Goud zoeken in data

Op National Geographic en Discovery Channel zijn tegenwoordig tal van programma's te zien waarin onder moeilijke omstandigheden naar goud wordt gezocht (Yukon gold, Goldfathers, Gold Rush). Laatst viel mij op hoe ver de vergelijking tussen het 'minen van goud' en het 'minen van data' opgaat. Beide zijn een zoektocht naar soms zeer moeilijk te vinden elementen met de belofte van grote winst. Aan het eind van de meeste afleveringen is dan ook de conclusie dat de opbrengst tegenvalt en dat de landeigenaren en investeerders zeer teleurgesteld zijn. Ook dat beeld komt mij als data scientist bekend voor… Aan de hand van deze vergelijking kom ik met een aantal praktische tips en tricks voor het analyseren van data. De meeste tips zijn zeer logisch maar toch vergeet ik ze zelf ook nog wel eens, net als de goudzoekers die keer op keer dezelfde fouten lijken te maken.

Prospectie

Voordat goudzoekers hun graafmachines en 'wash plants' in stelling brengen gaan ze eerste kijken of er wel goud te vinden is. Prospectie heet dat en is handmatig werk waarbij kleine hoeveelheden grond worden bewerkt om in te schatten hoeveel goud er te vinden is en waar het precies in de grond zit. Hierna kunnen ze de kosten afwegen tegen de baten en inschatten welk materieel ze precies nodig hebben om het goud uit de grond te halen.

Ook bij data analyse is 'prospectie' een goed idee. Neem als het mogelijk is een kleine maar representatieve sub-set van de data en kijk goed of de gezochte informatie wel te vinden is. Door de dataset klein te houden wordt het makkelijker en sneller om allerlei transformaties en analyses te proberen. Het is aan te raden om de data scientist te laten samenwerken met een domein expert die resultaten in de juiste context kan plaatsten. Door dit proces iteratief aan te pakken wordt ook de kans vergroot om 'onverwachte informatie' te vinden die heel waardevol kan zijn. Zeker door het combineren met andere data is veel mogelijk. Dit is vaak iets dat er door de wensen van de stakeholders bij inschiet maar dergelijke bijvangst kan ook veel opleveren.

Waar landeigenaren het meestal wel prima vinden dat er naar goud wordt gezocht is dat bij stakeholders rond data toch anders. Meestal kost het de nodige moeite om data te krijgen omdat eigenaren bang zijn iets heel waardevols 'weg te geven' en bang zijn dat de data ook negatieve informatie bevat. Meestal wegen de voordelen ruimschoots op tegen de mogelijke nadelen maar dat maakt de koudwatervrees er niet minder om. Gelukkig zijn er tal van voorbeelden waarbij data analyse door anderen een positieve bijdrage heeft geleverd. Zie bijvoorbeeld het artikel over de Goldcorp challenge waarbij middels vrijgegeven data meer goud werd gevonden.

Paydirt

Als er voldoende goud of andere mineralen zijn gevonden wordt het tijd om op te schalen. Voordat er grootschalig naar goud gegraven kan worden moet het werkterrein gereed worden gemaakt. Het gebied moet worden ontbost en meestal de toplaag aan aarde weggegraven om bij de paydirt (goudhoudende aarde) te komen. Deze fase wordt vaak sterk onderschat en zijn de goudzoekers gefrustreerd dat ze 'nog steeds niet' naar goud aan het graven zijn. Deze voorbereidende fase is sterk vergelijkbaar met het 'extract, transform and load' (ETL) proces van data analyse. Hierbij wordt data opgeschoond en omgezet in een bruikbaar formaat. Voordat data kan worden geanalyseerd moet de omgeving worden opgezet om dit daadwerkelijk te doen. De benodigde software moet worden geconfigureerd en data in databases of filesystemen worden geladen. Naar mijn ervaring gaat dit zelden zo makkelijk als verwacht. Meestal bevat de data naast waardevolle elementen ook een hoop 'ruis' die moet worden verwijderd. En elke conversie stap moet worden getest om te checken of alles nog wel klopt. Deze stappen worden vaak ook gezet tijdens de eerder genoemde 'prospectie' maar in het groot valt het vaak tegen. Uiteindelijk zorgt een gedegen voorbereiding ervoor dat de daadwerkelijke analyse makkelijker wordt maar toch levert dit tijdrovende proces vaak de nodige frustraties op.

Machines

Als de grond is geprepareerd kan aan de daadwerkelijke delven van goud worden begonnen. Afhankelijk van de grond en de verwachte opbrengst wordt gepaste machinerie naar de locatie gebracht. Voor veel goudhoudende grond is er een grote 'washplant' nodig en genoeg graafmachines en bulldozers om de grond bij de plant te krijgen.
Voor data analyse geldt natuurlijk hetzelfde; de benodigde infrastructuur en technologieën hangen af van de hoeveelheid data en complexiteit van de analyses. Hierbij wordt al snel gedacht aan allerlei big data technologieën zoals Hadoop, Storm, Cassandra etc. Maar als de hoeveelheid data meevalt en de eis rond schaalbaarheid geen must-have is het natuurlijk ook prima om klassieke technologieën zoals relationele databases te gebruiken. Uiteraard helpt het als je verschillende technologieën kent zodat je deze afweging kan maken. Naar mijn ervaring is het gat tussen een stukje software kennen en het effectief gebruiken echter vaak heel groot. Afhankelijk van iemands achtergrond zijn toch al snel honderden uren nodig om iets echt goed te leren kennen. Dit is dus echt een investering die goed overwogen moet worden gemaakt. Ditzelfde geldt voor goudzoekers; er moet worden geïnvesteerd in het juiste materiaal waarna er vele uren hard werken nodig zijn om iets effectief te bedienen. Een goede graafmachine machinist kan veel meer grond verplaatsen dan een beginner en dat betaald zich direct in goud uit.
Er lijkt ook een garantie te bestaan dat materiaal kapot gaat. In elke aflevering over goud zoeken die ik heb gezien moet er worden gesleuteld om de boel weer draaiende te krijgen. Bij data analyse is het net zo, hoe groter het aantal hardware/software componenten dat wordt gebruikt hoe groter de kans dat er iets kapot gaat. Het is aan te bevelen om hier rekening mee te houden. Dit kan natuurlijk door alles zo fault-tolerant en redundant mogelijk op te zetten maar dat kan niet altijd. Zorg er in elk geval voor dat de data en (tussentijdse) analyse resultaten goed zijn opgeslagen en dat er na failure kan worden doorgegaan met analyseren. Helemaal opnieuw moeten beginnen is zelden leuk om te doen.

De opbrengst
De goudzoekers in de genoemde programma's richten zich altijd op een specifiek aantal gram goud dat ze in een periode willen opgraven. De landeigenaar en investeerders krijgen meestal een deel van het goud en zijn dus direct gebaat bij een hoge opbrengst. Hoe de verwachte opbrengst precies wordt ingeschat is mij onduidelijk maar er lijkt vaak veel wishful thinking bij betrokken te zijn want het wordt zelden gehaald. Resultaat? Boze investeerders want ze hadden meer verwacht.
De parallel met verwachtingen bij de analyses is makkelijk gevonden. Vaak moet er behoorlijk wat worden beloofd om de stakeholders over de streep te trekken. Het waarmaken van de geschapen verwachtingen valt meestal tegen. Zeker als de te verwerken data nog niet is bekeken is de kans op te optimistische verwachtingen groot (meestal blijkt data toch onvolledig en vol met 'ruis'). De kunst is natuurlijk om zo min mogelijk te beloven en tijdens het proces zo goed mogelijk aan verwachtingsmanagement te doen. Klein beginnen (prospectie) en oog houden voor bijvangst komen hierbij goed van pas. Het is altijd makkelijker om naast het temperen van verwachtingen ook positief nieuws te kunnen brengen. "Het delven van goud blijkt toch moeilijker dan verwacht maar we hebben wel waardevolle edelstenen gevonden."

Dus…
De kern van het verhaal laat zich redelijk samenvatten in het woord 'klein'. Houd de verwachtingen beperkt en begin in kleine stapjes een kleine hoeveelheid data te doorgronden. Pas daarna kan worden bepaald welke waarde de data daadwerkelijk bevat (prospectie) en welk materieel het meest geschikt is om die waarde eruit te halen. Deze fase is erg belangrijk en mag best wel wat tijd kosten!
Gebruik de technologie die goed past bij de beoogde verwerking en dat hoeft geen Hadoop of NoSQL database te zijn. Ook ouderwetse technologie kan vaak prima worden gebruikt. Uiteraard loont het wel de moeite om nieuwe technologieën te leren en in te zetten maar houd hierbij rekening met een forse tijdsinvestering (easy to learn but hard to master).

Mocht het proces van goud zoeken totaal onbekend zijn dan is het wellicht leuk om één van die programma's eens te bekijken. Wellicht levert het kijken naar onze 'robuuste goudzoekende collega's' nog meer veel meer nuttige inzichten op.

corne.versloot@tno.nl
december 2014