Effectieve datamining heeft geen datawarehouse nodig

Resultaten en ROI vallen vaak tegen

Tom Breur

Datamining lijkt in veel opzichten op tienersex: het is leuk en spannend om over te praten, maar er zijn slechts weinigen die het werkelijk praktiseren. En als ze het al doen gebeurt het tamelijk klungelig. Voor datamining lijkt hetzelfde te gelden. Datamining was lange tijd omgeven door hype. Waarom komt het zo moeilijk van de grond?

Het is overduidelijk dat data-analyse een steeds belangrijker rol speelt in competitieve markten. Bestsellers zoals Competing on Analytics (2007, Davenport & Harris), Supercrunchers (2007, Ayres), en The Numerati (2008, Baker) hebben – opnieuw – het belang van data en datamining onder de aandacht gebracht van een grote groep managers. Maar de praktijk is weerbarstig.
De term datamining is vrij nieuw, ontstaan uit een amalgaam van statistische technieken, Artificiële Intelligentie en Machine Learning. Meer en meer gegevens werden digitaal opgeslagen, en de benodigde rekenkracht om deze oceanen van data te doorzoeken kwam binnen handbereik.
De hype leidde in veel gevallen tot vertwijfeling, toen het in de praktijk lang niet eenvoudig bleek om de techniek aan de praat te krijgen. Maar zelfs nu datamining software steeds gebruiksvriendelijker en volwassener wordt, blijkt het inbedden in de organisatie ook nog een hele uitdaging. In dit artikel zullen we de geschiedenis van datamining schetsen en praktische obstakels bespreken om er effectief mee te werken.

Geschiedenis en ontwikkeling

In het prille begin werd datamining veelal ingezet om procescontrole te optimaliseren. Feitelijk bestond de term datamining nog niet, maar de applicaties die werden ingezet maakten onder andere gebruik van Neurale Netwerken, een techniek die vandaag de dag door menigeen als het epitoom van datamining wordt beschouwd.
De opkomst van datamining in de jaren negentig was omgeven door onrealistische verwachtingen, zeg maar gerust een grote hype. Gooi een berg (ongeorganiseerde) data door een Neuraal Netwerk, dat dan 'als vanzelf' interessante patronen vindt en klaar is Kees. Een 'zelflerend' netwerk als metafoor spreekt natuurlijk aan; de computer leert voor ons, dan hoeven wij zelf niet meer na te denken.
De resultaten waren echter meestal teleurstellend. Wie kent niet de uitkomst van analyse waarbij de brandstofefficiency van vliegtuigen werd gemodelleerd? Na eindeloze series CPU's, en een moeizame rekenkundige convergentie, bleek *achteruit vliegen* veruit de meeste besparingen op te leveren: er kwam zelfs brandstof bij in plaats van dat deze op raakte! Of de test om op fotomateriaal tanks (van de vijand) te identificeren: aangezien alle foto's overdag waren gemaakt 'dacht' het Neurale Netwerk in het licht overal tanks te zien, en in het donker nergens.
Een ander klassiek broodje-aap verhaal is 'bier en luiers'. Dit verhaal is zó beroemd geworden, dat het een eigen leven is gaan leiden. Thomas Blischok van NCR/Teradata had een aansprekend voorbeeld nodig voor een conferentie speech, en vertelde over deze correlatie bij Osco. Ondertussen heeft deze 'case' de status van een mythe aangenomen. Degenen die wel eens supermarktdata over een langere periode hebben geanalyseerd, kunnen bevestigen dat luiers en bier *zelden* samen worden verkocht. In ieder geval niet vaker dan op grond van toeval mag worden verwacht. Overigens heeft Osco de schapindeling van bier en luiers nooit veranderd, dit was een klassiek voorbeeld van een interessante correlatie, maar niet een waarmee je ook iets kunt doen.

In het Financieel Dagblad viel op 5 november 2008 de volgende kop te lezen: "Datamining vaak onbetrouwbaar." Vervolgens betogen de auteurs (die overduidelijk bijzonder weinig ervaring in dit vakgebied hebben) dat in veel gevallen de gevonden samenhangen als feit en causaliteit worden gepostuleerd. Wat een onzin. Als mensen zonder rijbewijs in een auto stappen is de kans op ongelukken aanwezig – met meer dan 200 PK onder de motorkap al helemaal. Hoe komt het dat een gerespecteerde krant dergelijke tendentieuze journalistiek publiceert?
Het is interessant dat de auteurs in FD het gebruik van datami-

ning voor credit scoring als voorbeeld nemen: op basis van klantkenmerken kredietwaardigheid beoordelen. Hieruit blijkt hun gebrek aan deskundigheid nog duidelijker: als er één domein is waar geautomatiseerde acceptatie zijn waarde heeft bewezen is het hier wel. Amper 30, 40 jaar geleden werden kredieten uitsluitend 'handmatig' beoordeeld. Die situatie is totaal omgeslagen. Alhoewel soms tegen-intuïtief, bleek acceptatie met behulp van criteria die door datamining zijn verkregen zó superieur dat in korte tijd bedrijven ofwel overstapten, of anders eenvoudigweg uit de markt werden gedrukt. Vandaag de dag worden al deze kredietbeslissingen door datamining ondersteund.

Waarom is datamining 'moeilijk'?

Ondanks verwoede pogingen lijkt datamining in de business praktijk vaak maar moeilijk van de grond te komen. Hoe komt dat? De afgelopen vijf tot tien jaar zijn de beschikbare tools sterk verbeterd, zowel qua functionaliteit als gebruiksvriendelijkheid. Daar ligt het probleem dus niet, in hoofdzaak. Er zijn twee andere redenen, die er *samen* voor zorgen dat datamining maar moeilijk de 'tienerjaren' ontgroeit. Op de eerste plaats zijn betrouwbare data onontbeerlijk, en op de tweede plaats heb je een team met hoogwaardige kenniswerkers nodig. De data en infrastructuur moeten aan speciale eisen voldoen en dit soort van teams blijkt in de praktijk vaak een korte halfwaardetijd te hebben.

Om effectief datamining te kunnen toepassen heb je zeker geen datawarehouse nodig. Met de juiste kennis en toewijding is het mogelijk de benodigde gegevens bij elkaar te harken, om een *proof of concept* te doen. Wanneer je datamining in een productie(achtige) omgeving wilt gaan inzetten echter, zullen deze zelfde gegevens maand in maand uit klaar moeten staan. Datamining vereist data die in flat files worden weergegeven, een wezenlijk andere modellering dan de gebruikelijke sterschema's (en aanzienlijk meer redundantie). Er worden zo veel mogelijk attributen afgeleid, omdat je van tevoren niet kunt weten welke variabelen 'het beste' zullen werken in modellen. Je laat de data immers spreken, je maakt daar op voorhand zo min mogelijk aannames over.

Ook de kwaliteit van deze gegevens moet in orde zijn. Behalve de gebruikelijke 'garbage in, garbage out', speelt hier nog iets anders. Datamining zoekt naar patronen in *gedrag*, en daaraan liggen stochastische processen ten grondslag. Zelfs kleine aberraties in ETL leiden tot deterministische effecten, die een veel sterkere samenhang kennen. Gevolg hiervan is dat de sterkste, meest in het oog springende verbanden die door een datamining-algoritme worden gesignaleerd vaak artefacten zijn van datawarehouse-laadprocessen. Dergelijke 'effecten' zijn niet gebaseerd op willekeurige gedragingen van klanten! Dit type kwaliteitsproblemen is fnuikend voor je datamining-modellen,

> Succesvolle introductie van datamining brengt ook organisatieverandering mee

Beknopte geschiedenis datamining

Databases hebben grootschalig hun intrede gedaan en een explosieve groei van beschikbare gegevens veroorzaakt. Daarmee ontstond er een 'markt' voor tools die geautomatiseerd kunnen zoeken naar verbanden. Oorspronkelijk werd de term datamining door statistici gebruikt als een pejoratieve diskwalificatie: als je maar lang genoeg blijft zoeken in eindeloze hoeveelheden data vind je altijd wel ergens een verband, was hun bezwaar.

In 1989 organiseerde Gregory Piatetsky-Shapiro tijdens IJCAI-89 in Detroit de eerste 'Knowledge Discovery in Databases' (KDD), een workshop die onmiddellijk veel belangstelling trok. Vanaf dat moment heeft het veld een enorme groei doorgemaakt. Bedrijfsleven en de pers gebruikten vooral de term 'datamining' en tegenwoordig worden de termen als synoniemen gehanteerd, zoals in de titel van het tijdschrift Data Mining and Knowledge Discovery.

In de beginjaren had je (alleen) specifieke tools per algoritme die ook nog eens lastig te bedienen waren.

In twee richtingen heeft men grote vordering gemaakt:

1. Inbedding van dataminingapplicaties in de productieketen. Van data-extractie via preprocessing, scoren van het model, en deployment van scores in bestaande business logica;
2. Uitbreiding naar *suites* vanaf midden jaren negentig, waarbij meerdere algoritmen in één tool beschikbaar kwamen. SPSS Clementine introduceerde als eerste een grafische interface die de 'process flow' toont van extractie, preparatie, mining en deployment.

De grootste resource op het web is www.kdnuggets.com, die wekelijks een goede nieuwsbrief uitbrengt. In de afgelopen twintig jaar is het aanbod van één naar zo'n tien internationale conferenties gegroeid. Tools zijn (veel) gebruiksvriendelijker geworden, soms in horizontale (applicatiespecifieke) of verticale (industriespecifieke) suites geïntegreerd. Wat bleef is de vraag wanneer datamining het hypestadium zou ontgroeien.

Met speciale dank aan Gregory Piatetsky-Shapiro voor zijn bijdrage aan dit historisch overzicht.

want de ernst hiervan wordt pas duidelijk nadat je modellen in de praktijk hebt toegepast, en het resultaat tegen valt. Reden te meer waarom continu monitoren van de effectiviteit van datamining-modellen een must is (helaas in de praktijk doorgaans geen *usance*). De tweede grote uitdaging bij de implementatie van datamining is het samenstellen en bij elkaar houden van een team specialisten.
Er zijn wellicht genoeg statistische experts en databasespecialisten te vinden, en ook mensen die verstand hebben van strategie en ervaring in marketing. Helaas is die intersectie dun, heel erg dun bezaaid.
Dit soort medewerkers heeft dikwijls heel idiosyncratische motivatie, zodat 'reguliere' managementinstrumenten (beloning, promotie) opmerkelijk ineffectief zijn. In de afgelopen tien jaar hebben we een aantal van dit soort hele sterke clubs zien komen, en ook weer zien imploderen. Door de hoge employability van dit soort specialisten wordt instabiliteit versterkt door succes: zij worden soms 'weggekocht'. Dan hoeft er in het management niet veel mis te gaan om zulke mensen met bosjes te zien vertrekken.
Deze twee redenen samen maken dat veel datamining-initiatieven stranden in schoonheid. Soms loopt men tegen de problematiek van data aan.
En degenen met een lange adem die hun data op orde krijgen moeten ook nog een competent team bij elkaar zien te houden.
Hoe is deze impasse te doorbreken?

Where's the money, honey?

Hoewel dit natuurlijk voor alle BI-initiatieven geldt, moet datamining in het bijzonder gedreven worden door business cases. En wel om twee redenen. Op de eerste plaats brengt succesvolle introductie van datamining ook altijd een organisatieverandering met zich mee. Een model maken is (relatief) eenvoudig, er voor zorgen dat modelgedreven targeting gemeengoed wordt in een organisatie is de échte uitdaging. Dit vereist dat mensen hun gedrag veranderen, en vaak ook dat prestatiedoelen worden bijgesteld. Organisatieverandering is een taai onderwerp.

> Binnen credit scoring is gebruik van modellen al gemeengoed

De meest gebruikte toepassingen voor datamining zijn op dit moment aanwezig in de *credit scoring* en direct marketing. Binnen credit scoring is gebruik van modellen al gemeengoed zoals we hebben gezien. Voor direct marketing geldt dat gebruik

van modellen een andere mindset vergt van marketeers: zij moeten zich nu concentreren op vormgeven en testen van hun aanbod, en kwantitatieve kosten/baten-richtlijnen opstellen. De doelgroepselectie laten zij dan over aan de modelbouwer. Dit vereist een cultuur van 'marketing accountability', en dat is soms wennen. Marketing wordt behalve creatief-, plotseling ook heel zakelijk, ROI-gedreven. En juist daarom is een constante focus op de ROI van alle activiteiten zo belangrijk.

De tweede reden waarom een continue focus op business cases zo belangrijk is heeft te maken met nieuwe 'arbeidsverhoudingen' en effectiviteit. Het werken met dataminingmodellen vereist dat marketing en BI een duidelijke scheiding van competenties en verantwoordelijkheden hebben. Marketing is verantwoordelijk voor het aanbod en kiest de meest veelbelovende tests: alles draait om onophoudelijk testen, wat onvermijdelijk tot resourceproblemen leidt. BI wordt verantwoordelijk voor doelgroepbepaling, en idealiter zelfs de oplage van campagnes. Zonder deze afbakening van verantwoordelijkheden loop je het risico dat datamining niet of verkeerd wordt ingezet. Dat probleem doet zich helaas nogal vaak voor in de praktijk, en is waarschijnlijk de belangrijkste reden waarom getalenteerde dataminers 'weer om zich heen gaan kijken'. Dergelijke mensen zijn schaars en hun employability is hoog. Zonder leuk werk en resultaten zijn ze dan snel vertrokken.

Conclusie

Na een periode van hype is het enthousiasme rond datamining enigszins bekoeld, en wellicht geldt dit wel voor Business Intelligence in het algemeen. De resultaten vallen dikwijls tegen, de ROI ook. In sommige gevallen hebben spectaculaire *bloopers* datamining in een slecht daglicht geplaatst. Zoals een Engelsman ooit zei: "When a model looks too good to be true, it usually is".

Toch lijkt het tij niet meer te keren. Datagedreven beslisondersteuning zal in moderne, competitieve bedrijven steeds belangrijker worden. Datamining speelt daarbij een grote rol. Er is een reden waarom analyticskampioenen zoals Amazon, Google of Capital One zo'n spectaculaire groei hebben laten zien. En managers worden zich daar in toenemende mate van bewust. Het is de hoogste tijd dat dataminingspecialisten laten zien waar en hoe met deze technologie geld moet worden verdiend. En er is een bepaald soort van succes dat door alle managers, in alle lagen van de onderneming snel wordt begrepen, en gemakkelijk wordt onthouden. Dat zijn namelijk resultaten die zich laten uitdrukken in die ene universele dimensie die overal uitstekend wordt begrepen: Euro's.

Tom Breur

Drs. A.F. Breur is Principal bij XLNT Consulting en docent bij DNV-Cibit Academy.