Valkuilen bij het interpreteren van data

4orange, dec 2020

Hogehilweg 24
1101 CD Amsterdam Zuidoost
www.4orange.nl

Valkuilen bij het interpreteren van data

We worden dagelijks gebombardeerd met cijfers. Het nieuws bestaat voor een groot deel uit statistieken. Of het nou de Amerikaanse verkiezingen zijn, de economie of de coronapandemie: overal tabellen, grafiekjes en schema's. Maar het is, niet alleen voor de neutrale toeschouwer, maar ook voor journalisten, beleidsmakers, politici en zelfs wetenschappers, ontzettend moeilijk om tot een eenduidige interpretatie van die cijfers te komen. Ook binnen bedrijven speelt dit probleem.

Data-analyse helpt je met beslissingen nemen. Maar kijk je wel naar de juiste data? En heb je wel alle data die je nodig hebt? Wishful thinking en confirmation bias liggen op de loer, als je niet kritisch genoeg naar je analyses en modellen kijkt.

"Kijk je wel naar de juiste data?"

Niemand kan op basis van het R-getal en het aantal nieuwe besmettingen precies zeggen: deze maatregel moeten we nemen en dat gaat over twee weken een x aantal minder besmettingen opleveren. En dan heb je het in het geval van corona alleen nog maar over een relatief klein dashboard met een paar simpele, dagelijkse statistieken erop. Hoe moet dat dan met de enorme hoeveelheden bedrijfsdata die we dagelijks analyseren? Bedrijven willen heel graag aan het werk met data-analyse, machine learning en voorspelmodellen. Terecht, want als je het goed aanpakt is er veel winst te behalen. Maar bij het interpreteren en toepassen van data kan veel misgaan.

Zinnige resultaten

Ik ben ervan overtuigd dat je, met de tools die er tegenwoordig zijn, zo'n beetje iedereen kunt leren om een machine learning-model te bouwen. Maar wat veel moeilijker is, is om ook iets te bouwen dat zinnige resultaten oplevert. Je kunt een computer een rij cijfers laten uitspugen, maar heeft die rij cijfers ook echt iets te maken met jouw bedrijf of jouw klanten? Om dat te weten heb je menselijke intelligentie nodig.

Ter illustratie: Je kunt met een wiskundig model berekenen dat het opleidingsniveau van een persoon voornamelijk afhankelijk is van het opleidingsniveau van de ouders. Dat klopt ook met de sociologische kennis die we hierover hebben. En dat laatste is het belangrijkste. Want we kunnen wiskundig gezien met hetzelfde gemak bewijzen dat opleidingsniveau vooral afhankelijk is van leeftijd. Ga maar na: een gemiddelde kleuter heeft een lager opleidingsniveau dan een gemiddelde middelbare scholier en die is weer lager opgeleid dan een gemiddelde afgestudeerde. Kijk je puur naar de data, dan kloppen deze analyses allebei. Maar kijk je breder dan alleen je model, dan zie je meteen dat het tweede model volkomen nutteloos is.

Kijk je naar de goede data?

Maar zelfs als het model op zich wel zinnig is, kun je nog in de problemen komen. Public Health England meldde bijvoorbeeld begin oktober 16.000 coronabesmettingen te weinig. Wat bleek? De gebruikte Excel-sheet kon de hoeveelheid rijen niet meer aan. Een typisch voorbeeld van hoe een banale oorzaak een enorme impact kan hebben op de data die we zien, en dus op de beslissingen die

we nemen. Hier ging het om een technische misser, maar het is algemeen bekend dat er ook bewust met data 'geschoven' wordt om de dashboards van het management te beïnvloeden. Er zijn genoeg salesafdelingen die orders in een andere maand boeken of retentie-orders dubbel tellen om hun cijfers 'op te leuken'. Dagelijks worden belangrijke beslissingen genomen op basis van Excelsheets en/of datadashboards waarvan de beslisser niet helemaal weet hoe ze gevuld worden. Hoe weet zo'n beslisser dat die data kloppen, volledig zijn en ook echt de data zijn die er op dat moment toe doen?

Heb je wel de goede data?

Je kunt proberen politieke voorkeuren te voorspellen op basis van inkomen en leeftijd, omdat dat nou eenmaal is wat je weet. Maar als je via externe bronnen een koppeling kunt maken met postcode en culturele achtergrond, wordt je model veel preciezer. Combineer je je eigen data met openbaar beschikbare data van het CBS, dan krijg je inzichten die je anders niet zou hebben. Kom je tot dit soort inzichten voordat je modellen gaat bouwen, dan bespaar je jezelf veel tijd en moeite.

Perspectieven op je data

Toen Donald Trump op 4 november naar de verkiezingsuitslagen keek, zag hij een lappendeken van rode staten en concludeerde dat hij dik gewonnen had. Joe Biden zag een hele andere kaart, want hij keek naar de nog niet getelde stemmen en de data die zeiden dat daar heel veel stemmen voor hem bij zouden zitten. Zo zie je dat verschillende perspectieven op de data tot totaal verschillende conclusies kunnen leiden.

Confirmation bias en *wishful thinking* liggen daarbij altijd op de loer, dus is het zaak om met een breed en divers team naar je data te kijken. Veel te vaak is data-analyse het terrein van *hard core* beta's, die modellen maken die technisch kloppen en vanuit hún perspectief zinnig zijn. Maar je hebt een bredere blik nodig van mensen die vragen: wat is eigenlijk voor ons als organisatie eigenlijk zinnig en nuttig? Want binnen een organisatie start data-analyse altijd vanuit een doel. Er is een vraag die beantwoord moet worden, en met het antwoord moeten mensen ook echt aan het werk kunnen.

Meerdere perspectieven en een cultuur waarin kritische vragen worden aangemoedigd zorgen dat je minder snel de mist in gaat met het interpreteren van data. In het ideale datateam heb je, naast goede analisten en developers, ook mensen nodig die het vakgebied waar je in opereert goed kennen. Vakspecialisten die misschien minder van data-analyse weten, maar die wel vanuit hun ervaring weten wanneer iets onzin is of juist hout snijdt.

Verder is het fijn om er een econometrist bij te hebben, of in ieder geval iemand met een wetenschappelijke blik. Daarmee bedoel ik: iemand die een hypothese kan opstellen, testen en eventueel verwerpen als hij niet blijkt te kloppen. Je hebt daarnaast ook algemeen inzicht in menselijk gedrag nodig. Ik heb zelf sociologie gestudeerd en ik merk dat dat me helpt om met wat meer afstand naar modellen te kijken, omdat ik op kwalitatief niveau de logica van menselijk gedrag kan begrijpen.

Verlies van controle

Voor managers leveren data, machine learning en AI controle op, omdat ze voor het eerst echt zien wat er gaande is in hun bedrijf en daar direct op kunnen reageren. Maar tegelijkertijd verliezen ze een stuk controle, omdat er altijd vragen blijven waarop je het antwoord niet hebt. Ook dat wordt duidelijk als je met data aan het werk gaat. Een klant zei ooit tegen mij: “Jij bent ook een mooie! Bij iedere vraag die ik je stel, kom je met een antwoord maar ook met twee nieuwe vragen!” Dat klopt. Het analyseren van data levert je veel meer op dan antwoorden. Het levert je vragen op waarvan je helemaal niet wist dat je ze moest stellen. En hoe meer data je hebt, hoe meer vragen er tevoorschijn komen. Dat kunnen er zoveel zijn, dat je er niet aan toe komt om ze allemaal met analyses te beantwoorden. Ze zullen ook niet allemaal relevant zijn. Ook hier heb je weer, naast data-expertise, een dosis gezond verstand en zakelijk inzicht nodig. Blijf je altijd afvragen wat je echt moet weten om je bedrijf beter te laten draaien.

Bouw, test, herhaal

Het is onvermijdelijk dat je hiermee af en toe de mist ingaat. Soms bouw je een model dat weliswaar resultaten levert, maar dat uiteindelijk in je bedrijf niet veel impact heeft. Dat is niet erg. De enige manier om het echt te verprutsen is om in één keer iets heel groots te gaan bouwen, daar heel veel tijd en budget aan te besteden om er dan achter te komen dat het niet werkt. Als je in korte trajecten werkt en alles wat je maakt meteen test, kun je eigenlijk geen verkeerde beslissingen nemen. Kijk gewoon of dingen werken en pas dat principe niet alleen toe op het bouwen van je modellen, maar ook op het daadwerkelijk handelen op de uitkomsten. Natuurlijk is het nuttig om de conversiekans van bepaalde klantsegmenten te kennen. Maar wat gebeurt er als je de minst kansrijke segmenten ook echt niet meer belt? Blijft je algehele conversie dan op peil? Of heb je misschien iets over het hoofd gezien? Pas als je dat test, zet je de stap van model naar operationeel inzetbare kennis. De stap waar je leert of ‘0,34’ in jouw context een hoge of juist een lage conversiekans is. Als je dat weet en je kunt die kennis toepassen, heb je data succesvol ingezet.

Wat kan 4orange daarin betekenen?

4orange is onderdeel van de MarketResponse Groep en trekken meer als partner dan als leverancier met onze relaties op. We denken mee over oplossingen en als die er niet zijn dan maken we ze. Met onze Data Scientists en specifieke Customer Data Platform helpen wij organisaties (complexe) datavraagstukken te vertalen naar oplossingen die direct implementeerbaar zijn. Benieuwd naar de mogelijkheden voor jouw organisatie? Neem dan contact met mij op.

Richard van Meurs, Data Creative bij 4orange
E: richard.vanmeurs@4orange.nl | T: 020 – 750 4400