Data kunnen unieke bron van concurrentievoordeel zijn

Analytics – een blik op de toekomst

Tom Breur

Voor het gros van de BI community (maar ook daarbuiten) staat Business Intelligence synoniem aan het bouwen van rapporten en kubussen. Dan hebben we het over tools als Cognos, Business Objects, MicroStrategy, maar ook Pentaho, JasperSoft, of QlikView. Merk op dat dergelijke rapportages en OLAP-tools allemaal betrekking hebben op historische data.

En dat is dan de informatie waar organisaties mee worden 'bestuurd'. Als je de metafoor 'besturen' gebruikt, dan roept het associaties op met het rijden in een auto met een geblindeerde voorruit waarbij je 'stuurt' door in de achteruitkijkspiegel te kijken. En dan maar hopen dat je ongehavend op de plaats van bestemming komt. In de waardeketen van ruwe data naar informatie naar actie, biedt 'Analytics' het gereedschap voor de meest verfijnde en geavanceerde toepassingen van BI. Waarmee je dus ook af en toe *vooruit* wilt kijken.

Volgens Gartner is dat iets nieuws. Ze hebben er in ieder geval een nieuwe term voor uitgevonden: pattern based strategies. Het idee is dat bedrijven die beter in staat zijn om op veranderingen in hun omgeving te anticiperen (met behulp van Analytics) een duurzaam concurrentievoordeel kunnen verwerven.

Analytics software: SAS en SPSS

Voor SAS bestaat het leeuwendeel van het gebruik vooral uit SAS Base en SAS Stat licenties. Ze hebben ook een aantal nicheproducten, denk bijvoorbeeld aan SAS/IML voor matrix algebra, SAS/ETS voor tijdreeksanalyse, SAS/OR voor optimalisatievraagstukken (linear programming) enzovoort.
Daarnaast biedt SAS een 'vlaggenschip' aan: SAS Enterprise Miner. Dit is een volledig grafisch gestuurde, horizontale datamining suite. Deze draait overigens ook gewoon op SAS Base, zodat 'die-hard' SAS-hackers in de onderliggende code hun weg kunnen vinden.
Voor SPSS wordt het merendeel van gebruik bepaald door SPSS Base en SPSS Tables. Daarnaast zijn er modules zoals SPSS Missing Value Analysis voor de bewerking van missing data, of bijvoorbeeld SPSS Categories (in Leiden ontwikkeld) voor de analyse van data op 'lage' meetniveaus (ordinale en nominale meetgegevens).

Voor IBM/SPSS is SPSS Modeler (voorheen Clementine genaamd) het 'vlaggenschip'. Dit product werd oorspronkelijk door ISL ontwikkeld, dat in 1999 door SPSS werd overgenomen. Clementine wordt beschouwd als de eerste gebruiksvriendelijke, volledig grafisch gestuurde datamining suite.
SPSS heeft zo'n tien jaar eerder dan SAS de overstap gemaakt naar een volledig grafische gebruikersinterface. Alhoewel er 'onder de motorkap' nooit veel aan de kern van het product is veranderd, is er in de laatste 20 jaar enorm veel aan de gebruikersinterface gesleuteld. Zowel de aansturing aan de voorkant met commando's, als de bewerking van output kan volledig grafisch. Met Python en VB.net (en ook nog SaxBasic) kun je eindeloos automatiseren.

Deze voorsprong in de ontwikkeling van de user interface maakt dat IBM/SPSS gebruiksvriendelijker is, en dat de leercurves korter zijn. En doordat ze al jaren hun producten nagenoeg gratis ter beschikking stellen aan universiteiten (het zogenaamde SURF contract) is er op de arbeidsmarkt een veel grotere poel aan data-analisten met op zijn minst enige SPSS-kennis.

Analytics Software: de rest

Zoals gezegd, SAS en SPSS nemen het leeuwendeel van de Analytics-markt voor hun rekening. Wikipedia biedt een uitgebreid overzicht van het hele speelveld (http://en.wikipedia.org/wiki/List_of_statistical_packages). Behalve SAS en IBM/SPSS blijven er nog enkele kleine(re) vergelijkbare spelers over. Denk dan aan Statistica, JMP (eigendom van SAS), of STATA (op MIT ontwikkeld). STATA is superieur voor tijdreeksanalyse, maar vergt meer affiniteit met programmeren. Statistica en JMP zijn SPSS klonen, dus volledig grafisch te bedienen. SAS Enterprise Guide is ook een grafisch product dat binnen de SAS community echter nooit echt met enthousiasme is omarmd.

Statistica lijkt inhoudelijk op SPSS, hun minuscule marktaandeel ligt dus vooral aan een verschil in marketing, en niet (of nauwelijks) aan een verschil in functionaliteit.
Ook JMP is min of meer equivalent aan SPSS, het staat voor John's Macintosh Program. John Sall was één van de vier SAS-oprichters van het eerste uur (sinds 1973) en werkt er nog steeds. JMP was oorspronkelijk alleen beschikbaar voor de Apple Macintosh, maar later werd ook een versie voor de PC uitgebracht.
Er is ook een groep van mathematische pakketten zoals Matlab, S-Plus en Gauss. Je moet dan wel op zijn minst weten hoe uiteenlopende algoritmes 'werken' en geprogrammeerd moeten worden. En dan is er nog een scala aan open source tools zoals: R, Weka, RapidMiner, en vele, vele anderen. Het product R geniet de laatste tijd een warme belangstelling, zowel SAS als IBM/SPSS bieden extensies waarmee koppelingen naar R gelegd kunnen worden.

Het belangrijkste argument om met open source tools te werken is meestal geld: je kunt ze gratis downloaden. Voor commerciële toepassingen van Analytics vervalt dat argument grotendeels, omdat het aandeel van software licentiekosten in de total cost of ownership (TCO) relatief laag is. Opleidings- en personeelskosten zijn meestal een veelvoud.
Alleen datamining suites zoals IBM/SPSS Modeler en SAS Enterprise Miner hebben een stevig prijskaartje, de meeste overige tools koop je voor enkele duizenden euro's. Vandaar dat open source vooral in de academische wereld wordt gebruikt. Met uitzondering van SPSS, SAS, Statistica, Angoss, KXEN en JMP vallen de meeste andere tools in de categorie 'Meccano doos'. Dat wil zeggen dat je moet programmeren om überhaupt resultaten te krijgen. Dat geeft als voordeel maximale flexibiliteit, maar als nadeel dat de leercurve nog ietsje langer is, en de productiviteit lager. De eisen die zo'n product aan het (inhoudelijke) niveau van analisten stelt liggen ook hoger.

Supervised versus unsupervised technieken

Het analytische speelveld kent een tweedeling in 'supervised' en 'unsupervised' technieken (ook wel 'directed' en 'undirected' genoemd). Het verschil is de aanwezigheid van een specifieke doelvariabele die je probeert te voorspellen. Bij supervised technieken probeer je die te voorspellen *uit* een array van input variabelen. Bij unsupervised technieken probeer je samenhang te vinden *binnen* een array van (input) variabelen.
Voorbeelden van unsupervised technieken zijn clustering, associatie analyse (nearest neighbour algoritmes), of collaborative filtering. Meestal bedoeld om *inzicht* te verschaffen.
Bij clustering probeer je groepen klanten te vinden waarbij de overeenkomsten binnen het cluster zo groot mogelijk zijn. Tegelijkertijd probeer je clusters onderling zo veel mogelijk van elkaar te laten verschillen. Dit wordt bijvoorbeeld gebruikt voor marktsegmentatie.
Collaborative filtering en associatie analyse zijn procedures waar bijna dezelfde algoritmes aan ten grondslag liggen als bij clustering. Alleen worden ze gebruikt om de aanbevelingen te genereren zoals bekend van Bol.com of Amazon.
Supervised technieken hebben *wel* een expliciete doelvariabele. Voorbeelden zijn decision trees, regressie analyse, (de meeste soorten) neurale netwerken, of forecasting (wat meestal een speciaal soort regressie is). Mogelijke toepassingen: doelgroepbepaling bij direct marketing (DM), credit scoring, fraudedetectie, bepalen van lifetime value (LTV) enzovoort.

Voorspellen is niet meer wat het geweest is

Tot dusver heb ik het woord 'voorspellen' nogal lichtzinnig gebruikt. Maar wat bedoelen we daar nu eigenlijk mee? We hebben een dataset met een historisch klantbeeld, bijvoorbeeld alle attributen die bij de klant hoorden toen we een direct marketing pilot actie deden. Vervolgens stellen we vast wie wel en niet hebben gereageerd. Nadat we de samenhang tussen klantkenmerken en koopgedrag hebben bepaald, laten we dit verband los op het huidige klantbeeld. Zo zie je dan op individueel niveau de gelijkenis met respondenten in het verleden, en dus de kans dat iemand zal reageren.

Er zijn twee voetangels waar we even overheen gestapt zijn. Op de eerste plaats hebben we het verband tussen input variabelen en koopgedrag losgelaten op het huidige klantbeeld. De resultaten daarvan zijn natuurlijk alleen geldig als exact dezelfde samenhang vandaag de dag nog steeds geldig is. Alleen onder die aanname kunnen we 'voorspellingen' doen. Wat we hier een 'voorspelling' noemen, komt dus in wezen neer op het classificeren van het verleden. Als we stellen dat iemand 80 procent kans heeft om te zullen reageren, dan bedoelen we dat dit soort klanten ten tijde van de pilot mail in vier van de vijf gevallen heeft *gereageerd*. Anders kunnen we die kans niet berekenen.
De aanname dat de toekomst zal lijken op het verleden moeten we maken, maar het is ook een hele zware. Als deze niet opgaat, is onze 'voorspelling' ongeldig. Meestal is het verleden een redelijk goede voorspeller van de toekomst, denk maar aan het weer. Als het vandaag warm is, dan is de kans groot dat het morgen ook weer mooi weer zal zijn. En vice versa. Maar helaas geldt dit niet altijd, en in die gevallen gaan we dan ook de mist in.
De tweede reden waarom 'standaard' voorspellingen verraderlijk

Ontstaan SAS en SPSS

Sinds jaar en dag wordt de Analytics softwaremarkt gedomineerd door twee grote spelers, SAS en SPSS (deze laatste is in 2009 overgenomen door IBM). SAS werd 'geboren' in 1966, en vanaf 1968 was Jim Goodnight (huidige CEO) betrokken bij de ontwikkeling. Uiteindelijk werd SAS Institute Inc. in 1976 juridisch opgericht. SPSS werd in 1968 voor het eerst beschikbaar gesteld, net als SAS aanvankelijk vooral voor gebruik in de academische gemeenschap.

kunnen zijn is wat subtieler. We gebruiken een dataset met historische klantgegevens, en relateren die aan responsgedrag. Deze verbanden laten we los op actuele klantgegevens. De makke zit in actuele gegevens die buiten de range vallen die we in het verleden hebben geobserveerd. De wiskundige berekeningen om de voorspelling te doen werken nog steeds. Papier is geduldig, en formules gewillig. Maar hoe *geldig* zijn die bewerkingen? Voor 'vreemde' waardes ontbreekt elke empirische basis voor de berekening, maar een wiskunde formule hoor je niet klagen.
De traditionele aanpak van voorspellen draagt dus een aantal risico's in zich, en daar is sinds de jongste kredietcrisis dan ook al het nodige over geschreven. Er bestaan weliswaar alternatieven (bijvoorbeeld systeemtheorie), maar deze zijn nauwelijks bekend en worden ook weinig gebruikt. Maar ook dergelijke modellen kennen hun beperkingen. Abelson (Statistics as Principled Argument, 1995): "There ain't no such thing as a free *hunch*".

Supervised technieken

Er is een overmaat aan beschikbare technieken, en elk jaar worden er weer nieuwe algoritmes uitgevonden. De belangrijkste voor de praktijk zijn decision trees, regressie analyse en neurale netwerken. Rule based en fuzzy logic algoritmes lijken uit de gratie geraakt. Genetisch programmeren is een bijzonder geval van een rule based techniek, en genetische algoritmes zijn een optimalisatieprocedure en helemaal geen datamining algoritme (het wordt dus in combinatie met andere algoritmes ingezet). De laatste jaren komen *Bayesian networks* en support vector machines op, hun doorbraak zal afhangen van de vraag of er goede, gebruiksvriendelijke software beschikbaar komt. Forecasting is (meestal) een bijzonder geval van regressie analyse op tijdreeksgegevens.

Decision trees

Van deze technieken zijn decision trees het meest eenvoudig en toegankelijk. Software is relatief goedkoop. Vanaf gratis Excel 2007 plug-ins tot aan enkele duizenden euro's per licentie. Voordeel is dat de techniek erg transparant is, en daardoor voor iedereen gemakkelijk te begrijpen. Daarnaast kunnen decision trees probleemloos met missing data werken (dit in tegenstelling tot regressie analyse en neurale netwerken).
Decision trees werken op het principe van 'divide and conquer'. Nadat een doelvariabele is gekozen gaat het decision tree algoritme (waar er zeer vele van zijn, zie voor een overzicht: www.xlntconsulting.com/resources/decision-trees.htm) op zoek naar de variabele die het beste in staat is om de twee categorieën (eentjes en nullen) van elkaar te onderscheiden. Voor een continue doelvariabele werkt het algoritme vergelijkbaar, in dat geval probeer je hoge van lage waarden onderscheiden.

Nadat in de hele dataset de variabele is gevonden die het best de doelvariabele onderscheidt, wordt voor de twee of meer subdatasets deze procedure herhaald. Daarom praten we soms over algoritmes voor recursieve partitionering. Het splitsen blijft doorgaan tot er geen zinnige subgroepen meer te vinden zijn. Dit resulteert dan in een set van if-then-else regels met aflopende proporties enen.
Bijvoorbeeld; als je koopgedrag modelleert vind je wellicht dat geslacht de beste predictor is. Vrouwen blijken meer te kopen dan mannen (dit is puur hypothetisch, uiteraard). Binnen de deelverzameling van vrouwelijke klanten blijkt inkomen de beste voorspeller van koopgedrag. Vrouwen met een hoog inkomen kopen meer dan vrouwen met een laag inkomen. Als dit alles is (in de praktijk valt er meestal meer door te splitsen), ziet het model er als volgt uit:

```
IF geslacht = vrouw AND inkomen = hoog THEN
                                  koopgeneigdheid = 0,1
ELSE IF geslacht = vrouw AND inkomen = laag THEN
                                 koopgeneigdheid = 0,08
ELSE IF geslacht = man THEN koopgeneigdheid = 0,07
```

Regressie analyse

Bij regressie analyse worden variabelen niet *na elkaar* maar *tegelijkertijd* gebruikt om de doelvariabele te beschrijven. Net zoals bij decision trees zijn er ook heel veel smaken van regressie analyse. In veel gevallen zul je met een stapsgewijze procedure de selectie van variabelen (en aantal) bepalen die in het uiteindelijke model komen.
Een nadeel van regressie is dat ontbrekende waarnemingen (NULL values) moeten worden vervangen. Elke rij die ook maar één *missing* heeft wordt anders genegeerd. Een tweede nadeel is dat categorische variabelen (dus met *meerdere* discrete klassen) eerst moeten worden bewerkt voor deze meegenomen kunnen worden in de analyse.
Het uiteindelijke regressiemodel krijgt de vorm van een array aan input variabelen en bijbehorende gewichten (regressie coëfficiënten). Bijvoorbeeld:

```
Koopgeneigdheid = constante + 0,03*geslacht +
                              0,02*inkomen + error term
```

Neurale netwerken

Neurale netwerken zijn feitelijk een bijzonder soort van regressie analyse. Het verschil zit in de optimalisatieprocedure, de manier waarop regressie coëfficiënten worden bepaald. Een tweede verschil is dat door de complexiteit het model niet meer door mensen kan worden geïnterpreteerd. Dit kan een groot nadeel zijn, soms zelfs onoverkomelijk.
Door het gebruik van een zogenaamde 'hidden layer' (hulpvariabelen die afhankelijk zijn van de input variabelen) kunnen functies van arbitraire complexiteit worden gemodelleerd. Mits je over genoeg data beschikt, wel te verstaan. Met neurale netwerken kun je als regel de meest accurate voorspellingen doen. Net als bij regressie analyse moeten missings eerst vervangen worden, en moeten categoriale variabelen eerst bewerkt worden om ze mee te kunnen nemen in het model.

Kredietcrisis & voorspellen

Problemen met ongeldige 'voorspellingen' zijn een van de oorzaken van de kredietcrisis geweest. Men berekende default modellen (kans op betalingsachterstanden) in een tijd van lage werkloosheid, economische groei, en stijgende huizenprijzen. Maar die werkelijkheid veranderde, en daardoor verloren de oude modellen hun geldigheid. Het model zelf sputtert niet tegen, blijft 'trouw' zijn voorspellingen genereren. En als business verantwoordelijken deze beperkingen dan niet (snel genoeg) onderkennen, dan neem je onverantwoorde risico's als je daar beslissingen op baseert, zoals het al of niet verstrekken van hypotheken. Zo bezien is de kredietcrisis het bankroet van corporate governance. Als executives dit soort voorspellende instrumenten niet kunnen doorgronden (en hun beperkingen), hoe kun je dan verantwoorde bedrijfsvoering handhaven?

Voorbeeld model:

$$\text{Koopgeneigdheid} = \text{constante} + 0{,}03*\text{geslacht}^4 - 0{,}23*\text{geslacht}^3 + 0{,}16*\text{geslacht}^2 + 1{,}88*\text{geslacht}^4 * \text{inkomen}^3 - 0{,}02*\text{geslacht}^3*\text{inkomen}^2 + 2{,}01*\text{geslacht}^3 + 0{,}46*\text{geslacht}^2 - 0{,}02*\text{geslacht} + \text{error term}$$

Als je decision trees, regressie analyse, en neurale netwerken met elkaar vergelijkt dan zijn decision trees het gemakkelijkst in gebruik, en neurale netwerken het moeilijkst. Gemiddeld genomen krijg je met neurale netwerken de meest accurate voorspelling, gevolgd door regressie analyse en dan decision trees. Maar universeel is die regel zeker niet!

Waar gaan Analytics heen?

Het Analytics speelveld werd in het verleden gekenmerkt door software met ondoorgrondelijke user interfaces, waarvan het gebruik voorbehouden was aan mensen met diepgaande statistische kennis. Maar applicaties worden steeds gebruiksvriendelijker en laagdrempeliger. Hierdoor komen ze binnen ieders bereik.

Daarnaast is het aantal mogelijke business problemen waar Analytics een verdienstelijke rol bij kan spelen heel groot. Er zijn de laatste jaren veel boeken verschenen die de waarde van 'Analytics' onderschrijven, en minstens zo belangrijk: onder de aandacht van management hebben gebracht. Denk aan: 'Competing on Analytics' van Davenport en Harris (2007), 'Supercrunchers' van Ian Ayres (2007), 'Data Driven' van Redman (2008), of 'Analytics at Work' van Davenport, Harris en Morison (2010).

De vraag naar Analytics is veel groter dan het aanbod aan academici met een zware wiskundige achtergrond die de arbeidsmarkt kan leveren. Deze spanning heeft een nieuwe ontwikkeling gestimuleerd: Analytics toepassingen waarbij business gebruikers (Self-service BI) rechtstreeks toegang krijgen tot technologie die door middel van templates of wizards alleen nog maar om input vraagt over business parameters. De statistiek is helemaal onder de motorkap weggewerkt.

Net zoals het merendeel van de huidige computergebruikers nog nooit een DOS prompt heeft gezien, komt er wellicht nog een tijd dat Analytics toepassingen net zo gemeengoed zijn geworden als de grafische schil om een besturingssysteem, en dat we dus niet meer hoeven na te denken over de manier waarop gegevens in de database worden opgeslagen. Of over geldigheid van statistische bewerkingen. Maar daar zijn we nog niet.

Conclusie

In tal van populaire managementboeken is te lezen dat de toekomst behoort aan bedrijven die op een slimme manier gebruik weten te maken van hun data. En Gartner beweert hetzelfde met hun 'pattern based strategies'.

De softwaremarkt voor Analytics software wordt al heel lang gedomineerd door twee grote spelers: SAS en IBM/SPSS. Daarnaast is er nog een paar dozijn alternatieven waarvan het merendeel nogal specialistische toepassingen kent. Er zijn pakketten met een grafische gebruikersinterface, en software die meer als een programmeerplatform moet worden gezien. IBM/SPSS en SAS Enterprise Guide, maar ook JMP en Statistica, zijn grafische pakketten aan de 'onderkant' van deze markt. Aan de bovenkant zitten suites als IBM/SPSS Modeler en SAS Enterprise Miner.

Gespecialiseerde mathemathische pakketten zoals R, S-Plus, of MATLAB, maar ook het aanbod aan open source valt in de categorie programmeerplatforms. Je krijgt dan een library met statistische procedures waarmee een vaardige programmeur alle kanten op kan. Die flexibiliteit wordt door sommigen erg gewaardeerd, maar stelt nogal hoge eisen aan programmeervaardigheden en statistische kennis.

Doordat er veel meer vraag is naar competente data analisten dan er aanbod is aan academici met een stevige kwantitatieve achtergrond, lijkt de softwaremarkt vooral richting grafische, gebruiksvriendelijke tools te evolueren. Door met wizards en templates zo veel mogelijk complexiteit van gebruikers af te schermen neemt het aantal mogelijke gebruikers enorm toe. En dat is nodig ook. Management is zich steeds meer bewust van toepassingsmogelijkheden voor Analytics. Door betere beschikbaarheid van steeds meer data is een 'overspannen' arbeidsmarkt ontstaan.

Bijna al je activiteiten kunnen gekopieerd worden door de concurrentie. Je beste mensen kunnen ze proberen weg te kopen. Maar aan je data, *daar* zal de concurrentie nooit aan kunnen komen. Daarom bieden diezelfde data zo'n unieke bron van duurzaam concurrentievoordeel.

Tom Breur is eigenaar van XLNT Consulting.