projectsone Samen in één keer goed

Big Data-analyse

Big data is een containerbegrip voor een bonte verzameling van data-analyse technieken. Ze zijn bijzonder geschikt zijn om met grote datasets om te gaan. Deze datasets bevatten vaak verborgen informatie over mensen, processen, gebeurtenissen, etc.

Het resultaat van een big data-analyse is (vaak) een wiskundig model waarmee bijvoorbeeld gedrag, proces-resultaten of kansen op gebeurtenissen kunnen worden voorspeld.

Een voorbeeld ter verduidelijking.

Opdrachtgever: financiële instelling
Proces: verstrekken persoonlijke lening obv een complete aanvraag
Onderzoeksvraag: hoe kan worden voorspeld of een aanvraag incompleet is?

Aanpak in fasen:

1. *Verkenning van de onderzoeksvraag en achtergrondinformatie.*

Opdrachtgever heeft groot belang bij het zo snel mogelijk herkennen van aanvragen die mogelijk incompleet (missende documentatie) worden aangeleverd. Mensen vragen bij meerdere partijen een persoonlijke lening aan en wie het eerst kan beslissen over acceptatie krijgt ook vaak de opdracht. De onderzoeksvraag wordt daarom: "*wat is de kans dat een aanvraag voor een lening incompleet wordt ingediend door welke eigenschappen van die aanvraag*"

2. *Valideren en structureren van de beschikbare data set(s).*

Er wordt een dataset gebruikt welke ongeveer 8000 aanvragen bevat met per aanvraag 36 eigenschappen. De eigenschap of een aanvraag nu wel of niet compleet is aangeleverd, is onderwerp van stevige discussie. Ook wanneer dat in het proces definitief duidelijk is geworden en hoe betrouwbaar dat oordeel dan is. Verder wordt de data ontdaan van fouten, dubbelingen en co-lineaire eigenschappen. Daarnaast worden sommige kolommen samengevat (postcodes naar het eerste cijfer) en natuurlijk geanonimiseerd.

3. *Verkennen en visualiseren van de dataset.*

Er wordt een overzicht gemaakt van de dataset voor elk der eigenschappen; bijvoorbeeld hoeveel aanvragers zijn man en hoeveel zijn vrouw en wat is hun percentage incomplete aanvragen. Zo wordt ook duidelijk of er ergens informatie mist. Daarna worden allerlei grafieken gemaakt om duidelijk te krijgen hoe de kans op incompleetheid zich gedraagt. Bijvoorbeeld de verdeling van incompleetheid over de samengevatte postcodegebieden. Dit resulteert in een groot aantal grafieken welke allemaal een eigen inkijk in de kans op incompleetheid geven. Met de opdrachtgever worden deze grafieken gedeeld en besproken zodat nieuwe inzichten kunnen worden meegenomen in het onderzoek.

4. *Beantwoorden van de onderzoeksvraag d.m.v. analyse.*

Uit de grafieken van de dataset blijkt niet meteen welke factoren (eigenschappen) de kans op incompleetheid beïnvloeden. Daarom worden daar geen aannames op gedaan. Met behulp van een *Random Forest* analyse worden de belangrijkste eigenschappen (en hun interactie) voor de kans op incompleetheid zichtbaar. Random Forest verschaft ons echter geen wiskundig model waarmee de kans op incompleetheid kan worden berekend voor een specifieke aanvraag.

5. *Vereenvoudigen en modelleren van de uitkomsten.*

Vervolgens wordt een ander machine learning algoritme, *Multivariate Adaptive Regression Splines* oftewel MARS, ingezet om tot een model te komen waarmee de kans op incompleetheid voor een aanvraag a priori kan worden uitgerekend aan de hand van de eigenschappen van die aanvraag.

Om te voorkomen dat we overfitten (teveel naar de bestaande dataset toerekenen) splitsen we de dataset steeds opnieuw in een trainingset waarmee het model gemaakt wordt en een control set waarmee we de juistheid van het model controleren (*Repeated Cross Validation*).

Het resultaat is een wiskundig model dat in meer dan 80% van de aanvragen correct kan voorspellen of deze compleet zal zijn of juist incompleet.

6. *Uitleg, overdracht en rapportage aan de opdrachtgever.*

Al deze big data technieken kunnen worden toepast door gebruik te maken van een programmeer taal zoals R of Python. Hiervoor is kennis van programmeren en wiskunde/machine learning vereist. Om het werkbaar te maken voor de opdrachtgever is het wiskundig model (het antwoord op de onderzoeksvraag) vertaald naar invoervelden in excel. Hiermee kunnen scenario's worden nagerekend en het effect van beleid en/of maatregelen worden ingeschat. De eindrapportage vindt plaats door het sturen van het onderzoek van ruwe data tot en met eindmodel waarbij alle geprogrammeerde code regel voor regel kan worden uitgevoerd. Zo kan ook met nieuwe informatie een herijking van het bestaande model plaatsvinden. Ook kan de opdrachtgever de gebruikte onderzoeksmethodologie tot op de punt en komma nagaan.

projectsone Samen in één keer goed

Conclusie:

Met big data-analyse kunnen nieuwe inzichten worden verkregen die managers en beleidsmakers helpen in hun besluiten. Een gedegen, transparante werkwijze zorgt voor herhaalbaarheid van het onderzoek en hoge kwaliteit van uitkomsten.

ProjectsOne
De corridor 12L
3621 BZ Breukelen

www.projectsone.nl
info@projectsone.nl
t +31(0)23 707 8115

Bedrijfsopleidingen cedeo*erkend Maatwerk