Machine Learning en fiscaliteit

Machine Learning wordt steeds belangrijker in ons (dagelijks) leven. Spamfilters beschermen onze e-mail en Netflix bepaalt wat we zien op televisie. Hoe zit dat in de fiscaliteit? Welke rol speelt machine learning en welke kansen en risico's zijn er?

In dit thema krijgt u antwoord op de volgende vragen:

- Wat is machine learning?
- Welke vormen van machine learning zijn er?
- Wat zijn de aandachtspunten en risico's van het gebruik van machine learning?
- Wat zijn de mogelijkheden van machine learning binnen de fiscaliteit?

Wat is machine learning?

Machine Learning (hierna: ML) is een technologie waarbij een computer wordt getraind om op basis van historische data voorspellingen te doen op toekomstige data. Er zijn drie vormen van ML: supervised learning, unsupervised learning en reinforcement learning.

Supervised learning

Bij supervised learning wordt een ML-model getraind op basis van gelabelde data. Dit zijn data waarbij elke regel in de dataset specifieke kenmerken heeft en een vooraf gedefinieerde categorie. Op fiscaal gebied wordt dit bijvoorbeeld toegepast bij de classificatie van kosten voor de heffing van BTW. De categorieën zijn dan bijvoorbeeld: volledig aftrekbare BTW, pro rata aftrekbare BTW en niet-aftrekbare BTW. Relevante kenmerken van zo'n dataset kunnen zijn: de tekst van de grootboektransactie of factuur, de aanbieder van de diensten en voor welke vennootschap of afdeling binnen een onderneming deze kosten zijn bestemd.

Unsupervised learning

In het geval van unsupervised learning, hebben we wel de beschikking over de kenmerken van een specifieke dataset, maar deze heeft geen labels. Dat houdt dus in dat er geen voorspelling plaatsvindt naar vooraf gedefinieerde categorieën. Deze techniek wordt vooral gebruikt om data te clusteren in bepaalde categorieën.

Uitgaande van het hiervoor beschreven voorbeeld van de BTW , kan unsupervised ML worden toegepast als de labels 'volledige aftrek, 'pro rata aftrek' en 'geen aftrek' niet beschikbaar zijn in de dataset. De regels in de dataset kunnen dan met behulp van unsupervised ML aan de hand van de hiervoor genoemde kenmerken worden geclusterd in één van deze drie categorieën.

Reinforcement learning

Reinforcement learning kan het beste vergeleken worden met 'trial and error'. Een computer krijgt in dat geval geen data vooraf te zien, maar begint aan zijn taak en probeert zichzelf telkens te

verbeteren op basis van de ervaring die hij opdoet. Feitelijk ontstaan de data voor het learning proces gedurende het proces zelf.

Aandachtspunten en risico's

Bij het maken van een ML-model zijn twee zaken belangrijk: de kwaliteit van de trainingsdata en het doel waarvoor het model wordt gebruikt.

Trainingsdata vormen de basis van een ML-model. Om een goed ML-model te maken, is de kwaliteit van de data die daarvoor worden gebruikt daarom cruciaal. Het testen en verbeteren van de kwaliteit van de data heet 'pre processing'. Dit proces beslaat in de praktijk een groot deel zo niet het grootste deel van de totale tijd van een ML-project.

Kwaliteit

De vraag is wat dan goede data zijn. Hoewel één en ander afhankelijk is van de dataset, zijn er in het algemeen drie zaken die invloed hebben op de kwaliteit van de trainingsdata:

- Onvolledige data: Een veel voorkomend probleem is dat er data ontbreekt in datasets. Met pre-processing kan dit op verschillende manieren worden opgelost, bijvoorbeeld: het verwijderen van dataregels met ontbrekende data of het invullen van een vervangende waarde.
- Afwijkende formats: Een goed voorbeeld hiervan is een kolom in een dataset waarin een datum staat vermeld. Zeker op het moment dat de data uit verschillende databronnen komen (bijvoorbeeld verschillende ERP-pakketten), zal het format hiervan meestal verschillend zijn. Denk hierbij aan EU-notatie versus US-notatie en maanden uitgeschreven versus een maand vermeld in cijfers.
- Ongebalanceerde data: Ongebalanceerde data houdt in dat bepaalde kenmerken in een dataset oververtegenwoordigd zijn. Een categorie in een dataset komt bijvoorbeeld veel meer voor in die dataset dan andere categorieën of er is met name data uit een specifieke periode aanwezig. In het geval van ons voorbeeld, kan bijvoorbeeld de dataset voor 60% bestaan uit regels met het label 'volledige aftrek'. Het gevaar van ongebalanceerde data is dat het slecht werkt voor de ondervertegenwoordigde categorieën. Hierdoor kan een ML-model 'biased' worden. Dit houdt in dat het meer geneigd zal zijn nieuwe data te labelen naar een specifieke categorie ten opzichte van de andere categorieën.

Doel

Naast de kwaliteit van de data is het belangrijk vooraf te bepalen wat het doel is waarvoor het ML-model wordt gebruikt en wat een acceptabele uitkomst is. Stel een situatie voor waarbij op basis van historische röntgenfoto's een ML-model wordt getraind op het herkennen van tumoren. De wens daarbij zal zijn om een mogelijke tumor te ontdekken. In dat geval heeft men liever een vals positief (ten onrechte een tumor herkend) als uitkomst dan een vals negatief (ten onrechte geen tumor herkend). In de situatie van een ML-model om spam e-mail te ontdekken, is de situatie net andersom. Over het algemeen heeft men liever ten onrechte een spam e-mail in de e-mailbox (vals negatief) dan een belangrijke e-mail die in de spam-box terecht komt (vals positief).

Fiscale mogelijkheden

ML staat binnen de fiscaliteit nog redelijk in kinderschoenen. In de toekomst verandert dit mogelijk drastisch, maar ML zal het fiscale vakgebied niet volledig overnemen. Er zal altijd behoefte zijn aan menselijke interpretatie van feiten en situaties. ML heeft de meeste kans van slagen als we het zien als hulpmiddel bij ons werk. Te denken valt aan drie gebieden: compliance, audits en research.

Compliance

Op het gebied van compliance zijn er veel werkzaamheden waarbij een deel van het analyseren van de data kan worden voorbereid met behulp van ML. Een voorbeeld hiervan is het classificeren van data ter voorbereiding van een aangifte. Denk echter ook aan het onttrekken van tekst en tabellen van documenten (bijvoorbeeld: aanslagen, taxatieverslagen voor de WOZ, dividend vouchers, etc.).

Audits

Fiscale onderzoeken vinden veelal plaats op basis van steekproefsgewijze controles wat niet altijd een consistent beeld geeft van de feitelijke situatie. Het is mogelijk om met behulp van ML data te classificeren in specifieke categorieën. Denk naast de voorbeelden op het gebied van de BTW bijvoorbeeld ook aan classificaties voor de werkkostenregeling. In combinatie met data-analyse, kan dit tot betere inzichten leiden bij audits.

Research

Veel researchwerk in de adviessector vindt momenteel nog plaats op de klassieke manier: het doorzoeken van boeken (al dan niet digitaal) en databases naar de juiste jurisprudentie en literatuur. Dit voorbereidende werk kan voor een deel worden voorbereid met behulp van ML. Een zoekmachine die je suggesties geeft voor soortgelijke literatuur en jurisprudentie is hiervan een praktisch voorbeeld.

Belangrijke aandachtspunten

ML heeft in ons dagelijks leven veel impact en kan binnen de fiscaliteit van grote betekenis zijn, zeker in combinatie met andere technieken op het gebied van kunstmatige intelligentie. Dat er zeker ook negatieve effecten kunnen kleven, op fiscaal gebied uit de reeds veel aangehaalde toeslagenaffaire. Uit de diverse rapporten is bekend dat in de risicomodellen van de fiscus bijvoorbeeld ‘nationaliteit’ als kenmerk werd gebruikt. De algemene consesus is dat het onwenselijk is dergelijke kenmerken mee te nemen in het trainen van een ML-model. Ook uit dit schrijnende voorbeeld blijkt dat een goed begrip van de soorten ML, de voorwaarden voor een goed ML-model en inzicht in de kansen en de risico’s zullen van belang zullen zijn voor het slagen van deze technieken binnen onze sector.

Jan-Jaap Gobius du Sart

- Associate Partner Tax Technology & Transformation bij EY;

- Specialisaties: Robotic Process Automation, Data Analytics & Machine Learning.

- Gastdocent Rijksuniversiteit Groningen voor het vak Tax Risk Management & Tax Technology

> Meer over Jan-Jaap Gobius du Sart

Literatuur

- Hoe ICT dienstbaar kan zijn aan wetgeving en wetsinterpretatie, G.M. Nijssen & L.G.M. Stevens, *WFR* 2019/182

  Om fiscale wetgeving en wetsinterpretatie voor de toekomst beheersbaar te houden moet volgens de auteurs gebruik worden gemaakt van de mogelijkheden die de ICT kan bieden. Fundamentele samenwerking van fiscale professionals en ICT-deskundigen is daarvoor noodzakelijk.
- Will code be taxed? Blockchaintechnologie en kunstmatige intelligentie in de fiscaliteit, P.R. de Jong, *WFR* 2019/17

  De auteur bespreekt de gevolgen die de technologieën blockchain en kunstmatige intelligentie kunnen hebben voor de fiscale wetgeving.

Verwante onderwerpen

Thema: Digitale economie en belastingheffing

Digitalisering heeft invloed op op de meer formele kanten van het belastingrecht.

Thema: Blockchain, crypto-munten en fiscaliteit

Uitleg over de fiscale aspecten van blockchaintechnologie.

Naslag

- Vakstudie 08 - Nederlands Internationaal Belastingrecht - Fiscale Encyclopedie De Vakstudie Nederlands Internationaal Belastingrecht, Beschouwing

  Steeds meer belastingdiensten breiden de toepassing van innovatieve technologieën uit om compliance te waarborgen en belastingontduiking en -fraude op te sporen.