esculine
inzicht in zorgdata

Esculine Voorspelmodel uren – zorgintensiteit thuiszorg

Achtergrond voorspellingmodel wijkverpleging

Esculine helpt zorginstellingen hun data voor zich te laten werken. Laat dat wat je registreert bijdragen tot inzicht en grip op de toekomst. In 2019 is Esculine samen met zorgorganisaties een aantal ontwikkelprojecten gestart waarbij we Machine Learning techniek toepassen om patronen te herkennen uit jouw (big?) data en daarmee ook voorspellingen te kunnen maken.

Esculine is een data specialist op het gebied van (oa) de VVT.

We weten dus welke data VVT organisaties registreren en wat je daarmee kunt doen.

Voor het voorspellingsmodel zorgintensiteit wijkverpleging zien wij 2 toepassingen:
1. Feedback loop richting wijkverpleegkundigen en teams
2. Input voor de bespreking met financiers

In deze factsheet leggen we uit wat dit model doet, hoe het werkt en wat de achterliggende parameters zijn.

Voorspellingsmodel uren thuiszorg

Datasets

Allereerst de datasets. De data wordt in eerste instantie uit het ECD (primair is het model ontwikkeld op data uit ONS van Nedap) gehaald. Dit doen we met behulp van data-query's op de ruwe data waarbij we de benodigde informatie ophalen. Dit bevat informatie over: de cliënten, hun eerste zorgplan, de geregistreerde uren rond dat eerste zorgplan én de OMAHA-classificatie. Deze data wordt met behulp van de bij data scientists populaire programmeertaal R verder verwerkt. Uiteindelijk levert dit een dataset op met per cliënt:

- Het aantal uur zorg dat is geleverd voor dat het zorgplan is opgesteld
- Aantal dagen tussen het eerste zorgmoment en het maken van het zorgplan
- Leeftijd en geslacht van de cliënt
- Ingevulde scores op elk van de 42 OMAHA gebieden (indien ingevuld)
- Een variabele die het aantal OMAHA gebieden bevat waarbij de score 2 of hoger was, dus waarbij echt iets aan de hand was.
- Het aantal OMAHA aandachtsgebieden waarop een actie is uitgezet
- Het aantal uur zorg dat is geleverd in de eerste dertig dagen na het zorgplan*
- Het aantal uur zorg dat is geleverd in de tweede dertig dagen na het zorgplan*. Cliënten die binnen 30 dagen uit zorg gaan zullen hier 0 hebben, dit maakt het voorspellen van deze groep lastiger.

*=dit zijn de twee variabelen die we willen voorspellen

Hierbij worden de patiënten die terminaal zijn niet meegenomen. Dit zijn cliënten waarbij in de eerste zestig dagen tijd is geschreven op de vektiscode voor terminale zorg. De reden hiervoor is dat deze slecht voorspelbaar blijken en vaak (voor het model) onvoorspelbaar veel zorg krijgen. Dit verslechterd het model, en daarbij worden dan ook voorspellingen voor de andere cliënten slechter – het model zal immers proberen ook voor die terminale cliënten beter te voorspellen. Waarschijnlijk is bij deze cliënten bij het maken van het zorgplan al duidelijk dat ze terminaal zijn en kunnen dus ook in toepassing van het model worden uitgesloten.

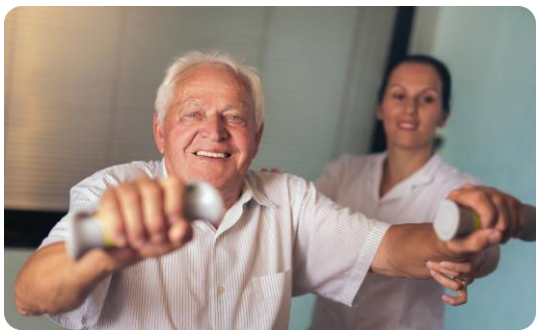
Voordat we gaan voorspellen doen we dan nog een aantal zaken:

- OMAHA gebieden die bij minder dan 20 cliënten zijn ingevuld nemen we niet mee. Op dermate weinig cliënten is lastig een trend te ontdekken.
- Cliënten waarbij minder dan 5 OMAHA gebieden zijn ingevuld nemen we niet mee; bij minder dan 5 gebieden is het lastig voorspellen en is het ook maar de vraag of de OMAHA volledig is ingevuld – er kan immers ook worden ingevuld dat er geen signalen zijn op een bepaald gebied wat dus zou betekenen dat er wel registratie plaatsvindt.
- We vullen alle dan nog lege OMAHA gebieden in met een score van 0 (geen signalen), omdat Machine Learning-modellen niet met lege data om kunnen gaan.
- Verder doen we een logaritme op het aantal uur in de 1e 30 of 2e 30 dagen en gaan we dit voorspellen. De reden is dat de verdeling van aantal uur zorg, ook na het verwijderen van terminale cliënten, logaritmisch verdeeld is en anders zal het model zich vooral richten op het voorspellen van de enkele cliënten met veel zorguren omdat het daar zwaar vanaf hangt. (NB formeel doen we log (uren+1) om negatieve waarde en vooral min oneindig te voorkomen.)

Vervolgens wordt het model geladen in diverse soorten AI modellen die met elkaar vergeleken worden:

- Linear model
- Random forest
- Gradient boosting
- Neural network
- Linear support-vector network

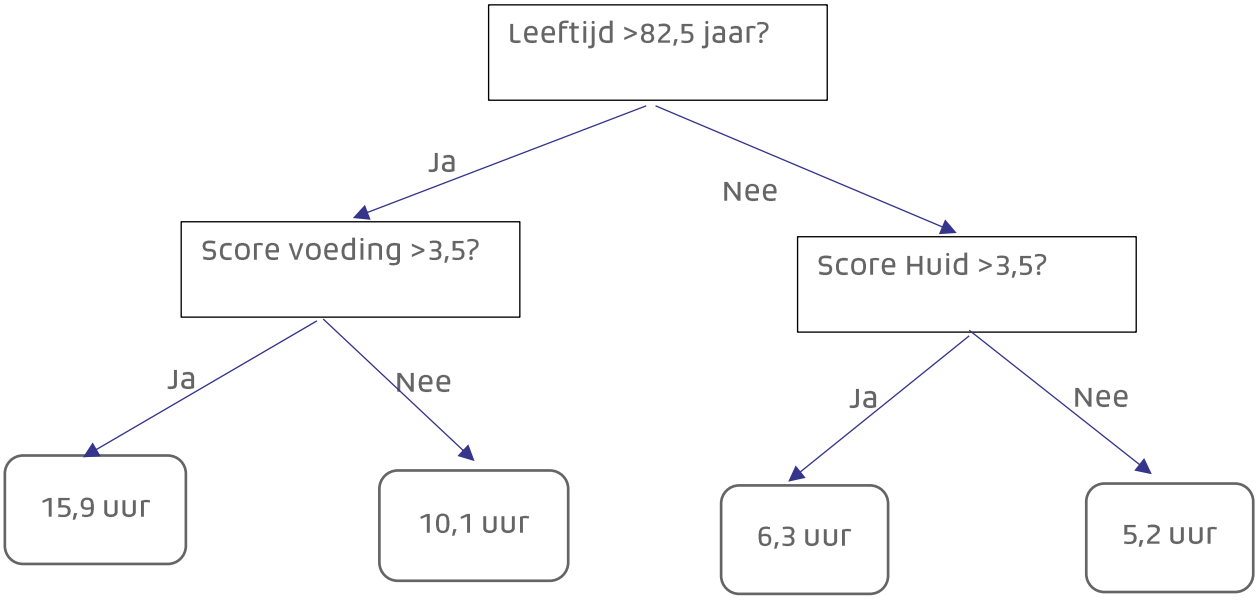
Voor velen zullen dit onbekende termen zijn. Er wordt voorlopig gebruik gemaakt van het Random Forest, dus laten we deze wat nader bekijken.



Ambachtweg 12, 2841 MA Moordrecht
i esculine.nl t +31 182 205 014

Random Forest

Een Random Forest bestaat uit een groot aantal (bij ons voorlopig 1000) beslisbomen. De computer maakt een beslisboom door eerst de variabele met de meeste invloed uit te zoeken (bijvoorbeeld leeftijd of OMAHA-gebied voeding of huid).
Vervolgens wordt gekeken naar waar het verschil het grootst is (bijvoorbeeld 82,5 jaar of score=3,5). Hierop wordt de populatie vervolgens gesplitst, waarna wordt gekeken naar welke variabele binnen elke groep de meeste invloed heeft etc. Dit gaat zo door tot alle variabelen op zijn en dan volgt er een voorspelling. Een versimpeld voorbeeld staat hieronder, waarbij er in de praktijk natuurlijk veel meer lagen zullen zijn.



Dit is dus één boom, maar zoals de naam al aangeeft worden er 1000 van dergelijke beslisbomen gegenereerd en voor elke cliënt wordt het gemiddelde van deze 1000 bomen gebruikt. Zonder ingrijpen zullen deze precies hetzelfde zijn, daarom wordt bij elke van deze beslisbomen een willekeurig aantal van de variabelen niet meegenomen en alleen met de rest een

Trainings- en testset

De set van cliënten wordt hierbij gesplitst voor 75% in de zogenaamde trainingsset en 25% in de testset (hold-out set). De beslisbomen worden gemaakt gebaseerd op de data uit de trainingsset, en daarna wordt gekeken hoe goed ze voorspellen op de testset. Zo kan worden gekeken of de voorspelling goed is op de set die als trainingsdata dient, maar bij andere cliënten een stuk minder presteert (dit heet bias). Daarmee kunnen ook de verschillende modellen en instellingen worden vergeleken.

Bij dit laatste kan gedacht worden aan:

- De beschreven variabelen: minimaal 20 cliënten bij een vraag (is 17 of 22 niet beter?) en minimaal 5 ingevulde vragen voor een voorspelling (zou het model niet ook al prima presteren bij minimaal 3?).
- Variabelen die specifiek over het model gaan. Niet alleen welk model, maar bijvoorbeeld ook hoe veel van deze beslisbomen gebruikt worden (1000).

Dit testen van verschillende parameters kunnen we nog gestructureerder uitvoeren en daarmee kan de voorspelling van het model zeker nog geoptimaliseerd worden. Andere manieren waarmee het model geoptimaliseerd kan worden zijn:

- Meer informatie uit de OMAHA aandachtsgebieden en domeinen gebruiken
- Meer data toevoegen
- Onderscheid maken tussen de verschillende soorten uren zorg

Methodes waarmee de score van een model worden vergeleken:

- Root-mean-square deviation (RMSE), waarbij voor elke voorspelling het kwadraat van het verschil tussen voorspelling en realisatie wordt genomen. Deze worden opgeteld en daarvan wordt de wortel genomen. Dit straft enkele voorspellingen die er ver naast zitten hard af.
- Mean-Absolute-Error (MAE), dit is simpelweg de absolute waarde van het verschil tussen voorspelling en realisatie. Hierbij is er vaak iets naast zitten vaak iets erger dan er een keer een groter stuk naast zitten.

Deze twee maten worden nog steeds op het logaritme van het aantal voorspelde en gerealiseerde uren toegepast. Daarin wordt het model ook getraind.
Daarnaast willen we ook evalueren of het model ook de vele cliënten met enkele of enkele tientallen uren zorg goed voorspelt. Zonder logaritme zouden de enkele cliënten met heel veel zorg dus erg zwaar mee tellen in deze voorspelling.

De uiteindelijke voorspelling wordt toegepast op de gehele set cliënten (zowel trainings- als testset) en hiervan wordt het exponent genomen zodat we het daadwerkelijke aantal uren hebben. Dit wordt eenvoudig weergegeven in

een BI-app, samen met de factoren die relatief belangrijk zijn. Dit is in de beta ontwikkelfase een handmatig proces. Het Esculine Cloud platform is klaar om dit te kunnen automatiseren zodat de voorspelling elke dag actueel is met nieuwe cliënten.

De roadmap van het model is beschikbaar voor geïnteresseerde VVT organisaties en kan worden gedeeld.