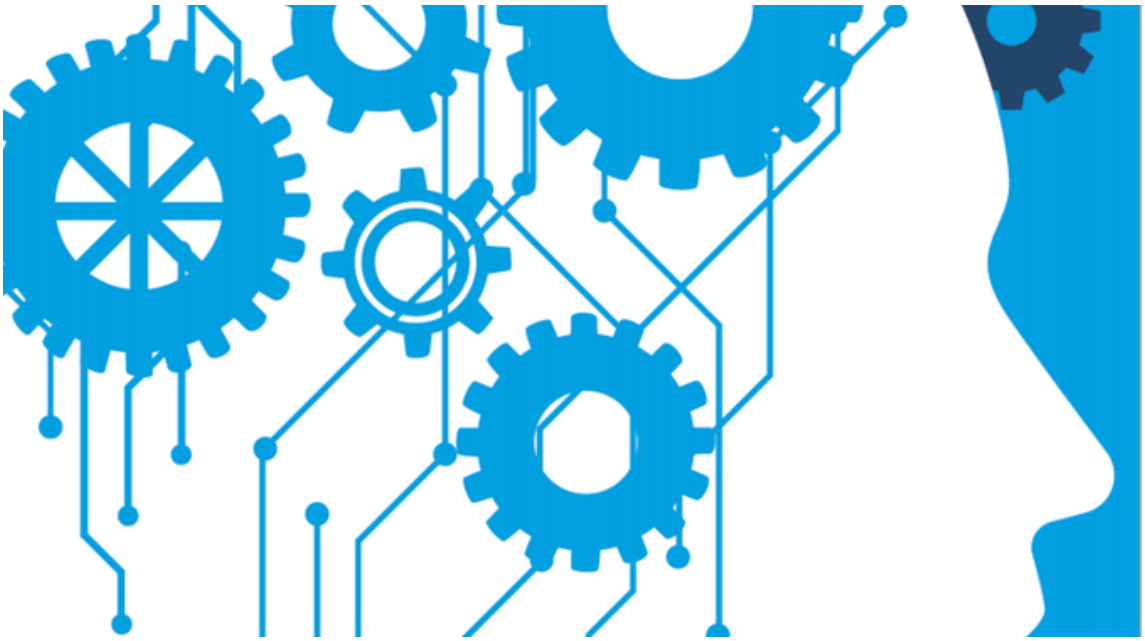


Handreiking user interface design van uitlegbare AI

Auteur
Stefan Leijnen
Henry Maathuis
Kees van Montfort (Hogeschool van Amsterdam)
Sieuwert van Otterloo
Danielle Sent
Marcel Stalenhoef
Koen van Turnhout
Raymond Zwaal (Hogeschool van Amsterdam)

Inlichtingen
Sieuwert.vanotterloo@hu.nl
Henry.maathuis@hu.nl

Datum
20 maart 2025

Versie
1.0

© Hogeschool Utrecht,
Utrecht, 20 maart 2025

Bronvermelding is verplicht.
Verveelvoudigen voor eigen gebruik
of interngebruik is toegestaan.

Inhoudsopgave

1 Waarom deze handreiking 3
2 Eisen aan uitleg 4
2.1 Op wetgeving gebaseerde eisen 4
2.2 Eisen uit de praktijk van de financiële sector 5
3 Gebruik van AI in de financiële sector 7
3.1 Achtergrond van AI-gebruik 7
3.2 Voorbeeldprobleem fraudebeoordeling 7
3.3 Voorbeeldprobleem beoordeling van leningaanvragen 8
4 Wat is uitleg van AI 10
4.1 Algemene principes voor uitleg 10
4.2 Overzicht soorten uitleg 10
5 Voorbeelden van uitleg 11
5.1 Voorbeeld van uitleg: feature importance 11
5.2 Voorbeeld van uitleg: vergelijkbare gevallen 12
5.3 Voorbeeld van uitleg: hypothetische tegenvoorbeelden 14
5.4 Voorbeeld van uitleg: Regelgebaseerde uitleg 15
6 Conclusies 18
7 Referenties 19

© Hogeschool Utrecht, 2025

1 Waarom deze handreiking

Deze handreiking is ontwikkeld voor designers en ontwikkelaars van AI-systemen, met als doel om te zorgen dat deze systemen voldoende uitlegbaar zijn. Voldoende betekent hier dat het voldoet aan de wettelijke eisen vanuit AI Act en AVG en dat gebruikers het systeem goed kunnen gebruiken.

De uitlegbaarheid van beslissingen is namelijk een belangrijke eis bij veel systemen en zelfs een belangrijk principes voor AI-systemen [HLEG19]. Bij veel AI-systemen is de uitlegbaarheid niet vanzelfsprekend. AI-onderzoekers verwachten dat de uitdaging om AI-uitlegbaar te maken alleen maar groter wordt. Dit komt enerzijds vanuit de toepassingen: AI zal steeds vaker ingezet zal worden, voor steeds grotere en meer gevoelige besluiten. Anderzijds maken organisaties steeds betere modellen, die bijvoorbeeld meer verschillende invoer gebruiken. Bij complexere AI-modellen is het veelal minder duidelijk hoe een besluit tot stand gekomen is.

Organisaties die AI gaan inzetten, moeten rekening houden met de behoefte van gebruikers aan uitleg. Systemen die AI gebruiken moeten zo ontworpen worden dat de gebruiker de juiste uitleg krijgt.

In deze handreiking leggen we ten eerste uit wat de eisen zijn die er wettelijk gelden voor uitlegbaarheid van AI-systemen. Deze zijn afkomstig uit de AVG en de AI-Act. Vervolgens leggen we uit hoe AI gebruikt wordt in de financiële sector en werken één probleem in detail uit. Voor dit probleem laten we vervolgens zien hoe de user interface aangepast kan worden om de AI uitlegbaar te maken. Deze ontwerpen dienen als prototypische voorbeelden die aangepast kunnen worden op nieuwe problemen.

Deze handreiking is gebaseerd op uitlegbaarheid van AI-systemen voor de financiële sector. De adviezen kunnen echter ook gebruikt worden in andere sectoren.

Dit whitepaper is geschreven in het* FIN-X onderzoeksproject[1]*. FIN-X is een project dat subsidie ontvangt van het Nationaal Regieorgaan Praktijkgericht Onderzoek SIA in het kader van de RAAK-mkb regeling september 2022.

Het whitepaper is mede tot stand gekomen dankzij kennisbijdrages van de projectpartners, waaronder: FRISS, Floryn, De Volksbank, Deeploy en BG.legal.

[1] https://www.hu.nl/onderzoek/projecten/fin-x

© Hogeschool Utrecht, 2025

2 Eisen aan uitleg

2.1 Op wetgeving gebaseerde eisen

Wie een AI-systeem ontwerpt moet er rekening mee houden dat gebruikers en betrokkenen recht hebben op uitleg. Dit volgt uit twee wetten, de **verordening Artificiële Inteligentie** (AI-Act) en de **Algemene Verordening Gegevens bescherming** (AVG). De AI Act is formeel in werking getreden op 1 augustus 2024[2] en organisaties moeten hier dus aan voldoen. Er wordt echter nog niet gehandhaafd omdat het toezicht is verdeeld over verschillende toezichthouders voor verschillende soorten bedrijven en toepassingen. In de loop van 2025 en 2026 zal er door de verschillende toezichthouders, gehandhaafd gaan worden.
In de AI Act is in artikel 86 (recht op uitleg bij individuele besluitvorming) vastgelegd dat individuen voor wie een besluit een aanzienlijk gevolg heeft, recht hebben op uitleg wanneer dat besluit voornamelijk op AI gebaseerd is. In overweging 171 is omschreven wat hiermee wordt bedoeld:

> *Getroffen personen moeten **het recht hebben om uitleg te krijgen** indien het besluit van een gebruiksverantwoordelijke voornamelijk is gebaseerd op de output van bepaalde AI-systemen met een hoog risico die binnen het toepassingsgebied van deze verordening vallen en indien dat besluit rechtsgevolgen of gelijkaardige aanzienlijke gevolgen heeft voor de gezondheid, veiligheid of grondrechten van die personen. Die uitleg moet **duidelijk en zinvol** zijn en moet de grondslag zijn waarop de getroffen personen zich kunnen baseren om hun rechten uit te oefenen.*

Deze eis geldt naar de getroffen persoon toe. Dit zijn bijvoorbeeld de mensen van wie transacties worden beoordeeld.

In een andere overweging zijn de eisen uitgelegd die gelden voor de gebruiker van het op AI gebaseerde systeem (bijvoorbeeld een medewerker van de financiële instelling). Dit is overweging 71:

> *AI-systemen met een hoog risico moeten op zodanige wijze worden ontworpen en ontwikkeld dat natuurlijke personen **toezicht kunnen houden op de werking ervan** en dat zij zoals beoogd worden gebruikt, en dat de gevolgen ervan gedurende de levenscyclus van het systeem worden aangepakt. Daartoe moeten passende maatregelen voor menselijk toezicht worden bepaald door de aanbieder van het systeem voordat het in de handel wordt gebracht of in gebruik wordt gesteld.*
>
> *Dergelijke maatregelen moeten, waar passend, met name waarborgen dat voor het systeem ingebouwde operationele beperkingen gelden die niet door het systeem zelf kunnen worden omzeild, **dat het systeem reageert op de menselijke operator** en dat de natuurlijke personen aan wie de taak van het menselijk toezicht is toegewezen, beschikken over de noodzakelijke competenties, opleiding en autoriteit om deze taak uit te voeren.*

Veel van de eisen die hieronder zijn geformuleerd, zijn gebaseerd op deze toezichthoudende functie. In deze overweging is ook expliciet genoemd dat aanbieders van AI-systemen rekening moeten houden met deze eisen. Aanbieders mogen systemen alleen aanbieden als er sprake is van uitleg die dusdanig duidelijk is dat gebruikers goed toezicht kunnen houden.

Naast de AI Act moeten bedrijven ook voldoen aan de algemene verordening gegevensbescherming (AVG, Engels: GDPR). De AVG is aanvullend: bedrijven moeten aan beide wetten voldoen. De AVG stelt eisen aan de verwerking van persoonsgegevens en geldt dus voor AI-systemen als daar persoonsgegevens in worden verwerkt. Dit is vaak het geval in de financiële sector. De volgende eis is dan gesteld in artikel 14:

> *Wanneer persoonsgegevens niet van de betrokkene zijn verkregen, verstrekt de verwerkingsverantwoordelijke de betrokkene de volgende informatie: **het bestaan van geautomatiseerde besluitvorming** en, ten minste in die gevallen, **nuttige informatie over***

[2] https://nl.wikipedia.org/wiki/Verordening_Kunstmatige_Intelligentie

© Hogeschool Utrecht, 2025

***de onderliggende logica**, alsmede het belang en de verwachte gevolgen van die verwerking voor de betrokkene.*

Deze eisen gaan over de uitleg die gegeven moet worden aan degene over wie het besluit gaat. Dit is vaak de klant van de financiële instelling en wordt in de AVG de betrokkene genoemd. In dit whitepaper gaan we ervan uit dat de gebruiker van het systeem een bankmedewerker is, en deze gebruiker eerst een uitleg te zien krijgt om dan later te delen met de klant/ betrokkene.

In de hierboven genoemde regelgeving worden er eisen gesteld aan uitleg als geheel. In de praktijk bestaat de uitleg van een AI-systeem uit twee delen:

- **Globale uitleg** is uitleg over de werking van het gehele AI-systeem, niet voor een specifieke beslissing. De globale uitleg kan toelichten hoe het systeem is ontwikkeld, hoe de data eruitziet waar het systeem op getraind is, welke AI-methodes zijn gebruikt, hoe het systeem is getest maar ook hoe het systeem te gebruiken. Deze globale uitleg kan in het AI-systeem zitten door middel van hulpschermen, maar ook in andere documenten waar de gebruiker bij kan.
- **Lokale uitleg** is uitleg specifiek voor één geval of beslissing en legt die ene beslissing uit. Deze uitleg geeft aan welk besluit genomen is welke eventuele deelscores hierbij gebruikt zijn, welke specifieke invoer is gebruikt en wat de invloed van verschillende factoren is.

Een voorbeeld van globale uitleg is: *"Het systeem bestaat uit 140 beslisregels die opgesteld zijn op basis van een interne dataset van 550 eerder gemaakte beslissingen. Het systeem maakt gebruik van 28 invoersoorten. Eén daarvan is de betaalgeschiedenis van gebruikers, inclusief het aantal maanden dat iemand al premie betaald heeft. Deze data is afkomstig uit systeem XYZ en wordt maandelijks opnieuw opgehaald."*

Een voorbeeld van lokale uitleg is: *"Je vraagt uitleg voor Claim 3720, ingediend op 30 jan 2025 en op 1 feb 2025 afgewezen. De indiener had op moment van beoordeling 7 maanden premie betaald, dat is minder dan gemiddeld."* De lokale uitleg moet door het systeem zijn gegenereerd worden. Het is niet mogelijk voor een groot AI-systeem om een apart document te maken met lokale uitleg voor alle mogelijke casussen die voorgelegd kunnen worden.

In de praktijk moet een AI-systeem zowel globale uitleg als lokale uitleg kunnen geven. De grootste ontwerp-uitdagingen zitten in de lokale uitleg, omdat hiervoor echt aanpassingen in het systeem nodig zijn. De gebruikersinstructies die onderdeel van de globale uitleg zijn, zullen vervolgens ook moeten toelichten hoe gebruikers de lokale uitleg kunnen vinden en gebruiken. In dit whitepaper gaan we daarom in op de eisen voor lokale uitleg, omdat dit belangrijk zijn voor het ontwerp van AI-systemen.

2.2 Eisen uit de praktijk van de financiële sector

Deze handreiking is gemaakt binnen het FIN-X onderzoeksproject. In het project zijn er eisen in kaart gebracht voor uitlegbaarheid van AI. Deze zijn gebaseerd op analyse van wetgeving en regelgeving en gevalideerd door betrokken experts die werken bij de praktijkpartners van het onderzoek.

De volgende eisen zijn hieruit naar voren gekomen als het meest belangrijk volgens de betrokken experts. We raden aan deze altijd mee te nemen in een ontwerp van het systeem. Het gaat hierbij om eisen voor de lokale uitleg die echt in het systeem moet zitten.

Lokale uitleg moet…

1. ..mensen helpen begrijpen hoe een uitkomst veranderd kan worden: wat is er nodig om van bijvoorbeeld een afwijzing naar een toelating te gaan
2. ..gebruikers aanmoedigen om actie te ondernemen, bijvoorbeeld aanvullende informatie zoeken of bepaalde invoer extra te controleren
3. ..gebruikers helpen om het risico op bias, discriminatie of profilering te begrijpen
4. ..de mogelijkheid bieden om feedback te geven over het systeem, bijvoorbeeld het corrigeren van een uitkomst.
5. ..een gebruiker helpen om de uitkomst uit te leggen aan een betrokkene

© Hogeschool Utrecht, 2025

6. ..gebruikers helpen om ‘automation bias’ te vermijden: gebruikers moeten niet teveel op de juistheid van de uitkomsten van het systeem vertrouwen
7. ..mensen die toezicht houden op een AI-systeem helpen om in specifieke gevallen het systeem niet te gebruiken

De volgende eisen zijn ook genoemd in het onderzoek en van toepassing op sommige systemen. Ze zijn niet altijd vereist en deels ook tegenstrijdig. We raden aan deze eisen door te nemen in een designreview, en dan te bespreken of het systeem er voldoende aan voldoet. Je kunt ook kijken of er kleine verbeteringen mogelijk zijn om meer aan deze eisen te voldoen, zonder concessies te doen aan andere eisen.

1. ..moet compact zijn
2. ..moet gebruikers meer vertrouwen geven in hun besluiten
3. ..moet gebruiker helpen om afwijkingen, uitzonderingen en onverwacht gedrag op te merken
4. ..moet interactief zijn (de uitleg moet op de gebruiker reageren)
5. ..moet helpen om de beperkingen van AI-systemen te begrijpen
6. ..moet inzicht geven in de gebruikte invoer, ontbrekende gegevens en niet toepasbare regels
7. ..moet inzicht geven of het besluit gebaseerd is op incorrecte gegevens
8. ..moet voldoende gedetailleerd zijn
9. ..moet visueel zijn in plaats van alleen uit tekst bestaan
10. ..moet inzicht geven in vergelijkbare gevallen
11. ..moet visueel en tekstueel zijn

Bij twijfel of de uitleg van een AI-systeem voldoet of discussies welke uitleg het best is, is het belangrijk om een evaluatie van uitleg te doen. In een eerder artikel (zie [KMD24] in de referenties) is een taxonomie gegeven van hoe men uitleg kan evalueren. De belangrijkste soorten evaluatie zijn:

- Gebruikers de uitleg laten beoordelen in termen van, bijvoorbeeld, begrijpelijkheid, bruikbaarheid of betrouwbaarheid (*kwaliteit van uitleg*)
- De tevredenheid van gebruikers met het systeem (waar de uitleg in zit) meten, bijvoorbeeld in termen van begrijpelijkheid van het systeem, gebruiksvriendelijkheid van het systeem, vertrouwen in het systeem (*mens-computer-interactie*)
- Meten hoe goed gebruikers bepaalde taken kunnen uitvoeren waarbij ze de uitleg moeten gebruiken, bijvoorbeeld het controleren van beslissingen of het maken van uitleg voor betrokkenen (*mens-computer-prestatie*).

© Hogeschool Utrecht, 2025

3 Gebruik van AI in de financiële sector

3.1 Achtergrond van AI-gebruik

Financiële instellingen gebruiken al veel IT-systemen, onder andere voor de volgende processen.

- Het beoordelen of een klant een bepaald product/dienst mag afnemen (bijvoorbeeld een krediet of lening) en tegen welke voorwaarden (hoeveelheid, rente). De uitkomst kan positief zijn (acceptatie) of negatief (afwijzing).
- Het beoordelen of een transactie nader onderzocht moet worden op mogelijke financiële fraude. Ook hier is er sprake van een positieve uitkomst (acceptatie zonder verder onderzoek) of negatieve uitkomst (nader onderzoek nodig).

Het mogelijke gebruik van AI bij deze toepassingen is al in 2021 reeds besproken tussen De Nederlandse Bank als toezichthouder en individuele banken [DNB21]. Vaak zijn het systemen gebaseerd op eenvoudige logica (bijvoorbeeld een klein aantal beslisregels die door experts zijn bepaald), waarbij de uitkomsten goed uitlegbaar zijn door de gebruikers. Financiële instellingen verwachten nog meer datagedreven te werken dan dan ook meer AI-vraagstukken te hebben, waarvoor zij besluitvorming moeten inrichten[Kr23]. Dit vraagt om veel veranderingen, waaronder meer aandacht dat systemen ontworpen worden om goede uitleg te geven aan gebruikers.

3.2 Voorbeeldprobleem fraudebeoordeling

Het volgende probleem is een voorbeeld van een veelvoorkomend beslisprobleem in de financiele sector, waar AI op toegepast kan worden. Het is gebaseerd op de echte cases uit het onderzoeksproject, maar is wel veralgemeniseerd en niet op één organisatie gebaseerd. Dit voorbeeld wordt gebruikt om uitlegvoorbeelden bij te geven.

Situatie

Een verzekeringsmaatschappij biedt verzekeringen aan tegen diefstal en schade van auto's. De verzekerde personen moeten bij schade of diefstal een claim indienen. De verzekeringsmaatschappij moet vervolgens inschatten of de claim direct geaccepteerd en uitgekeerd wordt of dat de verzekerde meer informatie moet leveren voor verder onderzoek en de claim dus in eerste instantie afgewezen wordt. Dit wordt gedaan op basis van een risico-inschatting. Als de kans op fraude laag is (bijvoorbeeld minder dan 5%) wordt de claim geaccepteerd. Bij een hogere kans op fraude wordt de claim in eerste instantie afgewezen.

Het is belangrijk om te beseffen dat het systeem alleen de kans op fraude inschat en niet beslist of er sprake is van fraude. Ook is een afwijzing niet definitief: de verzekerde mag dan meer informatie aanleveren en de claim wordt opnieuw beoordeeld. Dit is noodzakelijk, vanwege het in de AVG vastgelegde recht op een menselijke blik[3].

Informatie voor de beslissing

Om te komen tot een fraude-beoordeling, kan de volgende informatie worden gebruikt:

- Claim cause (reden voor de claim): Total theft, two-sided collision, one-side collision, damage-while-parked (diefstal hele auto, aanrijding met andere partij, aanrijding zonder andere partij, schade terwijl auto geparkeerd was).

[3] https://www.autoriteitpersoonsgegevens.nl/themas/basis-avg/privacyrechten-avg/recht-op-een-menselijke-blik-bij-besluiten

© Hogeschool Utrecht, 2025

- Location (locatie): de wijk waarin de schade is ontstaan, bijvoorbeeld Rotterdam-harbor, Kralingen, Rotterdam-centre, Amsterdam-centre.
- Claim-amount (claimbedrag): De hoeveelheid schade die geclaimd wordt in euro. Dit kan liggen tussen de € 250 en € 100.000.
- Car model (Automerk): Het merk en eventueel type van de auto, bijvoorbeeld audi, bmw, kia, volvo.
- Consecutive-months-paid (Maanden-achtereen-betaald): hoeveel maanden er al premie is betaald zonder onderbrekingen. Mogelijke waarden zijn 0,1,2, … tot 60 of meer maanden.
- Payments-overdue (betaalachterstand): of er sprake is van een betaalachterstand.

Er is een model ontwikkeld, dat de kans op fraude uitdrukt in punten. Vervolgens is er in de documentatie van het systeem vastgelegd hoe gebruikers dit moeten gebruiken. Als voorbeeld zou men het volgende kunnen gebruiken:

- Bij minder dan 75 punten is de kans op fraude klein en wordt de claim geaccepteerd.
- Bij tussen de 75 en 125 punten is er twijfel en zal de medewerker die het systeem gebruikt een besluit moeten nemen.
- Bij meer dan 125 punten is de kans op fraude groot en wordt de claim afgewezen.

Deze interpretatie van de score kan gebruikt worden in het ontwerp, bijvoorbeeld door voor bepaalde scores bepaalde kleuren te gebruiken: blauw of groen voor scores die passen bij accepteren, oranje en rood voor scores die passen bij twijfel en afwijzen.

3.3 Voorbeeldprobleem beoordeling van leningaanvragen

Machine learning kan op heel veel manieren worden toegepast in de financiële sector. Hieronder is een tweede voorbeeldsituatie uitgewerkt van een realistisch beslisprobleem. Ook van dit probleem in de invoer uitgeschreven. In dit voorbeeld is het de bedoeling dat de gebruiker met de betrokkenen op zoek gaat naar mogelijkheden. Anders dan bij een verzekeringsclaim mag de betrokkene de invoer wijzigen om hopelijk tot een acceptatie te komen. Dit probleem is niet verder uitgewerkt, maar kan in vervolgonderzoek gebruikt worden voor maken van meer voorbeelden. Ook dit probleem is gebaseerd op de echte cases uit het onderzoeksproject, maar is wel veralgemeniseerd en niet op één organisatie gebaseerd.

Situatie

Een onderneming vraagt een lening aan bij een kredietverstrekker voor financiering van een bedrijfspand. De kredietverstrekker moet beoordelen of de lening verstrekt kan worden. Hiervoor moet een risico-inschatting worden gemaakt, en bij hoog risico (bijv meer dan 15% kans op niet volledig aflossen) wordt de lening niet verstrekt.

Informatie voor de beslissing

De kredietverstrekker heeft de volgende kenmerken beschikbaar:

1. Naam van de onderneming
2. Omzet van onderneming vorig jaar
3. Winst van onderneming vorig jaar
4. Aantal jaar dat onderneming bestaat
5. Vestigingsplaats onderneming
6. Postcode van het aan te kopen pand
7. Omvang (m2) van het pand
8. Soort gebruik (kantoor, winkel, horeca, industrieel)
9. Vraagprijs van het pand

© Hogeschool Utrecht, 2025

10. Aankoopprijs (verwacht) van het pand
11. Gevraagd financieringsbedrag
12. Verwachte huurinkomsten van pand
13. Huurinkomsten van pand vorig jaar
14. Jaar waarin pand laatst gekocht is
15. Aankoopprijs van het pand bij vorige transactie

Het AI-systeem bestaat uit een scherm waarop men al deze kenmerken kan invullen, en het systeem vervolgens tot een ja/nee uitkomst komt.

Vragen waarbij uitleg moet helpen

Vragen die de gebruiker kan hebben in dit geval zijn:

- Zal de aanvraag wel toegekend worden bij een lagere aankoopprijs?
- Heeft de omzet en winst van de onderneming invloed gehad op de beslissing? Heeft het zin om de aanvraag vanuit een andere onderneming te doen?
- Heeft het zin een nieuwe aanvraag te doen voor een lager financieringsbedrag?
- Wat is de invloed van het adres van het pand?

Mogelijke beslisregels

Voor dit voorbeeldprobleem zouden de volgende regels kunnen gelden. Dit soort regels zijn gebruikelijk bij het beoordelen van kredieten. De exacte regels en getallen zijn fictief.

- **Leeftijd-bedrijf**: Als het bedrijf minder dan 36 maanden bestaat, moet er geen lening worden vertrekt
- **Toets-omzet**: Als het financieringsbedrag meer is dan 280% van de omzet, moet er geen lening worden vertrekt
- **Verhouding huur-financieringsbedrag**: Als het financieringsbedrag meer is dan 10.0 keer de verwachte huurinkomsten, moet er geen lening worden vertrekt
- **Verhouding aankoopprijs-huurinkomsten**: Als de aankoopprijs meer is dan 15.1 keer de verwachte huurinkomsten, moet er geen lening worden vertrekt
- **Plotselinge-waardestijging**: Als het pand minder dan 24 maanden geleden gekocht is de en aankooprijs (verwacht) is meer dan 1.40 keer de aankoopprijs van het pand bij vorige transactie, moet er geen lening worden verstrekt

© Hogeschool Utrecht, 2025

4 Wat is uitleg van AI

4.1 Algemene principes voor uitleg

In een beslissingsondersteunend systeem wordt er invoer gebruikt die hoort bij een specifieke casus, om te komen tot een uitkomst. De gebruiker van het systeem moet in ieder geval het volgende kunnen zien:

- De gebruikte invoer.
- De door het systeem gegeven uitkomst.
- De datum wanneer de uitkomst bepaald is.
- De naam en versie van het gebruikte AI-systeem. Veel organisaties passen hun AI-modellen vaak aan, en het kan soms lastig zijn om terug te vinden welk model precies gebruikt is, als dat niet bijgehouden wordt.
- Als het mogelijk is voor gebruikers om uitkomsten te corrigeren: de reeds eerder gedane correcties, inclusief wie wanneer welke correctie gedaan heeft en wat de eerdere uitkomst is en de uiteindelijke uitkomst.

Bij veel beslissingsondersteunende systemen in de financiële sector, is er niet alleen een ja/nee uitkomst, maar ook een score, vaak uitgedrukt in procenten: hoe hoger de score, hoe groter de kans dat de uitkomst positief is. Als dit zo is, dan is het belangrijk ook de score te tonen aan de gebruiker. De gebruiker weet dan of er sprake was van een twijfelgeval, of hoe zeker het AI-model is.

Bij veel AI-systemen is er sprake van voorbewerking van de data. De gebruiker geeft bijvoorbeeld postcode en huisnummer in, en het systeem vertaalt dit naar een wijk. Het is belangrijk dat het voor de eindgebruiker duidelijk is wat de echte invoer is (straat en huisnummer) en wat door systeem is toegevoegd of afgeleid is van andere data.

4.2 Overzicht soorten uitleg

Er zijn meerdere manieren om lokale uitleg te geven. Uit een uitgevoerd literatuur-onderzoek[KMS24] blijkt dat de volgende vier manieren veel gebruikt worden. Deze vier hoofdtypes worden verderop in detail toegelicht.

- Feature importance (invloed/zwaarte van kenmerken)
- Similar cases (vergelijkbare casussen uit de data)
- Counterfactuals (hypothetische tegenvoorbeelden)
- Rule based explanations (regelgebaseerde uitleg)

© Hogeschool Utrecht, 2025

5 Voorbeelden van uitleg

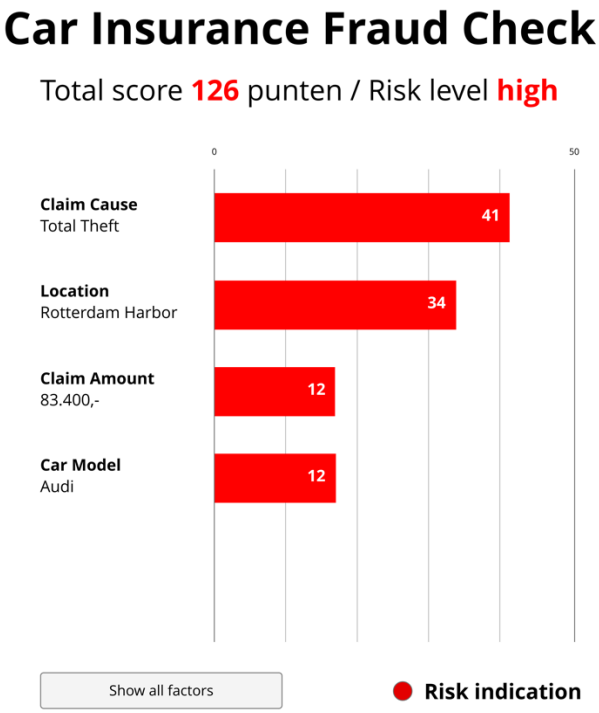
5.1 Voorbeeld van uitleg: feature importance

Feature importance (Nederlands: **Weging van Kenmerken** of **Zwaarte van Kenmerken**) is een belangrijke uitlegmethode die veel gebruikt wordt en waar ook veel bibliotheken voor ontwikkeld zijn. LIME[4] en SHAP[5] zijn bijvoorbeeld bekende technieken voor feature importance. De techniek feature importance betekent dat je laat hoe belangrijk de kenmerken zijn geweest bij een besluit. Feature importance is één van de meest gebruikte technieken voor uitleg en het is aanbevolen deze altijd toe te passen.

Feature importance-gebaseerde uitleg kan de gebruiker helpen om de volgende vragen te beantwoorden (zie [LGM01] voor voorbeeldvragen):

- Op welke kenmerken is dit besluit gebaseerd?
- Heeft de locatie een positieve of negatieve invloed gehad op het systeem?

Hieronder is een voorbeeld weergegeven van uitleg door middel van feature importance. Voor elke invoerwaarde is weergegeven hoeveel extra risico-punten deze factor heeft bijgedragen aan de score. Staafgrafieken zijn een veelgebruikte standaardmanier om dit weer te geven.



Figuur 1: Een interface voor uitleg met feature importance. De rode kleuren geven de bijdrage (in fraudepunten) van de getoonde waarde aan. De bijdrage van alle getoonde factoren is negatief (wijzen op fraude) omdat alleen de factoren die het meeste negatief bijdragen worden getoond. Bovenaan in tekst is de totale score (126) en eindoordeel (hoog risico).

De volgende tips helpen met het goed toepassen van deze manier van uitleg:

- Laat de gebruikte invoerwaarde als tekst zien voor het concrete geval waarover je een besluit neemt. Dit helpt bij het ontdekken van ontbrekende of incorrecte invoer.

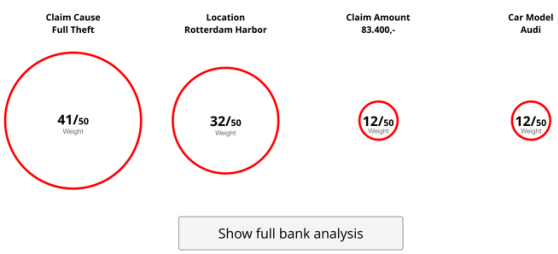
[4] https://lime.readthedocs.io/en/latest/
[5] https://shap.readthedocs.io/en/latest/

© Hogeschool Utrecht, 2025

- Laat alleen de top-kenmerken zien: Het is onoverzichtelijk om alle kenmerken te laten zien voor al bij modellen met tientallen kenmerken. Het tonen van alle kenmerken kan ook verwarrend zijn voor gebruikers, omdat er dan ook kenmerken worden getoond met verwaarloosbare invloed. Een top 5 is vaak voldoende.
- Maak duidelijk met kleur of de getoonde invoerwaarde een positieve of negatieve invloed had. Een relatief hoge winst van de onderneming zou bijvoorbeeld een positieve invloed moeten hebben op het besluit een lening toe te kennen.

Er zijn ook visueel andere manieren om hetzelfde weer te geven. Deze zijn hieronder getoond. Het afwisselen van stijl kan helpen om verschillende systemen er anders uit te laten zien, zodat de gebruiker de systemen niet door elkaar haalt. De informatiewaarde is verder nagenoeg gelijk.

Figuur 2: Voorbeeld van uitleg feature met een tree-map. Hoe groter de rechthoek hoe groter de (negatieve) invloed van de factor.



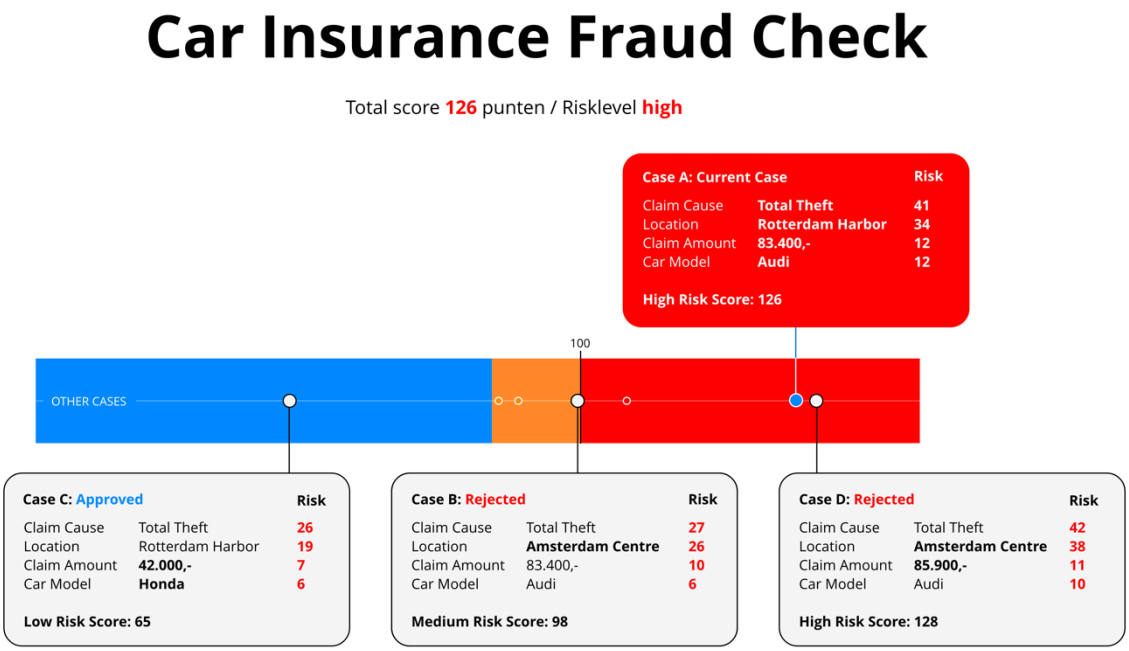
Figuur 3: Nog een andere manier om feature importance toe te passen. Voor de top 5 factoren met een negatieve invloed is de bijdrage in punten getoond. In tekst zijn de punten weergegeven en het theoretisch maximale aantal punten.

5.2 Voorbeeld van uitleg: vergelijkbare gevallen

Bij het tonen van vergelijkbare gevallen (Engels Similar Cases), ziet de gebruiker een aantal eerdere vergelijkbare situaties en het besluit wat daarbij genomen is. Dit kunnen casussen uit de trainingsdata van het systeem, waarvan de juistheid van het besluit al gecontroleerd is. Het is de bedoeling dat het systeem vergelijkbare gevallen toont met zowel een gelijke uitkomst maar ook andere uitkomst (een zogeheten contrasterend geval of een contrastive example). Dit helpt de gebruiker om te begrijpen wat er aan de invoer zou moeten veranderen om tot een andere uitslag te komen.

Hieronder is een voorbeeld getoond hoe dit eruit ziet in de praktijk. Er zijn twee andere gevallen getoond waarin een andere uitkosten geldt. Er is met kleur aangegeven of de andere waarde inderdaad tot een lagere score leidt.

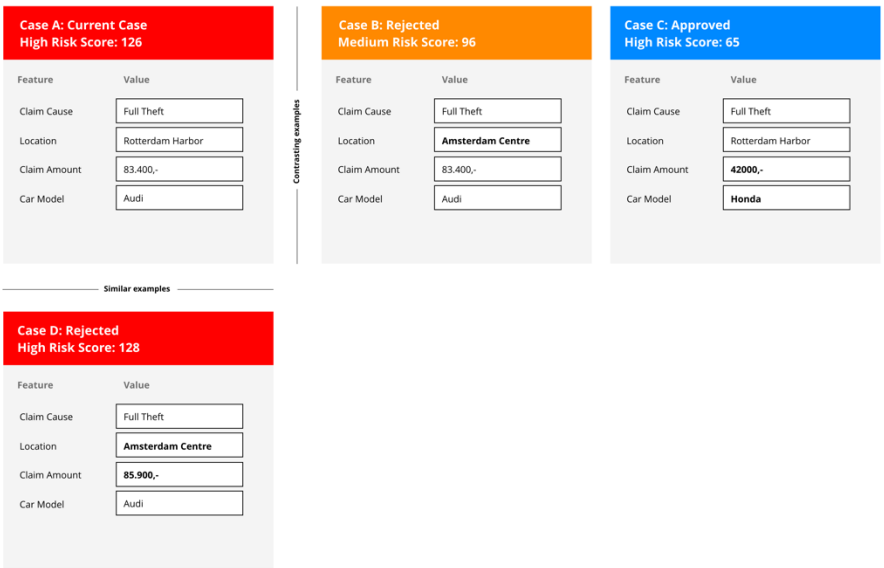
© Hogeschool Utrecht, 2025



Figuur 4: Een voorbeeld van uitleg door middel van vergelijkbare gevallen. Er is uitleg gegeven voor geval A (rood) en deze wordt vergeleken met drie andere gevallen (B,C,D).

In dit voorbeeld is de score van alle gevallen geprint maar ook visueel gemaakt door de gevallen op een score-balk te rangschikken. Met kleur is de voorgestelde uitslag ook duidelijk gemaakt. Rood geeft een hoge score en dus reject aan. Opvallend is dat er een geval met gelijke uitkomst (D, rejected) getoond wordt maar ook een geval B dat in feite een twijfelgeval is, en een geval C met een andere uitkomst.

Ook voor deze uitlegmethode zijn er meerdere manieren om deze weer te geven. Hieronder is een andere manier getoond met meer gevallen.



Figuur 5: Voorbeeld van uitleg door middel van één vergelijkbaar en twee contrasterende gevallen. Een kleurbalk benadrukt de voorgestelde uitkomst op basis van de score.

De volgende tips helpen met het goed toepassen van deze manier van uitleg:

© Hogeschool Utrecht, 2025

- Laat een beperkt aantal gevallen zien, bijvoorbeeld drie tot zes gevallen. Kies gevallen die het meeste lijken op het huidige geval.
- Geef eventueel de gebruiker de mogelijkheid om aan te geven voor welke kenmerken men graag vergelijkbare gevallen wil zien, bijvoorbeeld zelfde locatie of zelfde reden voor de claim.
- Laat altijd een aantal gevallen zien met hetzelfde uitkomst en enkele met de andere uitkomst. Dit helpt op de gebruiker bewust te maken van de kans op fouten door het systeem
- Vat de gevallen samen door de belangrijkste kenmerken te laten zien
- Zorg dat de gebruiker op de getoonde gevallen kan klikken om alle invoerwaardes te zien.

Houd er bij het toepassen van deze uitleg rekening mee dat de gegevens over betrokkene B niet altijd aan betrokkene A getoond kunnen worden. Soms bevatten gevallen persoonsgegevens die alleen met de betrokkene zelf gedeeld mogen worden . Houd hier rekening mee als je dit implementeert, bijvoorbeeld door bepaalde gegevens zoals naam en adres en antwoorden op open vragen niet te tonen, of door waardes af te ronden.

5.3 Voorbeeld van uitleg: hypothetische tegenvoorbeelden

Bij uitleg met hypothetische tegenvoorbeelden (Counterfactuals), toont het systeem niet alleen de uitkomst bij de gegeven invoer, maar wordt ook getoond welke minimale verschuivingen in de invoer leiden tot een andere uitkomst. Deze uitleg kun je interactief maken: de gebruiker kan dan zelf wijzigingen doen in alle invoer en het AI-systeem toont dan de uitkomst voor de aangepaste invoer. Vaak wordt zowel de oorspronkelijke invoer getoond en de gewijzigde invoer. Er wordt vervolgens getoond of de score wijzigt en daarmee of besluit anders zou uitvallen.

Hieronder is een voorbeeld van dit soort uitleg te zien. De huidige casus en uitkomst is links getoond, de gebruiker kan de invoer wijzigen en ziet dan een andere uitslag rechts.

Ook dit soort uitleg kan goed helpen met het bewust maken van gebruiker van eventuele beperkingen van het systeem, of de invloed van ontbrekende invoer.

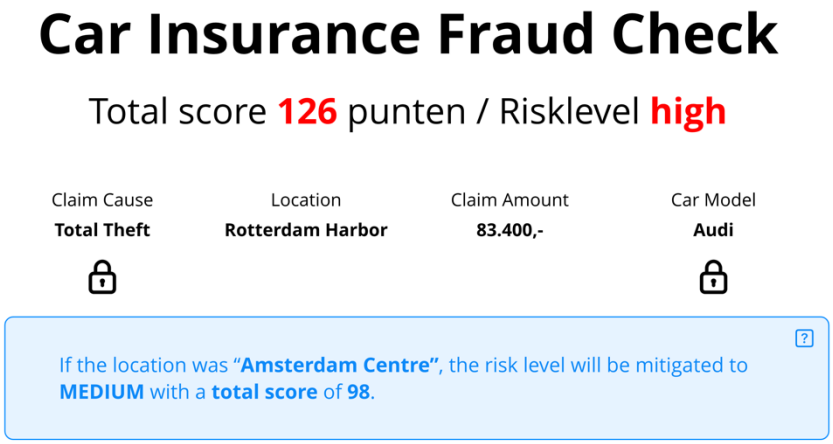
De volgende tips helpen met het goed toepassen van deze manier van uitleg:

- Zorg dat de gebruiker zelf invoerwaardes kan wijzigen en dus de uitleg interactief kan gebruiken.
- Zorg dat de gebruiker alle invoerwaardes kan zien en aanpassen door te klikken.

Deze vorm van uitleg kan wel aan eindgebruikers getoond worden, de gevallen zijn immers hypothetisch.Het is zelfs mogelijk om gebruikers ook controle te geven over de parameters van het model. Dit is in dit eerdere paper uitgewerkt [HNS22].

Een andere manier om dit soort uitleg te gebruiken, is om het systeem al suggesties te geven voor alternatieve invoer. Het systeem evalueert het model dan met telkens andere waardes en rapporteert aan de gebruiker of en hoe een andere uitslag bereikt kan worden.

© Hogeschool Utrecht, 2025



Figuur 6: Voorbeeld van uitleg met hypothetische gevallen, waarin het systeem een suggestie doet van een waarde die tot een andere score leidt.

5.4 Voorbeeld van uitleg: Regelgebaseerde uitleg

Binnen veel beslissingsondersteunende-systemen die gebruikt worden in de financiële sector, worden beslisregels gebruikt. Dit kan op meerdere manieren: het is bijvoorbeeld mogelijk om één beslisboom te maken, gebaseerd op trainingsdata, die de uitkomst bepaalt. Het is ook mogelijk om in het systeem meerdere regels op te nemen, en de uitkomst te laten bepalen door de combinatie-score van elke regel. In beide gevallen is het mogelijk om aan de gebruiker van het systeem de delen van de beslisboom of de gebruikte regels te tonen.

Voorbeelden van regels die voor dit systeem zouden kunnen gelden zijn:

- **Claim cause** (Reden voor de claim): de oorzaak *Full Theft* heeft hogere kans op fraude dan *two-side collision*, omdat er immers geen bevestiging van een tweede partij is.
- **Location**: bepaalde locaties kunnen als hoog risico bekend staan, bijvoorbeeld omdat er geen mensen wonen.
- **Claim value**: een claimbedrag van meer dan € 40.000 is een indicatie van mogelijke fraude.
- **Payments**: als het aantal maanden achtereen betaald (*Consecutive-months-paid*) minder is dan 18, dan geeft dat een hogere kans op fraude.

De meeste regels bevatten parameters (de getallen zoals 18 of 40.000) die aangepast kunnen worden. Vaak wordt dit gedaan om basis van trainingsdata met behulp van machine learning, om te komen tot regels die zo goed mogelijk eerdere besluiten nadoen. De variabelen kunnen ook door domeinexperts bepaald worden. Het is van belang voor de globale uitlegbaarheid om goed vast te leggen hoe de parameters bepaald worden, en wat voor data hiervoor gebruikt worden.

Soorten regelgebaseerde systemen

Er zijn in principe meerdere manieren om de regels te gebruiken: men kan het systeem zo maken dat als één regel een negatieve indicatie geeft, het besluit negatief is. Dit is een volledig regelgebaseerd systeem. De regels kunnen samengevoegd worden tot één beslisboom.

Het is echter ook mogelijk om een gewogen gemiddelde te nemen over alle uitkomsten om tot een score te komen. In dat geval wordt het besluit alleen negatief als het gemiddelde onder een bepaalde grenswaarde komt. Dit is een meer genuanceerde aanpak, die men vaker ziet bij financiële besluiten. Bij deze aanpak kan men ook met twee grenswaardes werken: één waarbij er twijfel ontstaat maar nog geen afwijzing, en één nog lagere grenswaarde waarbij er afwijzing aangeraden wordt.

© Hogeschool Utrecht, 2025

Het is in principe ook mogelijk om hybride systemen te maken, en vaak wordt dat gedaan om systemen te maken die bepaalde uitkomsten waarborgen: Als de toepassing van regels tot een indicatie leidt wordt deze gevolgd. Als er op basis van de regels geen uitkomst bepaald is, dan wordt de uitkomst door een tweede model bepaald, bijvoorbeeld logistische regressie[6].
Voor zowel volledig regelgebaseerde systemen, systemen die een gemiddelde score bepalen en voor hybride systemen kan men regelgebaseerde uitleg toepassen.

Uitlegmethodes gebaseerd op regels

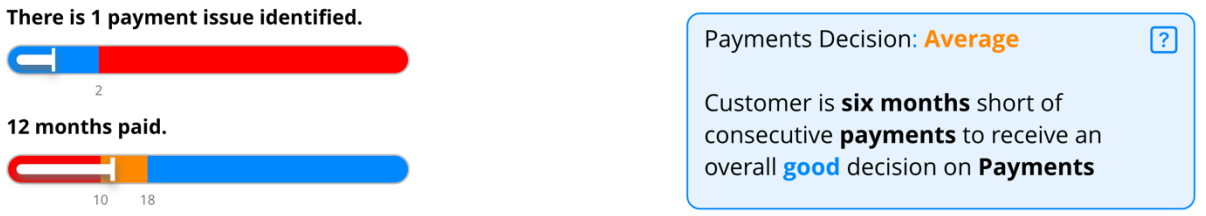
Als een systeem regelgebaseerd is, dan is een goede uitlegmethode om te tonen welke regels wel en niet toegepast konden worden (omdat de invoer wel/niet bekend was) en wat de indicatie is van elke regel. Dit kan in een lijst met de naam van de regel en de uitkomst, zoals hieronder getoond is. In dit voorbeeld is het resultaat van de regel getoond onder "*Decision Rule*". *Poor* betekent een negatieve indicatie (kans op fraude). In deze interface kan de gebruiker de score van een regel aanpassen, als de gebruiker daar aanleiding toe ziet. Dit wordt gedaan in de kolom "*User Decision"*.

Category	Decision Rule	User Decision
Claim Cause	Poor	Poor Average Good
Location	Average	Poor Average Good
Claim Value	Poor	Poor Average Good
Payments	Average	Poor Average Good

Figuur 7: Uitleg gebaseerd op het tonen van regels bij een regelgebaseerd systeem. De gebruiker ziet de score per regel en weet dus welke regels zijn gebruikt

Als een gebruiker dan een regel aanklikt, kan de volledige regel getoond worden met de waardes ingevuld. Dit is hieronder getoond. Er is visueel getoond welke invoerwaardes door de regel gebruikt worden en hoe deze vergeleken zijn met grenswaardes. Rechtsonder wordt er voor de volledigheid nog tekstuele uitleg gegeven, In dit voorbeeld is er een fraude-indicatie omdat de gebruiker nog langer verzekerd had moeten zijn.

Category	Decision Rule	User Decision
Claim Cause	Poor	Poor Average Good
Location	Average	Poor Average Good
Claim Value	Poor	Poor Average Good
Payments	Average	Poor Average Good



Figuur 8: Voorbeeld van op regelgebaseerde uitleg.

[6] https://nl.wikipedia.org/wiki/Logistische_regressie

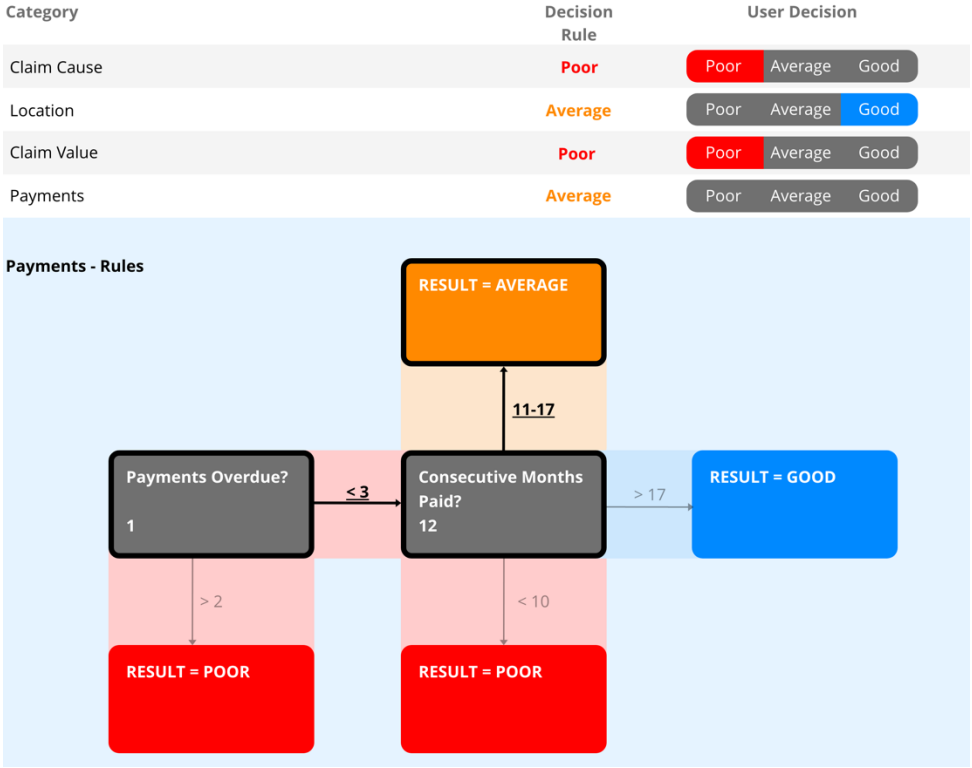
© Hogeschool Utrecht, 2025

De volgende tips helpen met het goed toepassen van deze manier van uitleg:

- Zorg dat een gebruiker altijd ziet of een regel wel of niet toegepast kon worden met de verstrekte invoer. Je wilt niet dat de gebruiker niet weet dat er sprake was van onvolledige invoer.
- Geef elke beslisregel een duidelijke naam die herkenbaar is voor gebruikers, en bijvoorbeeld aangeeft welke invoer gebruikt is
- Als de regels uitgaat van een getal en een grenswaarde, laat dan zien hoever het getal van de grenswaarde afligt. De gebruiker moet kunnen zien of een kleine verandering in invoer tot een verandering in de regel leidt. Maak eventueel drie uitkomsten: positieve indicatie, twijfel, één negatieve indicatie.
- Als regels een verschillend gewicht hebben, laat dan de regel met het meeste gewicht eerst zien.
- Zorg dat een gebruiker de volledige regel kan zien als deze dat wil, bijvoorbeeld door een regelnaam aanklikbaar te maken
- Als een volledige regel getoond wordt, maak dan met tekst en visueel duidelijk wat de gebruikte invoerwaardes zijn, en wat de uitkomst van de regel is.

Het is mogelijk om deze uitlegmethode interactief te maken. Je kunt het bijvoorbeeld mogelijk maken voor gebruikers om invoerwaardes aan te passen als deze getoond worden, om te zien wat het effect op de regel en de uitkomst is. Je krijgt dan een uitleg die zowel regel-gebaseerd is en gebruik maakt van hypothetische tegenvoorbeelden.

Ook bij deze uitlegmethode zijn er andere, meer grafische manieren denkbaar. Het is bijvoorbeeld mogelijk om beslisregels als een beslisboom weer te geven. Dit geeft meer inzicht in de logica, maar laat minder ruimte voor grafische weergave van de parameters. Hieronder is uitleg met een beslisboom te zien.



Figuur 9: Voorbeeld van uitleg van beslisregels in een beslisboom. De gebruiker heeft de onderste regel geselecteerd. Deze wordt als boom getoond.

© Hogeschool Utrecht, 2025

6 Conclusies

Als je een AI-systeem ontwerpt, zorg dan dat de gebruiker ook passende uitleg te zien krijgt. De gebruikers van systemen moeten uitleg krijgen, omdat dit geëist wordt in de AVG en in de AI Act. Deze uitleg moet aan bepaalde eisen voldoen: de uitleg moet inzicht geven aan de gebruiker en ook helpen om inzicht te geven aan de betrokkene, en moet de gebruiker helpen om fouten te kunnen opsporen. Er zijn gelukkige meerdere manieren om uitleg te geven. In dit paper zijn de vier hoofdmanieren uitgewerkt in verschillende ontwerpen. Uit de gebruikersstudies die in het project FIN-X zijn gedaan, is naar voren gekomen dat het verstandig is om meerdere soorten van uitleg te implementeren, zodat elke gebruiker de best passende uitleg kan kiezen en ook de soort uitleg interactief kan veranderen. Vervolgens is het belangrijk de manier van uitleg goed te testen met gebruikers, om zeker te weten dat de uitleg hen effectief helpt. Alleen op deze manier is zeker dat het model aan de wettelijke eisen voldoet.

Ook de ontwikkelaars van de onderliggende beslismodellen zullen de uitlegbaarheid moeten meenemen bij het ontwikkelen van modellen. Sommige soorten modellen zijn makkelijker uitlegbaar dan andere soorten, omdat ze bijvoorbeeld regel-gebaseerde uitleg mogelijk maken. Andere soorten uitleg vereisen extra functies bij het model, bijvoorbeeld om goede vergelijkbare gevallen te vinden of om invoerwaardes te vinden die tot een andere uitslag leiden. Breng vroegtijdig in kaart wat voor soorten uitleg nodig zijn en zorg dat de modellen deze vormen van uitleg ondersteunen.

© Hogeschool Utrecht, 2025

7 Referenties

[DNB21] *De Nederlandse Bank* - Perspectives on Explainable AI in
The Financial Sector An exploratory study between banks and supervisory authorities, 1 juli 2021
https://www.dnb.nl/nieuws-voor-de-sector/iforum-onderzoek-naar-uitlegbaarheid-bij-gebruik-van-ai-bij-banken/

[HNS22] Hekman, Nguyen, Stalenhoef – Towards a pattern library for algorithmic affordances. *CEUR Workshop Proceedings* (Vol. 3124, pp. 24-33), 2022.
https://dspace.library.uu.nl/handle/1874/421114

[HLEG19]: EU High Level Expert Group – Ethische richtsnoeren voor betrouwbare KI, 2019.
https://digital-strategy.ec.europa.eu/nl/library/ethics-guidelines-trustworthy-ai

[KMS24] *Kim, Maathuis, Sent* – Human centered evaluation of explainable AI applications: a systematic review, published in Front. Artif. Intell., 17 October 2024.
https://doi.org/10.3389/frai.2024.1456486

[Kr23] Krijger – Operationalising Ethics for AI in the Financial Industry: Insights from the Volksbank Case Study. Journal of Digital Banking. 8. 220-241, 2023.
https://doi.org/10.69554/YQZC2796

[LGM01] *Liao, Gruen, Miller* – Questioning the AI: Informing Design Practices for Explainable AI User Experiences, 2001. https://arxiv.org/pdf/2001.02478

© Hogeschool Utrecht, 2025