resultaten op verzekeraarsniveau waarbij de twee risicodragers die steeds de feitelijke bandbreedte bepalen buiten beschouwing zijn gelaten

[g] De GGARV is voor alle modellen, inclusief het uitgangsmodel 2020, vergeleken met M0 – OLS

Figuur 11: het financiële resultaat in euro's per verzekerdenjaar naar leeftijd in jaren



Figuur 12: het financiële resultaat in euro's per verzekerdenjaar voor 24 risicodragers.



5 Relevantie van uitkomsten voor het huidige vereveningsmodel

In dit hoofdstuk kijken we met een OLS-bril naar de uitkomsten van de machine learning algoritmes. Dat wil zeggen dat we de uitkomsten van de algoritmes interpreteren om te bepalen welke lering eruit getrokken kan worden voor het toevoegen of aanpassen van vereveningscriteria en risicoklassen. Daarmee kunnen de getoetste algoritmes in dit onderzoek als proeftuin voor OLS gebruikt worden in de onderzoeksfase.

Paragraaf 5.1 laat van alle modellen zien wat het relatieve belang is van de risicokenmerken binnen het betreffende model. Daarmee geeft het enerzijds een indicatie van de toegevoegde waarde van het vervangen van OLS door een machine learning algoritme, bijvoorbeeld omdat een model minder belang toekent aan niet-morbiditeitscriteria zoals MHK en MVV. Anderzijds geeft het een indicatie van de waarde voor de huidige OLS van nieuwe risicokenmerken die toegevoegd zijn aan de modellen. Paragrafen 5.2 en 5.3 gaan dieper in op de risicoklassen met een groot belang uit respectievelijk Decision Tree en Piecewise Regressie. Deze twee algoritmes zijn het meest vergelijkbaar met het huidige OLS en bieden handvatten om relevante nieuwe risicoklassen af te leiden voor OLS. Tenslotte vat paragraaf 5.4 samen welke risicokenmerken en riscioklassen verder onderzoek waard zijn binnen OLS.

5.1 Relatief belang van risicokenmerken

Tabel 6 geeft van ieder van de modellen weer wat de 10 belangrijkste risicoklassen zijn. Bijlage J bevat deze resultaten voor de aanvullende modellen op OT-data (M3a, M4a, M5a).

Tabel 6: 10 belangrijkste risicoklassen per model, genummerd van 1 (belangrijkste) tot 10

Risicoklasse	M0 OLS	M1 Decision Tree	M2 Piecewise lineaire regressie	M3 Random Forest	M4 Gradient Boosting Machine	M5 Artificial Neural Network
FKG	1	5	1	4	6	3
MVV	2	3	4	2	2	2
MHK	3	1	3	1	1	1
pDKG	4	4	5	3	3	5
Leeftijd-geslacht	5	6	9	9		4
sDKG	6	7	6	6	10	9
PPA	7	2	8	7	7	6
HKG	8	9	7	8		
FDG	9		10			
AVI	10	8		10		
SES		10				
Leeftijd in jaren	nvt	nvt	2	5	4	nvt
Aantal verpleegdagen	nvt	nvt	nvt	nvt	5	10
ddd's in FKG 37	nvt	nvt	nvt	nvt	8	7
ddd's in FKG 33	nvt	nvt	nvt	nvt	9	
Aantal zorgproducten	nvt	nvt	nvt	nvt	nvt	8

NB: wanneer in bovenstaande tabel een cel leeg is betekent dit dat het risicokenmerk niet voorkomt in de top 10. Nvt = niet van toepassing, want niet beschikbaar in de onderhavige dataset.

Hiertoe zijn per model alle risicoklassen gerangschikt van meest zwaarwegend naar minst zwaarwegende klasse. Om te kunnen rangschikken is *permutation feature importance* toegepast op de testset. Dit is een gangbare methodiek waarvoor binnen de gehanteerde pakketten standaard routines aanwezig zijn. Met deze methode wordt steeds één risicokenmerk binnen de

dataset gehusseld wordt tussen de verzekerden om vervolgens opnieuw een R^2 te bepalen. Omdat zodoende een risicokenmerk niet meegenomen wordt zal dit leiden tot een lagere R^2. De mate waarin R^2 afneemt beschouwen we als maat voor het belang van het betreffende kenmerk voor het betreffende model. De 10 belangrijkste nieuw toegevoegde risicoklassen zijn opgenomen in Tabel 7.

Tabel 7: 10 belangrijkste toegevoegde risicoklassen per model, aangeven met *x*

Risicoklasse	M4 Gradient Boosting Machine	M5 Artificial Neural Network	M0d1 OLS op M4-data	M0d2 OLS op M5-data
Leeftijd in jaren	x	nvt	x	nvt
Aantal verpleegdagen	x	x	x	x
ddd's in FKG 37	x	x	x	x
ddd's in FKG 33	x			
Aantal operatief	x			
Dx 175 (dialyse)	x	x		x
ddd's in FKG 36	x			
ddd's in FKG 13	x			
Dx 1731 (chemo- of immunotherapie)	x	x	x	x
ddd's in FKG 24	x			
Aantal zorgproducten	nvt	x	nvt	x
Dx 1732 (chemo- én immunotherapie)		x		
Aantal diagnoses	nvt	x	nvt	x
ddd's in FKG 26		x	x	x
Dx 400002 (kinderoncologie)		x	x	
Dx 21099 (leukemie)		x	x	x
Dx 176 (beademing)			x	x
Dx 41013[a]			x	
Dx 21015[b]			x	
ddd's in FKG 8				x

NB: wanneer in bovenstaande tabel een cel leeg is betekent dit dat het risicokenmerk niet behoort top de 10 belangrijkste nieuwe features. Nvt = niet van toepassing, want niet beschikbaar in de onderhavige dataset. a = Maligne neoplasma tractus respiratorius en overige neoplasma, b = Maligne neoplasma lymfatisch en bloedvormend weefsel

Uit Tabel 6 blijkt dat er slechts beperkte verschuiving plaatsvindt van de belangrijkste risicokenmerken. De kenmerken MHK en MVV behoren in ieder model tot de top vier van kenmerken – zij blijven onverminderd belangrijk om tot een goede voorspelling van zorgkosten te komen. Dit effect is het grootst bij de *tree-based* modellen M1, M3 en M4. Zij kennen meer waarde toe aan MHK doordat deze modellen uit minder risicokenmerken kunnen kiezen. MHK is in deze modellen slechts één risicoklasse, terwijl het in M0, M2 en M5 negen binaire risicoklassen zijn.

Leeftijd in jaren is een belangrijke bijdragende nieuwe feature in 3 modellen waarin deze is toegevoegd. Ditzelfde geldt voor het aantal verpleegdagen en ddd's in FKG 37.

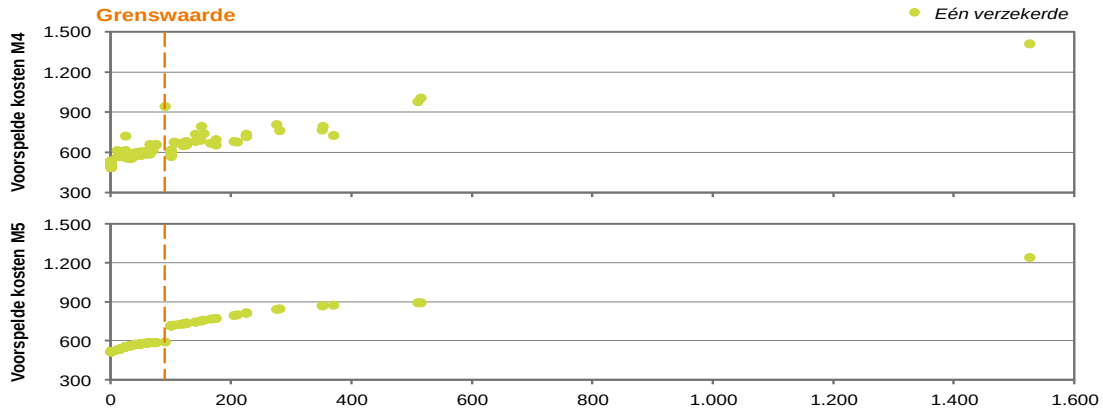
Uit Tabel 7 volgt dat meerdere toegevoegde risicoklassen van invloed zijn op het uiteindelijke voorspellend vermogen van het betreffende model. Zo leidt het toevoegen van Dx-risicoklassen bovenop de reeds bestaande pDKG- en sDKG-klassen aan de modellen tot hoger voorspellend

vermogen - in M4 en M5 maar ook in het OLS-model.[30] Ook vier volledig nieuwe risicokenmerken worden door de modellen gebruikt voor de voorspelling: het aantal verpleegdagen, het aantal operatieve verrichtingen, het aantal zorgproducten en het aantal unieke diagnoses.

Overigens geven bovenstaande tabellen alleen aan wat er gebeurt als in de *reeds getrainde* modellen bepaalde features worden weggelaten. Het zou interessant zijn om te zien hoe bijvoorbeeld een neuraal netwerk zou presteren en welke features het zwaarwegend zou maken wanneer het specifiek getraind wordt op een dataset waarin MHK niet als feature wordt meegenomen.

Model M4 en M5 gebruiken continue ddd's per FKG naast de bestaande binaire FKG's die voor iedere verzekerde met aantal ddd's in de betreffende FKG boven een grenswaarde gelijk is aan 1. Door de relatie die deze modellen leggen tussen aantal ddd's en zorgkosten te analyseren kunnen we hypotheses ontwikkelen en verder toetsen over de passendheid van de gekozen grenswaarde en de impact van meer of minder ddd's op zorgkosten. Figuur 13 geeft een voorbeeld hiervan voor FKG 28 astma.[31] Voor deze analyse keken we naar een zo groot mogelijke groep verzekerden die op alle risicokenmerken gelijk zijn behalve op het kenmerk FKG 28. Vervolgens keken we binnen deze subset naar de relatie tussen aantal ddd's op FKG 28 en de voorspelde zorgkosten. Vooral het artifical neural network laat een duidelijk verband zien dat met een lineaire vergelijking benaderd kan worden. Voor deze specifieke FKG zijn er duidelijke meerkosten voor verzekerden die minder dan de grenswaarde van 90 ddd's gebruikt hebben ten opzichte van geen ddd's. Maar ook boven de grenswaarde geeft het gebruik van meer ddd's hogere gemodelleerde zorgkosten. Interessant is ook het patroon dat model M5 laat zien. Bij 90 ddd krijgen de voorspelde zorgkosten een 'boost' omdat daarboven ook de binaire variabele voor FKG 28 'aan' staat. We zien echter zowel boven als onder de grenswaarde een duidelijke (lineaire) relatie tussen aantal ddd's en voorspelde zorgkosten.

Figuur 13: relatie tussen het aantal ddd's (op de x-as) en gemodelleerde zorgkosten in euro's (op de y-as) voor een groep verzekerden waarvan alle risicoklassen gelijk zijn behalve FKG 28 (astma)



[30] Het zij benadrukt dat de dx-groepen zijn toegevoegd als *aanvullende* features, dus niet ter vervanging van pDKG en sDKG. Daarvoor is gekozen om zoveel mogelijk de additionele waarde te laten zien ten opzichte van bestaande model- en featurekeuzes in het huidige risicovereveningsmodel. Uiteraard is het ook goed mogelijk om dx-groepen de huidige DKG-structuur te laten vervangen, zie bijvoorbeeld het recente groot onderhoud DKGs (WOR 988) waarin enkele mogelijkheden daarvoor zijn verkend.

[31] Dit voorbeeld is gekozen omdat het een relatief grote groep verzekerden betreft waardoor het nog mogelijk was om een voldoende grote groep te identificeren die op *alle* kenmerken gelijk was behalve op deze FKG. De analyse dient puur ter illustratie.

5.2 Duiding resultaten Decision Tree voor OLS

Zoals Figuur 7 in het vorige hoofdstuk laat zien karakteriseert de Decision Tree enkele grote groepen verzekerden met een beperkt aantal kenmerken. Zie Tabel 8 voor een overzicht van de vijf grootste groepen verzekerden die de Decision Tree clustert[32].

Uit de gebruikte kenmerken van deze vijf groepen blijkt dat met name relatief gezonde verzekerden goed gevangen kunnen worden met minder kenmerken dan het huidige OLS hanteert.

Tabel 8: vijf grootste verzekerdengroepen gedefinieerd door Decision Tree; beschreven zijn de kenmerken van deze groepen.

	Groep 1 19% van verzekerden	**Groep 2** 9% van verzekerden	**Groep 3** 7% van verzekerden	**Groep 4** 6% van verzekerden	**Groep 5** 4% van verzekerden
Geslacht	Man	Vrouw	Vrouw	Man	Man
Leeftijd	0_{t-1} – 49	0_{t-1} – 24	35 – 59	0 – 49	50 – 59
MHK	0	0	0	1	0
MVV	0, 1 of 2	0, 1 of 2	0, 1 of 2	0, 1 of 2	0, 1 of 2
sDKG	0, 1 of 2	0, 1 of 2	0, 1 of 2	0, 1 of 2	0, 1 of 2
AVI	Student, hoogopgeleid, zelfstandig of referentie	-	Hoogopgeleid, zelfstandig of referentie	Student, hoogopgeleid, zelfstandig of referentie	-
FKG	-	-	FKG0	FKG0	-
PPA	-	-	Eenpersoons-huishouden of overig	-	-
pDKG	-	-	-	pDKG = 0	-
FDG	-	-	-	FDG = 0	-

Dit roept de vraag op of het huidige model voor deze groepen verzekerden overfit. Hiertoe is de voorspelling van het OLS vergeleken met de voorspelling van de decision tree op deze vijf groepen. De maatstaven van OLS voor deze vijf grootste verzekerdengroepen zijn samengevat in Tabel 9; de betreffende maatstaf van de Decision Tree zijn per definitie nul – de gemodelleerde zorgkosten op de subset zijn immers gelijk aan de gemiddelde kosten van de subset.

Tabel 9: de voorspelling van OLS op de vijf grootste verzekerdengroepen van de decision tree.

Maatstaf	**Groep 1**	**Groep 2**	**Groep 3**	**Groep 4**	**Groep 5**
R^2 x 100%	-0,4%	-0,7%	-0,1%	0.0%	0,4%
CPM x 100%	7,9%	3,9%	3,1%	-8,7%	-1,6%

Groene *waarden geven aan dat OLS een beter resultaat geeft op deze maatstaf voor deze groep;* - Oranje *waarden geven aan dat het decision tree model een beter resultaat geeft op deze maatstaf voor deze groep*

Bovenstaande tabel geeft geen eenduidige resultaten op de relatieve kwadratische afwijking (R^2) of de relatieve absolute afwijking (CPM). Vaak presteert OLS beter op een maat maar slechter op de andere maat. Wel blijkt uit bovenstaande resultaten dat het aannemelijk is dat het OLS op

[32] Op basis van de 70% verzekerden in de trainingsset.

sommige van deze groepen overfit – het gebruik van méér vereveningscriteria voor deze verzekerden verslechtert de voorspelling.

Het OLS-model kan van dit inzicht profiteren door rekening te houden met de interactie door parameters die de decision tree niet gebruikt (zoals FKG voor groep 1) op 0 te zetten voor de betreffende groep in de OLS.

Onderzoek van Buchner (2017) past deze tweetraps aanpak van het doorrekenen van een decision tree om relevante interactietermen te vinden voor een OLS toe op een dataset van ongeveer 2,9 miljoen Duitse verzekerden [17]. Het toevoegen van de 95 gevonden interactietermen verbetert de R^2 van 25,43% naar 25,81%. Van Veen (2018) past dezelfde methodiek toe op het Nederlandse vereveningssysteem en vindt vergelijkbare resultaten: het toevoegen van 145 gevonden interactietermen verbetert de R^2 van 25,56% naar 27,34% [18]. Door het toevoegen van enkel 3^e order interactietermen, dat zijn er 7 in het onderzoek, verbetert de R^2 slechts 0.08-procentpunt. Het is dus waarschijnlijk dat het toevoegen van de in dit onderzoek gevonden interactietermen slechts beperkt tot verbetering van de verevenende werking zal leiden.

5.3 Duiding resultaten Piecewise Regressie voor OLS

In aanvulling op de algemene resultaten van hoofdstuk 4 geeft Bijlage K een overzicht van de resultaten van het Piecewise Regressie model per leeftijdsegment in vergelijking met het OLS model. Op alle segmenten behalve de eerste (0-8 jaar) doet Piecewise Regressie het beter. Met name op segment 17-26 jaar en segment 34-41 jaar is het verschil opvallend groot. Blijkbaar voegt het hebben van een separaat model per leeftijdssegment wel echt iets toe, en is het vooral de problematiek in de groep 0-8-jarigen welke het algehele resultaat van Piecewise Regressie wat omlaag trekt.

De normbedragen van model M2 zijn lastig te interpreteren, omdat er een gewicht hangt aan de continue variabele voor leeftijd. Daarnaast hanteert het model niet de restricties zoals in het OLS-model gebruikelijk zijn, waardoor er geen afslagklasse is. Om die reden hebben we als aanvullende analyse een regulier OLS toegepast binnen de segmenten die door het Piecewise Regressie algoritme bepaald zijn. Dat heeft beperkte, maar positieve, impact op de maatstaven (zie voor details Bijlage L).

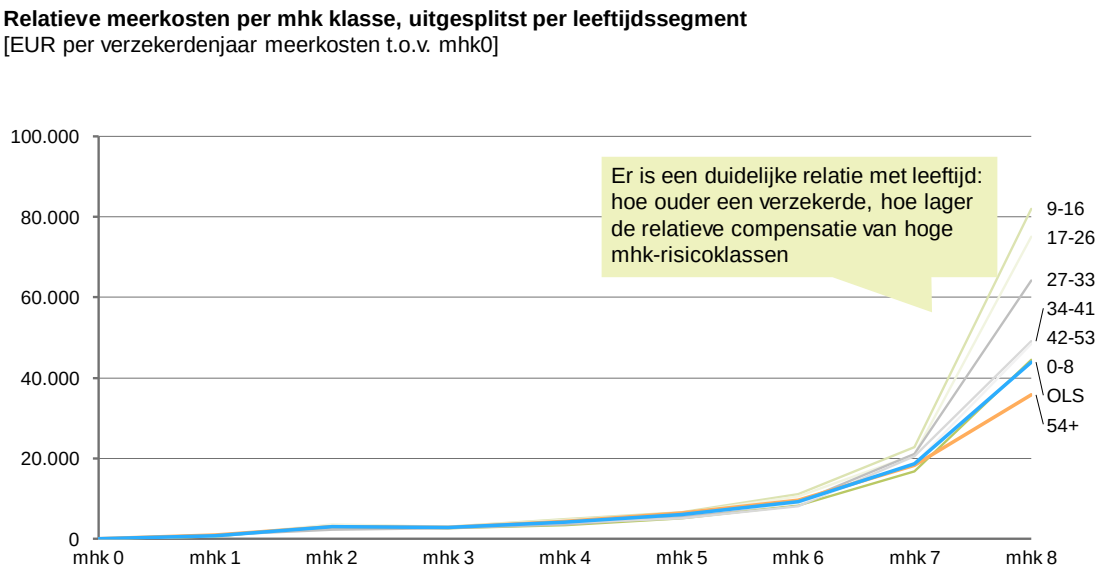
Met deze aangepaste normbedragen is de interpretatie van uitkomsten veelal intuïtief. Zo zijn bijna alle normbedragen positief (uitgezonderd de afslagklassen), variëren ze veelal rondom de normbedragen van het uitgangsmodel en lopen de normbedragen over het algemeen op bij zwaardere DKG's, MHK's en MVV's. In Bijlage M zijn voor alle morbiditeitscriteria de normbedragen per leeftijdssegment opgenomen.

Uit deze normbedragen per leeftijdssegment volgt een interessante observatie: de relatieve meerwaarde van een hoge risicoklasse, bijvoorbeeld MHK8, ten opzichte van de referentieklasse, MHK0 neemt af bij oudere verzekerden. Dit effect is weergegeven in Figuur 14. Het ontvangen van het MHK8-normbedrag conform de reguliere OLS zou hebben geleid tot een overschatting van de zorgkosten.

Of dit echt zo is valt op basis van dit onderzoek niet met zekerheid te zeggen. Wel zien we dat op het merendeel van verzekerdengroepen op basis van MHK en leeftijdssegment M2 de resultaten beter voorspelt dan het uitgangsmodel (Tabel 10). Vooral opvallend zijn de betere resultaten onder verzekerden ouder van 54 (voorspelling van M2 ligt EUR 50 dichter bij werkelijke kosten dan M0) en verzekerden met MHK8 (voorspelling van M2 ligt EUR 6.713 dichter bij werkelijke kosten dan

M0). Vervolgonderzoek zou nader op deze observatie in kunnen gaan en bepalen of bijvoorbeeld segmentatie van kenmerken als MHK naar leeftijd het vereveningsresultaat kan verbeteren.

Figuur 14: de relatieve meerkosten van de verschillende mhk-risicoklassen per leeftijdssegment t.o.v. de referentieklasse mhk0. Oudere verzekerden ontvangen een relatief lagere bijdrage voor een hoog mhk-kenmerk dan jongere verzekerden.



Tabel 10: Financieel resultaat M2 t.o.v. M0. Negatieve bedragen zijn een verbetering op financieel resultaat door M2 te gebruiken.

	Leeftijdssegmenten:							
	0-8	9-16	17-26	27-33	34-41	42-53	54+	**Totaal***
Geen MHK	18	-26	-10	1	4	1	-52	**-15**
1	10	-137	-53	-104	16	-14	-46	**-38**
2	-1.464	-2.008	-59	-587	-190	35	-26	**-193**
3	57	-67	320	-149	130	-70	-26	**-42**
4	423	-268	-88	-555	-246	61	78	**60**
5	731	206	-346	-544	430	-35	-31	**20**
6	833	842	349	-54	-549	-447	22	**64**
7	-1.631	-3.007	-640	-1.005	1.930	581	-158	**-467**
8	1.576	-11.217	-38.827	11.506	-3.041	-5.039	-7.161	**-6.713**
Totaal*	**9**	**-61**	**-29**	**-45**	**7**	**-9**	**-50**	**-30**

** Totaal is geen optelsom van onderliggende componenten doordat een gewogen gemiddelde is genomen op basis van aantal verzekerdenjaren.*

5.4 Nieuwe risicoklassen voor OLS

In dit hoofdstuk zijn op verschillende wijzen potentiele verbeteringen van het OLS geïdentificeerd op basis van de uitkomsten van de machine learning algoritmes. Dit is geen uitputtende lijst, omdat het voor dit onderzoek slechts 'bijvangst' betreft.

We beoordelen de mogelijke aanpassingen aan het huidige OLS aan de hand van de beoordelingscriteria uit het toetsingskader (WOR 871): verevenende werking, doelmatigheid,

beheersbare complexiteit en validiteit en meetbaarheid [19]. Aanvullend hanteren we het criterium toepasbaarheid: de mate waarin de aanpassing hanteerbaar is binnen de huidige vereveningssysteem en bestaande risicoklassen. Tabel 11 vat de resultaten samen.

Tabel 11: Overzicht resultaten beoordelingscriteria voor mogelijke aanpassingen aan OLS.

	Toepasbaar-heid	Verevenende werking	Doelmatig-heid	Beheersbare complexiteit	Validiteit en meetbaarheid
Leeftijd in jaren	+ / -	+/-	+	+	+
Dx-groepen	++	+	-	-	+
Aantal ddd's per FKG	-	+	- -	-	+
Aantal zorgprod.	+ / -	++	-	+	+ / -
Aantal verpleegd.	+ / -	++	-	+	+
Aantal diagnoses	+ / -	++	-	+	+
Kenmerken, bijv. MHK, opdelen in leeftijdssegmenten	++	+	+	+/-	+
Interactietermen voor specifieke groepen[33]	++	+	+	+/-	+/-

- De **toepasbaarheid** is beoordeeld als zeer positief voor de aanpassingen die een nieuwe binaire risicoklasse betreffen, omdat dit goed aansluit bij het huidige OLS. De beoordeling is neutraal voor leeftijd in jaren omdat deze als continue variabele of ongeveer 100 binaire variabelen toegevoegd wordt. Hetzelfde geldt voor aantal zorgproducten, aantal verpleegdagen en aantal diagnoses. Deze risicokenmerken zouden ook als risicokenmerk met een aantal binaire klassen toegevoegd kunnen worden aan OLS. De beoordeling is negatief indien een continue variabele toegevoegd wordt aan het model.
- De **verevenende werking** van leeftijd in jaren is neutraal beoordeeld omdat Tabel 7 laat zien dat dit weinig toevoegt voor OLS. Dit in tegenstelling tot aantal zorgproducten, aantal verpleegdagen en aantal diagnoses dat aan OLS respectievelijk 0,6, 0,6 en 0,4 procentpunt R^2 toevoegt. Voor overige potentiële aanpassingen is de beoordeling positief, omdat er duidelijke indicaties zijn dat deze aanpassingen de verevenende werking kunnen verbeteren. Dat is niet voor alle aanpassingen in dit onderzoek getoetst.
- De **doelmatigheid** van aanpassingen waarvoor geen nieuwe informatie aan het model is toegevoegd is beoordeeld als positief, omdat er geen verandering is in de doelmatigheidsprikkels van verzekeraars en aanbieders. Ook het toevoegen van meer gedetailleerde informatie over leeftijd heeft geen invloed op de doelmatigheid. Voor het toevoegen van een of meerdere Dx-groepen[34] aan het model geldt dat de doelmatigheid negatief beïnvloed wordt doordat er prikkels kunnen ontstaan in de registratie van diagnoses die leiden tot het classificeren van een verzekerde in een Dx-groep. Dit geldt ook voor aantal zorgproducten, aantal verpleegdagen en aantal diagnoses, wat kan leiden tot ongewenste prikkels. Door het toevoegen van een continue variabele met aantal ddd's per FKG wordt de doelmatigheid zeer negatief beïnvloedt, omdat zonder ondergrens een

[33] Bijvoorbeeld zoals de groepen gedefinieerd in Tabel 8.
[34] In recent groot onderhoud DKG (WOR 988) en in de pre-OT 2020 (WOR 990) is deze mogelijkheid reeds verkend en is een fijnmaziger diagnosecategorisering opgenomen. Deze uitkomsten zijn uiteraard zwaarwegender dan de eerste verkenning die we hier uitvoeren.

verzekerde direct een hogere vereveningsbijdrage ontvangt bij het voorschrijven van een medicijn uit de referentietabel.

- De **beheersbare complexiteit** van de aanpassingen die geen interactietermen toevoegen is beoordeeld als positief omdat het model nauwelijks aan eenvoud en transparantie inboet. Voor aanpassingen die wel interactietermen toevoegen is de beheersbare complexiteit neutraal, omdat het aantal klassen zeer beperkt toeneemt. Echter zorgen de interacties wel voor een afname in inzichtelijkheid van de classificatie van de vereveningscriteria. Bij dx-groepen en FKG-ddd scoren we een "-" op complexiteit, omdat het daar gaat om zeer veel extra variabelen, die wellicht transparant zijn maar die de uitvoering wel complexer maken (denk aan regulier onderhoud, beslissingen wel/geen trend, beoordeling plausibiliteit, vragen van verzekeraars over specifieke klassen).
- De **validiteit** van de meeste aanpassingen is beoordeeld als hoog, omdat deze risicoklassen een logische voorspellende waarde hebben voor toekomstige zorgkosten. Dit is in mindere mate het geval voor aantal zorgproducten, omdat het aantal ook afhankelijk is van de maximale doorlooptijd van een zorgproduct en de eerste maand waarin een zorgproduct geregistreerd wordt. Ook voor de interactietermen voor specifieke groepen geldt dat de relatie tussen de nieuw geconstrueerde risicoklassen en toekomstige zorgkosten minder voor de hand liggen. De **meetbaarheid** van alle aanpassingen is hoog, aangezien de benodigde data niet toeneemt ten opzichte van het huidige OLS.

6 Consequenties van het toepassen van machine learning technieken in de reguliere risicovereveningscyclus

Naast de positieve resultaten op R^2 van de inzet van machine learning modellen (resultaat), zijn de consequenties van het toepassen van machine learning (haalbaarheid) ook een belangrijk criterium in de aantrekkelijkheid van machine learning voor de risicoverevening. In dit hoofdstuk verkennen we de consequenties van het toepassen van machine learning technieken in de reguliere risicovereveningscyclus, wanneer een van deze technieken de huidige OLS zou vervangen. We beginnen het hoofdstuk met een samenvatting van onze bevindingen, welke we in de volgende paragrafen nader toelichten. Paragraaf 6.1 verkent de te organiseren randvoorwaarden bij elk van de modellen. Vervolgens beschrijven paragraaf 6.2 en 6.3 de implicaties voor de verbeter- en onderhoudscyclus, respectievelijk. In paragraaf 6.4 worden de consequenties voor de uitvoering uiteengezet, waarna paragraaf 6.5 afsluit met de impact van de modellen op interpretatie en prikkelwerking.

Wanneer een van de machine learning technieken het huidige OLS-model zou vervangen heeft dit consequenties op alle aspecten van de risicovereveningscyclus. Tabel 12 laat zien dat vooral de implementatie van modellen M3, M4 en M5 een grote impact zou hebben op de cyclus zoals deze nu is ingericht. Voor toepassing ter vervanging van de huidige OLS is wellicht aanpassing van wet- en regelgeving nodig, en ook het uitvoeren van onderzoeken in de verbeter- en onderhoudscycli wordt complexer. Piecewise lineaire regressie (M2) lijkt in veel opzichten op het huidige OLS-model. Implementatie daarvan zou in dat opzicht dus laagdrempeliger zijn.

Tabel 12: consequenties van toepassing machine learning op risicovereveningscyclus.

		M1 Decision Tree	M2 Piecewise lineaire regressie	M3 Random Forest	M4 Gradient Boosting Machine	M5 Artificial Neural Network
Randvoorwaarden	Wet- en regelgeving	3	3	1	1	1
	Betrokkenen	3	3	3	3	3
	Infrastructuur	3	3	2	2	3
Verbetercyclus	Uitvoer verbetercyclus	2	2	2	1	2
	Interpretatie verbeteronderzoeken	3	3	1	1	1
Onderhoudscyclus	Gegevensfase	3	3	3	3	3
	OT	2	2	1	1	1
	Normbedragenfase	2	3	2	1	1
Uitvoering en afrekening	Afrekening	2	3	2	1	1
	Ex-post mechanismen	2	3	1	1	1
Interpretatie en prikkelwerking	Validiteit	2	3	1	1	1
	Stabiliteit en homogeniteit	2	3	3	3	3
	Transparantie en eenvoud	3	3	2	2	2
	Prikkelwerking	2	3	2	1	1

Legenda: 1 = grote consequenties, 2 = enige consequenties, 3 = geen of beperkte consequenties

Bij de interpretatie van Tabel 12 zijn enkele kanttekeningen van belang:

- De tabel laat de consequenties zien wanneer een van deze technieken de huidige OLS zou vervangen. Vergelijking met de huidige werkwijze is niet het hoofddoel van deze analyse, maar helpt wel om consequenties van het gebruik van deze modellen inzichtelijk te maken.
- Het feit dat consequenties in sommige gevallen groot zijn is niet een waardeoordeel of een teken van onuitvoerbaarheid. Het is wel een teken dat de implementatiedrempel waarschijnlijk hoger zal zijn.

In de volgende paragrafen gaan we nader in op de aspecten zoals samengevat in Tabel 12.

6.1 Organiseren randvoorwaarden

Wet- en regelgeving

In artikel 32 en artikel 34 van de Zorgverzekeringswet wordt ingegaan op de risicoverevening. In artikel 32 staat dat “jaarlijks bij algemene maatregel van bestuur regels omtrent de berekening van de vereveningsbijdragen worden gesteld”. Deze regels bepalen ten minste dat de hoogte van de vereveningsbijdrage wordt berekend op basis van bij die maatregel te bepalen, voor alle zorgverzekeraars gelijke criteria, waaronder in ieder geval het aantal verzekerden bij een zorgverzekeraar en een aantal verzekerdenkenmerken. Tevens wordt statistisch onderbouwd aan elk criterium een bijdrage gekoppeld.

Artikel 34 specificeert vervolgens dat uiterlijk op 1 april van het vierde jaar volgende op het kalenderjaar waarvoor de bijdragen, bedoeld in artikel 32 en 33, zijn toegekend, het Zorginstituut de bijdragen definitief vaststelt. De vaststelling van een vereveningsbijdrage houdt in ieder geval in een herberekening van de vereveningsbijdrage op basis van het werkelijke aantal verzekerden dat de zorgverzekeraar in het desbetreffende jaar had en de werkelijke verdeling van de verzekerdenkenmerken over die verzekerden, voor zover de daartoe benodigde gegevens tijdig bij het Zorginstituut zijn aangeleverd.

In de zorgverzekeringswet zelf is dus vastgelegd dat verevening moet plaatsvinden op basis van verzekerdenkenmerken, en dat aan elk kenmerk een bijdrage gekoppeld moet zijn. In het Besluit Zorgverzekering (artikel 3.6, lid 2) wordt dit nog eens nader gespecificeerd: “Onze Minister kent aan alle klassen van de genoemde criteria gewichten toe”.

Normbedragen voortkomend uit de OLS voldoen aan deze formulering. Modellen M1 en M2 voldoen ook aan deze formulering: bij deze modellen is immers nog steeds sprake van een weging per criterium (in model M2) of per groep van criteria (in model M1) die voor alle verzekeraars gelijk zijn. Bij modellen M3-M5 is daarentegen niet langer sprake van een ‘statistisch onderbouwde bijdrage aan elk criterium’. Onder deze modellen wisselt de bijdrage per criterium immers tussen individuen omdat deze afhangt van het totaal aan kenmerken van ieder individu.

Deze analyse is een eerste verkenning, niet een uitgebreide juridische toets. Deze verkenning laat zien dat voor modellen M3-M5 ten minste een grondige juridische toets, en mogelijk een wetswijziging noodzakelijk is. Daarom scoren we deze modellen op dit aspect als ‘grote consequenties’.

Betrokkenen

Wanneer modellen M1-M5 gehanteerd zouden worden in plaats van M0 zijn bij gegevensverzameling of uitvoer geen aanvullende partijen noodzakelijk. We scoren dit aspect in Tabel 12 daarom voor alle modellen als ‘geen consequenties’.

Infrastructuur

Voor het maken van een OLS-model op 17 miljoen verzekerden en ongeveer 200 risicokenmerken is geen bijzondere rekenkracht vereist, maar voor sommige machine learning modellen is wel degelijk veel rekenkracht nodig. Machine learning modellen worden veelal ontwikkeld op cloud-infrastructuur welke gemakkelijk op- en afschaalbaar is. Gegeven de gevoeligheid van de gegevens benodigd voor risicoverevening is deze manier van werken niet wenselijk. Voor dit onderzoek maakten wij gebruik van onze eigen infrastructuur, welke is ingericht op het verwerken van grote hoeveelheden zorggegevens in complexe analyses.

We ondervonden bij het uitvoeren van dit onderzoek met name bij de ontwikkeling van modellen M3 en M4 merkbare beperkingen van deze infrastructuur (Tabel 13). Het trainen van deze modellen duurde enkele dagen, en ook ondervonden we zeer sterk oplopende run-tijden wanneer we meer data beschikbaar stelden aan deze modellen (zie bijvoorbeeld het onderscheid tussen model M3a, op OT data, vs. M3 op OT data plus continue leeftijd). Het is belangrijk hierbij te realiseren dat deze runtijden sterk beïnvloed worden door de hoeveelheid modellen die gelijktijdig worden getraind. In dit onderzoek, waarin in korte tijd veel modellen gelijktijdig werden ontwikkeld en modelberekeningen vaak parallel binnen dezelfde servercapaciteit draaiden was dit een factor van belang.

In de basis is dit een oplosbaar probleem: bijschakelen van extra infrastructuur lost het namelijk op. Ook voor dit onderzoek konden we het probleem ondervangen door het tijdelijk bijschakelen van extra servercapaciteit. Het is daarmee vooral een factor om rekening te houden in de onderzoeksopzet en -budgettering.

Dit probleem speelt uiteraard alleen bij het trainen van het model. Het *toepassen* van een getraind machine learningmodel is eenvoudig en snel, net zoals bij een getraind OLS-model.

Tabel 13: run-tijden voor het trainen van modellen per fold en totaal (in uren)[35]

	M0	M1	M2	M3	M4	M5	M3a	M4a	M5a
UREN *(per fold en totaal)*	OLS	Decision Tree	Piecewise lineaire regressie	Random Forest	Gradient Boosting Machine	Artificial Neural Network	Random Forest	Gradient Boosting Machine	Artificial Neural Network
Min	0,24	0,03	0,13	17,94	4,41	1,05	7,27	0,27	0,79
Max	0,38	0,05	0,15	18,34	9,44	1,09	11,60	0,30	0,82
TOTAAL	**3,38**	**0,39**	**1,38**	**181,21**	**62,59**	**10,76**	**93,35**	**2,82**	**8,08**

Minimum- en maximumtijd per fold, en totale tijd voor de 10 folds (in uren). Nota bene: deze tijden zijn sterk afhankelijk van de hardware waarop het model getraind werd, en in het bijzonder ook hoeveel andere modellen op hetzelfde moment getraind werden. Tijdens dit onderzoek, waarvoor in korte tijd veel modellen werden ontwikkeld, was dit een factor van significant belang.

Naast benodigde fysieke infrastructuur is het ook nodig om met andere softwarepakketten te werken. Analyses in de risicoverevening worden veelal uitgevoerd in het softwarepakket SAS (alhoewel uiteraard ook andere pakketten mogelijk zijn). Machine learning daarentegen wordt vooral ontwikkeld binnen R en Python. Betrokken partijen, zoals Zorginstituut Nederland (ZINL), universiteiten en onderzoeksbureaus zullen wellicht door een learning curve moeten om deze software voor dit doel te implementeren, medewerkers te trainen, etc.

Al met al scoren we het aspect “infrastructuur” als “enige consequenties” voor modellen die de grootste rekenkracht vereisten in ons onderzoek. De impact is er zeker, maar het is in principe eenvoudig oplosbaar. De learning curve aspecten zijn wel degelijk belangrijk, maar niet van blijvende aard en wegen daarom maar beperkt mee in de scores in Tabel 12.

6.2 Verbetercyclus

Deze paragraaf beschrijft de situatie waarin machine learning modellen de huidige OLS zouden *vervangen*. Wellicht ten overvloede merken we op dat dit geen inschatting is van de *meerwaarde* die deze modellen mogelijk hebben in *huidige* verbetercyclus (zie daarvoor Hoofdstuk 7).

[35] Deze modellen maken gebruik van de hyperparameters zoals gerapporteerd in de bijlage van dit document. Verschillen in run tijd tussen M3, M4, M5 en M3a, M4a en M5a, respectievelijk, zijn derhalve een gevolg van een andere dataset én andere modelinstellingen.

Uitvoer verbetercyclus

Het doel van onderzoeken in de verbetercyclus is het testen van mogelijke verbeteringen aan het risicovereveningsmodel. Verbeterideeën worden inhoudelijk onderbouwd, hetgeen bijvoorbeeld kan leiden tot nieuwe of herziene risicokenmerken. Deze kenmerken worden vervolgens getest door toepassingen in een risicovereveningsmodel, waarbij het toetsingskader uit WOR871 wordt gebruikt om de verbetering te toetsen op aspecten als verevenende werking, meetbaarheid, stabiliteit en validiteit [19].

Wanneer een van de modellen M1-M5 het huidige vereveningsmodel zou vervangen wordt het uitvoeren van verbeteringsonderzoeken complexer, omdat nu ook hyperparameters geoptimaliseerd moeten worden. Ook in dit onderzoek ondervonden wij dat, net als bij de selectie van features, hier een bepaalde mate van 'kunst' bij komt kijken die zal verschillen tussen verschillende modelleurs: bij veel modellen is 'optimale' selectie van hyperparameters (nagenoeg) onmogelijk. Daarmee zal steeds weer de vraag ontstaan of het vinden van een beter model nu het resultaat is van betere features of door 'betere' modelleurs die betere sets van hyperparameters vinden. Vooral bij de complexere modellen, in het bijzonder model M4, is de keuze van hyperparameters zeer bepalend voor het resultaat.

Een beperktere, maar wel aanwezige, consequentie is het feit dat sommige modellen meer vragen van infrastructuur dan andere modellen. Bij de opzet van verbeteronderzoeken zal men hier rekening mee moeten houden, zeker wanneer veel verschillende modelvarianten getest dienen te worden. Dit aspect is hierboven al eerder benoemd en nemen we hier niet nogmaals mee in de score.

Samenvattend heeft vooral het feit dat bij het maken van machine learning modellen ook hyperparameters geoptimaliseerd moeten worden enige consequenties voor het uitvoeren van verbeteronderzoeken. Omdat vooral bij model M4 veel hyperparameters bestaan, en de keuze van deze parameters in dit onderzoek ook zeer bepalend voor het resultaat leken, scoren we vooral de consequenties van het toepassen van model M4 als groot.

Interpretatie verbeteronderzoeken

Naast de uitvoer van verbeteronderzoeken heeft het hanteren van machine learning modellen ook consequenties voor de interpretatie van verbeteronderzoeken. Uiteraard is het nog steeds mogelijk om op basis van een inhoudelijk onderbouwde hypothese te evalueren of toevoeging van een bepaald kenmerk leidt tot een vereveningsmodel met een betere verevenende werking. Maar doordat de complexere modellen, vooral M3-M5, meer rekening houden met interactie en doordat bij iedere modelwijziging ook hyperparameters gewijzigd kunnen worden kan het wel lastiger zijn om precies aan te wijzen of een eventuele verbetering ook daadwerkelijk veroorzaakt wordt door toevoeging van dat kenmerk. Daarnaast is het wellicht lastiger om aspecten uit het toetsingskader, zoals stabiliteit (in het toetsingskader verwoord als: "Het vereveningscriterium vertoont stabiliteit indien het systematische verband tussen vereveningscriterium en vervolgkosten jaarlijks is terug te vinden") en validiteit (in het toetsingskader verwoord als: "Het vereveningscriterium is valide als het een systematisch verband vertoont met de op grond van de eigenschap te verwachten kosten voor verzekeraars.") volgens de nu gangbare definities aan te tonen.

Samenvattend denken we dat vooral het toepassen van de complexere modellen (M3-M5) grote consequenties heeft voor de interpretatie van verbeteronderzoeken.

6.3 Onderhoudscyclus

Gebruik van de hier geteste modellen hoeft niet noodzakelijkerwijs tot een andere tijdlijn van activiteiten in de onderhoudscyclus te leiden. Wel zullen gedurende de cyclus een aantal activiteiten substantieel anders verlopen:

- Gegevensfase:
    - Machine learning modellen, meer dan OLS, kennen een risico op overfitting. Het is mogelijk dat daarbij een model ontwikkeld wordt dat zeer goed in staat is random variatie in een dataset samen te vatten, maar dat totaal niet goed presteert bij het voorspellen op een nieuwe, ongeziene dataset. Om dit te voorkomen is het belangrijk om het model te trainen via een techniek zoals 10-fold cross-validation, en om conclusies te kunnen trekken over hoe goed het model op een ongeziene dataset zal werken is het belangrijk om het uiteindelijke model te testen op een ongeziene testset. In deze gegevensfase moet daarom, na constructie van de nieuwe totaaldataset, een testset en een trainingsset geconstrueerd worden. De testset mag niet meer bekeken of gebruikt worden tot aan het eind van de OT, om enige schijn te voorkomen dat modelleurs onbewust beïnvloed zouden kunnen worden door kenmerken van de testset bij het ontwerpen van hun model (een principe vergelijkbaar met het 'dubbelblind' werken in gerandomiseerde klinische trials).
- OT: in deze fase wordt het model ontwikkeld op de trainingsset.
    - De eerder afgescheiden testset kan aan het eind van de OT gebruikt worden wanneer meerdere modelvarianten naast elkaar getest dienen te worden, bijvoorbeeld wanneer na de pre-OT nog meerdere varianten mogelijk zijn.
    - Na de OT staan de hyperparameters vast en worden deze niet verder aangescherpt, want het is niet uitvoerbaar om in de korte tijd die beschikbaar is voor de normbedragenfase nog een verdere verkenning te doen naar hyperparameters. Om te zorgen dat dit in de Normbedragenfase niet tot problemen zal leiden zou men kunnen overwegen om in de OT meerdere modelopties te maken die rekening houden met meerdere scenario's voor hoe het macrobudget er in september uiteindelijk uit zal zien. Op deze manier bestaat in de normbedragenfase nog wel enige vrijheid om te kiezen voor de set aan hyperparameters die het meest geschikt is.
    - Voor modellen M3-M5 kunnen ook geen normbedragen (per criterium of per combinatie van criteria) meer worden bepaald. In plaats daarvan wordt het model in zijn geheel *open source* gepubliceerd. Ook is het mogelijk om het getrainde model als eenvoudige applicatie beschikbaar te stellen aan betrokken partijen, zodat bijvoorbeeld verzekeraars eenvoudig zelf de vereveningsbijdrage voor hun verzekerden kunnen bepalen.
    - Met name modellen M3 en M4 hebben relatief veel rekentijd nodig. Bij gebruik van deze modellen kan tijdsdruk mogelijk een belangrijke rol gaan spelen.
- Normbedragenfase:
    - In deze fase wordt het model definitief getraind. Alleen de definitieve gewichten binnen het model worden nog bepaald, de hyperparameters staan al vast vanuit de OT-fase.
    - De totale OT-set wordt geschaald naar de relevante macrobedragen, en verzekerdenaantallen worden herwogen zodanig dat ze in lijn zijn met de verzekerdenraming van ZINL. Het detailniveau waarop deze raming nog

betekenisvol kan plaatsvinden hangt niet af van specifieke machine learning technieken, maar van het detailniveau van de *features* die worden meegenomen in het model. In dit onderzoek maken modellen M4 en M5 gebruik van gedetailleerde onderliggende datasets, waarvoor een precieze verzekerdenraming vermoedelijk niet mogelijk is. Het belang van een goede verzekerdenraming in deze fase is evident – anders moeten conclusies immers gebaseerd worden op sterk verouderde gegevens – maar het effect van het *detailniveau waarop* deze raming plaatsvindt is niet geheel duidelijk. Op basis van gevoerde interviews en literatuuronderzoek vonden wij geen duidelijk bewijs dat zeer gedetailleerd ramen waardevoller is dan iets minder gedetailleerd ramen. Aanvullend onderzoek op dit gebied zou nuttig zou zijn om hierover preciezere conclusies te kunnen trekken. Het niet uitvoeren van een verzekerdenraming is vermoedelijk pas een optie wanneer de verevening kan worden uitgevoerd op veel recentere data. Vooralsnog lijkt dit niet realistisch.

Samenvattend heeft het hanteren van machine learning modellen grote consequenties voor de OT fase en de Normbedragenfase. Het feit dat hyperparameters in de normbedragenfase niet meer kunnen worden aangescherpt leidt tot onzekerheid waarvan de impact nog onderzocht moet worden. Wanneer modellen ook gebruik maken van meer brongegevens is het belangrijk om te realiseren dat een zeer gedetailleerde verzekerdenraming vermoedelijk niet langer mogelijk is, en zal eerst beter onderzocht moeten worden of het iets minder precies ramen nog tot betrouwbare uitkomsten leidt.

6.4 Uitvoering

Het gebruik van alternatieve modellen heeft ook consequenties voor de daadwerkelijke uitvoer van de risicoverevening door het Zorginstituut. Hieronder geven we de belangrijkste aspecten weer.

Afrekening

Voor de verschillende afrekenmomenten (van ex-ante tot definitieve vaststelling) leveren verzekeraars, de belastingdienst, UWV en Vektis brongegevens bij ZINL aan. ZINL controleert die bronbestanden, en gebruikt deze voor de indeling in kenmerken, het ramen t.b.v. het schatten van de normbedragen, het bepalen van ex-post correcties, etc. Modellen M4-M5 maken gebruik van aanzienlijk meer brongegevens, hetgeen de uitvoer van de afrekening zoals hier beschreven complexer zal maken.

Bij het daadwerkelijk inzetten van het uiteindelijke vereveningsmodel voor het afrekenen is het belangrijk dat dit tot een zo goed mogelijke ex-ante vaststelling per verzekeraar leidt die vervolgens niet onverklaarbaar schommelt over de verschillende momenten. Stabiliteit van de vereveningsresultaten over de verschillende vaststellingen *binnen een vereveningsjaar* is dan ook zeer belangrijk. Het is onwenselijk dat grote verschuivingen optreden, zeker wanneer dit als gevolg van modeleigenschappen is.

Bij machine learning modellen, net als bij OLS, staat het model na de normbedragenfase vast – het model zelf wisselt dan niet meer. Verschuivingen in budgetten (behoudens de ex-post correcties) worden dan veroorzaakt door steeds definitiever wordende verzekerdenaantallen en kenmerken. Omdat de complexere machine learningmodellen, vooral modellen M3-M5, uitgebreide interacties modelleren is echter niet met zekerheid te zeggen dat het effect van veranderende inputs niet tot grotere schommelingen in uitkomsten zal leiden dan bij het gebruik van OLS. In dit onderzoek is dit niet getest, en het verdient aanbeveling om dit in de toekomst wel te doen.

Voor zorgverzekeraars is het belangrijk dat zij de bijdragen volledig zelf kunnen narekenen. In essentie zal dit met het gebruik van machine learningmodellen niet anders zijn. Verzekeraars kunnen met het getrainde definitieve model zelf de bedragen precies narekenen op basis van kenmerken van hun verzekerden. Het is vooral belangrijk dat er voldoende vertrouwen bestaat in de werking van het model – door een combinatie van transparantie, begrip en ervaring.

ZINL krijgt niet alleen geregeld vragen over hoogte van de bijdragen, maar vooral ook over de indeling van verzekerden in kenmerken. De noodzaak tot het betrouwbaar, uitlegbaar vaststellen van kenmerken per verzekerde blijft onveranderd groot – ongeacht het model waarin deze gebruikt worden.

Samenvattend denken we dat voor het succesvol gebruiken van machine learningmodellen vooral moet worden gewerkt aan een vertrouwensbasis. Stabiliteit, vooral van de complexere modellen, is nog niet aangetoond en het is belangrijk dat bij zowel verzekeraars als ZINL vertrouwen ontstaat in de werking van het model. Daarom beoordelen we de consequenties van het gebruik van vooral de complexere machine learning modellen, zeker de modellen die ook nog eens meer data gebruiken, op dit moment als groot. Uitzondering hierop is model M2, omdat dit model zeer veel lijkt op de huidige OLS.

Ex-post mechanismen

Criteriumneutraliteit, waarbij ZINL achteraf enkele gewichten opnieuw bepaalt omdat van tevoren het aantal verzekerden niet goed voorspelbaar bevonden wordt, is niet toe te passen wanneer een model wordt gehanteerd dat niet leidt tot normbedragen per criterium of per groep van criteria (modellen M3-M5). Voor deze criteria zou dan een ander ex-post correctiemechanisme moeten worden gezocht en getest. De consequenties van het gebruik van deze modellen beoordelen we op dit aspect dan ook als groot. Andere ex-post mechanismen die niet afhankelijk zijn van modeluitkomsten op kenmerkniveau, zoals hoge kostencompensatie en flankerend beleid blijven wel technisch uitvoerbaar.

6.5 Interpretatie en prikkelwerking

Het gebruik van een ander model zal ook consequenties hebben voor de manier waarop resultaten geïnterpreteerd kunnen worden, en voor de prikkelwerking die uitgaat van het model. In deze paragraaf vatten we de belangrijkste consequenties samen.

Validiteit

Validiteit wil, volgens het huidige toetsingskader, zeggen dat er een logische, systematische relatie is tussen de kenmerken waarop het model is gebaseerd en de zorgkosten. Op basis van de analyses uitgevoerd in deze studie, met name in hoofdstuk 5, kunnen we concluderen dat de belangrijkste kenmerken in het OLS-model een min of meer gelijke rol spelen bij de modellen M1-M5: de top 10 kenmerken zijn min of meer gelijk over de verschillende modellen. De modellen drijven dus grotendeels op dezelfde kenmerken die al eerder valide bevonden zijn.

Wat echter niet is getoetst is, is of de kenmerken nog steeds in alle gevallen een valide relatie vertonen met zorgkosten. Bij OLS, en ook bij M2, is dit eenvoudig na te gaan. Bij OLS geldt immers eenvoudigweg een normbedrag per criterium waarvan de (relatieve) hoogte beoordeeld kan worden. Ook bij een decision tree is het effect van een individueel criterium nog redelijk eenvoudig na te gaan. Bij complexere modellen is dit, door de uitgebreide interactiemogelijkheden tussen features, vermoedelijk wel (grotendeels) het geval maar niet meer eenvoudig te toetsen. Het is bijvoorbeeld mogelijk dat een bepaalde DKG over het algemeen tot hoge gemodelleerde kosten leidt, maar bij sommige combinaties van verzekerdenkenmerken niet.

Het is de vraag in hoeverre dit belangrijk zou moeten zijn. Men kan immers besluiten dat kenmerken valide zijn wanneer er een duidelijke (medisch-)inhoudelijke grond voor is, en er bijvoorbeeld in verkennende multivariate regressie een duidelijk verband aantoonbaar is tussen het kenmerk en zorgkosten. Wanneer een machine learningmodel op basis van dit kenmerk vervolgens tot goede, stabiele verevenende werking leidt, maar niet in alle gevallen de verwachte relatie tussen inputvariabele en uitkomst laat zien is dit laatste dan misschien van ondergeschikt belang.

Duidelijk is evenwel dat de *consequenties* van het hanteren van dergelijke modellen groot zijn op dit aspect: het is ofwel noodzakelijk om uitvoerige validiteitstesten te doen, ofwel om een andere definitie van het begrip validiteit te gaan hanteren.

Stabiliteit en homogeniteit

Stabiliteit en homogeniteit, zowel binnen een vereveningsjaar als over verschillende jaren is in dit onderzoek niet getest. We kunnen daarom geen uitspraken over stabiliteit doen op basis van kwantitatieve bevindingen.

Conceptueel valt wel te concluderen dat model M1 mogelijk minder stabiel is bij veranderingen in data. Veranderingen in inputs kunnen namelijk leiden tot een andere volgorde en combinatie van vertakkingen, hetgeen dan al snel tot een heel ander model leidt. Bij andere modellen, waarbij features parallel werken in plaats van serieel, is dit effect vermoedelijk minder substantieel.

Binnen een vereveningsjaar speelt dit effect niet – het vereveningsmodel staat immers vast na de normbedragenfase. Wel kan het zijn dat het in de normbedragenfase ontwikkelde model suboptimaal is voor de uiteindelijke gegevensset bij de definitieve vaststelling. Die onzekerheid speelt potentieel bij ieder model, maar wellicht in iets grotere mate bij modellen die vatbaarder zijn voor overfitting. Daarom is het werken met een trainingsset vs. een testset zo belangrijk.

Mogelijk is het feit dat modellen M3-M5 gebruik maken van randomness ook van invloed op stabiliteit. Bijvoorbeeld, het algoritme voor neurale netwerken gebruikt een *random number generator* voor het kiezen van initiële waarden. Voor het reproduceren van uitkomsten is hiermee gemakkelijk om te gaan, namelijk het gebruiken van een 'random seed'-waarde. Deze waarde helpt om binnen dezelfde gegevensset dezelfde uitkomsten te reproduceren. Het kiezen van een andere random seed zal wel leiden tot subtiel andere uitkomsten. Van jaar-op-jaar helpt deze oplossing niet, maar in de praktijk zal het effect hiervan zeer beperkt zijn – het effect van het gebruik van andere features en nieuwere gegevens jaar-op-jaar is naar alle waarschijnlijkheid veel groter. Daarom beoordelen we consequenties van het gebruik van deze modellen op het aspect 'stabiliteit' als zeer beperkt.

Samenvattend zal vooral het gebruik van model M1 tot enige consequenties op het gebied van stabiliteit leiden waarmee men rekening moet houden. Voor de overige modellen schatten we de consequenties in als beperkt.

Transparantie en eenvoud

Uitlegbaarheid van modellen is van groot belang voor het vertrouwen dat partijen erin hebben. Het Zorginstituut krijgt jaarlijks veel, zeer gedetailleerde vragen van verzekeraars over de vastgestelde vereveningsbijdrage: over de hoogte van de bijdrage en over wijzigingen daarin over de jaren. Een transparant, eenvoudig model helpt bij het beantwoorden daarvan. Overigens hoeft een model niet noodzakelijk 'eenvoudig' te zijn. Net als in het toetsingskader verwoord moet dit uiteindelijk een

gebalanceerde afweging zijn met andere aspecten: een complexer model dat wel een duidelijk betere verevenende werking heeft kan alsnog zeer wenselijk zijn.

De modellen M1 en M2 zij eenvoudig intuïtief te doorgronden. M2 is even uitlegbaar als het OLS-model, maar leidt wel tot meer normbedragen omdat de OLS op meerdere leeftijdssegmenten wordt uitgevoerd. M1 is als beslisboom zelfs uit te tekenen, en bevat minder combinaties van kenmerken dan het OLS-model.

Daar waar modellen meer ruimte bieden om interacties tussen features te modelleren worden ze ook complexer om te doorgronden. Model M3 is conceptueel nog goed uitlegbaar, maar de doorwerking van het model is minder intuïtief. Voor modellen M4 en M5 geldt dit in nog grotere mate. Nota bene, het is wel degelijk mogelijk om een relatie te leggen tussen een kenmerk en een uitkomst – maar de precieze samenhang tussen combinaties van kenmerken en de exacte voorspelling is niet meer zonder aanvullende analyse van de uitkomsten te bevatten.

Hoewel deze modellen lager scoren op eenvoud is het wel degelijk mogelijk om zeer transparant te zijn, ook over modellen M3-M5. Niet alleen de broncode, maar ook het getrainde model kan gepubliceerd worden, analoog aan de normbedragen van de OLS. Hiermee kan iedere verzekeraar voor zich gemakkelijk de vereveningsbijdrage voor haar eigen populatie bepalen. Zoals eerder opgemerkt is het ook mogelijk om het getrainde model als eenvoudige applicatie beschikbaar te stellen aan betrokken partijen, zodat bijvoorbeeld verzekeraars eenvoudig zelf de vereveningsbijdrage voor hun verzekerden kunnen bepalen.

Al met al schatten we in dat het hanteren van met name modellen M3, M4 en M5 “enige consequenties” zal hebben. Het is mogelijk om dezelfde transparantie te bieden als bij andere modellen, maar het vereist meer analysewerk omdat relaties niet altijd meer intuïtief zijn. Het feit dat modellen minder eenvoudig zijn weegt slecht beperkt mee in onze afweging, omdat eenvoud wel wenselijk is, maar geen op zichzelf staand doel.

Prikkelwerking

Alle modellen maken gebruik van dezelfde features als het huidige model. Modellen M2 en M3 gebruiken continue leeftijd, hetgeen niet een noemenswaardig effect op prikkelwerking heeft. Modellen M4 en M5 gebruiken daarnaast evenwel ook fijnmaziger features, bijvoorbeeld het aantal ddd onderliggend aan de FKG’s, het aantal operaties en het aantal ligdagen.

Deze features kunnen leiden tot een ondoelmatigheidsprikkel. Voor ddd’s geldt dit zeker: ook de verkennende analyse in paragraaf 5.1 laat zien dat een hoger aantal ddd’s zal leiden tot een hogere vereveningsbijdrage in zowel model M4 als model M5. In praktijk zal men dus liever werken met fijnmaziger grenswaarden dan dat daadwerkelijk de continue variabele wordt meegenomen. Het aantal operaties en het aantal ligdagen zijn gebaseerd op normaantallen per zorgproduct, waardoor de prikkel enigszins beperkt wordt. Ook hier ligt het hanteren van grenswaarden meer voor de hand.

Bij alle modellen m.u.v. model M2 is het mogelijk dat bepaalde features niet ‘opgepikt’ worden omdat deze niet significant genoeg bijdragen aan het optimaliseren van de doelfunctie, in het geval van dit onderzoek een betere R^2. Dit kan tot onvoorziene en/of ongewenste prikkelwerking leiden. Bijvoorbeeld, de groep ‘hoger opgeleiden’ is met opzet toegevoegd aan de OLS om de prikkel tot risicoselectie te beperken, maar heeft als extra ‘feature’ binnen AVI slechts beperkte invloed op de R^2. Het is evenwel een oplosbaar probleem: het is mogelijk om modellen constraints mee te geven zodanig dat het resultaat op bepaalde groepen expliciet op nul komt. In dit onderzoek is dit niet verder onderzocht.

Tegenover het potentieel negatieve effect van modellen op prikkelwerking staat ook dat de potentiële waarde van voorop blijven lopen. Wanneer op landelijk niveau niet de best mogelijke technieken gehanteerd worden valt niet uit te sluiten dat individuele verzekeraars deze wel toepassen. Als op deze manier de ene verzekeraar beter de winstgevendheid van bijvoorbeeld een collectiviteit kan voorspellen dan de andere verzekeraar kan het *niet implementeren* van technieken juist leiden tot risicoselectie.

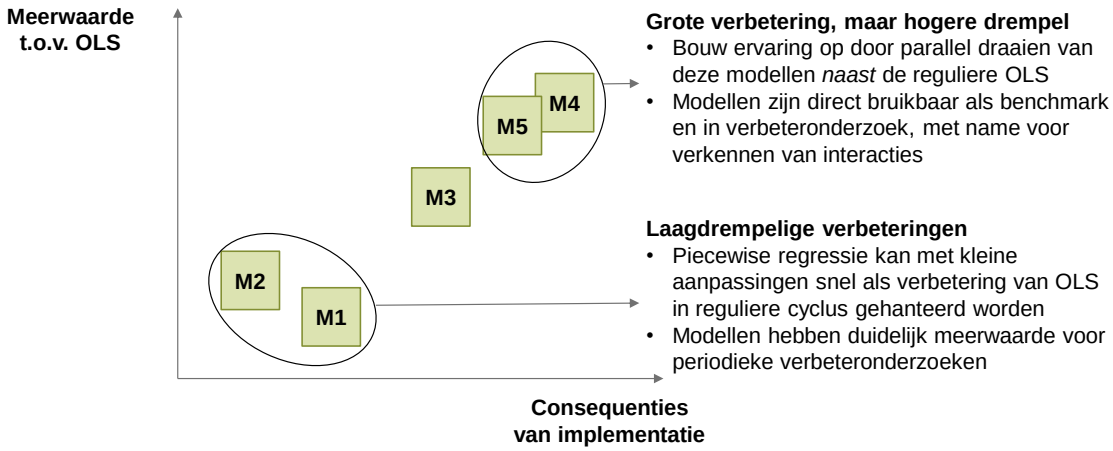
7 Conclusies en aanbevelingen

Dit verkennende onderzoek geeft veel aanknopingspunten voor het toepassen van machine learning in de risicoverevening. In paragraaf 7.1 beschrijven we waar op basis van dit onderzoek laagdrempelige toepassingskansen liggen, en waar kansen liggen op grote maar wel hoogdrempeliger verbeteringen. In paragraaf 7.2 gegeven we concrete aanbevelingen voor aanvullend onderzoek, en in paragraaf 7.3 vatten we de belangrijkste 'bijvangst-'aanbevelingen voor de reguliere risicoverevening samen.

7.1 Hoe verder met machine learning in de risicoverevening?

Op basis van een afweging van impact op verevenende werking enerzijds, en consequenties van implementatie anderzijds geven we in dit hoofdstuk laagdrempelige verbeterkansen en kansen met grote meerwaarde maar ook een hogere implementatiedrempel. De aanbevelingen zijn samengevat in figuur 15.

Figuur 15: hoe verder met machine learning in de risicoverevening?



Een aantal modellen getest in dit onderzoek zijn laagdrempelig toepasbaar en zouden al snel meerwaarde kunnen bieden:

- Model M2 (Piecewise Lineaire Regressie) is een logische extensie van de huidige OLS en zou, met beperkte doorontwikkeling, potentieel snel geschikt zijn om de OLS te vervangen, met name om beter om te kunnen gaan met interacties tussen leeftijd en reguliere vereveningskenmerken zoals FKG of DKG. Het model is voor dit doel ook goed inzetbaar in de verbetercyclus.
- Model M1 (Decision Tree) is gemakkelijk uit te voeren, en heeft het ook in dit onderzoek geleid tot identificatie van grote, kostenhomogene groepen. Het periodiek uitvoeren van dergelijk onderzoek kan helpen om interacties tussen kenmerken op eenvoudige wijze inzichtelijk te maken, en vast te stellen of bijvoorbeeld interactietermen van toegevoegde waarde kunnen zijn.

Modellen M4 (Gradient Boosting Machine) en M5 (Artificial Neural Network) zijn zeer veelbelovende modellen. Zelfs op precies dezelfde data laten deze modellen enigszins betere resultaten zien dan OLS. Daarbij is van belang dat de machine learning modellen op de beperkte

datasets niet uitgebreid geoptimaliseerd zijn, en verdere verbetering wellicht nog mogelijk is. Ook wanneer deze modellen voorlopig de OLS nog niet zullen *vervangen* kunnen ze als aanvullende onderzoekstechniek al snel van meerwaarde blijken als benchmark voor het 'best haalbare' vereveningsresultaat (bijvoorbeeld als onderdeel van de OT), en als onderdeel van partiële onderzoeken in de verbetercyclus.

Vooral wanneer meer informatie wordt toegevoegd laten deze modellen duidelijk meerwaarde zien. Het lijkt dat het vermogen van deze modellen om voorspelkracht te halen uit interacties daarmee nog meer uit de verf komt. Daarbij was dit slechts een verkennend onderzoek, en het is mogelijk dat in de toekomst met verdere optimalisatie nog betere resultaten behaald kunnen worden. Daarentegen zijn dit ook de modellen waarvan implementatie grotere consequenties zal hebben voor de huidige manier van werken, wat maakt dat de implementatiedrempel hoger is.

7.2 Aanbevelingen m.b.t. vervolgonderzoek op dit gebied

Naast de specifieke aanbevelingen voor implementatie van de verschillende modellen leidt dit onderzoek ook tot een aantal algemenere aanbevelingen voor vervolgonderzoek naar machine learning in de risicoverevening:

- Onze belangrijkste aanbeveling is om deze machine learning modellen enkele jaren parallel te draaien *naast* de reguliere OLS om geleidelijk ervaring op te bouwen met robuustheid en implementatie-aspecten van deze modellen, en om de belangrijkste onzekerheden te adresseren. Ook zouden de best presterende modellen retrospectief getest kunnen worden over meerdere jaren.
- Daarnaast zou het interessant zijn om te onderzoeken hoe machine learning modellen presteren wanneer inhoudelijk minder wenselijke aspecten van het huidige model worden weggelaten. In dit onderzoek zijn modellen getraind op *minimaal dezelfde features* als in het huidige model beschikbaar zijn, maar niet op een *beperktere* set aan features. Hoe waardevol zijn deze modellen wanneer bijvoorbeeld MHK of MVV niet meegenomen wordt als feature?
- In dit onderzoek is alleen het somatisch model verkend. Bij toekomstig onderzoek zou het ook nuttig zijn om de andere modellen, met name het GGZ-model, te verkennen. Ook is het zinvol om te verkennen in hoeverre machine learning kan helpen om te komen tot minder modellen in het algemeen.
- Het zou goed zijn om in vervolgonderzoek ook te kijken naar andere doelfuncties. Belangrijkste doel van een risicovereveningsmodel is om prikkels op de juiste manier in te richten. Goede doelfuncties houden wellicht niet alleen rekening met R^2.
- Regularization en dimensionality reduction worden in machine learning gebruikt om overfitting op de data te voorkomen. Middels een penalty op het aantal mee te nemen variabelen of een principal component analysis worden variabelen met veel ruis minder zwaar gewogen in het model – of zelfs verwijderd – dan variabelen met een hoog voorspellend vermogen. Uit literatuur volgt dat dit het voorspellend vermogen van algoritmes op nieuwe records kan doen toenemen. In bijvoorbeeld het OLS-model met aanvullende data (zoals M0b en M0d) in dit onderzoek is deze techniek niet toegepast, dit kan derhalve nog een aanknopingspunt zijn voor toekomstig onderzoek.

7.3 Aanbevelingen voor de reguliere risicovereveningscyclus

7.3.1 Lessen op basis van het "machine learning paradigma"

Machine learning is niet alleen een collectie van tools en technieken, het is ook een ander paradigma. Binnen dit paradigma zijn sommige zaken vanzelfsprekend terwijl dat buiten het veld minder vanzelfsprekend is. Hieronder doen we een aantal aanbevelingen die in eerste instantie vanuit het machine learning-paradigma komen, maar waarvan wij denken dat ze in het algemeen nuttig kunnen zijn, ook binnen de risicoverevening.[36]

- Het is van belang om het totaal van alle modellen en bewerkingen te testen: van het samenstellen van de dataset tot en met het bepalen van de bedragen in de definitieve vaststelling. Hiervoor moeten alle handmatige handelingen geautomatiseerd of tenminste in een ondubbelzinnige procedure met strikte termen en definities gevangen worden. Alle stappen die worden genomen, inclusief bijvoorbeeld het uitvoeren van de RAS methode en het maken van een verzekerdenraming, zouden (althans volgens het machine learning paradigma) geprotocolleerd moeten worden uitgevoerd en als één geheel worden getest op de gegevens van 2 of meer jaren eerder. Zo kan de totale voorspelling beoordeeld worden.[37]
- Forceer het gebruik van één uitkomstmaat. Tijdens dit onderzoek liepen de modelleurs met enige regelmaat aan tegen het gegeven dat waar gevraagd werd om R^2 te optimaliseren impliciet ook andere uitkomsten, zoals resultaten op subgroep- of verzekeraarsniveau belangrijk bleken. Discussies over het relatieve belang worden dan achteraf gevoerd, en het is mogelijk dat uiteindelijk een model ontwikkeld is dat niet optimaal past bij de *impliciet* gewenste mix van uitkomsten. Het vooraf expliciet maken van de uitkomstmaat dwingt onderzoekers om het probleem van tevoren voldoende te doorgronden en optimale uitkomsten zo precies mogelijk te definiëren. Wanneer meerdere uitkomstmaten belangrijk zijn, kunnen deze het best via een vooraf bepaalde weging worden gecombineerd. Ook is het mogelijk om vooraf een ondergrens vast te stellen voor belangrijke maar niet-geoptimaliseerde uitkomstmaten.
- Houdt een testset apart en geheim, en test daarop eenmalig. Dit is belangrijk omdat anders altijd het risico van overfitting bestaat, waarbij random variatie wordt gemodelleerd. In de praktijk is dit risico bij OLS niet groot, maar het zou, getuige ook de werkwijze in dit onderzoek, ook in reguliere vereveningsonderzoeken prima werkbaar zijn om volgens deze best practice te werken.

7.3.2 Bijvangst: mogelijke aanknopingspunten voor regulier vervolgonderzoek

Tenslotte heeft dit onderzoek geleid tot inzichten die mogelijk relevant kunnen zijn bij vervolgonderzoek binnen de reguliere risicovereveningscyclus.

---

[36] Voor lezers geïnteresseerd in de verschillen in paradigma's bevelen we de zeer leesbare paper van Leo Breiman aan: "Statistical Modeling: The Two Cultures (2001)"

[37] Dit zou neerkomen op het volgende: starten met de kenmerken van 2014 en de kosten van 2015. Doorloop vervolgens alle stappen volgens een vastgelegd protocol (dus ook de RAS-methode en het aanvullen van ontbrekende waarden). Als laatste zouden dan de totale kosten voor 2018 (inclusief eigen risico en ggz) voorspeld worden adhv de kenmerken van 2017. Deze voorspelde kosten zouden dan vergeleken moeten worden met de kosten die daadwerkelijk in 2018 gemaakt zijn.

- Het is nuttig om bij evaluatie van de huidige risicoverevening ook andere stappen dan de modelontwikkeling zelf te toetsen. In dit onderzoek kwam bijvoorbeeld de vraag voorbij in hoeverre het schalen van de OT-dataset naar de verzekerdenraming invloed heeft op de kwaliteit van de verevening. Hoe erg is het als we dit minder precies of zelfs helemaal niet doen? De raming kan er immers ook naast zitten, en de RAS-procedure is op zichzelf ook een aanname over de werkelijke verzekerdenaantallen per subgroep. Daarnaast kan bijvoorbeeld het samenvoegen van verschillende modellen (somatisch, GGZ en eigen risico) voor een totaalresultaat per verzekerde nuttig zijn om apart te testen. Ook kleinere stappen die worden toegepast in de verevening lijken niet altijd volledig doorleefd. Een verkennende analyse laat bijvoorbeeld zien dat de praktijk van het opschalen van kosten voor verzekerden die slechts een deel van het jaar verzekerd zijn kan leiden tot een bias in de analyse (zie Bijlage N). Uiteraard zijn er goede, volstrekt logische redenen om deze opschaling te doen, maar de historische referenties [20]–[22] waarin deze opschaling voor het eerst werd toegepast hebben de introductie van een dergelijke bias niet uitgebreid verkend.
- De Decision Tree analyse identificeerde enkele grote groepen gezonde verzekerden waarvoor het bieden van meer informatie aan het model ongunstig lijkt. Het is het onderzoeken waard of het aanpassen van de OLS zodat deze rekening houdt met de groepen geïdentificeerd in de Decision Tree analyse kan leiden tot een beter vereveningsresultaat in het huidige vereveningsmodel.
- Piecewise Regressie laat zien dat vooral MHK mogelijk interactie vertoont met leeftijd. Of dit echt zo is valt op basis van dit onderzoek niet met zekerheid te zeggen. Wel zien we dat op het merendeel van verzekerdengroepen op basis van MHK en leeftijdssegment model M2 de resultaten beter voorspelt dan het uitgangsmodel. Vervolgonderzoek zou nader op deze observatie in kunnen gaan en bepalen of segmentatie van kenmerken als MHK naar leeftijd het vereveningsresultaat kan verbeteren.
- Het uitsplitsen van FKG o.b.v. aantal ddd's lijkt van toegevoegde waarde voor de voorspelkracht, ook bij de reguliere OLS. Het is, bijvoorbeeld bij groot onderhoud, zinvol te verkennen of een fijnmaziger model, of in ieder geval nog specifiekere drempelwaardes per FKG, kan leiden tot een beter vereveningsmodel.
- Leeftijd in jaren, aantal verpleegdagen, aantal zorgproducten, aantal diagnoses en aantal operaties zijn allen mogelijke *features* die de voorspelkracht van het OLS-model kunnen vergroten, en zijn daarmee het onderzoeken waard. Daarbij is vooral ook belangrijk om prikkelwerking mee te nemen in de evaluatie.
- Het uitsplitsen van DKG naar dx-groepen lijkt van toegevoegde waarde voor de voorspelkracht. Bij het recente groot onderhoud DKG (WOR 988) en in de pre-OT 2020 (WOR 990) is deze mogelijkheid reeds verkend, en is een fijnmaziger diagnosecategorisering opgenomen. Wij doen daarom nu geen aanvullende aanbeveling op dit gebied.

Referenties

[1] I. Ismail, “Improving risk equalization through machine learning: a comparative evaluation of Random Forests and Gradient Boosted Machines to OLS regression,” 2018.

[2] S. Rose, “A Machine Learning Framework for Plan Payment Risk Adjustment,” *Health Services Research*, vol. 51, no. 6, pp. 2358–2374, 2016.

[3] “WOR 856: Onderzoek ‘gezonde verzekerden’: verbetering van de compensatie voor chronisch zieken in het somatisch vereveningsmodel,” 2017.

[4] “WOR 902: Huisartsenzorg in de risicoverevening,” 2018.

[5] “WOR 973: Onderzoek risicoverevening 2020: Overall Toets,” 2019.

[6] W. van de Ven *et al.*, “Preconditions for efficiency and affordability in competitive healthcare markets: Are they fulfilled in Belgium, Germany, Israel, the Netherlands and Switzerland?,” *Health Policy*, vol. 109, no. 3, pp. 226–245, 2013.

[7] “WOR 929: Onderzoek risicoverevening 2019: Overall Toets,” 2018.

[8] I. Duncan, M. Loginov, and M. Ludkovski, “Testing Alternative Regression Frameworks for Predictive Modeling of Health Care Costs,” *North American Actuarial Journal*, vol. 20, no. 1, pp. 65–87, 2016.

[9] M. Morid, K. Kawamoto, T. Ault, J. Dorius, and S. Abdelrahman, “Supervised Learning Methods for Predicting Healthcare Costs: Systematic Literature Review and Empirical Evaluation,” *AMIA Annual Symposium proceedings*, vol. 2017, pp. 1312–1321, 2017.

[10] S. Sushmita *et al.*, “Population cost prediction on public healthcare datasets,” *ACM International Conference Proceeding Series*, vol. 2015-May, pp. 87–94, 2015.

[11] D. Bertsimas *et al.*, “Algorithmic prediction of health-care costs,” *Operations Research*, vol. 56, no. 6, pp. 1382–1392, 2008.

[12] C. Yang, C. Delcher, E. Shenkman, and S. Ranka, “Machine learning approaches for predicting high cost high need patient expenditures in health care,” *BioMedical Engineering Online*, vol. 17, no. S1, pp. 1–20, 2018.

[13] B. Panay, “Predicting Health Care Costs Using Evidence,” pp. 1–13, 2019.

[14] C. Yang, C. Delcher, E. Shenkman, and S. Ranka, “Machine learning approaches for predicting high utilizers in health care,” *International Conference on Bioinformatics and Biomedical Engineering*, pp. 382–395, 2017.

[15] H. Kan, H. Kharrazi, H. Chang, D. Bodycombe, K. Lemke, and J. Weiner, “Exploring the use of machine learning for risk adjustment : A comparison of standard and penalized linear regression models in predicting health care costs in older adults,” pp. 1–8, 2019.

[16] L. Breiman, J. Friedman, R. Olsehn, and C. Stone, *Classification and Regression Trees*. Wadsworth, 1984.

[17] F. Buchner, J. Wasem, and S. Schillo, “Regression Trees Identify Relevant Interactions: Can This Improve the Predictive Performance of Risk Adjustment?,” *Health Economics (United Kingdom)*, vol. 26, no. 1, pp. 74–85, 2017.

[18] S. van Veen, R. van Kleef, W. van de Ven, and R. van Vliet, “Exploring the predictive power of interaction terms in a sophisticated risk equalization model using regression trees,” *Health Economics (United Kingdom)*, vol. 27, no. 2, pp. 1–12, 2018.

[19] “WOR 871: Toetsingskader 2017,” 2017.

[20] R. P. Ellis, B. Martins, and S. Rose, “Risk adjustment for health plan payment,” *Risk Adjustment, Risk Sharing and Premium Regulation in Health Insurance Markets: Theory and Practice*, pp. 55–104, 2018.

[21] A. Ash, F. Porell, L. Gruenberg, E. Sawitz, and A. Beiser, “Adjusting Medicare capitation payments using prior hospitalization data.,” *Health Care Financing Review*, vol. 10, no. 4, pp. 17–29, 1989.

[22] R. P. Ellis and A. Ash, “Refinements to the diagnostic cost group (DCG) model,” *Inquiry*, vol. 32, no. 4, pp. 418–429, 1995.

Bijlage A: Beschikbare data voor de modellen

Tabel 14 geeft een overzicht van de beschikbare databronnen en de velden die gebruikt zijn voor de verschillende modellen. M0 en M1 gebruiken enkel de OT-dataset. M2 en M3 gebruiken daarnaast de leeftijd in jaren. M4 en M5 gebruiken aanvullende informatie over morbiditeitscriteria uit bronbestanden.

Tabel 14: overzicht beschikbare databronnen.

Bronbestand	Naam	Omschrijving	Gebruikt voor modellen M4 en M5
Persoonskenmerken 2016 en 2017	Leeftijd	Leeftijd in hele jaren op peildatum 30 juni van betreffende jaar	ja, tevens in modellen M2 en M3 voor 'leeftijd_continu'
	BTL_ident	Identificatie woonachtig buitenland (1 = Nederland / 2 = buitenland)	nee; alleen verzekerden die voorkomen in OT-bestand meenemen
Declaratiebestand farmaceutische zorg 2016	ATC_code	ATC-code in 2016 (xxxxxxx = niet gevuld) o.b.v. G-Standaard Z-Index	ja, voor feature 'FKG_DDD'
	PRGR	Product-code in 2016 (xx = niet gevuld) o.b.v. G-Standaard Z-Index	nee
	HPK_ddd	DDD-factor per HPK-basiseenheid	ja, voor feature 'FKG_DDD'
	HPK_eh	HPK-eenheid (xx = niet gevuld) o.b.v. G-Standaard Z-Index	ja, voor feature 'FKG_DDD'
	IK_eh	Eenheid afgeleverde hoeveelheid (xx = niet gevuld)	ja, voor feature 'FKG_DDD'
	HOEVEEL	Afgeleverde hoeveelheid	ja, voor feature 'FKG_DDD'
	ddd_aantal	*DDD construeren met HPK_ddd en HOEVEEL*	ja, voor feature 'FKG_DDD'
	ZI_nr	ZI-artikelnummer o.b.v. G-Standaard Z-Index	nee
	AFL_datum	Datum van aflevering (EEJJMMDD)	nee
	schade_FAR	Schadebedrag farmaceutische zorg	nee
	Dcind	Indicatie debet (= D) / credit (= C)	nee
Declaratiebestand add-on geneesmiddelen 2016	diag_code_add	Diagnosecode, uitsluitend voor nivolumab (decl_code_add = 194610)	ja, voor feature 'FKG_DDD'
	decl_code_add	DBC-declaratiecode, uitsluitend ADD-ON duur- en weesgeneesmiddelen	nee
	hoeveel	Aantal gebruikte eenheden	ja, voor feature 'FKG_DDD'
	schade_ADD	Schadebedrag ADD-ON geneesmiddelen	nee
	Dcind	Indicatie debet (= D) / credit (= C)	nee
Declaratiebestand DBC-somatisch 2016	DBC_code	Declaratiecode DBC somatisch	nee
	ZP_code	Zorgproductcode DBC somatisch	ja, voor features 'Dx-groepen', 'Ligdagen_aantal', 'Operaties_aantal' en 'Zorgproducten_aantal'
	DIAG_code	Diagnosecode DBC somatisch	ja, voor features 'Dx-groepen' en 'Diagnoses_aantal'
	SPEC_code	Specialismecode van poortspecialisme	ja, voor features 'Dx-groepen', 'Diagnoses_aantal' en 'Specialismen_aantal'
	INST_code	AGB_code instelling volgens declaratie	nee
	MND_open	Maand van opening van subtraject DBC somatisch	nee
	schade_DBC	Schadebedrag DBC-somatisch	nee
	DCind	Indicatie debet (= D) / credit (= C)	nee

Bijlage B: Decision Tree

Voor dit onderzoek is gekozen voor de *rpart* implementatie van het CART-algoritme in R, versie 4.1-15. Voor het risicovereveningsmodel zijn de volgende hyperparameters toegepast (niet weergegeven hyperparameters zijn conform de standaardinstellingen van de betreffende R-implementatie).

Hyperparameter	Instelling	Toelichting
Method	'Anova'	Standaard splitsingscriterium voor een regression tree.
Control.minsplit	60 (default = 20)	Alleen bij sets > 60 zoekt het algoritme naar een nieuwe splitsing
Control.minbucket	9 (default = minsplit/3))	Iedere subset bevat minimaal 9 verzekerden
Control.cp	0.000015 (def = 0.01)	Een groep wordt alleen gesplitst indien de R^2 met de splitsing minimaal 0.000015 punt toeneemt.
Control.xval	0 (default = 10)	We gebruiken geen interne cross-validation
Control.maxdepth	Default (30)	Niet nodig om te definiëren – met 2^30 mogelijkheden zijn control.cp en control.minsplit leidend in de diepte van de keuzeboom.

Bijlage C: Piecewise Regressie

Voor dit onderzoek is gekozen voor de *partykit* implementatie in R, versie 1.2-5. Voor het risicovereveningsmodel zijn de volgende hyperparameters toegepast (niet weergegeven hyperparameters zijn conform de standaardinstellingen van de betreffende R-implementatie).

Hyperparameter	Instelling	Toelichting
Alpha	Default = 0.05	Het algoritme splits een dataset indien dit leidt tot een significante verbetering ($p < 0.05$).
Minsize	10.000 (Default = 10)	Iedere subset bevat minimaal 20.000 verzekerdenjaren.
Maxdepth	Default = inf	Er staat geen limiet op de diepte van de boom. De boom wordt gebruikt om de spltsing van de data in segmenten te maken.
Model	Model = aangepaste OLS-regressie	Het algoritme dat wordt toegepast in het bepalen van de splitsingen is een bewerkte versie van de reguliere OLS-regressie

Bijlage D: Random Forest

Voor dit onderzoek is gekozen voor de *ranger* implementatie, versie 0.12.1, van het Random Forest algoritme in R. Voor het risicovereveningsmodel zijn de volgende hyperparameters toegepast (niet weergegeven hyperparameters zijn conform de standaardinstellingen van de betreffende R-implementatie).

Hyperparameter	Instelling	Toelichting
Num.trees	200	Het algoritme traint 200 onafhankelijke keuzebomen
Mtry	20	Ieder boom heeft beschikking over 20 risicokenmerken
Importance	"impurity"	Iedere gemaakte splitsing minimaliseert de *impurity index*.
Min.node.size	16 (op OT-data) 30 (op OT-data met leeftijd)	Iedere boom overweegt alleen een splitsing indien de betreffende subset gegevens van minimaal 16 of 30 verzekerden bevat.
Max.depth	NULL	De bomen hebben geen limiet op het aantal achtereenvolgende splitsingen.
Replace	TRUE	Verzekerden kunnen meerdere keren worden geselecteerd om dezelfde boom te trainen.
Sample.fraction	1	Iedere boom maakt gebruik van evenveel willekeurig geselecteerde verzekerden als het aantal verzekerden dat de trainings-set bevat.

Bijlage E: Gradient Boosting Machine

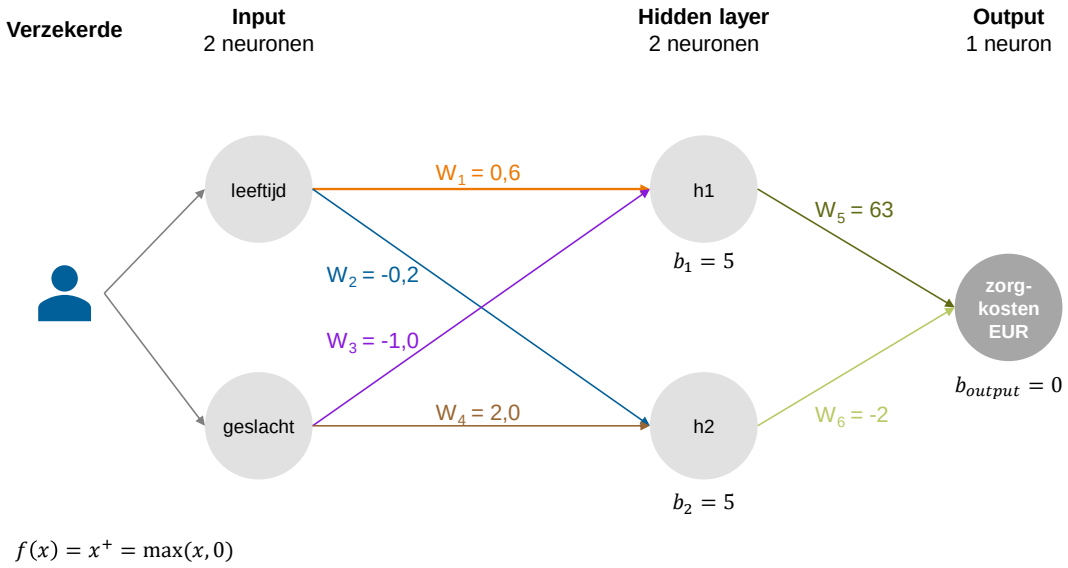
Voor dit onderzoek is gekozen voor de *XGBoost* implementatie, versie 0.90.0.2, van het Gradient Boosting Machine algoritme in R. Voor het risicovereveningsmodel zijn de volgende hyperparameters toegepast (niet weergegeven hyperparameters zijn conform de standaardinstellingen van de betreffende R-implementatie).

Hyperparameter	Instelling	Toelichting
alpha	128	De regularizatieparameter vermindert de invloed van risicokenmerken met weinig voorspellend vermogen.
Base_score	0	De startwaarde van het algoritme is gelijk aan 0 euro.
Colsample_bytree	0,9	Het algoritme selecteert per boom willekeurig 90% van de beschikbare vereveningskenmerken om de keuzebomen op te trainen.
Colsample_bynode	0,16	Iedere splitsing van de boom gebruikt willekeurig 16% van de vereveningskenmerken die reeds voor de betreffende *laag* zijn geselecteerd.
Eta	0.5 (op OT-data) 0.4 (op OT- en brondata)	Iedere boom in het algoritme heeft een weegfactor van 0,5 of 0,4 op de totale voorspelling.
Gamma	1	Iedere splitsing vermindert de *loss function* met minimaal 4,8.
Lambda	1	De regularizatieparameter vermindert de invloed van risicokenmerken met weinig voorspellend vermogen.
Max_depth	200 (op OT-data) 10.000 (op OT- en brondata)	Iedere boom bevat maximaal 200 of 10.000 opeenvolgende splitsingen. (dit wordt zeker niet gerealiseerd)
Min_child_weight	16 (op OT-data) 1,5-32 (op OT- en brondata)	Simpel gezegd: nodes die OT-data bevatten van minder dan 16 verzekerden worden niet verder gesplitst. Het model voor OT- en brondata varieert dit aantal tussen 1,5 en 32, afhankelijk van hoeveel opvolgende bomen al zijn gemaakt.
Subsample	0.632	Het algoritme gebruikt per boom de helft van de verzekerden om de betreffende boom te trainen.
Tree_method	"Exact"	Het algoritme overweegt alle beschikbare risicokenmerken om een splitsing op te maken.
Nround	9	Het algoritme maakt gebruik van 9 opeenvolgende keuzebomen die ieder de restfout van de voorgaande boom minimaliseren.

Bijlage F: Eenvoudig voorbeeld Artificial Neural Network

Stel we hebben een neural network met één *hidden layer* met daarin twee neuronen en de *input layer* heeft ook twee neuronen. We modelleren een eenvoudig risicovereveningsmodel met slechts twee kenmerken: geslacht en leeftijd. Figuur 16 geeft het reeds getrainde model weer.

Figuur 16: voorbeeld van eenvoudig neural network dat reeds getraind is.



$f(x) = x^+ = \max(x, 0)$

De gewichten tussen de input layer en de hidden layer vertalen door middel van een functie de inputwaarden naar een signaal voor de neuronen in de hidden layer. Dat kunnen we weergeven als:

$$h_1 = f(x_1 * w_1 + x_2 * w_3 + b_1)$$

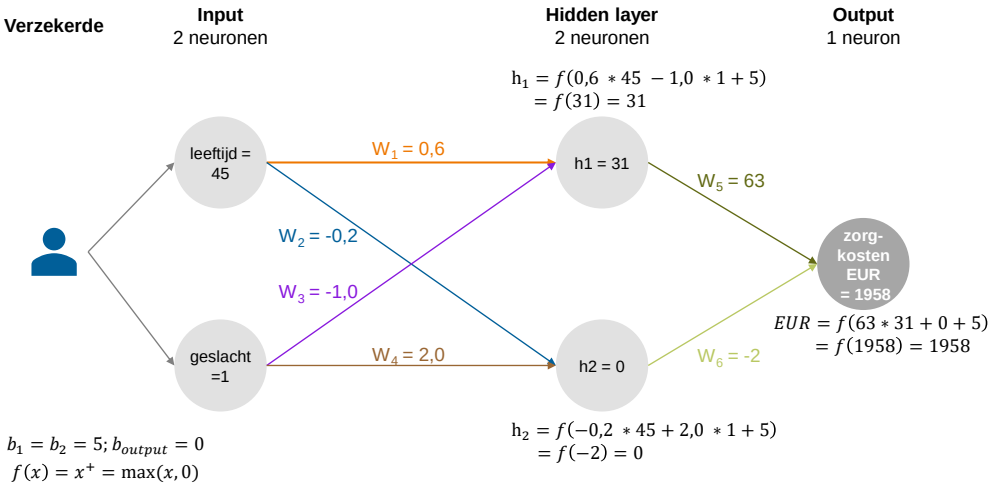
$$h_2 = f(x_1 * w_2 + x_2 * w_4 + b_2)$$

Waarbij (x_1, x_2) de kenmerken (geslacht, leeftijd) zijn van een verzekerde; (w_1, w_2, w_3, w_4) de weegfactoren tussen de neuronen, b_1, b_2, b_{output} zijn de bias van de neuronen en $f(x)$ de activatiefunctie. De activatiefunctie wordt gebruikt om de input in een output te vertalen die ook *niet-lineaire* verbanden kan leren – zonder een dergelijke activatiefunctie zou een neuraal netwerk niet meer kunnen dan OLS. We gebruiken in dit voorbeeld een bias van 5 voor elk neuron, al zou dit in werkelijkheid natuurlijk niet zo hoeven zijn; $b_1 = b_2 = 5; b_{output} = 0$. Verder hanteren we voor het voorbeeld (en ook in het onderzoek) als activatiefunctie de *rectifier*: $f(x) = x^+ = \max(0, x)$.

Voor de vertaling van de hidden layer naar de output layer wordt dit proces herhaald. Voor complexere neural networks kunnen meerdere hidden layers met een groot aantal neuronen gebruikt worden. Het principe blijft echter hetzelfde.

Stel dat we voor een 45-jarige mannelijke verzekerde zorgkosten willen voorspellen. Deze zijn volgens het model 1958 euro; zie Figuur 17.

Figuur 17: voorbeeld van voorspellen zorgkosten met neural network voor individuele verzekerde (man van 45 jaar).



Bij het trainen van het model worden de weegfactoren (w_1, w_2, w_3, w_4) vastgesteld zodanig dat deze voor de trainingsset van verzekerden, waarvan de kenmerken (x_1, x_2) en de zorgkosten O bekend zijn, de R^2 gemaximaliseerd wordt. Om dat te bereiken worden de verzekerden uit de trainingsset in batches meerdere keren aan het model gegeven om de weegfactoren steeds verder te optimaliseren.

Bijlage G: Artificial Neural Network

Voor dit onderzoek is gekozen voor de *keras* implementatie, versie 2.2.5.0, van het Artificial Neural Network algoritme in R. Voor het risicovereveningsmodel zijn de volgende hyperparameters toegepast (niet weergegeven hyperparameters zijn conform de standaardinstellingen van de betreffende R-implementatie).

Hyperparameter	Instelling	Toelichting
Layers 1, 2 en 3	256, 128 en 64 neuronen (op OT-data) 512 en 256 neuronen (op OT- en brondata)	In het model op OT-data bevat de eerste *hidden layer* 256 neuronen, de tweede 128 en de derde 64 (op OT-data). In het model op OT- en brondata bevat de eerste hidden layer 512 neuronen en de tweede 256.
Dropout	0 (op OT-data) 0,35 (op OT- en brondata)	Het model plaatst 0% of 35% van neuronen op non-actief om overfitting te voorkomen.
Activation	'relu' [38]	De neuronen (exclusief het outputneuron) gebruiken een niet-lineaire activatie functie om input te converteren naar output. Dit vertaalt negatieve waarden naar 0 – positieve waarden worden zonder bewerking doorgevoerd naar de volgende laag.
Optimizer	"Adam"	Het *adam* algoritme past de weegfactoren en biases na iedere batch aan.
Batch_size	1.024 (op OT-data) 2.048 (op OT- en brondata)	Iedere batch bestaat uit de gegevens van 1.024 of 2.048 verzekerden. Na iedere batch worden de weegfactoren aangepast.
Epochs	20 (op OT-data) 45 (op OT- en brondata)	De volledige dataset wordt 20 of 45 keer toegepast om het algoritme te trainen.
Shuffle	True	Na iedere *epoch* wordt de data willekeurig verdeeld over nieuwe *batches*.

[38] Rectified linear unit

Bijlage H: Vergelijking maatstaven M1-M5 met OLS op dezelfde datasets

In deze bijlage vergelijken we de maatstaven van alle doorgerekende modellen met de maatstaven van OLS op exact dezelfde dataset. Zo is het effect van het toepassen van een ander predictiemodel goed te beoordelen, zonder dat uitkomsten vertroebeld worden door het effect van het verrijken van de data die modellen tot hun beschikking hebben om de predictie te maken. Alle maatstaven in deze bijlage zijn berekend op de testset.

Met uitzondering van M1 (decision tree) en M2 (Piecewise Regressie) scoren alle modellen beter op de individumaatstaven dan het OLS-model op dezelfde data. Op subgroep- en verzekeraarsniveau is het beeld wisselender. M4 (gradient boosting machine) en M5 (artificial neural network) doen het op de aanvullende data op bijna alle maatstaven beter dan OLS op dezelfde data. Vooral de substantiële verbetering van R^2 voor modellen M4 en M5 ten opzichte van OLS valt op. De aanvullende data verbetert OLS slechts met 1,0-procentpunt (M0 versus M0d1 en M0d2), terwijl dezelfde data M4 en M5 verbetert met respectievelijk 2,2-procentpunt en 1,9-procentpunt.

Tabel 15: maatstaven van relevante modellen in testset op OT-data. Met groen gemarkeerde waarden zijn verbeterd t.o.v. M0; met oranje gemarkeerde waarden zijn verslechterd

Niveau	Maatstaf		M0 OLS	M1 Decision Tree	M3a Random Forest	M4a Gradient Boosting Machine	M5a Artificial Neural Network
Individu	R^2 x 100%		35,1%	35,0%	36,3%	36,3%	36,3%
	CPM x 100%		33,6%	33,6%	34,1%	34,0%	34,1%
	GGAA		1.984	1.983	1.969	1.971	1.968
	Standaarddeviatie resultaten		6.909	6.915	6.845	6.843	6.843
	# met negatieve normkosten		4.412	-	-	19	-
Subgroepen	GGAA op alle subgroepen[a]		1.132	1.124	1.080	1.141	1.099
	Res. 15% laagste kosten in t-3		112	171	159	159	150
	Res. 15% hoogste kosten in t-3		-129	-147	-191	-124	-158
Verzekeraar	R^2 x 100%		98,9%	98,3%	98,8%	98,3%	98,9%
	GGAA		29	38	31	32	30
	Bandbreedte van resultaten	Allen	310	347	315	434	323
		Excl. 2[b]	180	191	167	165	148
		Klein	279	319	265	409	259
		Middel	157	192	172	168	178
		Groot	63	99	79	60	67
		Niet-concern	186	209	167	183	148
		Concern	310	347	315	434	323
	GGARV[c]		-	14	6	6	6

[a] Per model zijn de subgroepen gedefinieerd op basis van de subgroepen uit het uitgangsmodel (1,85 miljoen subgroepen)

[b] Op deze regel staat de bandbreedte van de resultaten op verzekeraarsniveau waarbij de twee risicodragers die steeds de feitelijke bandbreedte bepalen buiten beschouwing zijn gelaten

[c] De GGARV is voor alle modellen vergeleken met M0 – OLS

Tabel 16: maatstaven van modellen in testset op OT-data inclusief leeftijd in jaren. Met groen gemarkeerde waarden zijn verbeterd t.o.v. M0; met oranje gemarkeerde waarden zijn verslechterd

Niveau	Maatstaf		M0b OLS	M2 Piecewise Regressie	M3 Random Forest
Individu	R^2 x 100%		35,1%	35,3%	36,3%
	CPM x 100%		33,6%	32,9%	34,2%
	GGAA		1.984	2.004	1.965
	Standaarddeviatie resultaten		6.909	6.898	6.843
	# met negatieve normkosten		4.411	62.137	-
Subgroepen	GGAA op alle subgroepen[a]		1.131	1.176	1.091
	Res. 15% laagste kosten in t-3		113	140	144
	Res. 15% hoogste kosten in t-3		-132	-105	-131
Verzekeraar	R^2 x 100%		98,9%	98,6%	98,8%
	GGAA		29	31	30
	Bandbreedte van resultaten	Allen	311	394	302
		Excl. 2[b]	180	176	138
		Klein	280	335	258
		Middel	158	179	160
		Groot	63	64	77
		Niet-concern	186	176	138
		Concern	311	394	302
	GGARV[c]		-	3	4

[a] Per model zijn de subgroepen gedefinieerd op basis van de subgroepen uit het uitgangsmodel (1,85 miljoen subgroepen)

[b] Op deze regel staat de bandbreedte van de resultaten op verzekeraarsniveau waarbij de twee risicodragers die steeds de feitelijke bandbreedte bepalen buiten beschouwing zijn gelaten

[c] De GGARV is voor alle modellen vergeleken met M0 – OLS

Tabel 17: maatstaven van modellen in testset op OT-data inclusief brondata. Met groen gemarkeerde waarden zijn verbeterd t.o.v. M0; met oranje gemarkeerde waarden zijn verslechterd

Niveau	Maatstaf		M0d1 OLS	M4 Gradient Boosting Machine	M0d2 OLS	M5 Artificial Neural Network
Individu	R^2 x 100%		36,1%	38,5%	36,1%	38,2%
	CPM x 100%		34,1%	35,7%	33,8%	35,7%
	GGAA		1.969	1.920	1.977	1.920
	Standaarddeviatie resultaten		6.854	6.726	6.855	6.741
	# met negatieve normkosten		4.390	58	3.176	544
Subgroepen	GGAA op alle subgroepen[a]		1.114	1.047	1.108	1.058
	Res. 15% laagste kosten in t-3		108	107	206	111
	Res. 15% hoogste kosten in t-3		-149	-118	-457	-90
Verzekeraar	R^2 x 100%		98,9%	99,0%	97,8%	99,0%
	GGAA		28	26	44	21
	Bandbreedte van resultaten	Allen	294	302	353	235
		Excl. 2[b]	185	138	228	96
		Klein	274	258	333	230
		Middel	160	160	231	138
		Groot	84	77	135	57
		Niet-concern	185	138	250	146
		Concern	294	302	353	235
	GGARV[c]		-	4	-	23

[a] Per model zijn de subgroepen gedefinieerd op basis van de subgroepen uit het uitgangsmodel (1,85 miljoen subgroepen)

[b] Op deze regel staat de bandbreedte van de resultaten op verzekeraarsniveau waarbij de twee risicodragers die steeds de feitelijke bandbreedte bepalen buiten beschouwing zijn gelaten

[c] De GGARV is voor alle modellen vergeleken met M0 – OLS

Bijlage I: Maatstaven op trainingsset via 10-fold cross validation

Tabel 18 geeft de maatstaven die berekend zijn middels 10-fold crossvalidation op de trainingsset. Deze uitkomsten zijn ondergeschikt aan de uitkomsten op de testset, maar geven een extra inzicht in de stabiliteit van de resultaten.

Het uitgangsmodel 2020 is getraind en getest op de volledige dataset en daarom niet representatief voor de vergelijking met de modellen M1-M5. De maatstaven in de kolom M0 – OLS zijn wel via 10-fold cross-validation berekend en daarmee vergelijkbaar met de maatstaven van M1-M5.

Uit de berekende maatstaven van de vijf ontwikkelde modellen en het huidige model is een aantal conclusies te destilleren. De opvallendste uitkomsten binnen deze trainingsset zijn:

- Model M1 - Decision Tree scoort iets minder goed op nagenoeg alle maatstaven dan het huidige model. Op zowel individu-, subgroep-, als verzekeraarsniveau worden de maatstaven negatief beïnvloed.
- Model M2 – Piecewise lineaire regressie behaalt een verbetering op R^2 ten opzichte van het huidige model. Hier staat echter tegenover dat er substantieel meer verzekerden zijn met een negatief normbedrag. Dit hoge aantal wordt veroorzaakt doordat het model een lineair verband veronderstelt tussen leeftijd in jaren en zorgkosten. Voor het leeftijdssegment 0-8 jaar geldt echter dat verzekerden geboren in het vereveningsjaar significant duurder zijn, waardoor dit verband niet bestaat. De gemodelleerde zorgkosten van een deel van de 8-jarigen worden hierdoor negatief. Ook op subgroep- en verzekeraarsniveau is een verslechtering waarneembaar op het merendeel van de maatstaven.
- Model M3 – Random Forest scoort op alle individuele maatstaven een superieur resultaat ten opzichte van het huidige model. In het bijzonder laat de R^2 een substantiële verbetering zien, van 34,0% naar 35,3%. Op subgroep- en verzekeraarsniveau laten verschillende maatstaven zowel een verbetering als een verslechtering zien. Het toevoegen van leeftijd in jaren aan de dataset geeft een lichte verbetering van de meeste maatstaven (M3 t.o.v. M3a). Het zorgt vooral voor een sterke verbetering van het financiële resultaat op de subgroep van 15% duurste verzekerden in t-3. Hier staat echter wel een verslechtering van de bandbreedte van verzekeraars – ook t.o.v. M0 – tegenover.
- Model M4 – Gradient Boosting Machine behaalt het meest gunstige resultaat op de verschillende maatstaven, met vooral op het niveau van individuen een sterke verbetering t.o.v. M0. Opvallend is dat dit gepaard gaat met een sterke ondercompensatie op de subgroep 15% hoogste kosten in t-3. Mogelijk verschuift het model de aandacht naar meer recente informatie, waardoor maatstaven die over het komende jaar gaan verbeteren, terwijl maatstaven die verder terugkijken verslechteren. Er is een duidelijk positief effect te zien van het toevoegen van aanvullende brondata (M4 t.o.v. M4a) op bijna alle maatstaven, met bijvoorbeeld een sterke verbetering van R^2 van 35,3% naar 37,3%.
- Model M5 – Artificial Neural Network behaalt eveneens vooral gunstige resultaten op de verschillende maatstaven. Op individu-, subgroep- en verzekeraarsniveau zijn positieve ontwikkelingen zichtbaar. In het bijzonder valt op dat de bandbreedte van het financieel resultaat op verzekeraarsniveau op bijna alle maatstaven het beste is van alle onderzochte modellen. Ook voor M5 is er een duidelijk positief effect te zien van het toevoegen van aanvullende brondata (M5 t.o.v. M5a) op bijna alle maatstaven verbeteren sterk. Het voorspellend vermogen gaat omhoog van 35,4% naar 37,0%, maar ook het resultaat op de subgroepen verbetert sterk.

Tabel 18: maatstaven van modellen in trainingsset via 10-fold cross validation. Met groen gemarkeerde waarden zijn verbeterd t.o.v. M0; met oranje gemarkeerde waarden zijn verslechterd

Niveau	Maatstaf		Uitgangs-model 2020[a]	M0 OLS	M1 Decision Tree	M2 Piecewise Regressie	M3 Random Forest	M4 Gradient Boosting Machine	M5 Artificial Neural Network	M3a[b] Random Forest	M4a[b] Gradient Boosting Machine	M5a[b] Artificial Neural Network
Individu	R^2 x 100%		34,4%	34,0%	34,0%	34,2%	35,3%	37,3%	37,0%	35,3%	35,3%	35,4%
	CPM x 100%		33,5%	33,5%	33,5%	33,0%	33,9%	35,9%	35,8%	33,8%	34,0%	34,4%
	GGAA		1.980	1.979	1.979	1.994	1.967	1.907	1.912	1.969	1.966	1.953
	Standaarddeviatie resultaten		6.906	6.907	6.909	6.897	6.840	6.735	6.748	6.842	6.842	6.834
	# met negatieve normkosten[c]		15.054	10.773	-	113.548	-	113	836	-	78	-
Subgroepen	GGAA op alle subgroepen[d]		1.011	1.042	1.038	1.079	944	922	947	933	1.026	960
	Res. 15% laagste kosten in t-3		107	107	167	138	142	93	97	155	153	124
	Res. 15% hoogste kosten in t-3		-124	-122	-141	-113	-84	-183	-83	-153	-123	-145
Verzekeraar	R^2 x 100%		99,1%	99,1%	98,5%	98,7%	99,1%	99,1%	99,1%	99,0%	98,6%	99,1%
	GGAA		26	26	34	30	25	33	21	26	27	28
	Bandbreedte van resultaten	Allen	302	297	340	388	320	296	235	292	417	286
		Excl. 2[e]	115	103	182	97	82	102	96	89	104	99
		Klein	284	286	336	348	289	274	230	268	415	258
		Middel	154	153	185	174	150	163	138	158	159	158
		Groot	64	65	91	65	66	70	73	77	58	63
		Niet-concern	167	160	191	151	135	150	146	145	161	145
		Concern	302	297	340	388	320	296	235	292	417	286
	GGARV[f]		3	-	13	6	13	23	17	9	6	9

[a] Als gerapporteerd in WOR 973 en gereproduceerd voor dit onderzoek; het kleine verschil op resultaat 15% laagste kosten t-3 mogelijk door afronding of zeer beperkte verschillen in data

[b] Per model zijn de subgroepen gedefinieerd op basis van de subgroepen uit het uitgangsmodel (1,85 miljoen subgroepen)

[c] Op deze regel staat de bandbreedte van de resultaten op verzekeraarsniveau waarbij de twee risicodragers die steeds de feitelijke bandbreedte bepalen buiten beschouwing zijn gelaten

[d] De GGARV is voor alle modellen, inclusief het uitgangsmodel 2020, vergeleken met M0 – OLS

Tabel 19 beschrijft de R^2 per model en per fold. Dit varieert van 32,5% (OLS op fold 4) tot 38,4% (GBM en Artificial Neural Network op fold 10). Op 8 van de 10 folds behaalt de Gradient Boosting Machine de beste voorspelling. Het Artificial Neural Network algoritme voorspelt het best op de overige twee folds, maar ook op deze folds behaalt de Gradient Boosting Machine slechts 0,2-procentpunt lagere score. Dit laat zien dat modellen M4 en M5 stabiel beter presteren dan de andere modellen. Het is echter geen absoluut bewijs van de 'superioriteit' van deze modellen.

Tabel 19: R^2 van voorspelling per model en per fold; in groen dikgedrukt is de beste voorspelling op een fold; in rood dikgedrukt de minst goede voorspelling

	R^2 x 100% van voorspelde t.o.v. werkelijke zorgkosten; per model en per fold					
Fold	**M0 - OLS**	**M1 - DT**	**M2 - PLR**	**M3 - RF**	**M4 - GBM**	**M5 - ANN**
1	33,6%	**33,3%**	33,6%	34,9%	**36,7%**	36,3%
2	**33,4%**	33,6%	33,9%	34,8%	**36,9%**	36,5%
3	**34,4%**	34,7%	34,5%	35,9%	**37,8%**	37,3%
4	**32,5%**	32,9%	32,9%	34,0%	**35,9%**	35,8%
5	34,2%	**34,2%**	34,4%	35,2%	**37,2%**	37,1%
6	34,4%	34,5%	**34,4%**	35,7%	**37,6%**	37,1%
7	33,8%	**33,0%**	33,1%	34,2%	36,3%	**36,5%**
8	**35,0%**	35,0%	35,4%	36,2%	**38,1%**	37,9%
9	**34,4%**	34,9%	35,2%	35,8%	37,9%	**38,1%**
10	35,0%	**34,9%**	35,4%	36,3%	**38,4%**	38,4%

Bijlage J: Permutation feature importance M3-M5 op OT-data

Tabel 20 geeft de 10 belangrijkste risicoklassen per model op de OT-data, geordend van meest belangrijk naar minst belangrijk. Deze zijn bepaald met behulp van permutation feature importance. Dat wil zeggen dat steeds één risicoklasse gehusseld is in de dataset om te bepalen welk effect dit heeft op het voorspellend vermogen (R^2).

Tabel 20: 10 belangrijkste risicoklassen per model, weergegeven met *x*

Risicoklasse	M0 OLS	M3a Random Forest	M4a Gradient Boosting Machine	M5a Artificial Neural Network
FKG	x	x	x	x
MVV	x	x	x	x
MHK	x	x	x	x
pDKG	x	x	x	x
Leeftijd-geslacht	x	x	x	x
sDKG	x	x	x	x
PPA	x	x	x	x
HKG	x	x	x	x
FDG	x	x		
AVI	x	x	x	x
SES			x	x

Bijlage K: Piecewise Regressie vs. OLS – resultaten per segment

Segment	R^2 M0 (OLS)	R^2 M2 (Piecewise Regressie)	Gewicht (verzekerden in testset)
0 – 8 jaar	16,4%	16,0%	490.607
9 – 16 jaar	39,8%	40,4%	467.937
17 – 26 jaar	46,9%	49,6%	607.153
27 – 33 jaar	34,1%	34,1%	428.693
34 – 41 jaar	38,9%	39,4%	474.640
42 – 53 jaar	38,9%	39,0%	874.041
54+ jaar	36,8%	37,0%	1.708.976
TOTAAL	**35,1%**	**35,3%**	**5.052.047**

Opmerking: het totaalresultaat van de modellen is weergegeven ter referentie – dit is niet een gewogen gemiddelde van de individuele segmenten

Bijlage L: Maatstaven OLS – Piecewise Regressie en OLS op leeftijdssegmenten

Onderstaande tabel geeft de maatstaven op individuniveau binnen de testset van het reguliere OLS-model (M0), Piecewise Regressie (M2) en het aangepaste model dat normbedragen volgens regulier OLS bepaald binnen de segmenten die volgen uit M2.

Het model met OLS op segmenten doet het op alle individuele maatstaven beter dan het Piecewise Regressie model.

Niveau	Maatstaf	M0 OLS	M2 Piecewise Regressie	M2 OLS op segmenten PLR
Individu	R^2 x 100%	35,1%	35,3%	35,4%
	CPM x 100%	33,6%	32,9%	33,8%
	GGAA	1.982	2.004	1.979
	Standaarddeviatie resultaten	6.909	6.898	6.888
	# met negatieve normkosten	4.411	62.137	6.713

Bijlage M: Normbedragen morbiditeitscriteria per leeftijdssegment

Nota bene: voor zowel het uitgangsmodel als het Piecewise Regressie model betreffen het uitkomsten binnen de trainingsset. De normbedragen per leeftijdssegment zijn gebaseerd op de segmenten zoals bepaald door het Piecewise Regressie algoritme en berekend met het reguliere OLS-model pér leeftijdssegment.

Normbedragen in euro's per verzekerdenjaar

	Uitgangsmodel	Piecewise Regressie met als leeftijdssegmenten:						
	2020	0-8	9-16	17-26	27-33	34-41	42-53	54+
Geen pDKG	-212	-79	-43	-40	-57	-69	-134	-478
1	603	711**	456*	640	759	725	625	366
2	1,259	1.069*	784*	1.174	1.086	861	1.050	1.117
3	1,336	1.125	754	1.030	938	882	992	1.287
4	1,689	2.860	2.135	1.678	1.757	1.545	1.473	1.443
5	2,659	2.336*	1.569*	2.218	2.609	2.529	2.335	2.449
6	2,132	3.641	1.900	1.927	2.176	1.829	1.888	1.802
7	4,790	4.996	2.037	1.815	2.760*	4.267	4.935	4.993
8	5,790	6.617	3.247	1.633	3.071*	3.898	5.950	6.573
9	5,863	10.061*	9.485*	2.964*	437*	642*	4.305	5.247
10	7,427	19.045**	2.640**	5.675*	9.499*	7.307*	5.734*	7.687
11	11,862	nvt	nvt	9.509*	5.950*	11.335*	11.344	12.314
12	13,869	15.917*	2.468*	8.959*	13.569*	10.954*	13.657*	16.192
13	6,987	75.091*	27.564*	8.357*	3.798*	3.133*	3.808*	5.414
14	67,149*	22.815*	67.992*	61.107*	89.398**	43.433**	93.876*	58.056*
15	48,298	31.056***	nvt	39.075**	41.803**	39.563*	50.166*	51.953

** Gemarkeerde normbedragen zijn gebaseerd op minder dan 1000 verzekerdenjaren*
*** Gemarkeerde normbedragen zijn gebaseerd op minder dan 100 verzekerdenjaren*
**** Gemarkeerde normbedragen zijn gebaseerd op minder dan 20 verzekerdenjaren*

Normbedragen in euro's per verzekerdenjaar

	Uitgangsmodel	Piecewise Regressie met als leeftijdssegmenten:						
	2020	0-8	9-16	17-26	27-33	34-41	42-53	54+
Geen sDKG	-89	-18	-13	-10	-19	-33	-68	-203
1	932	2.909	952	957	1.097	1.139	1.109	751
2	2,369	3.140*	1.184**	1.077*	3.582*	2.939	2.342	2.204
3	3,892	-242*	5.937*	-1.471*	3.578*	4.185	4.428	4.155
4	7,189	3.958*	3.862*	4.098*	6.802*	10.371*	8.999	6.732
5	13,458	8.137*	7.151*	16.814*	12.225*	15.768*	12.908*	13.654
6	17,150	10.462*	716*	9.546*	15.170**	14.289**	19.656*	19.493
7	67,221*	28.107*	82.936**	52.083**	35.252***	62***	74.519***	33.709**

** Gemarkeerde normbedragen zijn gebaseerd op minder dan 1000 verzekerdenjaren*
*** Gemarkeerde normbedragen zijn gebaseerd op minder dan 100 verzekerdenjaren*
**** Gemarkeerde normbedragen zijn gebaseerd op minder dan 20 verzekerdenjaren*

Normbedragen in euro's per verzekerdenjaar

	Uitgangsmodel	Piecewise Regressie met als leeftijdssegmenten:						
	2020	0-8	9-16	17-26	27-33	34-41	42-53	54+
Geen FKG	-272	-42	-58	-56	-84	-127	-210	-588
1	25	746*	166	244	366	241	87	-202
2	250	349*	2.487*	1.359*	356*	453	611	1
3	117	693***	465	178	96	72	86	-8
4	388	1.351**	320	415	454	304	328	323
5	565	4.075	1.132	666	607	582	330	406
6	773	6.484**	-321**	891*	801	624	1.000	481
7	1,373	8.133*	1.363*	1.365	1.020	918	1.078	1.470
8	2,022	239***	-1.117***	-1.267**	2.950**	1.748*	1.689	1.814
9	1,745	31.945**	14.536**	976*	3.418*	3.344*	2.422	1.446
10	835	9.250***	3.510*	1.521	831	911	815	629
11	1,488	16.223***	3.981***	3.178*	2.088*	2.598	1.443	1.273
12	399	nvt	349**	463*	502*	710	517	95
13	761	nvt	-456***	5.347**	627*	1.080*	787	506
14	1,579	6.466*	2.958	1.622	1.886	1.731	1.448	1.115
15	1,927	20.736***	-5.201**	2.305*	3.571*	2.121	2.171	1.644
16	3,781	15.336*	22.225*	13.166*	9.733*	8.910*	4.127*	1.511
17	2,496	6.707*	3.414	-2.406*	1.214*	1.573*	1.211*	624*
18	2,797	19.255***	7.974**	1.277*	800*	1.837*	3.312	2.949
19	4,009	nvt	5.886***	6.195*	6.017*	5.576	4.116	2.150
20	4,191	5.500**	4.673**	5.794*	5.698	5.389	4.316	3.280
21	524	-2.288**	359*	707*	746*	804	646	266
22	663	-2.849***	1.984*	1.321	1.185	691	679	424
23	684	-838**	-1.475*	1.272*	1.211	1.138	502	483
24	4,823	14.310**	8.567*	5.097	5.340	5.820	4.919	4.285
25	7,329	20.113***	20.540***	1.442**	1.004**	9.458*	8.306*	7.265
26	11,491	25.015***	1.391***	12.706***	24.161**	20.033**	14.897*	10.111
27	8,079	-2.677**	20.823**	8.341**	6.398**	6.630**	7.704*	7.955*
28	405	565	362	374	319	385	393	306
29	1,583	13.592**	3.950**	2.097*	2.397*	2.057	1.658	1.288
30	11,673*	nvt	14.612***	11.484**	11.582**	8.642**	12.245*	12.119*
31	743	-3.290**	1.073*	1.878**	376*	1.139	-123	608
32	1,469	-19.289**	-9.497**	1.699*	2.657*	1.688*	1.816	1.097
33	11,390	51.976*	46.502*	22.112*	9.565*	8.927	9.676	10.441
34	19,845	11.151**	17.471**	26.887**	11.380**	22.314**	25.084*	18.969*
35	120,671*	114.344***	101.367***	93.169***	75.206***	130.923***	121.094**	125.891**
36	242,336**	42.744***	107.595***	305.349***	228.917***	259.941***	202.539**	318.950***
37	386,189**	320.395***	591.782***	427.178***	367.875***	402.387***	348.732***	297.490**

** Gemarkeerde normbedragen zijn gebaseerd op minder dan 1000 verzekerdenjaren*

*** Gemarkeerde normbedragen zijn gebaseerd op minder dan 100 verzekerdenjaren*

**** Gemarkeerde normbedragen zijn gebaseerd op minder dan 20 verzekerdenjaren*

Normbedragen in euro's per verzekerdenjaar

	Uitgangsmodel	Piecewise Regressie met als leeftijdssegmenten:						
	2020	0-8	9-16	17-26	27-33	34-41	42-53	54+
Geen HKG	-48	-30	-17	-10	-13	-21	-32	-94
1	219	-8.511***	910**	206*	191	455	320	103
2	428	100***	1.531**	203	64	369	248	383
3	1,267	2.173*	833*	1.220*	1.822*	1.895*	2.374	1.032
4	1,276	257	635	1.524	1.307*	1.405	1.042	1.398
5	1,808	3.565*	684*	537	1.379	2.191	2.094	1.833
6	2,169	3.479	6.278*	1.988	1.819	1.759	1.606	2.059
7	4,164	14.226*	4.065**	4.316*	2.364*	3.084*	4.187	4.165
8	6,987	7.809	7.579*	4.242*	4.743*	7.388*	7.248*	5.897
9	17,471*	52.209*	13.573**	4.928**	3.771***	-5.052***	7.564**	9.661*
10	9,099	21.455*	19.499*	8.839*	8.574*	10.480*	12.331	6.692

** Gemarkeerde normbedragen zijn gebaseerd op minder dan 1000 verzekerdenjaren*

*** Gemarkeerde normbedragen zijn gebaseerd op minder dan 100 verzekerdenjaren*

**** Gemarkeerde normbedragen zijn gebaseerd op minder dan 20 verzekerdenjaren*

Normbedragen in euro's per verzekerdenjaar

	Uitgangsmodel	Piecewise Regressie met als leeftijdssegmenten:						
	2020	0-8	9-16	17-26	27-33	34-41	42-53	54+
Geen FDG	-20	-43	-13	-4	-4	-4	-9	-36
1	580	686	367	381	1.679	809	636	735
2	1,903	851*	246*	1.245*	2.385*	829*	1.755	1.962
3	1,253	2.577*	-291*	2.215	919*	1.149	1.307	1.294
4	8,302*	4.399**	-431*	-92*	820**	5.386**	6.215**	10.637*

** Gemarkeerde normbedragen zijn gebaseerd op minder dan 1000 verzekerdenjaren*
*** Gemarkeerde normbedragen zijn gebaseerd op minder dan 100 verzekerdenjaren*
**** Gemarkeerde normbedragen zijn gebaseerd op minder dan 20 verzekerdenjaren*

Normbedragen in euro's per verzekerdenjaar

	Uitgangsmodel	Piecewise Regressie met als leeftijdssegmenten:						
	2020	0-8	9-16	17-26	27-33	34-41	42-53	54+
Geen MVV	-185	-35	-20	-11	-14	-21	-46	-491
1	1,226	1.467*	-616*	2.340*	1.120*	1.298*	1.760	871
2	1,860	11.551*	626*	1.302*	1.900*	2.009*	2.345	1.469
3	3,213	5.514*	2.844*	2.170*	3.256*	3.478*	3.610	2.880
4	5,706	5.769*	967*	2.946*	4.477*	5.075*	6.003	5.459
5	8,541	14.321*	13.063*	5.735*	7.762*	8.784*	8.155	8.254
6	12,196	6.314*	11.193*	9.665*	8.283*	12.171*	13.538	11.977
7	17,533	-2.493**	23.724*	11.756*	20.439*	14.122*	18.494*	17.353
8	29,665	-2.643*	5.173*	21.660*	29.212*	31.823*	33.253*	29.910
9	60,336*	54.412*	42.813*	48.439**	nvt	nvt	nvt	nvt

** Gemarkeerde normbedragen zijn gebaseerd op minder dan 1000 verzekerdenjaren*
*** Gemarkeerde normbedragen zijn gebaseerd op minder dan 100 verzekerdenjaren*
**** Gemarkeerde normbedragen zijn gebaseerd op minder dan 20 verzekerdenjaren*

Normbedragen in euro's per verzekerdenjaar

	Uitgangsmodel	Piecewise Regressie met als leeftijdssegmenten:						
	2020	0-8	9-16	17-26	27-33	34-41	42-53	54+
Geen MHK	-578	-268	-243	-268	-424	-422	-467	-1.052
1	143	235	302	384	423	321	240	-211
2	2,400	2.972	3.142	3.198	1.885	2.313	2.511	1.912
3	2,218	2.297	2.730	2.757	2.216	2.374	2.529	1.835
4	3,562	3.068	4.624	4.371	3.192	3.482	3.602	3.248
5	5,470	4.859	6.245	6.113	5.055	4.728	5.498	5.239
6	8,685	8.004	10.942	10.239	8.514	7.632	8.314	8.612
7	18,003	16.539*	22.548*	20.664*	20.575*	20.010*	18.173	17.142
8	43,292	44.202*	81.732*	74.790*	63.756*	48.686*	47.950*	34.722

** Gemarkeerde normbedragen zijn gebaseerd op minder dan 1000 verzekerdenjaren*
*** Gemarkeerde normbedragen zijn gebaseerd op minder dan 100 verzekerdenjaren*
**** Gemarkeerde normbedragen zijn gebaseerd op minder dan 20 verzekerdenjaren*

Bijlage N: Verkennende analyse naar het effect van het opschalen van kosten bij verzekerden die niet het hele jaar verzekerd waren

De huidige praktijk binnen de risicoverevening is om altijd te werken met volledige verzekerdenjaren, waarbij een verzekerdenjaar één verzekerde bij één verzekeraar gedurende een geheel kalenderjaar representeert. Om verschillende redenen kan het echter voorkomen dat personen niet een geheel kalenderjaar verzekerd zijn, of niet het gehele jaar bij dezelfde verzekeraar zijn verzekerd. In die gevallen worden de gemaakte kosten, voordat ze gebruikt worden om het model te maken, lineair geëxtrapoleerd naar een geheel jaar. Wanneer een persoon bijvoorbeeld 4 maanden bij een verzekeraar ingeschreven heeft gestaan, dan worden de gemaakte kosten bij die verzekeraar verdrievoudigd.

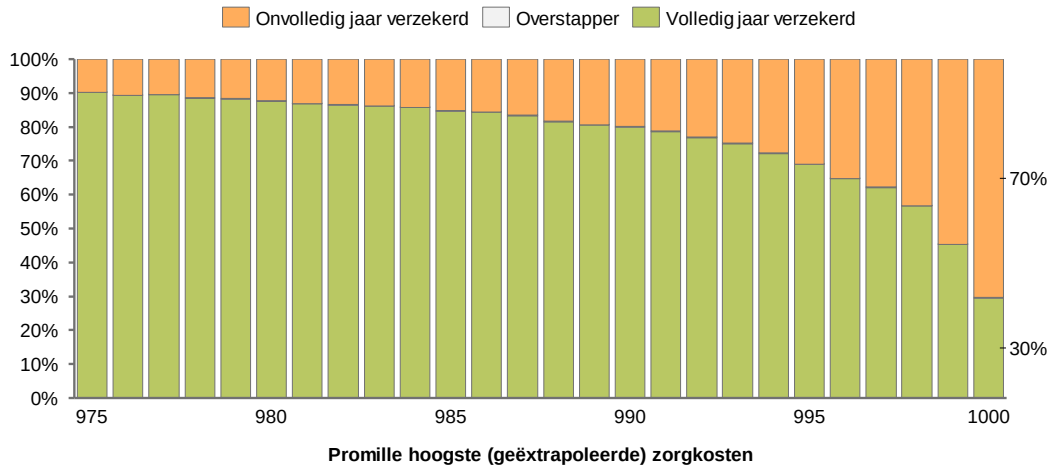
Voor herkenbaarheid van de resultaten sluiten we in dit onderzoek aan bij deze gangbare praktijk van extrapoleren. Voor de geïnteresseerde lezer volgt in deze bijlage een beschrijving van een verkennende analyse waaruit blijkt dat deze methodiek mogelijk tot ongewenste bias kan leiden.

In de OT-dataset onderscheiden we 3 soorten records:

- records van volledige verzekerdenjaren:
- records van personen die zijn overgestapt en waarvoor dus geldt dat het totaal van alle aanwezige records wel een geheel jaar vormt. Deze categorie bestaat voornamelijk uit mensen die 18 zijn geworden, want de meeste andere mensen mogen alleen overstappen bij aanvang van een nieuw kalenderjaar.
- overige records. Deze laatste categorie bestaat bijvoorbeeld uit nieuwgeborenen, overledenen en migranten. Dit zijn records die naar een volledige jaar worden gextrapoleerd.

De totale dataset (na de ultieme test van de machine learning algoritmen) blijkt sterke uitschieters te bevatten in de kosten (tot 11 miljoen euro). Zeer opvallend daarbij is dat de geëxtrapoleerde datapunten sterk oververtegenwoordigd zijn onder deze uitschieters. Figuur 18 hieronder toont hoe de recordtypes zijn verdeeld over de hoogste kosten (overstappers zijn nagenoeg onzichtbaar, omdat deze weinig voorkomen en niet vaak hoge kosten maken).

Figuur 18: aandeel verzekeringsduur voor de 25 promille verzekerden met hoogste (geëxtrapoleerde) zorgkosten



Dit doet vermoeden dat de extrapolatie naar verzekerdenjaren kan leiden tot onrealistisch hoge kosten. Bijvoorbeeld, wanneer een persoon in het begin van het jaar een zware operatie ondergaat, daarna een week op de intensive care doorbrengt en vervolgens overlijdt leidt dit tot zeer hoge kosten wanneer deze naar een volledig jaar worden geschaald. Dit is heel anders dan bijvoorbeeld kosten van medicijngebruik bij chronische aandoeningen, waarvoor het *wel* redelijk lijkt om te veronderstellen dat een persoon die in een jaar 2 keer zo lang leeft, ook 2 keer zoveel kosten maakt.

De vraag is nu of situaties zoals die hierboven beschreven, van zeer hoog (onevenredig) oplopende kosten in de laatste levensmaanden, een uitzonderingen zijn met een verwaarloosbaar effect of niet. Om dit verder te onderzoeken hebben we gekeken naar records van het type 'overig', van mensen van 70 jaar en ouder. We veronderstellen dat de overgrote meerderheid van deze mensen is overleden.[39] Als het effect van de onevenredig oplopende zorgkosten verwaarloosbaar zou zijn zou een lineaire regressie van werkelijke kosten op het aantal verzekerde dagen ongeveer door de oorsprong moeten gaan. In onze verkennende analyse blijkt dit niet het geval, maar vinden we bij een verzekeringsduur van 0 dagen al ongeveer € 4500 zorgkosten. Dit wijst erop dat de kosten van overledenen inderdaad niet geëxtrapoleerd kunnen worden op de wijze zoals nu gebruikelijk.

[39] Deze omweg is nodig omdat overlijden niet als kenmerk in onze dataset aanwezig is