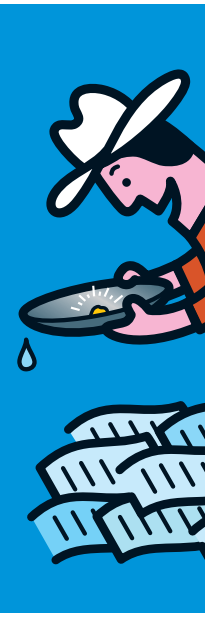

SAS al jarenlang toonaangevend met Enterprise Miner

Belangstelling voor datamining groeit

Hans Lamboo

SAS komt van oudsher uit de statistische hoek. De processen rondom statistiek zijn in de loop der tijd gegroeid, met data-integratie en datakwaliteit aan de ene en rapportage aan de andere kant. SAS is meegegroeid en staat al jarenlang aan de top als het om datamining gaat.

"BI bestaat grofweg uit drie componenten: techniek, rapportage en analyse. In de afgelopen jaren is binnen de BI heel veel nadruk komen te liggen op rapportages", begint Rens Feenstra, Senior Technology Solution Consultant bij SAS. "Je zult immers eerst goed moeten kunnen rapporteren voordat je verder kunt gaan kijken. Levert de BI-omgeving eenmaal goede, betrouwbare rapporten, dan heb je inzicht in wat er in het verleden gebeurd is. Dan kom je op het moment dat de vraagt rijst: kan er verklaard worden waarom zich een bepaalde ontwikkeling heeft voorgedaan, bijvoorbeeld waarom een bepaalde regio een afname van de verkoop laat zien. Dan kom je op het terrein van de statistische analyse. "

Dat vakgebied is de specialiteit van Feenstra. Al jarenlang helpt hij bedrijven bij het analyseren van hun data. Niet dat dat direct de diepte ingaat. "Meestal gaat het in het begin om simpele statistiek zoals bijvoorbeeld gemiddeldes en soms standaarddeviaties. De business, bij monde van een marketeer of sales manager, heeft dan meestal al een goed verhaal om de afwijkende ontwikkeling in een bepaalde regio, of met een bepaald product te verklaren. Dat kan te maken hebben met zaken als leeftijdssamenstelling of inkomensgroepen in een bepaalde regio. Dat soort gegevens kun je zo al boven water krijgen."

Wisselwerking

De groeiende belangstelling voor de analysecomponent van BI komt niet uit de lucht vallen. Vele bedrijven hebben inmiddels hun operationele systemen en rapportage op orde gebracht. De bedrijven bereiken een min of meer gelijk niveau van groei, van evolutie. Ze realiseren zich dat ze allemaal over dezelfde marktinformatie beschikken, en dat dus hun eigen data het verschil moeten maken, dat ze zich door slimme analyse van eigen data kunnen onderscheiden van de business. De essentie van boeken als 'Competing on Analytics' van Davenport begint door te dringen en realiteit te worden.
"Het is dus van belang om die ene kleine afwijking in de business processen beter te begrijpen, doordat je weet waarom het anders is en er adequaat op in te spelen. Daarbij kan datamining een rol spelen", zegt Feenstra. "In het proces van het verzamelen van gegevens, de problemen zien en de problemen benoemen, bereik je op een gegeven moment een punt dat je naar verbanden wilt zoeken die niet voor de hand liggen, die nauwelijks te ontdekken zijn door mensen. Zeker door de groei van de hoeveelheid gegevens, wordt het voor mensen steeds moeilijker bevatbaar. Dan ga je de mogelijkheden van datamining onderzoeken. "

> Datamining is dus feitelijk een proces en zeker niet alleen een techniekje of software tool

Het kan zijn dat de reguliere BI-rapportages zijn uitgerust met een kolom 'voorspelling' of 'verwachting'; dat ene cijfer is verkregen door dataminingtechnieken en helpt bij de bijsturing van de processen."

Bij het toepassen van datamining moet door de gebruiker zelf een model worden ontwikkeld. "We hebben modellen in onze datamining tool zitten, maar wat een model specifiek maakt is de toepassing van het model op de data van de organisatie die het gebruikt. Want ondanks het feit dat bedrijven steeds meer op elkaar gaan lijken, zijn er toch nog zoveel verschillen dat het toe-

Afbeelding 1: Verschillende segmenten die in data worden herkend.

passen van een model op de ene organisatie heel anders kan uitpakken dan bij een ander bedrijf", zegt Feenstra. "Het is de combinatie van technieken en de eigen data, en – eigenlijk nog belangrijker – de eigen kennis van het ecosysteem van de organisatie, die ervoor zorgt dat datamining een oplossing kan zijn."
Binnen SAS Enterprise Miner wordt een methodiek gebruikt met de naam SEMMA, acroniem van Sample, Explore, Modify, Model en Assess. "Dat is een cirkel. Je neemt een sample uit de bak met beschikbare data. Dan onderzoek je hoe de verschillende variabelen zijn vastgelegd. Stel dat je de leeftijd van de personen in de dataverzameling wilt gebruiken. In de data ligt echter alleen de geboortedatum vast. Dan moet je de data dus transformeren. Vervolgens bepaal je welke van de variabelen verder te gebruiken zijn voor datamining, waar je bijvoorbeeld de leeftijd niet als zodanig gaat gebruiken, maar transformeert naar een leeftijdsgroep, een indeling die al binnen de organisatie wordt gebruikt."

Monitoring van de modellen tijdens de toepassing is erg belangrijk. "Door het gebruik van het model verandert je klantenbestand en zullen dus ook de data veranderen waarop de modellen worden gemaakt. Een model heeft dus in principe een beperkte houdbaarheid, " legt Feenstra uit.
Het hebben van een datawarehouse is geen aanbeveling voor de toepassing van datamining. "Een datawarehouse is opgezet met een bepaald doel, meestal rapportage. De data liggen daarin zo vast dat dat goed is voor die rapportages, misschien wel geaggregeerd. Terwijl je bij datamining juist wilt kijken naar de individuele details. Een datawarehouse kán handig zijn, maar hoeft dat dus helemaal niet te zijn", zegt Feenstra. "Eigenlijk wil je je helemaal niet beperken tot bepaalde gegevens. Meestal worden dus de data voor datamining in de originele vorm uit de onderliggende operationele systemen gehaald. Vervolgens gebruik je tools die op zoek gaan naar variabelen die iets zeggen over die data. Wat niets zegt gooi je er uit; dan houd je van de oorspronkelijke 3000 variabelen er bijvoorbeeld nog maar 700 over. Vervolgens zijn er diverse explore-technieken om grafisch in de data te kunnen rondkijken. Vaak zie je dan al bepaalde opvallende zaken, kun je op het oog vaststellen waar zich bijzondere zaken openbaren die nader onderzoek behoeven. Of je zet verschillende variabelen tegen elkaar uit."

Mijnwerkers

Bij het horen van de term 'datamining' doemt al snel het beeld van een enorme berg data met mannetjes met mijnwerkershelmen op. To mine betekent in het Engels ontginnen, het ontginnen van data. Daar zit analyse eigenlijk al in. Dat beeld kent Feenstra wel. "Maar datamining begint al veel eerder. Die berg data bijvoorbeeld, waar komt die vandaan? Zijn dat alle beschikbare data? Een selectie daaruit? Het begint eigenlijk met het

formuleren van een probleemstelling. Je kunt wel een stel technieken neerzetten, maar op welke vraag zoek je antwoord? Dat is een belangrijke fase. Stel, je wilt een marketingcampagne opzetten om meer lezers te vinden voor een bepaald blad. Die vraag richt de blik op bepaalde data waar je je technieken op los kunt laten. Daarmee construeer je een model waarvan de uitkomsten statistisch gezien correct zullen zijn, maar waarvan een business persoon moet bepalen of het valide en bruikbaar is binnen de context en kennis binnen de organisatie. Vervolgens moet het model getest worden in de praktijk en voortdurend worden gemonitord."
Datamining is dus feitelijk een proces en zeker niet alleen een techniekje of software tool die door een nerd in de kelder wordt gebruikt. "Datamining is geen specifieke IT-zaak", stelt Feenstra. "Degene die achter de knoppen zit moet een business persoon zijn. Die moet de vraagstelling formuleren en vervolgens samen met IT bepalen welke data er in huis zijn die bij de oplossing een rol kunnen spelen. En welke data er eventueel van buiten moeten worden aangekocht. Stel, dat er vermoed wordt dat de verkoopcijfers verband houden met het weer, dan is het nodig om meteorologische gegevens in huis te halen. Dat zijn geen gegevens die normaal zitten opgeslagen in het operationele systeem van een bedrijf, en moeten dus worden ingekocht. Dat zal niet altijd zo zijn, maar vaak komt het bij marketingvraagstukken aan de orde. Datamining is een wisselwerking tussen een aantal mensen: het is de business persoon die de probleemstelling formuleert, het is de IT'er die weet welke data beschikbaar zijn, en de statisticus die het probleem analyseert en bepaalt in welke vorm hij over welke data zou willen beschikken. Hier zie je ook de belangrijkheid van de instelling van een BICC waar al die disciplines bij elkaar komen en waar naar onze mening zeker ook een analytisch persoon in moet zitten."

> Als het er niet inzit haal je het er ook niet uit

"Ik ken een bedrijf in Nederland dat de opbrengsten van marketingcampagnes heeft verdrievoudigd door gebruik te maken van datamining", vertelt Feenstra. "En ik ken een mooi voorbeeld met kwaliteitsproblemen in een staalfabriek. Die zijn opgelost met een dataminingtoepassing. Er kwam uit dat er verband bestond tussen de kwaliteit en de werkwijze van verschillende productieploegen. Op het eerste gezicht lijkt dat verband voor de hand te liggen. Maar staalfabricage is een vastliggend productieproces: de verschillende ploegen bedienen immers dezelfde machines en gebruiken dezelfde parameters. Wat was dan de oorzaak van het kwaliteitsverschil?
Op een bepaald moment moet elke ploeg de rollen waar het staal overheen loopt verwisselen. De ene ploeg deed dat anders dan de andere. Waardoor de luchtstroom van buiten vuil aanvoerde dat op de wals kwam en vervolgens op het staal. Met datamining is geconstateerd dat er verschil bestond tussen de ploegen. Maar het echte antwoord kwam natuurlijk uit de praktijk".

Conclusie

Een vorm van datamining wordt ook toegepast in het fraudedetectieplatform van SAS, dat real-time controleert of er sprake zou kunnen zijn van een frauduleuze handeling met bijvoorbeeld credit cards. Hoe ziet de toekomst van datamining eruit? Feenstra's SAS-collega Rein Mertens vertelt over de SAS Summer Academy in 2008. "De meeste aanwezigen hadden een achtergrond in de econometrie. Dat is precies de groep van mensen die potentieel een analytische rol in organisaties kan gaan vervullen. We zien dat ook partnerorganisaties er over denken; ze realiseren zich dat ze veelal te traditioneel met data-integratie en BI bezig zijn en dat er veel meer uit de beschikbare data te halen valt. We hebben bijvoorbeeld een partner die er bewust voor gekozen heeft om een van de deelnemers aan de genoemde Summer Academy aan te nemen om bij klanten in analyseprojecten te laten meelopen. Initieel zonder dat de klant daarvoor hoeft te betalen, enerzijds om kennis op te doen over BI in het algemeen, maar ook om tegelijkertijd meer met die data te doen. Om aan te kunnen tonen aan de klant dat er meer te halen is uit de data dan de mooie rapportages, bijvoorbeeld beter voorspellen. Ik hoor dat dit binnen de partner community steeds vaker gebeurt."
Datamining zal vaker worden ingezet in de toekomst, daarvan zijn Feenstra en Mertens overtuigd. "We bevinden ons momenteel in een economische terugval. Maar in de afgelopen twee maanden hebben toch weer meer klanten gekozen voor de dataminingoplossing van SAS", meldt Feenstra. "Daar zit bijvoorbeeld een grote autodealer bij. Men realiseert zich dat er juist nu maatregelen genomen moeten worden om overeind te blijven. Onze klanten zien dat in, dus dat is ook een van de redenen waarom we bij SAS niet zoveel merken van de crisis."
Datamining is niet per se verbonden aan één SAS-product, Enterprise Miner. Ook met andere tools kan het dataminingproces worden ingevuld. Het voordeel van een product als Enterprise Miner is dat er stapsgewijs wordt gewerkt naar een toepasbaar model, en de analist dus niet het wiel hoeft uit te vinden. Mertens: "Wij leveren een complete oplossing en niet alleen een tool. Dat kán wel, zoals bijvoorbeeld Predictive Asset Maintenance. Dat is een oplossing waar datamining in zit om te kunnen voorspellen hoe er slimmer onderhoud kan worden gepleegd."

Feenstra besluit met een relativerende opmerking. "Datamining is zo goed als de data waar je over beschikt. Het is geen wondermiddel dat altijd antwoord geeft. Als het er niet inzit, haal je het er ook niet uit."

Hans Lamboo is hoofdredacteur van Database Magazine.