Data science for tax administration

Pijnenburg, M.G.F.

Samenvatting

Datawetenschap voor Belastingdiensten

Dit proefschrift gaat over toepassingen van technieken uit de datawetenschappen ('data science') voor het verbeteren van heffings- en inningsprocessen binnen belastingdiensten. De toepassing van data science technieken kan leiden tot effectievere en/of efficiëntere processen of de compliantie (d.w.z. het vrijwillig naleven van regelgeving) bij burgers en bedrijven verhogen. Enerzijds worden bestaande technieken beoordeeld op hun toepasbaarheid in dit domein, anderzijds zijn nieuwe technieken ontwikkeld.

Na een inleidend hoofdstuk, staat in Hoofdstuk 2 de vraag centraal welke fiscale processen zich lenen voor verbetering door middel van data science technieken. Omdat het hele belastingdomein omvangrijk is, beperken we ons tot een belangrijk deelgebied: *het toezicht*. Door de toezichtsactiviteiten te categoriseren en voor elke categorie actviteiten na te gaan welke data science technieken waarde kunnen toevoegen, ontstaat een genuanceerd beeld van de (on-)mogelijkheden van data science. Het hoofdstuk geeft ook een overzicht van technieken die ingezet kunnen worden en is als zodanig bruikbaar voor belastingdiensten die verbeteringen willen realiseren in hun toezichtsactiviteiten met behulp van data science.

In Hoofdstuk 3 wordt een bestaande data science techniek, *logistische regressie*, uitgebreid. Logistische regressie heeft namelijk moeite met een bepaald type gegevens, de zogenaamde 'categoriale variabelen met veel waarden'. Denk bijvoorbeeld aan gegevens als 'woonplaats' of 'beroep'. Omdat deze variabelen regelmatig voorkomen bij belastingdiensten en omdat ook logistische regressie vaak wordt ingezet, is deze uitbreiding een nuttige aanvulling. Ook buiten het domein van belastingen komen categoriale variabelen met veel waarden voor en ook daar kan deze bijdrage nuttig zijn. Waarschijnlijk zijn de technieken om categoriale variabelen in te bedden

ook toe te passen bij andere voorspellende modellen dan logistische regressie, maar dit is niet verder onderzocht.

In Hoofdstuk 4 wordt gekeken naar anomalie detectie technieken (d.w.z. technieken om uitschieters in de data te vinden). Binnen belastingdata komen vaak een typische soort anomalieën voor die niet zo goed wordt gedetecteerd door bestaande anomalie technieken. In het hoofdstuk worden deze soort uitschieters beschreven en is er een nieuwe techniek ontwikkeld die deze anomalieën vindt. Binnen belastingdiensten is de techniek zinvol om in te zetten om belastingfraude te ontdekken of voor fouten in (handmatige) data-invoer op te sporen.

In Hoofdstuk 5 wordt ingegaan op een principe van rechtsgelijkheid: 'vergelijkbare gevallen dienen vergelijkbaar behandeld te worden'. Voor processen binnen een grote organisatie als een belastingdienst is het niet eenvoudig om te bewerkstelligen dat vergelijkbare gevallen een vergelijkbare behandeling krijgen, met name voor die processen die een menselijke beoordeling bevatten. Een eerste stap om te komen tot een vergelijkbare behandeling is het *meten* of gevallen vergelijkbaar behandeld worden en zo nee, waar in het proces deze ongelijkheid voorkomt. Hiervoor zijn in Hoofdstuk 5 twee nieuwe procedures ontwikkeld en beschreven. Deze nieuwe technieken behoren bij het vakgebied *process mining*.

In Hoofdstuk 6 wordt een ander aspect van de informatica-wetenschap opgepakt. Wetenschap bestaat niet enkel uit het ontdekken van nieuwe kennis, maar ook het ordenen van die kennis in logische categorieën. In dit hoofdstuk worden drie weinig gebruikte / nieuwe anomalie detectie technieken getest op een benchmark, recentelijk gepubliceerd door Goldstein en Uchida [43]. De vergelijking helpt om de sterke en zwakke punten van de verscheidene technieken te begrijpen. Verrassend blijkt dat de drie nieuwe technieken in 30% van de datasets in staat zijn om de benchmark te evenaren of te verbeteren. Daarnaast is het interessant om te zien hoe de nieuwe methoden zich verhouden tot de meer klassieke anomalie detectie methoden en welk soort anomalieën gevonden worden.

In Hoofdstuk 7 worden een viertal kleinere onderwerpen besproken. Dit hoofdstuk geeft een indruk van de verscheidenheid van data science toepassingen binnen het belastingdomein. De eerste paragraaf behandelt een onderwerp uit *HR Analytics*, een deelgebied binnen data science dat zich richt op toepassingen binnen Human Resources (HR). In de volgende paragraaf komt een toepassing van Reinforcement Learning technieken aan bod binnen het inningsproces van belastingdiensten. Reinforcement Learning is de tak van Artificiële Intelligentie die zich bezighoudt met het leren van strategieën bij bordspelen en computer games. Het derde onderwerp gaat over het uitleggen van risicomodellen aan eindgebruikers. Het laatste onderwerp betreft het toepassen van ideeën uit de Fuzzy Set gemeenschap op de belastingheffing.

Het proefschrift wordt afgesloten met conclusies en een korte vooruitblik.