Interpreteerbaarheid van Machine Learning

Steeds meer van onze dagelijkse beslissingen worden gemaakt door machinelearningmodellen. Deze modellen zijn getraind op historische data en kunnen mogelijk een ongewenste bias vertonen. Wanneer zulke modellen worden ingezet voor kritieke besluitneming is het belangrijk om het model te kunnen verklaren. Ook vanuit het oogpunt van eerlijkheid, privacy, betrouwbaarheid en transparantie is interpreteerbaarheid een erg belangrijk speerpunt. Mocht dit nog niet genoeg redenen zijn, onder nieuwe Europese wetgeving (de Digital Services Act) wordt van online platforms verwacht dat ze aan hun gebruikers kunnen uitleggen hoe hun algoritmen beslissingen maken en in de recente AI Act is een voorstel gedaan voor nieuwe regulatie omtrent AI modellen. Kortom, het kunnen interpreteren en uitleggen van machinelearningmodellen is niet iets wat alleen beperkt is tot het domein van Finance & Risk, maar iets waar veel bedrijven straks mee bezig zullen moeten zijn.

Omdat er niet een eenduidige definitie is van wat interpreteerbaarheid en uitlegbaarheid precies inhoudt, is het goed om wat dieper in te gaan op de taxonomie. We sluiten het artikel af met wat praktische tips voor bedrijven en datawetenschappers.

M. Knobbout PhD is Lead Data Scientist bij Just Eat Takeaway.com.

LOKAAL VERSUS GLOBAAL

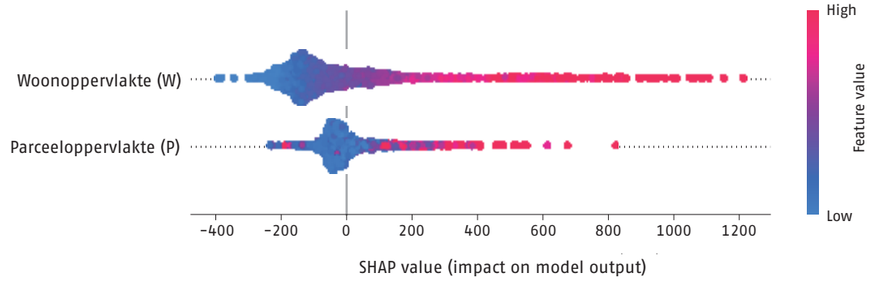
Laten we beginnen met een simpel voorbeeld. Stel onze taak is om de verkoopprijs van een woning te schatten op basis van de woonoppervlakte W en perceeloppervlakte P. Het meest simpele model dat je zou kunnen bedenken is een lineair model dat het volgende verband aanneemt:

$$\text{Prijs} = a \cdot W + b \cdot P + c + \varepsilon$$

Het doel van lineaire regressie is nu om geschikte waarden voor de parameters ‘a’, ‘b’ en ‘c’ te vinden die de historische data het beste verklaart. In dit geval worden ‘a’ en ‘b’ ook wel de coëfficiënten genoemd, ‘c’ de intercept en ‘ε’ de ruis. Het toepassen van lineaire regressie en het vinden van het beste model wordt veelal afgehandeld door één van de vele *libraries* die gemaakt is voor machine learning en statistiek. Maar stel we zijn nu geïnteresseerd in het interpreteren van het model. We willen bijvoorbeeld weten welk van de variabelen W of P het meest indicatief zal zijn voor de vraagprijs van het huis. Dit kunnen we doen door de geleerde coëfficiënten ‘a’ en ‘b’ met elkaar te vergelijken: een positieve waarde zal een positief effect hebben en vice-versa voor een negatieve waarde, en hoe hoger de absolute waarde van een coëfficiënt, hoe meer het zal bijdragen aan de uitkomst. Deze manier van interpreteren noemen we ook wel globaal: we zijn geïnteresseerd naar het algehele gedrag van een model onafhankelijk van de specifieke data waarop we het model willen gaan toepassen. Wanneer we willen weten waarom we een specifieke prijs hebben geschat voor een huis met oppervlakte W en P kunnen we kijken naar de kwantiteiten ‘a·W’ en ‘b·P’. Dit worden ook wel de effecten genoemd en via zogenaamde effect-plots kunnen we vervolgens in kaart brengen welke variabele het meeste effect had op de prijs voor een bepaalde dataset. Deze manier van interpreteren noemen we lokaal: de interpretatie is afhankelijk van specifieke data waarop we het model toepassen.

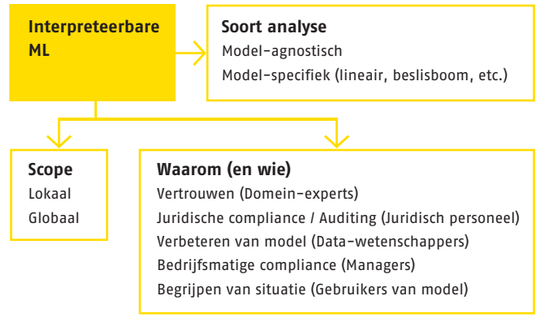
AGNOSTISCH VERSUS SPECIFIEK

In het eerdergenoemde voorbeeld keken we naar een vrij simpel lineair model voor het schatten van de vraagprijs van een huis. Eén van de problemen die we verwachten is dat er een hoge mate van collineariteit tussen de variabelen W en P zal zijn. Immers een huis met een groter perceeloppervlakte zal vaker een groter woonoppervlakte hebben. Een ander probleem is dat de relatie tussen oppervlakte en prijs waarschijnlijk niet lineair zal zijn met name voor erg kleine en grote waardes van W en P. Als we deze relaties willen modelleren volstaat een simpel lineair model niet meer. Tegenwoordig worden neurale netwerken op grote schaal toegepast om dit soort complexe verbanden te kunnen modelleren. Dit soort modellen zijn succesvol in staat om verbanden te leren in situaties waar we een grote kwantiteit van hoog-dimensionale data hebben en zulke modellen kunnen in sommige gevallen wel bestaan uit miljoenen parameters. Simpelweg kijken naar de individuele parameters om het model te kunnen begrijpen lijkt totaal hopeloos, omdat het vaak compleet onduidelijk is wat de contributie is van een individuele parameter op de voorspelling. Er is tegenwoordig veel onderzoek naar het uitlegbaar maken van neurale netwerken, maar een aanpak die erg effectief lijkt komt vanuit het vakgebied van coöperatieve speltheorie in de vorm van Shapley-waarden. In deze aanpak beschouwen we elke input variabele als een speler in een coalitie die coöperatief hun winst willen maximaliseren. De Shapley waarden geven ons een formule waarin we de individuele contributie van elke speler, en dus elke variabele, kunnen bepalen. Dit wordt gedaan door de uitkomsten inclusief de speler te vergelijken met de uitkomsten exclusief de speler. Wat deze aanpak zo aantrekkelijk maakt is dat het model-agnostisch is: het model kunnen we beschouwen als een black-box. Dit staat tegenover de model-specifieke aanpak uit de vorige paragraaf waar we echt naar de interne parameters van het model keken.



Figuur: Een overzicht van de individuele effecten van elk datapunt met behulp van de Shapley waarden. Te zien is dat Woonoppervlakte (W) de meeste informatie gaf. Deze ‘beeswarm’ plot is gegenereerd d.m.v. de Python library ‘shap’ en is toegepast op een niet-lineair model.

Het probleem met deze aanpak is dat het alleen goed werkt als er relatief weinig inputvariabelen zijn. Als we het bijvoorbeeld hebben over modellen voor imageclassificatie wordt het wat lastiger; een plaatje bestaat vaak uit duizenden inputvariabelen (namelijk elke pixel), dus wordt het lastig om te praten over de individuele contributie van een pixel. Dit is op dit moment nog steeds een actief onderzoeksgebied waar maandelijks nieuwe ontwikkelingen plaatsvinden.



Figuur: Overzicht van de taxonomie van interpreteerbare ML.

DE TOEKOMST

In dit artikel hebben we gekeken naar een taxonomie van uitlegbaarheid. Een uitgebreid technisch overzicht van de meest recente ontwikkelingen kan bijvoorbeeld gevonden worden in [1]. Ondanks de snelle groei is het naar mijn mening nog steeds geen volwassen veld, dat lijdt aan een gebrek aan formaliteit en consensus van definities. Hoewel er in de academische wereld veel technieken worden ontwikkeld, zijn ze vaak nog niet onderdeel van workflows en pipelines voor machine learning. Ondanks dat er waarschijnlijk meer ingezet gaat worden op algoritmes voor de taak van uitlegbaarheid, lijkt het niet alsof dit alle uitdagingen gaat oplossen. Het begrijpen van de data en hoe een model past binnen alle processen die plaatsvinden binnen een bedrijf zijn eveneens van cruciaal belang voor uitlegbaarheid. Dit lijkt niet iets wat puur opgevangen kan worden door algoritmes. De tijd zal het leren, maar voor nu heb ik alvast een aantal praktische tips op een rijtje gezet waar bedrijven en datawetenschappers mee aan de slag kunnen gaan.

Tips voor bedrijven

1. In eerste instantie is het voor veel bedrijven belangrijk om je eigen data goed te begrijpen. Maak gebruik van datadefinities. Op deze manier is het duidelijk op wat voor data een model getraind wordt.

2. Kies voor het meest simpele model dat nog steeds nauwkeurig genoeg is. Een white-box model loont omdat het ons in staat stelt om makkelijker en vroegtijdiger ongewenste bias te detecteren [2].

3. Definieer processen voor het kunnen uitleggen van modellen, wat vaak begint bij een goede documentatie van een model en het toewijzen van eigenaren van dit proces. Neem in de model governance expliciet op dat er aandacht besteed dient te worden aan de uitlegbaarheid en interpreteerbaarheid van modellen. Zorg dat dit ook geborgd wordt voor (machine learning) modellen die momenteel buiten de scope van de model governance vallen.

Tips voor datawetenschappers

1. Er bestaan verscheidene Python libraries die het proces van interpreteren kunnen automatiseren zoals ‘shap’, ‘lime’ en ‘eli5’. In R is er ook voldoende keuze zoals ‘fastshap’, ‘lime’ en ‘iBreakDown’.

2. Het in kaart brengen van ‘feature importance’ kan ook gebruikt worden voor modelverbetering. Als we bijvoorbeeld zien dat de meest recente activiteit van een gebruiker het meest bepalend is voor een uitkomst kunnen we meer features gaan creëren voor recentheid.

3. We kunnen model interpretatie ook gebruiken om een advies terug te geven. Als we bijvoorbeeld zien in een model dat een bepaalde promotie een zeer positief had op de klantretentie, kan het advies gegeven worden om meer gebruik te maken van een promotie. Probeer gebruik te maken van features waar we nog aan kunnen ‘draaien’. ■

Referenties:
[1] https://www.mdpi.com/1099-4300/23/1/18
[1] https://www.nature.com/articles/s42256-019-0048-x