DOOR RAYMOND VAN ES
EN HENNY VERHEUGEN

‘Fit for use’ – een pragmatisch AI-beoordelingskader voor verzekeraars

Kunstmatige intelligentie is allang geen vergezicht meer voor de verzekeringssector. Prijsstelling, schadebehandeling, fraudedetectie, kapitaal-modellering en zelfs klantinteractie maken steeds meer gebruik van Machine Learning (ML) en Large Language Models (LLM’s).

R. van Es (links) is Principal bij Milliman.

H. Verheugen is Principal & Consulting Actuary bij Milliman.

Deze toepassingen brengen, naast nieuwe kansen en mogelijkheden, ook een aantal risico’s met zich mee:

- Juridische claims, bijvoorbeeld door het overtreden van de AI-act of de AVG;
- Reputatieschade door een bevooroordeeld algoritme;
- Strategische, financiële en operationele risico’s door foutieve voorspellingen;
- Technologische risico’s door het lekken van informatie of het van buitenaf manipuleren van het algoritme.

Veel van de risico’s van AI-modellen worden gelopen na de implementatie van de modellen en door onjuiste ontwikkeling en beheer van de AI-modellen. Denk aan falende systemen en onvoldoende beveiliging. Onjuist gebruik leidt ook zeker tot risico’s die vooral gerelateerd zijn aan misinformatie en onjuiste beslissingen.

Om AI veilig en gecontroleerd in te blijven zetten is het daarmee noodzakelijk voor deze modellen, net als voor actuariële modellen, een validatiebeleid te ontwikkelen waarmee kan worden getoetst of een AI-model (nog steeds) doet wat het zou moeten doen en of het voldoet aan de wetgeving en eisen van de toezichthouders.

Om AI-modellen, of modelaanpassingen, gedurende hun levenscyclus te toetsen hebben wij een AI-validatie aanpak ontwikkeld. Deze kan worden toegepast op zowel Machine Learning modellen als op GenAI gebaseerde modellen en bestaat uit twee dimensies om te bepalen of een model ‘fit for use’ is:

(1) Compliance; voldoet het model aan de wetgeving en de eisen van de toezichthouders? Door de diversiteit aan AI-modellen en de daaraan verbonden risico’s, bestaat er geen one size fits all-set aan eisen. Afhankelijk van het type model en de risico’s geldt een passende set aan eisen en toetsingscriteria.

(2) Correct functioneren; werkt het model accuraat, betrouwbaar en verklaarbaar?

Om het functioneren van een AI-model te toetsen volgen we de ontwikkelcyclus van een AI-model. Daartoe gebruiken we een 9-stappenmodel. Per stap zijn er meerdere criteria waaraan moet zijn voldaan om uiteindelijk het label ‘fit for use’ te verkrijgen. Deze 9 stappen zijn:

1. *Businessvertaling* – is het doel helder en ethisch verdedigbaar?
2. *Datavoorbereiding* – zijn de databronnen relevant en betrouwbaar?
3. *Feature-engineering* – zijn de gebruikte features relevant?
4. *Data-analyse* – hoe zijn de train/test/validatie sets tot stand gekomen en is de mogelijke aanwezigheid van bias getoetst?
5. *Modellering* – past het algoritme bij de use-case en hoe is de modelkeuze tot stand gekomen?
6. *Evaluatie* – zijn performance, robuustheid en fairness gemeten ten opzichte van een baseline?
7. *Implementatie* – is de code veilig, is er adequaat versiebeheer en hoe is het model procesmatig ingebed?
8. *Gebruik* – begrijpen eindgebruikers de uitkomsten en beperkingen van het model en loggen zij incidenten?
9. *Hertraining & Review* – zijn er triggers voor data en model drift, wetswijziging of nieuwe data?

Net zoals bij de ontwikkeling en het gebruik van actuariële modellen wordt gesteund op adequate modelgovernance en documentatie.

Voor een goede validatie is het essentieel om kwalitatieve vereisten te vertalen naar kwantitatieve criteria. Termen als performance, robuustheid, fairness moeten getoetst kunnen worden aan objectieve en zoveel mogelijk kwantificeerbare normen. Bij het vaststellen van de normen wordt rekening gehouden met onder andere het type model, het gebruik en de risico’s van het model. Statistische kennis en kennis van data science zijn daarvoor essentieel.

FAIRNESS EN ETHIEK

Speciale aandacht bij valideren dient te worden gegeven aan het begrip fairness. Zowel vanuit compliance oogpunt als reputatierisico is het essentieel om te toetsen of een AI-algoritme bepaalde groepen bevoordeeld of benadeeld. Eerlijkheid kent echter een verschillende definities. Gelijkheid behandelt bijvoorbeeld iedere polishouder identiek terwijl gelijkwaardigheid juist corrigeert voor achterstand. Bias sluimert meestal in data en kan veroorzaakt worden door onder- of overrepresentatie van bepaalde groepen, historische vooroordelen of proxy-variabelen die correleren met beschermde kenmerken. Dit kan door statistische testen als demographic parity (krijgen verschillende groepen dezelfde uitkomst), equalised odds (is de modelnauwkeurigheid hetzelfde voor verschillende groepen) of predictive parity (is de voorspelling even betrouwbaar voor verschillende groepen).

Het is van belang bij de ontwikkeling van een AI-toepassing vooraf te bepalen hoe eerlijkheid wordt gedefinieerd en tijdens de validatie te toetsen of aan deze definitie wordt voldaan. Het validatieraamwerk checkt ook of de model ontwikkelaars passende maatregelen tegen bias hebben genomen (bijvoorbeeld het balanceren van de data bij de preprocessing, of het aanpassen van de lossfunctie bij het trainen van het model, of het weergeven van de resultaten in de postprocessing) en deze goed hebben gedocumenteerd.

VALIDEREN VAN GENAI-MODELLEN

Hoewel het valideren van AI-toepassingen die op GenAI-modellen gebaseerd specifieke uitdagingen kent, kan dezelfde validatie-systematiek worden gehanteerd. Ook voor deze toepassingen zijn de data waarmee het foundationmodel wordt bijgetraind essentieel, net als de keuze van het GenAI-model, de instellingen van de hyperparameters (bijvoorbeeld temperatuur) en het opbouwen van de prompt. Daarnaast zijn er specifieke criteria om te testen of een GenAI-model voldoende nauwkeurig is, bijvoorbeeld vergeleken met een door mensen gegenereerde tekst in geval van een Large Language Model. Hiervoor kan bijvoorbeeld gebruik worden gemaakt van een Bertscore (heeft de AI-gegenereerde tekst dezelfde betekenis als de door mensen gegenereerde tekst) of ROUGE (komen de gebruikte woorden in de AI-gegenereerde tekst en de door mensen gegenereerde tekst overeen).

MODELGOVERNANCE

Voor wat betreft de modelgovernance is het aan te raden zoveel mogelijk aan te sluiten bij de bestaande risicoraamwerken en organisatie, terwijl we ons tegelijkertijd ervan bewust moeten blijven dat AI-modellen een andere aanpak vereisen dan traditionele modellen. Gezien de benodigde ML en AI-kennis bij het valideren van AI-modellen is het wel wenselijk om de beoordeling van deze modellen te laten doen door experts op dit gebied die, net als bij actuariële modellen, onafhankelijk zijn van de ontwikkelafdeling en rapporteren aan het ARC (Audit & Risk Committee) of Model Committee onder verantwoordelijkheid van de CRO.

Het ontwikkelen van AI-modellen zal de komende jaren verder toenemen waarbij zowel de techniek als de regelgeving nog in ontwikkeling zijn. Hoewel het nog vroeg is om de meest voorkomende valkuilen, over het hoofd geziene risico's of het hoogste rendement te identificeren, blijkt uit onze eerste ervaringen dat belangrijke verbeterpunten in het ontwikkelen van AI-modellen zijn:

- Het verbeteren van de data(kwaliteit) waarmee een model is getraind. Hoogwaardige data—nauwkeurig, volledig, tijdig en representatief—stellen het model in staat om de werkelijke onderliggende patronen te leren.

- Het, vanwege het experimentele karakter van data science, documenteren van de stappen in het ontwikkelproces om de traceerbaarheid of reproduceerbaarheid te verbeteren.

- De business en strategie meer leidend maken ten opzichte van de technologie. Te veel focus op de technologie kan leiden tot te complexe modellen of verkeerd gebruik van AI.

- Het investeren in AI-geletterdheid en uitlegbaarheid van modellen. Een gebrek hieraan kan leiden tot een beperkt gebruik van en vertrouwen in AI binnen de organisatie.

- Meer balans aanbrengen in het risicomanagement. Hierin wordt vaak te veel de nadruk gelegd op compliance en minder op het functioneren van het model.

Onze eerste indicatie is dat het hoogste rendement bij validatie behaald kan worden door te focussen op de datavalidatie.

IMPLICATIES VOOR ACTUARISSEN

Actuarissen hebben een natuurlijke voorsprong door kennis van statistiek en het documenteren en valideren van actuariële modellen. Toch zijn ook voor een actuaris nieuwe vaardigheden van belang, zoals kennis van ML-architecturen, bias en fairness technieken en GenAI-technieken. Bedenk hierbij vooral dat AI geen statische technologie is, maar een snel ontwikkelend vakgebied.

CONCLUSIE

AI kan de waardeketen van een verzekeraar op veel vlakken verbeteren maar alleen als stakeholders de modellen kunnen vertrouwen. Een gestructureerd, proportioneel en transparant beoordelingskader maakt dat vertrouwen toetsbaar en audit-proof. Voor actuarissen is het beheersen van dit vakgebied geen optie, maar de volgende stap in professionele relevantie. ■