Verslag Introductie Data Science

Student: Wesley Clarijs

Studentnummer: 2142924

**101: Wat is Data Science?**

Tijdens het maken van de assignments voor het vak Introductie Data Science heb ik een duidelijker beeld gekregen van wat Data Science nu precies is. Wanneer het vak begon dacht ik dat het enkel data analyseren was en een conclusie trekken uit de data die aangeleverd wordt aan de Data Scientist. Tijdens het maken van de assignments is het duidelijk geworden dat er ook gegevens met elkaar vergeleken moet worden en daadwerkelijk dieper ingegaan moet worden op de verschillende plots en inzichten die gemaakt worden.

Tijdens de eerste assignments was de opdracht om enkel data te analyseren en hier niet verder op in te gaan, als voorbeeld assignment 5, waar de opdracht was om de 5 landen met de hoogste en de 5 landen met de laagste levensverwachting weer te geven. Dit was enkel het sorteren van de data en hieruit de conclusie op te maken van welke landen dit betrof.

Vanaf assignment 10 moest er dieper ingegaan worden op de data die mijn eigen dataset bevatte en moesten er ook daadwerkelijk vergelijkingen gemaakt worden en relaties onderzocht worden die de data van mijn eigen gekozen dataset kon bevatten. Vanaf dit punt werd het duidelijker dat Data Science ook veel dieper op de data ingaan betekende en dat er daadwerkelijk relaties in de dataset aanwezig waren die zonder dieper onderzoek niet zichtbaar waren. Natuurlijk heb je wel verwachtingen van welke data kolommen relaties hadden tot andere data kolommen, maar zonder verder onderzoek en diepgaande analyses, zou je de dieper liggende relaties nooit achterhalen en zou je nooit op andere resultaten kunnen komen die je voorheen nooit zou verwachten.

Als voorbeeld van wat hierboven is beschreven, neem assignment 14. Persoonlijk had ik niet verwacht dat leeftijd een duidelijkere relatie zou hebben voor het krijgen van een beroerte dan bijvoorbeeld roken.

Het laatste wat Data Science ook inhoudt in mijn ogen is het daadwerkelijk bedenken en onderzoeken of de data die een Data Scientist aangeleverd krijgt ook daadwerkelijk allemaal bruikbaar is. In mijn eigen dataset ben ik er achter gekomen dat ook het verkleinen van de scope en het veranderen van de verhouding van de data in een dataset belangrijk kan zijn voor het verkrijgen van goede resultaten en dat niet alle data die een data Scientist aangeleverd krijgt, beschouwd kan worden als compleet bruikbaar.

Kortweg, Data Science houdt in mijn ogen in dat aangeleverde data onderzocht wordt op verschillende onderdelen, zoals diepere relaties vinden, en dat de data getransformeerd moet worden totdat er bruikbare resultaten naar boven komen die zonder het werk van een data scientist niet naar boven zouden komen. Dieper liggende relaties en uitkomsten die niet verwacht zouden worden, dat is wat Data Science in mijn ogen inhoudt.

**102: De taken en het proces van een Data Scientist**

In mijn ogen is de eerste taak van een data scientist het vaststellen van wat het doel is dat hij moet bereiken met de aangeleverde dataset. In de dataset die ik voor de persoonlijke opdrachten heb gebruikt, heb ik als doel het volgende vastgesteld: *Welke factoren zorgen ervoor dat de kans op het krijgen van een beroerte groter wordt?*

Als dit is vastgesteld, kan de data scientist de aangeleverde dataset gaan onderzoeken en vaststellen welke gegevens hij moet gaan gebruiken om het eerder vastgestelde doel te behalen. De volgende stap is het gaan bepalen welke data binnen de dataset bruikbaar is voor het behalen van het doel. Als voorbeeld neem ik assignment 14 van mijn portfolio. Hier heb ik namelijk zelf ook gekeken welke data binnen mijn dataset in mijn ogen waardevolle data bevat en welke data minder goed bruikbaar is. De data die binnen mijn dataset waardevolle en bruikbare informatie bevat heb ik daarom dieper onderzocht, namelijk het verschil van een beroerte krijgen qua geslacht, woonomgeving en of iemand gerookt heeft. Als laatste heb ik gekeken of het werk type van iemand invloed heeft op het krijgen van een beroerte, maar na het bekijken van de uitkomsten kwam ik er achter dat de verhoudingen tussen de hoeveelheid mensen per beroep zeer ongelijk was en heb ik dit niet verder onderzocht omdat dit in mijn ogen geen duidelijke toevoeging zou geven op het uiteindelijke antwoord op het doel.

Wanneer de bruikbare data is vastgesteld, voert de data scientist diepgaande analyses uit en stelt hij zijn verwachtingen vast. Wanneer de data scientist de analyses gaat uitvoeren, gaat hij ook kijken of zijn verwachtingen kloppen met het uiteindelijke resultaat en of er eventuele uitkomsten zijn die hij niet zou verwachten. Als voorbeeld neem ik hier wederom assignment 14 voor. Hier is namelijk gebleken dat mensen die gerookt hebben een grotere kans hebben op het krijgen van een beroerte dan mensen die momenteel nog steeds roken. Persoonlijk had ik verwacht dat mensen die nog steeds roken een grotere kans zouden hebben op het krijgen van een beroerte dan mensen die gestopt zijn.

Een ander voorbeeld zijn de relaties die in assignment 10 onderzocht moesten worden. Hieruit kwam naar voren dat leeftijd de grootste factor was die de kans op het krijgen van een beroerte vergrootte. Persoonlijk had ik verwacht dat een hoge bmi of een hoog glucose gehalte de kansen zouden vergroten, maar uit dit onderzoek bleek dus dat leeftijd de grootste relatie had tot het krijgen van een beroerte. Tijdens het maken van dit assignment ben ik er ook achter gekomen dat het verkleinen van de scope en het veranderen van de verhoudingen binnen een dataset kan zorgen voor het krijgen van andere en bovendien duidelijkere resultaten. Zo bleek het vanaf assignment 10 dat het veranderen van de verhouding er voor zorgde dat dieper liggende relaties duidelijker werden en hierdoor kwam er ook in assignment 16 een andere Decision Tree naar boven die duidelijker was dan wanneer de volledige dataset gebruikt zou worden.

Nadat dit alles is gedaan, is het de taak van de data scientist zijn conclusies die hij opgemaakt heeft uit zijn onderzoeken en analyses te delen met de opdrachtgever. In mijn geval doe ik dit door eerst mijn verwachtingen op te schrijven in het notebook, dan de analyses uit te voeren en hierna mijn verwachtingen te vergelijken met de uitkomsten en dit te onderbouwen.