De tien Geboden van Datamining

Aldus Sentient

I. Gebruik automatische dataminingtechnieken

Datamining is een breed begrip en komt neer op het graven in databases naar informatie. De meest bruikbare informatie zit verborgen ergens in het totaal aan beschikbare gegevens, vaak in een complexe vorm. Om die informatie te achterhalen is het noodzakelijk gebruik te maken van technieken die automatisch patronen kunnen ontdekken in data. Handmatig alle gegevens doorlopen is niet meer haalbaar.

II. Maak datamining beschikbaar op de werkvloer.

Een grote algemene misvatting is dat datamining alleen door goed opgeleide en ervaren specialisten kan worden gedaan. Zij zijn immers de enigen die weten hoe modellen op de juiste wijze moeten worden ingesteld, hoe de data moet worden klaargezet, welke resultaten significant zijn en hoe die vervolgens moeten worden geïnterpreteerd. Er bestaan tegenwoordig echter genoeg technieken die automatisch modellering- en clustertechnieken kunnen instellen, de data kunnen klaarmaken en de gebruiker automatisch behoeden voor bekende valkuilen, bijvoorbeeld het trekken van conclusies op basis van niet significante afwijkingen. Dankzij deze ontwikkelingen is het tegenwoordig voor iedereen mogelijk om met datamining aan de slag te gaan.

Wanneer het beantwoorden van informatievragen wordt uitbesteed aan IT- of dataminingspecialisten ontstaan wachttijden en bottlenecks in de communicatie. Bovendien zijn de opgeleverde resultaten lang niet altijd relevant, omdat deze analisten soms moeilijk kunnen schatten wat belangrijk is. De meest effectieve werkwijze is dat degene met de vraag deze zelf kan beantwoorden. Groot bijkomend voordeel is dat mensen op de werkvloer de resultaten direct kunnen vertalen naar de praktijk.

III. Gebruik alle beschikbare data

Voor organisaties die enorme hoeveelheden gegevens verzamelen, is het met datamining haalbaar geworden om alle beschikbare data te analyseren. Omdat dataminingtechnieken automatisch de relevantie van informatie bepalen, is het mogelijk om met meer informatie meer inzichten te verkrijgen, zonder overstelpt te worden. Hoe meer data, des te meer bruikbare afhankelijkheden kunnen worden gevonden. Zo zorgt meer data voor betere voorspellende modellen.

IV. Behandel het bereiken van inzicht en het maken van voorspellingen als twee afzonderlijke doelstellingen.

Doelen die met datamining bereikt worden vallen in twee verschillende categorieën: ze kunnen de gebruiker meer inzicht opleveren, waardoor betere beslissingen kunnen worden gemaakt, of ze leveren individuele voorspellingen, door middel van scoremodellen. Vaak willen organisaties beide doelen bereiken. Als deze twee echter tegelijk worden nagestreefd leidt dat of tot een goed voorspellend maar onbegrijpelijk model, of een inzichtelijk model dat maar matig voorspelt. Immers, begrijpelijkheid gaat ten koste van de kwaliteit van een model en vice versa. Formuleer daarom eerst elk doel apart en plan daarna het beste datamining-traject om dit te bereiken.

Gaat het om het inzichtelijk maken van beschikbare data? Dan werken clustertechnieken en beslisbomen het best.

Moet er voor elk individu het gedrag worden voorspeld of een score worden berekend? Dan voldoen minder inzichtelijke modellen zoals associatieve geheugens en neurale netwerken. De kwaliteit van de voorspellende modellen wordt vervolgens met steekproeven getest.

V. Maak analyses interactief.

Antwoorden op vragen leveren altijd weer nieuwe vragen op. Het idee dat in een organisatie elke vraag beantwoord kan worden door het uitvoeren van een vooraf bepaalde procedure is achterhaald. Het is daarom van belang dat met tools wordt gewerkt waarmee dynamische analyses kunnen worden uitgevoerd. De gebruiker moet op elk antwoord gelijk kunnen inzoomen om de achterliggende verklaringen te vinden en dóór te analyseren. Daarbij is van belang dat de onderdelen van de toolset geïntegreerd zijn zodat snel gesprongen kan worden van analysefunctie naar analysefunctie, zonder ingewikkelde import- of exporthandelingen.

VI. Betrek de vrager zoveel mogelijk bij de analyse.

Interactief analyseren maakt het mogelijk om direct de dialoog met de data aan te gaan. Door de persoon met de informatiebehoefte te betrekken in het interactieve analyseproces wordt de vraag duidelijker en kan tijdig worden aangegeven of het

proces de goede richting op gaat. Dit soort sessies zorgen voor een zeer efficiënt analyseproces. Voorwaarde is wel dat de gebruikte tools in staat zijn direct inzichtelijke resultaten op te leveren.

VII. Werk met een vaste analysedatabase.

Het grote voordeel van één grote analysedatabase met alle beschikbare gegevens is dat niet voor elke vraag nieuwe data vanuit verschillende plekken moet worden verzameld en klaargemaakt. Dit is immers 80% van het werk en vereist andere kwaliteiten, zoals specifieke kennis van databases en het opschonen van de data. De vaste analysedatabase verandert natuurlijk wel, met beleid, naar gelang de behoeften veranderen.

VIII. Gebruik clusteren voor het vinden van combinaties.

Clusterfuncties zoals K-means, hiërarchisch clusteren en DataDetective's associatieve segmentatiefunctie, zijn krachtige datamining-technieken om automatisch elementen te groeperen en deze overzichtelijk weer te geven. Hierdoor kunnen veel voorkomende combinaties van eigenschappen snel worden ontdekt. Omdat het totaal aantal mogelijke combinaties veel te groot is om handmatig te testen, is dit met conventionele technieken niet haalbaar.

IX. Baseer modellen op het individu, niet op segmenten.

Vaak worden segmenten (bijvoorbeeld jonge mannen in grote steden) gebruikt voor voorspellingen, omdat een statistische uitspraak over een minimaal aantal personen moet gaan en omdat segmenten of 'hokjes' kunnen worden begrepen. Echter, dergelijke segmenten zijn gebaseerd op een beperkt aantal variabelen zoals geslacht, inkomen en leeftijd, terwijl veel meer informatie met voorspellende kracht binnen handbereik is. Datamining, met name associatieve technieken, baseert zich op alle gegevens die over het individu bekend zijn, waardoor voorspellingen nauwkeuriger worden.

X. Besteed niet te veel moeite aan het kiezen van de juiste modelleringtechniek.

Het aantal verschillende soorten voorspellingsmodellen op de dataminingmarkt is inmiddels groot, waardoor het maken van een keuze lastig lijkt. Toch is deze keus minder belangrijk dan het lijkt. Wanneer het gaat om voorspellend vermogen doen beslisbomen en lineaire modellen het wat minder dan non-lineaire modellen zoals neurale netwerken, maar verder doen de technieken nauwelijks voor elkaar onder als het gaat om voorspellend vermogen. Het verschil zit vooral in hoe deze technieken omgaan met de data en in praktische aspecten, zoals snelheid en de moeite die het kost om het model in te stellen. Het is daarom beter een techniek te kiezen op basis van deze aspecten.