SCRIPTIE

Cluster-gedreven Risicoclassificatie

Optimalisatie van Autoverzekeringsrisicomodellen door middel van Postcode- en Kentekenclustering

Clusteranalyse is een veelgebruikte techniek in statistische data-analyse en Machine Learning die groepsstructuren binnen datasets kan achterhalen. Deze methode groepeert objecten zodat zowel de heterogeniteit tussen de resulterende clusters, als de homogeniteit onder datapunten binnen hetzelfde cluster wordt gemaximaliseerd. Dit heeft actuariële toepassingen. Zo kunnen clusteringmethodes waardevol zijn voor het creëren van groepen van polishouders, waardoor de risicoclassificatie kan worden verbeterd. Echter worden clustering-technieken nog steeds onderbenut in de actuariële sector. Dit is grotendeels te wijten aan de gemengde data types (numeriek, categorisch en ordinaal) die in dit vakgebied worden gebruikt, terwijl veel clusteringtechnieken afhankelijk zijn van de Euclidische afstand tussen numerieke datapunten.

Mijn scriptie had als doel om de risicoclassificatie van de claimfrequentie modellen in autoverzekeringen van Achmea te verbeteren door clusteringtechnieken toe te passen op data gekoppeld aan postcodes (bijvoorbeeld urbanisatieniveau en gemiddeld inkomen) en kentekens (bijvoorbeeld automerk en gewicht van de auto).

A.J. Wijker MS is Junior Capital Model Specialist bij Rabobank.

METHODE

Voor mijn scriptie heb ik gefocust op de polisdata van twee verschillende dekkingen:
de wettelijke aansprakelijkheidsdekking (WAM), die de schade dekt die de bestuurder aan andermans auto aanricht; en de aanrijdingsdekking (ARD) dat onderdeel uitmaakt van de volledige casco dekking en de schade dekt die de bestuurder aan diens eigen auto aanricht.
Voor het berekenen van de premies van deze dekkingen wordt de verwachte claimfrequentie van de polishouder berekend met behulp van een GLM (Generalized Linear Model). Dit model gebruikt alleen risicofactoren waarvan in het verleden al is vastgesteld dat ze een significant effect hebben op de claimfrequentie. Dit betekent dat variabelen die geen directe impact hebben op de claimfrequentie, niet worden meegenomen als risicofactoren en dus dat het effect van de combinaties van deze variabelen verloren gaat. Door de resulterende clusters op te nemen als risicofactoren in de claimfrequentie GLM's, worden variabelen combinaties wel meegenomen wat zorgt voor een persoonlijkere premie prijsstelling.

Er zijn twee clusteringmethodes onderzocht;
- K-prototypes: clustering gebaseerd op de gelijkenis in afstand tot de middelpunten van de clusters. Gekozen omdat dit één van de meest gebruikte technieken is en de implementatie nodig is voor spectral clustering.
- Spectral clustering: clustering gebaseerd op het spectrum (oftewel de eigenwaardes) van de Laplacian matrix. Gekozen voor de effectiviteit van de techniek met betrekking tot grote, niet-lineair scheidbare datasets. Echter eist deze techniek veel opslag, dus is er gekozen voor U-SPEC als observatiereductietechniek. Dit betekent dat alle kentekens/postcodes worden geclusterd, maar dat deze clustering gebaseerd is op een kleinere selectie van de gehele dataset.

Voor dit artikel wordt niet ingegaan op de technische werking van deze methodes, maar dit is terug te vinden in mijn scriptie.

De twee methodes zijn aangepast zodat ze met gemengde datatypes om kunnen gaan. Zo is er een aangepaste afstandsmaat ontwikkeld die de Euclidean (voor numerieke data), Hamming (voor categorische data), en Gower's (voor ordinale data) afstanden combineert.

Bovendien onderzocht ik in mijn thesis de implicaties van de observatiereductietechniek in spectral clustering met betrekking tot high dimensional clustering. Hierbij is ook een nieuwe manier ontwikkeld om het aantal clusters te bepalen aan de hand van het aantal informatieve eigenvectoren in de Laplacian matrix.

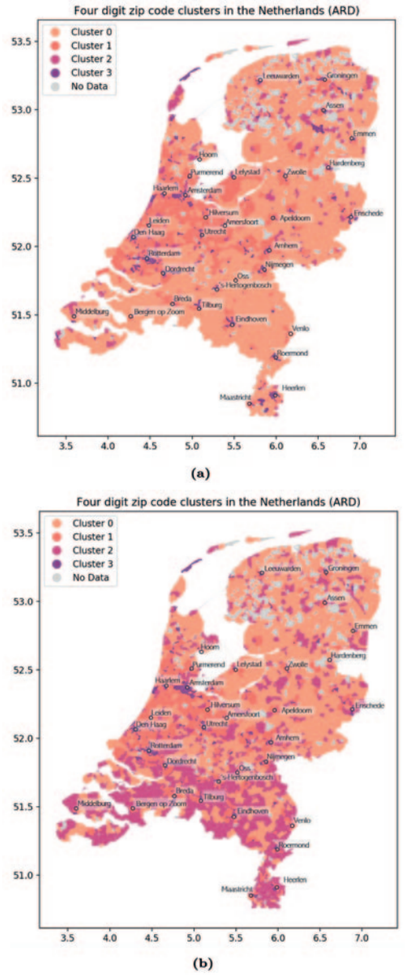
RESULTATEN

Beide clusteringtechnieken zijn toegepast op de kenteken- en postcodedata voor zowel de WAM als ARD dekking. Om deze acht clusteringresultaten te evalueren, analyseren actuariële experts de mate van logica van de clusters. Bovendien worden de clusters meegenomen als risicofactoren in de huidige GLM om de impact op modelmetrieken zoals de deviance, AICc, en BIC te achterhalen.

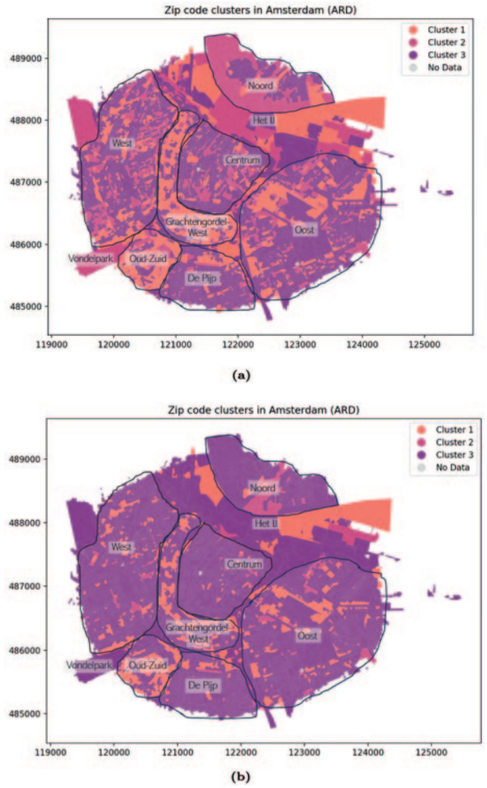
Spectral clustering presteerde beter dan K-prototypes (in deze context) en verbeterde de risicoclassificatie van de ARD dataset als de kenteken- en postcodeclusters worden meegenomen in de claimfrequentie GLM. Dit laatste bleek uit een verlaagde AICc en BIC, wat duidt op een betere fit van de GLM zonder het risico van overfitting.

In Figuur 1 zijn, per viercijferige postcode, de ARD clusters met a.) K-prototypes en b.) Spectral Clustering te zien. Hoe donkerder de kleur, hoe hoger de claimfrequentie volgens de clustering. Het valt op dat de ARD claimfrequentie in beide gevallen hoger is in de steden. Bovendien is de frequentie in het zuiden hoger volgens spectral clustering dan volgens K-prototypes. Dit is een fenomeen dat in de werkelijkheid ook wordt waargenomen. Dit betekent dat spectral clustering logischere resultaten oplevert. Bovendien zijn de resulterende clusters homogener aangezien ze minder ruis bevatten. Dit is te zien in Figuur 2 dat de ARD-postcodeclusters in Amsterdam weergeeft. Ook valt hier op dat welvarende buurten zoals Oud-Zuid en Grachtengordel West lagere gemiddelde claimfrequenties hebben volgens beide clusteringmethodes.

De clustering verbeterde niet de huidige GLM voor de WAM dataset, maar de spectral clusteringtechniek liet potentie zien voor toepassingen op polisdata van andere verzekeringen.



Figuur 1: De ARD clusters per viercijferige postcode met a.) K-prototypes en b.) Spectral Clustering. Hoe donkerder de kleur, hoe hoger de claimfrequentie volgens de clustering.



Figuur 2: De ARD clusters per postcode in Amsterdam met a.) K-prototypes en b.) Spectral Clustering.

CONCLUSIE

Kortom, de huidige risicoclassificatie voor autoverzekeringen van Achmea is verbeterd door het toepassen van clusteringtechnieken, waarbij gebruik werd gemaakt van variabelenselectie, een geavanceerde afstandsmaat en spectral clustering. Hoewel Machine Learningmethoden zoals clustering complex en moeilijk te doorgronden kunnen zijn, is het mogelijk ze effectief in te zetten bij pricing zonder dat het proces een 'black box' wordt. Hiervoor is het essentieel om de balans te vinden tussen technologische complexiteit en transparantie. Door aandacht te besteden aan het interpreteerbaar houden van de resultaten, kan de inzet van geavanceerde Machine Learningmethoden, zoals clustering, ervoor zorgen dat verzekeraars niet alleen nauwkeuriger prijzen (en hiermee laagrisicoklanten behouden), maar ook dat ze hun beslissingsprocessen op een begrijpelijke manier kunnen uitleggen aan hun polishouders. ■

Literatuur

Wijker, A.J. (2024). Cluster-Driven Risk Classification: Adapting Car Insurance Risk Models through Zip Code and License Plate Clustering.