



AI en Data Science: een nieuwe dimensie aan datakwaliteitcontroles

En een kans voor de actuaris

Datakwaliteit staat momenteel hoog op de agenda van financiële instellingen. Zo ook bij pensioenfondsen, waar dit mede is ingegeven door de transitie naar het nieuwe pensioenstelsel en het belang van goede datakwaliteit voor de individuele deelnemer. Niet voor niets heeft toezichthouder DNB dit als een van haar toezichtthema's voor 2021 – 2024 gedefinieerd¹ en heeft de AMF aandacht voor het onderwerp in relatie tot kwaliteit van communicatie². Het gevolg hiervan is dat ook u, als actuariel professional, steeds meer te maken krijgt met dit onderwerp.

Daarbij ziet u wellicht met verbazing dat pensioenfondsen al enige jaren bezig zijn met allerlei datakwaliteitsonderzoeken, hiervoor 'assurance verklaringen' worden afgegeven en dat toch dezelfde pensioenfondsen nu wéér bezig zijn met soortgelijke datakwaliteitsonderzoeken en het zoeken naar assurance. Dat roept de vraag op: is deze aanpak wel (kosten)efficiënt of is hier sprake van dweilen met de kraan open? Liggt de nadruk te veel op het meten van datakwaliteit *achteraf* en niet genoeg op het meten van datakwaliteit *vooraft*? In dit artikel gaan we nader in hoe opgekomen trends in data science en kunstmatige intelligentie een nieuwe dimensie toevoegen aan datakwaliteitcontroles aan de voorkant en hoe deze trends leiden tot een heel ander speelveld met alle bijbehorende uitdagingen én kansen voor u als actuariel professional. Met deze inhoudelijke baggage kunt u als actuariel professional de juiste vragen stellen als het gaat over het onderwerp datakwaliteit.

M. Verschuren AAG MSc (links) is partner en A. Jeurninck MSc is senior management consultant bij strategisch adviesbureau AethiQs.



ONZE VISIE

Onze visie is dat de beheersing op de datakwaliteit van 'nieuwe' data een grote impact heeft op de datakwaliteit. Immers, pensioenfondsen kunnen er alles aan gedaan hebben om een juiste stand van zaken te hebben, maar processen die niet goed lopen en/of onjuiste invoer van nieuwe data kunnen leiden tot vervuiling van deze stand. Om dit te voorkomen worden *business rules* ingezet om op voorhand foutieve data op te sporen. Deze business rules zijn vaak gebaseerd op ervaringen en incidenten uit het verleden en/of op de kenmerken van de pensioenregeling.

Nadelen van het huidig denkmodel

Business rules hebben echter twee nadelen. Allereerst zijn ze ingericht op basis van *bekende* fouten en kunnen ze daarmee *nieuwe* fouten doorlaten. Ten tweede is het inrichten van nieuwe regels tijdsintensief en daarmee duur. Juist op deze twee punten kunnen data science en kunstmatige intelligentie algoritmes helpen en ook het meenemen van kennis van andere sectoren.

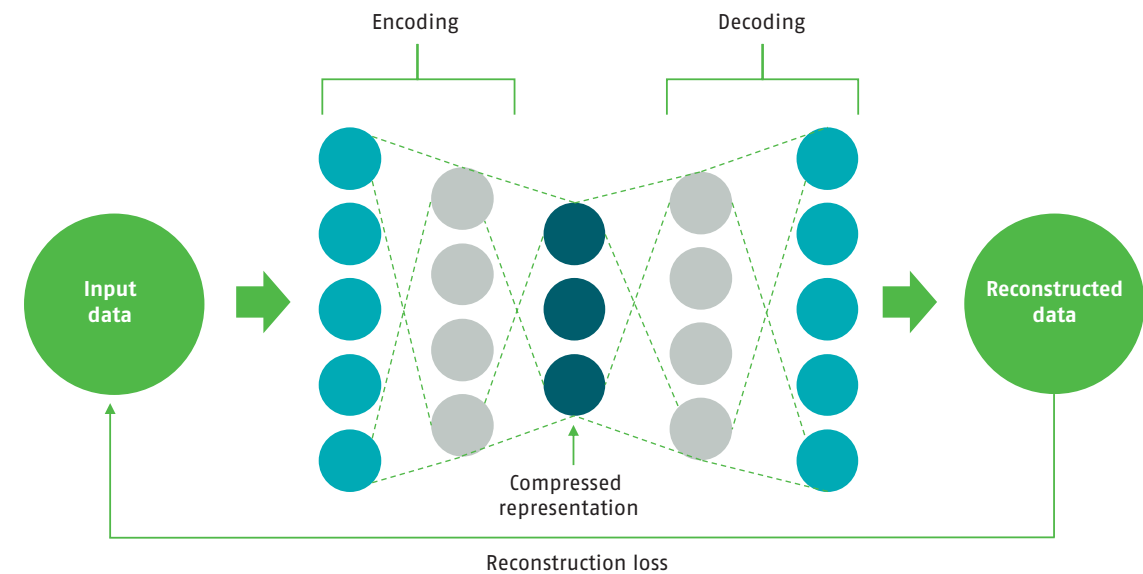
CROSS-SECTORAAL VERBINDEN

Door als actuariel professional met 'lenigheid van geest' over de grenzen van de eigen sector heen te kijken, komt u erachter dat dergelijke algoritmes al in andere sectoren ingezet worden. En waarom zou wat in de ene sector werkt ook niet in de andere kunnen werken? Immers, ook andere sectoren hebben te maken met grote hoeveelheden data waarbij slim ingerichte datakwaliteitcontroles essentieel zijn voor het bestaansrecht van deze bedrijven. Denk hierbij aan alle data die wij zelf achterlaten met smartphones, maar ook bijvoorbeeld data uit sensoren in de industrie en in productieprocessen. Deze data worden gebruikt om prestaties te optimaliseren, maar anderzijds ook om afwijkingen op voorhand te detecteren. Zo kunnen bijvoorbeeld beeldopnames van voedingsproducten door een beeldherkenning-algoritme gehaald worden om te analyseren of de producten aan de benodigde kwaliteits- en veiligheidsstandaarden voldoen, vóórdat deze naar de winkels rollen³.

Zoek hierin de parallellen met de pensioensector en u komt tot het inzicht dat dergelijke technieken ook ingezet kunnen worden om 'nieuwe' fouten in de data te voorkomen en te blijven leren van nieuwe inzichten, zonder tijdsintensieve inregelkosten. U bent als actuariel professional de ideale schakel tussen techniek en sector inhoudelijke kennis en u heeft de baggage om dergelijke verbindingen te leggen. Hieronder maken wij deze verbinding voor twee toepassingsmogelijkheden.

1: Nieuwe invoer. Het detecteren van afwijkingen met deep learning

Via de pensioenaangifte komt nieuwe data de systemen binnen. Nieuwe data die ook nieuwe fouten kan bevatten, maar het is tijdsintensief om deze grote hoeveelheden data te controleren. Om dit schaalbaar te doen, putten we in dit voorbeeld uit Autoencoder (deep learning) modellen⁴.



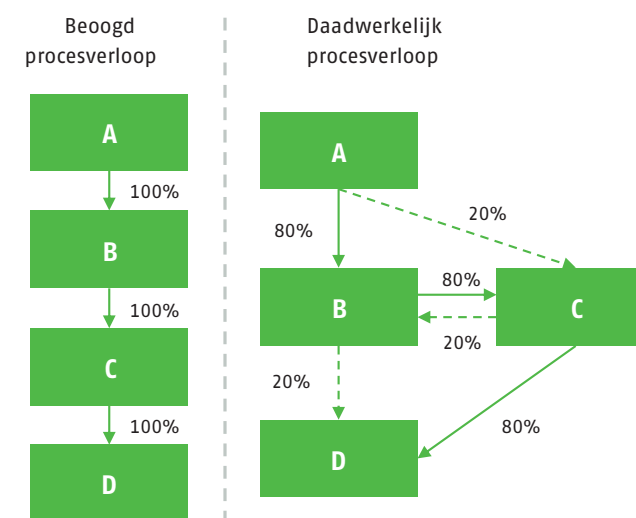
Figuur A: visuele impressie Autoencoder model.

De Autoencoder wordt getraind op de input data en bestaat uit twee onderdelen, namelijk een encoding en een decoding onderdeel. Binnen het encoding onderdeel wordt de input data, met vaak vele data attributen/variabelen, getransformeerd naar een lagere *compressed representation* met daarin de belangrijkste patronen van de data. Deze transformatie bestaat uit vele duizenden tot miljoenen non-lineaire (functie)combinaties van de input variabelen. Vervolgens wordt in het decoding onderdeel de input data weer gereconstrueerd vanuit deze gecomprimeerde representatie. De datapunten die door het model het minst goed kunnen worden gereproduceerd (met een hogere reproduction loss), zijn daarbij afwijkend van de belangrijkste 'normale' patronen van de data.

Door data vanuit de pensioenaangifte keten door het model te laten analyseren, worden afwijkingen met een hoge reproductie loss eruit gefilterd. Dit is een schaalbaar concept dat geautomatiseerd op een datastroom kan worden toegepast.

2: Processen. Analyse van niet goed lopende processen met process mining

Naast fouten door data input kunnen er ook fouten ontstaan door processen die verkeerd lopen. Door het toepassen van process mining krijgt u zicht op de processen zoals deze daadwerkelijk verlopen, hoe vaak het proces verkeerd verloopt en waar de bottlenecks zitten⁵. In andere sectoren wordt dit bijvoorbeeld toegepast op online bestel- en betalingssystemen⁶. Maar hoe werkt dit in de praktijk?



Figuur B: visuele impressie process mining.

Op papier volgt het proces de volgorde A > B > C > D. Echter ontvangt u signalen uit een steekproef (bijvoorbeeld als onderdeel van een ISAE audit) dat in de praktijk stap C in een aantal gevallen voor stap B wordt uitgevoerd. Dan is dit een verhoogd risico in het proces dat kan leiden tot datakwaliteit issues. Door de event log data te analyseren kan achterhaald worden hoe de processen daadwerkelijk lopen. Eén van de algoritmes daarvoor is het Alpha-algoritme, dat een matrix bouwt waarin alle relaties tussen processtappen zijn opgenomen. Vervolgens worden de processen in een zogenaamd petrinet (zie figuur B) gevisualiseerd, waarmee het daadwerkelijke verloop van processen inzichtelijk wordt.

Op basis hiervan kunnen concrete oplossingsrichtingen worden geformuleerd, waarmee de datakwaliteit van de toekomstige data aanvullend beheerst wordt.

WRAP UP

Data science en kunstmatige intelligentie brengen een nieuwe dimensie aan datakwaliteitcontroles. Daarbij kunnen we leren van hetgeen in andere sectoren al toegepast wordt. Het is een realistisch streven om deze technieken ook in te zetten in de pensioensector, om de beheersing op datakwaliteit aan de voorkant in te richten en foute data en verkeerd lopende processen te weren uit de administratie. Wanneer geeft u een nieuwe dimensie aan uw datakwaliteit controles? ■

1 – https://www.dnb.nl/media/43cnkbox/visie_op_toezicht_2021_2024.pdf

2 – <https://www.afm.nl/nl-nl/professionals/nieuws/2022/februari/pensioenadministratie-pensioencommunicatie>

3 – <https://www.sciencedirect.com/science/article/pii/S2665927121000228>

4 – https://www.researchgate.net/publication/329428050_Using_Autoencoder_to_Generate_Data_Quality_Tests

5 – <https://link.springer.com/book/10.1007/978-3-662-49851-4>

6 – <https://www.linkedin.com/pulse/value-process-mining-procure-pay-djalma-britto>