Kies Kundig. VERDONCK KLOOSTER & ASSOCIATES

Aan de slag met data kwaliteit?

Hier is je top 10

Auteurs: Sabine den Daas / Frank van Vonderen - VKA
Dit artikel is tot stand gekomen met hulp van Bluefield / Totta data lab

Wat moeten we toch met data gedreven werken. Aan de ene kant zijn we sceptisch over de rol van data en dol op de uitdrukking 'Lies, damned lies, and statistics'. Anderzijds zoeken we allemaal de kansen van datagedreven werken. Wat is de regenkans morgen? Met welk matchingspercentage beveelt Netflix mij die Deense film aan? Hoe kunnen we voorspellen wat er gaat gebeuren?

Maar er zijn ook kritischer geluiden, want hoe geloofwaardig is eigenlijk de data. Dagelijks keken we naar de besmettingscijfers en het aantal sterfgevallen als gevolg van Corona. Maar achteraf bleek dat er foutenpercentages waren van 6% tot 20% (klik hier voor artikel). Collectieve verontwaardiging alom, maar geen nieuws voor data specialisten.

Want weinig dingen zijn zo onderschat, randvoorwaardelijk en cruciaal als datakwaliteit. Want ja, garbage in is garbage out. En tegelijkertijd is kwaliteit een lastig begrip, omdat het enkele objectieve kenmerken kent, maar ook minder objectieve. Net als 'smaak'. Eten kan lekker zijn door de kwaliteit van de ingrediënten en de combinatie van dat alles, maar de perceptie van 'lekker' is persoonlijk. In dit artikel concentreren we ons de 'objectieve' kenmerken van datakwaliteit en schetsen we de top 10 problemen (én oplossingen) bij datakwaliteit.

De top 10 problemen (én oplossingen) bij datakwaliteit:

1. Data inconsistentie

Probleem - data worden aangeleverd uit verschillende bronsystemen, ingevuld door verschillende medewerkers die allemaal verschillende schrijfwijzes hanteren. Dit is met name zichtbaar bij data input op basis van open velden. De consequenties van de foutieve invoer zijn in eerste instantie niet zichtbaar.

Voorbeeld - op hoeveel manieren kan je een naam schijven. Neem de naam van een van de auteurs: Sabine den Daas. Schrijf je dit als 'Sabine den Daas', als 'Den Daas, Sabine', als 'S. den Daas', als 'S. Daas, den' en ga zo maar door.

Maatregel - hanteer vaste conventies voor schrijfwijzes en zet de informatie die je uit nieuwe bronnen ontsluit om in de conventie die je zelf aanhoudt. Controleer bij het ontsluiten van nieuwe bronnen hoeveel werk het is om dit te doen en controleer of in de nieuwe data er input is waar je nieuwe conventies voor moet gaan maken.

2. Data definitie is niet helder

Probleem - De beschikbare data betekenen voor een ieder iets anders, omdat de definitie niet is afgesproken en daardoor inconsistent wordt toegepast. Toch zijn de data waardevol voor iedereen omdat zij hun eigen definitie succesvol toepassen. Pas bij het vergelijken van de verschillende definities blijkt het niet meer mogelijk om de analyse toe te passen.

Voorbeeld - Een Inspectie voert onderzoeken uit. Bij het uitvoeren van de onderzoeken gebruikt men datums om te laten zien wanneer bepaalde onderzoeken hebben plaatsgevonden. De een gebruikt de datum van het moment van bezoeken van een locatie, de ander gebruikt de datum van het moment dat het rapport is afgerond. Ze hebben allebei gelijk, omdat ze in hun verantwoording benoemen waar de datum voor staat en ze deze vrijheid hebben. Echter bij de analyse gebruikt men alleen de datums en is niet meer duidelijk wat de eenduidige definitie is en wanneer de onderzoeken daarmee plaatsvonden.

Maatregel - Het maken van afspraken welke definities er gelden voor data - data conventie / naamgevingsconventies. Hierdoor ontstaat vergelijkbaarheid, consistentie in de data en daarmee de mogelijkheid tot data-analyse.

3. Lastig om conclusies te trekken

Probleem - Je hebt wel data, je weet ook wat deze betekenen, maar wat kan je er dan eigenlijk mee?

Voorbeeld - Stel: in 94% van alle zedenmisdrijven is de dader een man. Wat betekent dat dan? Dat in even zoveel gevallen een vrouw het slachtoffer is? Dat in 6% van de gevallen de dader een vrouw is? Dat heel veel mannen potentiele zedendelinquenten zijn? Of is één man verantwoordelijk voor meerdere zedenmisdrijven? Worden minderjarige jongens als mannen meegerekend?

Maatregel - Dit vraagt veel maatregelen, allereerst een toets of de data juist, actueel en volledig is. Even weer een voorbeeld met maar weer onszelf als voorbeeld

- Frank van Vonderen staat in een zedendossier - voor de volledigheid en de beeldvorming rondom deze persoon is het dan wel aardig om te weten of hij er in staat als dader, slachtoffer, getuige, melder, …
- Frank van Vonderen staat in een zedendossier - voor de actualiteit is het wel aardig om te weten of dit dossier nog actief is en tot vervolging heeft geleid of niet.
- Frank van Vonderen staat in een zedendossier - voor de juistheid is het wel aardig om te weten of Frank feitelijk een delict heeft begaan dat onderdeel is van het zedendossier of dat het een verkeersovertreding betrof die hij heeft gemaakt tegenover een pand waar het zedendossier betrekking op heeft.

Een andere maatregel is dat de waarde van data met inhoudelijke specialisten (geen data experts, maar mensen die het werkproces goed kennen) wordt besproken: wat kunnen zij op basis van hun ervaring zeggen over de constateringen die op basis van de data lijken te bestaan.

4. Data mutaties worden niet bijgehouden

Probleem - Alleen de meest recente status wordt opgeslagen (geen historie bekend). Een specifiek probleem wanneer deze gegevens gebruikt worden om patronen en causaliteiten met de gegevens te herkennen.

Voorbeeld - Sabine heeft het abonnement op haar favoriete tijdschrift opgezegd, hiermee wordt ze ook automatisch uitgeschreven voor het ontvangen van de wekelijkse nieuwsbrief. Uit de data-analyse blijkt dat er is een (significant) verband is tussen de opzegging en het wel of niet ontvangen van de nieuwsbrief. Maar er is hier geen sprake van causaliteit: het niet ontvangen van de nieuwsbrief was niet een van de kenmerken die er toe heeft geleid dat Sabine haar abonnement heeft opgezegd.

Maatregel - houd voortaan wél de mutaties bij.

5. Informatie wordt vastgelegd aan de hand van verschillende identificatoren

Probleem - In de ene tabel heeft een klant een gepseudonimiseerd nummer. Er is echter GEEN vertaal-tabel beschikbaar waarmee de gegevens uit de diverse bronnen aan elkaar kunnen worden gekoppeld. Voor analyses over bijvoorbeeld de klantreis kunnen de gegevens uit deze tabel niet worden gebruikt.

Voorbeeld - Frank belt met de klantenservice. Helaas weet hij zijn klantnummer niet. De call center agent is heel vriendelijk, maar kan Frank zonder het klantnummer niet vinden tussen alle klanten. Dit geldt natuurlijk ook voor het samenvoegen van gegevens over Frank. Wanneer we niet weten welke informatie bij welke Frank hoort, is het niet mogelijk om deze verschillende gegevens te gebruiken voor diepgaande analyses.

Maatregel - zorg dat elke tabel een duidelijk identificatie heeft en zorg dat deze juist gerefereerd wordt in andere tabellen. In de praktijk wordt ook wel gezocht naar een ‘golden record’, een uniek identificerend kenmerk, dat in alle bronnen wordt gebruikt en op basis waarvan data kan worden gekoppeld. Bij objecten kunnen dit coördinaten zijn, of een fysiek adres. Bij mensen bijvoorbeeld een BSN of ander uniek nummer. Alhoewel aan de verwerking van BSN duidelijke beperkingen bestaan (zie direct hierna).

6. Geen juridische grond voor gebruik persoonsgegevens

Probleem - de databron bevat (in)direct herleidbare persoonsgegevens. Op voorhand is niet geregeld dat deze gegevens ook voor data-analyse kunnen worden gebruikt.

Voorbeeld - bij een verzekeringsmaatschappij worden gegevens verwerkt over claims van personen in relatie tot de reisverzekering. De afdeling fraudebestrijding wil graag weten welke mensen frauderen en wil bestanden van verschillende collega verzekeraars koppelen om te kijken of mensen niet dubbel claimen bij meerdere verzekeraars. Super zinvol en super efficiënt, maar… hoe zit het ook alweer met die privacy wetgeving…

Maatregel - er zijn twee mogelijke maatregelen: mogelijkheid 1: regel vooraf welke gegevens je kunt gebruiken en mogelijkheid 2: gebruik de gegevens niet of reduceer de gevoeligheid.
Mogelijkheid 1 is de structurele oplossing, waar je een privacy specialist bij nodig hebt. Ga met deze specialist na wat er wel kan en wat niet. Vaak is het geen kwestie van ‘kan wel’ of ‘kan niet’, maar moet je op zoek naar de mogelijkheden. Sleutelwoorden zijn daarbij ‘wat zijn de grondslag en doelbinding van de oorspronkelijke verwerking’ oftewel: om welke redenen is deze data oorspronkelijk verzameld en is jouw gebruik daar een logisch gevolg van. Zo ja: dan moet dat ook duidelijk zijn (‘transparantie’) voor de personen wiens gegevens het betreft en zo nee: kijk of de grondslag/doelbinding kan worden aangepast zodat ook data analyse kan plaatsvinden.

Mogelijkheid 2 is dat je de gevoeligheid en herleidbaarheid van persoonsgegevens reduceert. Dat doe je door bij het ontsluiten van de bron of uiterlijk bij staging specifieke persoonsgegevens of herleidbare kenmerken verwijdert, pseudonimiseert of anonimiseert. Vervang bijvoorbeeld een BSN of naam door een geanonimiseerd gegeven. Verwijder het geslacht en geboortedatum uit je dataset. Dit wordt ook wel 'Privacy by Design' genoemd.

7. Te weinig variatie in de data

Probleem - de variatie in de gegevens komt niet overeen met de werkelijkheid. Dit zorgt voor een zwarte vlek/tunnelvisie in je data analyse of voorspelmodel, omdat de gegevens niet voldoende de weerbarstigheid van de werkelijkheid weerspiegelen om verantwoord gebruik te maken van deze gegevens.

Voorbeeld - zoals het ontbreken van een (representatieve) controlegroep. Alleen mensen met klachten worden getest op 't coronavirus. Hier zijn twee data problemen te constateren: 1. niet iedereen met klachten laat zich testen, en 2. je kunt ook drager van het virus zijn en geen klachten vertonen.

Maatregel - zorg altijd voor een voldoende grote controlegroep wanneer je aannames test. Als vuistregel kun je hanteren dat bij een statistische test op twee categorieën (wel/niet) er voor de testgroep en controlegroep minstens 100 observaties groot zijn. Echter, bij het testen van gebeurtenissen die relatief weinig voorkomen moeten deze groepen groter zijn, minimaal de mogelijkheid om minimaal 50 'wel' en 'niet' gevallen te identificeren.

Hetzelfde geldt voor modellen, bij een simpeler algoritme met weinig variabelen gaat bovenstaande op, maar hoe complexer het algoritme wordt hoe meer data er nodig is, zie punt 8.

8. Onvoldoende observaties

Probleem - Complexere algoritmen hebben meer data nodig omdat deze ook onderliggende patronen herkennen.

Voorbeeld - Vooral bij beeldherkenning zijn veel voorbeelden (cases/observaties) nodig, omdat hier zoveel combinaties bestaan in pixels. De grote hoeveelheid combinaties zorgt voor zoveel mogelijke patronen die ontdekt moeten worden door het algoritme om tot een goede herkenningsmodule te komen.

Maatregel - Even wachten totdat je genoeg observaties hebt. Dit kun je doen door data aan te kopen, data te genereren en/of meer data verzamelen.

9. Onvoldoende betekenis toegepast

Probleem - De variabelen die je gebruikt moeten relevant zijn in context met de uitkomst. Een model maken op persoonskenmerken leidt tot 'discriminatie' van groepen, immers zoeken we naar overeenkomsten en verschillen in relatie tot de uitkomst. Er worden niet altijd voldoende kenmerken vastgelegd en/of toegepast om tot de gewenste uitkomst te kunnen komen, omdat meer kenmerken nodig blijken te zijn om tot de uitkomst te komen.

Voorbeeld - Een verzekeringsmaatschappij wil graag de premies voor een woonverzekering individualiseren, maar beschikt alleen over basiskenmerken zoals leeftijd, geslacht en gezinssamenstelling. Het premiemodel generaliseert schade uit het verleden en alle jong volwassen krijgen een hele hoge premie toe berekend.

Maatregel - Modellen presteren beter als er ook gekeken wordt naar wijzigingen in gedrag: zoals verhuizen, verandering in gezinssituatie en recente contact-onderwerpen. Andere manieren van premiestellingen bijvoorbeeld op basis van rijgedrag (snelheid, remgebruik), geven jonge bestuurders de mogelijkheid om minder premie te betalen bij goed rijgedrag.

10. Systemen geven vrijheid of restricties

Probleem - Systemen zijn ingericht om bepaalde waarden toe te laten. Als deze vrijheid groot is ontstaan inconsistenties, als deze vrijheid klein is ontstaan beperkingen en ontbreekt data.

Voorbeeld - Men wil omzetrapporten van een restaurant invoeren, er zijn twee rapporten van dezelfde locatie op dezelfde dag. De een maakt hier één rapport van (dagrapport) en de ander maakt er twee rapporten van 'lunchrapport' met de omzet van 12.00 – 17.00 en 'dinerrapport' met de omzet na 17.00. Hier maakt men gebruik van de vrijheid. Ook zijn er veel velden die niet worden ingevuld omdat deze in een ander systeem zijn opgenomen. Verderop in het systeem is het verplicht om een veld in te vullen dat in sommige gevallen wel maar in dit geval niet van toepassing blijkt te zijn. Omdat er toch iets moet komen te staan vult men onzinnige data in, de andere data worden hierdoor in eerste instantie minder bruikbaar. Daarnaast is er een vinkje dat altijd aanstaat maar niet altijd uitgezet wordt als het wel nodig is. Hierdoor lijkt het alsof dit vinkje vaker geldt dan het daadwerkelijk het geval is.

Maatregel - Blijf met de gebruikers in gesprek om een passende mate van vrijheid en restricties in systemen te hanteren. Doorloop het proces regelmatig gezamenlijk om een juiste invoer te hebben. De invoer gaat gelden voor alle daaropvolgende handelingen, goede data begint bij goede invoer.

Met dank aan Lotte Meindertsma en Douwe Horst.

Wil je meer weten over data kwaliteit? Neem dan contact op!

Sabine den Daas: sabine.dendaas@vka.nl
Frank van Vonderen: frank.vanvonderen@vka.nl