DOOR EMMA KREMER, TOM PETERS
EN FRANK VAN BERKUM

Voorspellingen verbeteren dankzij postcodeclusters

Goede risicoclassificatie kan alleen met goede data. Ook als een modelontwerp statistisch juist of voldoende gedetailleerd en complex is, is het van weinig waarde als het niet kan worden gevoed met juiste, relevante, voorspellende data op voldoende detailniveau. In de praktijk is dit een grote uitdaging. Vaak is relevante data beperkt beschikbaar, van slechte kwaliteit of van gevoelige aard, waardoor deze in mindere mate geschikt is voor de ontwikkeling van een model. In dit artikel beschrijven we een methode om bestaande datasets te verrijken met extra variabelen die zijn bepaald aan de hand van postcode om daarmee de kwaliteit van de modelvoorspellingen te verbeteren.

MOTIVATIE GEBRUIK POSTCODECLUSTERING

Voor het sluiten van een verzekeringspolis hoeft een verzekerde meestal slechts beperkt informatie op te geven. De beschikbare informatie over een polishouder bestaat dan uit bijvoorbeeld leeftijd en geslacht van de polishouder, informatie over het verzekerd object (in geval van een schadeverzekering) of zelf-gerapporteerde informatie over gezondheid (leven en zorg), en postcode. Aanvullende objectieve informatie die meer kan vertellen over het onderliggende risico wordt vaak niet uitgevraagd. Wanneer er weinig karakteristieken beschikbaar zijn, bijvoorbeeld in het geval van levensverzekeringen, is het op basis van de beschikbare variabelen zowel lastig om te differentiëren in prijsstelling als ook om het onderliggend risico nauwkeurig te schatten.

De postcode van de polishouder is nagenoeg altijd beschikbaar, maar die is niet direct bruikbaar als variabele in prijsstelling. De zes karakters van de postcode hebben immers geen betekenis en ook lengte- en breedtegraad bevatten doorgaans weinig informatie over het risico dat we modelleren. Wel is het mogelijk om aan de hand van de postcode extra eigenschappen over de polishouder en diens woonadres en -omgeving te verkrijgen. Het toevoegen van alle afzonderlijke variabelen aan het regressiemodel leidt tot een veelvoud aan keuzes. Ook zijn dan de uitkomsten niet altijd eenduidig interpreteerbaar. Daarom vatten we deze informatie samen in variabelen die gebaseerd zijn op clusters van postcodes met vergelijkbare karakteristieken, waarna we deze clusteringen gebruiken als risicofactoren in regressiemodellen. Verschillende typen clusters zijn bijvoorbeeld gebaseerd op eigenschappen gerelateerd aan socio-economische klasse, veiligheid, milieu of voorzieningen rond het woonadres. Een enkel regressiemodel kan meerdere typen clusters gebruiken om te differentiëren tussen polishouders.

BESCHIKBARE DATA

Het CBS[1] publiceert jaarlijks kerncijfers per postcode, buurt, wijk en gemeente. Deze kerncijfers bevatten informatie over verschillende onderwerpen, waaronder: demografische informatie over de inwoners, informatie over inkomen en gebruik van sociale voorzieningen van de inwoners, en gegevens over de beschikbare voorzieningen en gebouwen op de postcode.

Postcode-informatie van verschillende granulariteit kan worden gecombineerd. Kenmerken op basis van de volledige postcode (vier cijfers en twee letters, kortweg PC6) kunnen worden verrijkt met bijvoorbeeld buurtinformatie. Op deze wijze is veel informatie beschikbaar om postcodes te clusteren. Ook kan informatie van andere bronnen worden gebruikt in het clusterproces.

E. Kremer MSc (links) is consultant, T. Peters MSc (midden) is senior consultant en F. van Berkum PhD is senior manager, allen werkzaam bij PwC Risk Modelling Services.



POSTCODES CLUSTEREN

Om per aspect dat we mee willen nemen in het model tot één categorische risicofactor te komen, worden alle postcodes geclusterd op basis van de geselecteerde groep aan kerncijfers die bij dat betreffende aspect horen. De hiervoor gebruikte clustering is een toepassingsveld binnen Machine Learning met als doel om een verzameling datapunten in een beperkt aantal zo homogeen mogelijke groepen in te delen, die onderling zo veel mogelijk van elkaar verschillen. De homogeniteit binnen de clusters wordt bepaald aan de hand van de inputvariabelen. Deze set dient dan ook zorgvuldig gekozen te worden voorafgaand aan het clusteren.

Voor het clusteren van de postcodes zijn verschillende methodes beschikbaar. Wij gebruiken een hiërarchisch clusteralgoritme. Voor dit algoritme dient vooraf het finale aantal clusters bepaald te worden, en het algoritme werkt als volgt:

0) Ieder datapunt (postcode) start als een eigen cluster;
1) Tussen ieder paar van clusters wordt de Euclidische afstand bepaald, op basis van de gekozen set aan variabelen;
2) De twee clusters die het meest op elkaar lijken (en dus de kortste onderlinge afstand hebben) worden samengevoegd tot één cluster;
3) Wanneer het nieuwe aantal clusters nog hoger is dan vooraf gedefinieerde uiteindelijke aantal, worden de stappen vanaf 1) herhaald.

Het resultaat is dat iedere postcode is ingedeeld in één van de clusters op basis van gelijkenis in de geselecteerde kerncijfers.

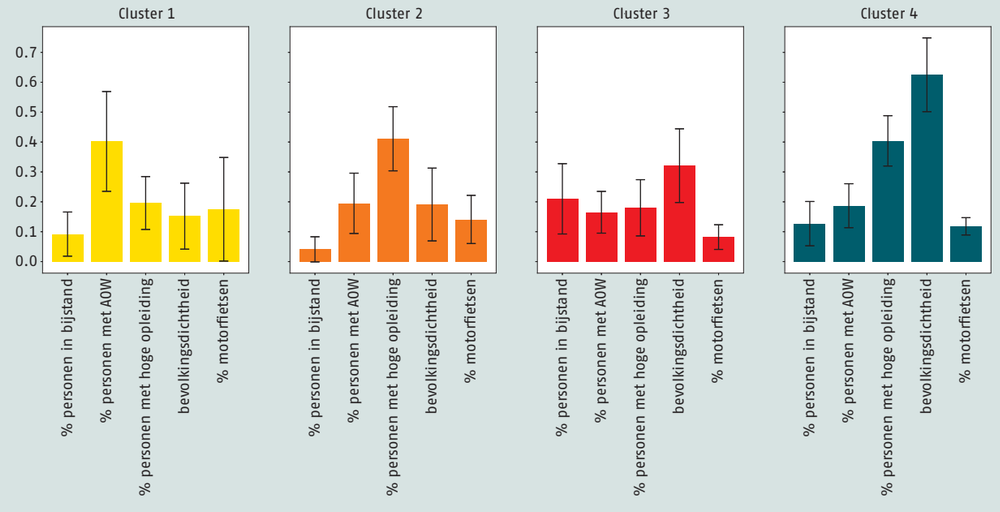
Gelijkenis tussen postcodes en clusters wordt berekend aan de hand van de Euclidische afstand. Het is daarom belangrijk om de variabelen te schalen tussen 0 en 1. Wanneer dit niet wordt gedaan, dan wordt de gelijkenis tussen postcodes gedomineerd door variabelen met grote absolute variantie. De ongeschaalde variabele '*gemiddeld inkomen in Euro*' krijgt dan bijvoorbeeld meer gewicht dan de variabele '*percentage koopwoningen*'.

VOORBEELD

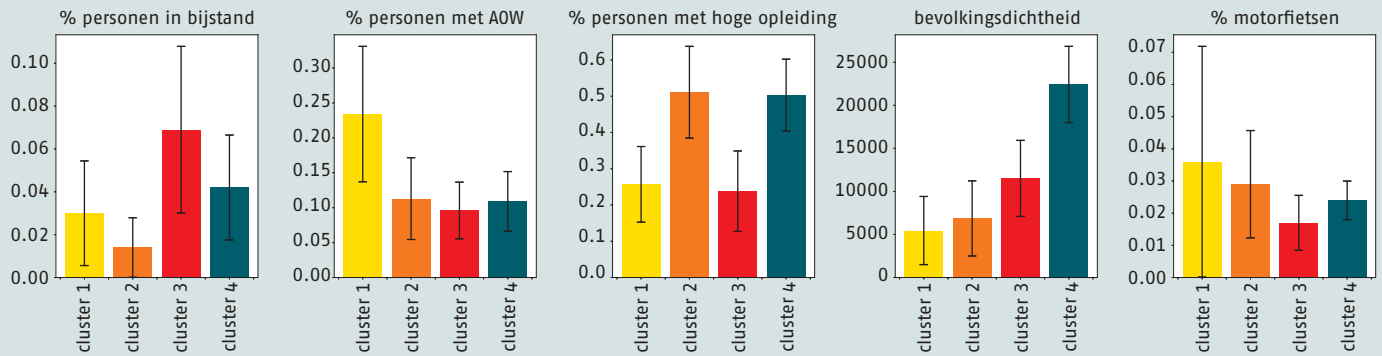
We illustreren de werking en uitkomsten van de clustermethode aan de hand van een vereenvoudigd voorbeeld. Ten behoeve van visualisatie clusteren we buurten in plaats van postcodes; gebruik van PC6 postcodes leidt tot nog specifiekere clustering.

In dit voorbeeld is onze intentie is om postcodes dusdanig te clusteren dat ze informatie bevatten over de socio-economische klasse van de bewoners. We gebruiken hiervoor vijf kenmerken: *% personen in de bijstand*, *% personen met A0W*, *% personen met een hoog opleidingsniveau*, *bevolkingsdichtheid* en *% motorfietsen*. Met uitzondering van *% motorfietsen* zijn deze kenmerken overwegend sociaaleconomisch.

Door toepassing van het hiërarchisch clusteralgoritme hebben we de buurten onderverdeeld in vier clusters. In Figuur 1 tonen we de verdeling van de (geschaalde) kenmerken per cluster. De hoogte van de balk geeft het gemiddelde weer, en de verticale lijn het 95% betrouwbaarheidsinterval. We zien dat er in de buurten in cluster 1 veel ouderen wonen, in cluster 2 met name veel hoogopgeleiden, in cluster 3 de meeste mensen in de bijstand, en cluster 4 bevat de buurten met de hoogste bevolkingsdichtheid.

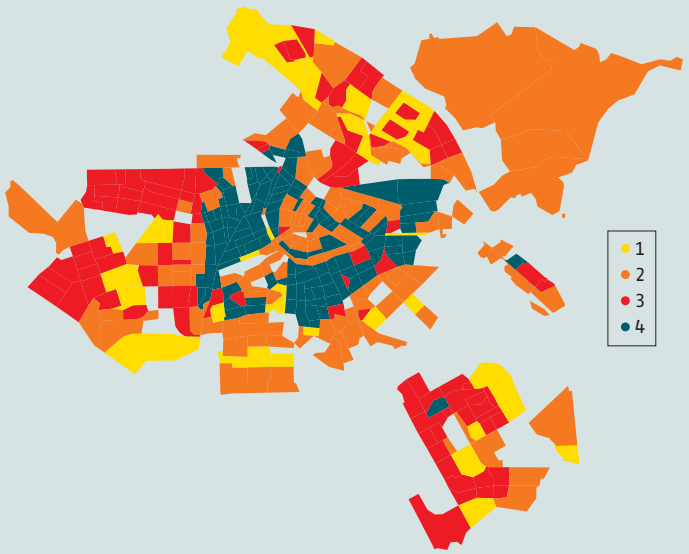


Figuur 1 Kenmerken per cluster (op basis van geschaalde data)



Figuur 2 Verdeling van de variabele over de clusters (op basis van ongeschaalde data)

Figuur 2 laat zien hoe de verdeling van de kenmerken verschilt tussen clusters, wat helpt bij het beoordelen van de relevantie van de variabelen. Zo zien we dat *% personen met een hoog opleidingsniveau* duidelijk verschilt tussen clusters 1 en 3 en clusters 2 en 4. De variabele *% motorfietsen* daarentegen blijkt geen belangrijke bijdrage te leveren aan het onderscheid tussen clusters. Het gemiddelde van de variabele tussen de clusters verschilt weinig, en binnen een cluster zijn de verschillen groot. Dit suggereert dat de variabele *% motorfietsen* een slechte indicator is voor de risicofactor sociaaleconomische status, en deze variabele kan daarom beter niet worden meegenomen in dit voorbeeld.[1]



Figuur 3 Clustering van de gemeente Amsterdam

In Figuur 3 is weergegeven hoe de gevonden clusters in de gemeente Amsterdam zijn verspreid. Volgens verwachting vallen buurten in Oud-Zuid veelal in cluster 2 (veel hoog opgeleid, lage bevolkingsdichtheid en weinig bijstand), waar de buurten in de Bijlmer grotendeels in cluster 3 terecht komen.

ETHISCHE VRAAGSTUKKEN

Bij het gebruik van innovatie, en met name bij nieuwe manieren van dataverwerking, is het van belang stil te staan bij de ethische kant van de toepassing. Discriminerende algoritmes zijn niet toegestaan en toezichthouders DNB en AFM kijken kritisch naar dit aspect. Postcode is mogelijk gecorreleerd met onderliggende, gevoelige en onbewuste kenmerken. Daardoor is het mogelijk dat door postcode zelf mee te nemen in de kalibratie van een model er onbewust gediscrimineerd wordt op gevoelige kenmerken.

Voor een ethisch verantwoorde clustering is het daarom noodzakelijk om kritisch te kijken naar welke variabelen worden meegenomen in het definiëren van de clusters. Hoewel een variabele een duidelijke correlatie kan hebben met een risicofactor, zoals '*aantal inwoners met niet-westerse migratieachtergrond*' met sociaaleconomische klasse, is het vanwege gevoeligheid voor discriminatie aan te bevelen deze variabele niet mee te nemen in het clusteren. Mogelijk zijn er alternatieve variabelen die minstens net zo belangrijk zijn voor de clustering die geen discriminerend karakter hebben, zoals *% sociale huur*.

Een ander aandachtspunt is in hoeverre informatie op postcodeniveau daadwerkelijk iets zegt over het risico van elke verzekerde met dezelfde postcode. Hiervoor is het wenselijk dat binnen een postcode de gebruikte informatie homogeen is. Wanneer dat niet het geval is, dan kan een polishouder worden 'gestraft' voor de eigenschappen van de buren.

MOGELIJKE TOEPASSINGEN

De resulterende postcodeclusters kunnen vervolgens als categorische variabelen in regressiemodellen worden gebruikt. Binnen een regressiemodel kunnen verschillende postcodeclusters worden gebruikt, bijvoorbeeld een sociaaleconomische cluster en een cluster die aangeeft hoe groen de omgeving is.

Postcodeclustering kan breed worden toegepast binnen de verzekeringssector. Een voor de hand liggende toepassing is in de prijsstelling van levensverzekeringsproducten. Sociaaleconomische klasse en sterfte zijn sterk gecorreleerd, en de postcode van een polishouder kan worden gebruikt om de sociaaleconomische klasse van de polishouder te schatten. Een andere toepassing is ten behoeve van preventie in het zorgdomein. Postcodeclusters gebaseerd op het aantal nabije sportfaciliteiten, wandel- en fietspaden en het aantal snackbars in de buurt, geven waardevolle informatie waar preventiemaatregelen waarschijnlijk het meest doeltreffend zijn. Tot slot, in de marketing kunnen specifieke postcodes worden geïdentificeerd op basis van gewenste eigenschappen van de inwoners.

CONCLUSIE

Datasets van verzekeraars bevatten vaak weinig relevante risicofactoren. Op basis van het adres van de polishouder is veel aanvullende informatie beschikbaar. Door vergelijkbare postcodes te groeperen in clusters kunnen regressieanalyses worden uitgebreid met nieuwe variabelen die eenvoudig te interpreteren zijn. Hiermee is meer granulaire differentiatie tussen polishouders mogelijk. Door de postcodes zelf te clusteren, zijn de resulterende clusters zowel goed uitlegbaar naar management en toezichthouders als ethisch verantwoord, iets wat in de huidige tijd een belangrijk goed is.■

Referenties

(1) https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/gegevens-per-postcode

1 – Voor een ander type clustering kan % motorfietsen uiteraard wel relevant zijn.