PREDICTIVE ANALYTICS

JAN MESKENS

1. Introductie

De meeste bedrijven en overheden hebben de voorbije jaren een enorme schat aan data verzameld. Deze data is echter vaak zo groot en divers dat het moeilijk wordt om ze te interpreteren en er iets uit te leren. Met behulp van ***data mining*** wordt het echter wel mogelijk om kennis uit grote en complexe datasets te halen. Data mining destilleert deze kennis uit data met behulp van technieken uit disciplines als machine learning, artificial intelligence en statistiek.

Data mining kan men op twee verschillende manieren aanwenden: om data uit het verleden te verklaren door middel van beschrijvende statistiek (***descriptive analytics***) of om toekomstige trends en evenementen te voorspellen (***predictive analytics***). In deze infosessie geven we een overzicht van predictive analytics technieken, hoe we deze kunnen toepassen alsook enkele praktische cases en tools.

2. Predictieve modellen

Predictive analytics gaat men vooral toepassen om het gedrag van één of meerdere afhankelijke variabelen Y te voorspellen aan de hand van wijzigingen in onafhankelijke variabelen X. Zo zou het bijvoorbeeld mogelijk zijn om te voorspellen of iemand een bepaald product wil kopen (afhankelijke variabele) afgaande op een reeks van parameters (onafhankelijke variabelen) zoals de leeftijd van deze klant, burgerlijke stand, etc.

Om het gedrag van een afhankelijke variabele te voorspellen maakt predictive analytics gebruik van predictieve (voorspellende) modellen. Een predictief model kan men zien als een functie $f(x_1,..,x_n) = y$ dat als input de onafhankelijke variabelen krijgt en als output de waarde van de afhankelijke variabele geeft. Hier dient wel opgemerkt te worden dat een predictief model niet volledig betrouwbaar is. Deze betrouwbaarheid wordt vaak uitgedrukt als de accuraatheid van een predictief model. Een model met accuraatheid 0.92 zal in 92 % van de gevallen een correcte voorspelling maken. In alle andere gevallen zal de voorspelling foutief zijn.

Een predictief model wordt opgebouwd met behulp van een trainingsdataset. Deze dataset bevat onafhankelijke en afhankelijke variabelen die waargenomen werden in het verleden. Een trainingsalgoritme overloopt deze dataset met als doel een predictief model te bouwen dat de relatie tussen onafhankelijke en afhankelijke variabelen zo goed mogelijk probeert te voorspellen.

Er bestaan verschillende soorten predictieve modellen. In deze infosessie bespreken we vijf veelgebruikte modeltypes:

- **Regressieanalyse** bestaat uit het vinden van een lijn of curve die de items in de trainingsset kan benaderen. De meest gebruikte vorm van regressieanalyse is de lineaire regressie, waar naar een lineair verband gezocht wordt tussen de items in een trainingsdataset.
- **Clustering** gaat zoeken naar verschillende sterk gekoppelde groepen in een dataset. Deze samenhorende delen worden 'clusters' genoemd. Zo kan achteraf bepaald worden tot welke groep de afhankelijke variabelen behoren op basis van de onafhankelijke variabelen.
- **Associatieregels** beschrijven de verbanden tussen verscheidene items in de dataset. Een voorbeeld associatie tussen variabelen A, B en C is **{A,B}=>{C}**. Deze associatie wil zeggen dat als A en B samen voorkomen, C vervolgens ook gaat voorkomen. Elke associatie heeft ook een bepaalde confidence en support: de confidence wil zeggen hoe zeker het is dat de rechterkant van de associatie volgt op de linkerkant, de support duidt op het aantal keer dat de volledige regel voorkomt in de dataset.
- **Beslissingsbomen** beschrijven een dataset als een boom. Een beslissingsboom overloopt men van de top richting de bladeren. Deze eindnodes bevatten tenslotte de juiste waarde van de afhankelijke variabele.
- **Neurale netwerken** is een techniek uit de artificiële intelligentie dewelke het verband tussen afhankelijke en onafhankelijke variabelen voorstelt als een verzameling intern verbonden nodes.

3. Predictive analytics tools

Predicitive analytics wordt vandaag ondersteund in verscheidene types tools. We onderscheiden hier drie categorieën:

- **Scripting tools**, dewelke als input een script vereisen dat geschreven is in een specifieke taal. De meest bekende tools hier zijn R, S+, Matlab en Octave.

- **Form-based tools**, waar men in een pre-designed formulier de inputs en parameters van een predictive analytics algoritme kan ingeven. Hier onderscheiden we Weka als meest mature tool.

- **Visual programming tools** zijn tools waar men blokjes op een visuele manier met elkaar kan verbinden. Elk blokje stelt een bepaalde data operatie of predictive analytics algoritme voor. Voorbeeld blokjes zijn dataselectie, clustering, tabelvisualisatie, etc. De meest voorname tools hier zijn grote spelers als SPSS Modeler, SAS Enterprise Miner, Oracle Data Miner en Tibco Spotfire Miner.

4. POC: Bestrijding van de sociale fraude

We hebben predictive analytics praktisch toegepast in een Proof Of Concept (POC) project ter bestrijding van de sociale fraude. Het doel van deze POC is om te onderzoeken of de RSZ-inspecties beter gestuurd kunnen worden. Dit kan gaan in drie fases:

- via predictive analytics worden mogelijke fraudegevallen opgespoord;
- een inspectie onderzoekt deze fraudegevallen;
- de feedback van deze inspecteurs wordt teruggekoppeld naar het predictief model. Het doel hiervan is om het model altijd maar accurater te maken.

Als concreet soort fraudegeval hebben we bedrieglijke faillissementen onderzocht. Vaak zijn bedrijven die bedrieglijk failliet draaien aan elkaar gelinkt. Deze links proberen we op te sporen door middel van een predictief model dat gebruikmaakt van association rules. Uit de eerste beperkte resultaten en de terugkoppelingen van de dienst TADT[1] blijkt dat we op deze manier wel degelijk mogelijke fraude kunnen opsporen.

5. Wat hebben we geleerd

Gedurende de studie rond predictive analytics en tijdens het uitvoeren van de POC hebben we verschillende lessen getrokken voor de toekomst:

[1] TADT staat voor Team Analyse - Detectie Team en is een onderdeel van de inspectiediensten van de RSZ.

- **Voorbereiden van data neemt veel tijd in beslag.** Bij het opstellen van een predictief model vertrekken we van een dataset. Het kost echter veel tijd om deze dataset te destilleren uit de beschikbare dataformaten, om de inhoud van de data goed te begrijpen en om de juiste variabelen te kiezen. Dit alles gebeurt tijdens de datavoorbereidingsfase. Wij adviseren om tijdens deze fase met verscheidene rollen de beschikbare data te bekijken en om te vormen tot een dataset. Uit onze ervaring raden wij samenwerking aan met specialisten op het vlak van de business, data quality, databases, predictive analytics en datawarehouses.

- **Predictive analytics is geen "black box".** In veel recente literatuur en verkoopspraatjes wordt predictive analytics omschreven als een techniek die voornamelijk wordt uitgevoerd door tools. Tools kiezen het juiste model, parametriseren dit correct en passen vervolgens het model toe op de data. Dit moeten we echter nuanceren. Tools maken het makkelijker om predictive analytics projecten tot een goed einde te brengen, maar er is nog steeds nood aan de nodige analytics kennis. Bepaalde keuzes voor tools – zoals het juiste model – kunnen dagen duren, terwijl een analytics expert bepaalde keuzes onmiddellijk kan uitsluiten.

- **Bouw predictive analytics projecten gradueel uit.** Wanneer men een predictive analytics project eerst kleinschalig begint, kan men deze eerste resultaten gebruiken om meer medewerking te verkrijgen van andere stakeholders in het project. Deze positieve sfeer draagt zeker bij tot het vlotter slagen van een project.

- **Er is nood aan een nieuwe rol: de (predictive) analyst.** Een predictive analyst is een uniek profiel dat verscheidene skills combineert. Eerst en vooral is de basisvereiste kennis hebben van analytics en IT. Maar bovenop deze skills verwachten wij van een predictive analyst om de businessnoden te begrijpen en om vlot met iedereen die betrokken is bij een project te communiceren.