**Samenvatting**

HoB blijkt vanaf haar 3e jaar een competatief salaris te bieden waarin het meer dan 50% van de markt verslaat. In de eerste 2 jaren verdienen werknemers boven het modale loon. Het blijkt echter dat House of Bèta nog ruimte heeft om de hard skills van haar personeel te ontwikkeling. Het belangrijkste zijn skills m.b.t de cloud, big data, en geavanceerde data science skills. Op het moment van schrijven staan de volgende potentiële opdrachtgevers boven aan de lijst met openstaande vacatures:

**Inleiding**

Dit onderzoek richt zich op de vraag hoe competitief House of Bèta is op de markt voor data professionals. Hierin wordt gekeken naar zowel het salaris (dus wat House of Bèta biedt aan het personeel) als de mogelijkheden voor House of Bèta om een groter deel van de markt te bedienen (gevraagde vaardigheden en potentiële opdrachtgevers).

House of Bèta heeft al langer de wens om inzichten te krijgen in de reden van uitstroom van haar consultants, gezien dit een belangrijke asset is van de organisatie. Wegens juridische en technische beperkingen is het echter niet mogelijk om hierop diepgaand onderzoek op te doen.

Een tweede wens is om de markt beter te kunnen bedienen en te groeien als de leverancier van IT personeel. Na overleg is gebleken dat beide wensen gekoppeld kunnen worden in dit onderzoek. Het onderzoek helpt House of Bèta inzichten te krijgen in een ontwikkelende markt, haar eigen arbeidsvoorwaarden op waarden te schatten en biedt een basis voor verdere analyse aangaande de churn van haar personeel.

**House of Beta**

House of Bèta (HoB) is een detacheerder op het gebied van (business) IT. HoB probeert voornamelijk starters op de arbeidsmarkt aan te trekken en probeert deze te matchen aan passende opdrachtgevers. De consultants hebben de mogelijkheid zich na verloop van tijd te specialiseren waarin HoB de trainingen en opleiding faciliteert of betaald.

HoB is een relatief nieuwe naam op de markt maar voert al sinds 2015 haar werkzaamheden uit onder het moederbedrijf Talent&Pro. De uitstroom van consultants die T&P heeft gezien over2021 is de aanleiding geweest om meer grip op deze uitstroom te willen krijgen (de eerste wens).

Tijdens de een recente informatieavond van HoB zijn nieuwe doelstellingen voor 2022 gepresenteerd waar ook de ontwikkeling en groei van data professionals naar voren komt. Dit is dan ook de aanleiding voor de tweede wens van HoB.

**Businessvraag**

Dit onderzoek en de wensen van HoB kunnen worden vastgelegd in een centrale vraag: “Hoe positioneert HoB zich ten opzichten van de Nederlandse arbeidsmarkt op het gebied van data professionals kijkend naar loon, vaardigheden en potentiële opdrachtgevers?”

Deze vraag kan verder worden opgedeeld in 3 deelvragen:

1. Hoe verhoudt het salaris van HoB zich tot de nederlandse markt?

2. In hoeverre sluit het “curriculum” van HoB zich aan bij de ontwikkelende markt?

3. Welke opdrachtgevers actief op de markt worden nog niet bediend door HoB?

**data gebruik**

Binnen HoB vinden nog geen datamining activiteiten op dit gebied plaats. Buiten de resultaten van dit onderzoek wordt er dan ook een een script aangeleverd die het mogelijk maakt om de datamining voor dit onderzoek voort te zetten. Dit script is beperkt tot het zoekwoord “data” op de vacaturesite Indeed, maar kan uitgebreid worden om op andere woorden te zoeken.

**Business Success criteria**

De volgende kwalitatieve eisen worden gesteld: er moet er een analyse komen van HoB versus markt inkomen, een analyse over de vraag verdeling over bedrijven en een over skills. Daarnaast moet er een bruikbare scraper voor vervolg onderzoek worden opgeleverd

**Methoden**

Dit hoofdstuk behandelt de data en het gebruik hiervan. In “data understanding” worden basale kenmerken van de dataset besproken, “data preperation en modelling” gaat in op het schonen een verrijken van de data. Als laatste gaat “reproduceren van het onderzoek” over de mogelijkheid om dit onderzoek na te bootsten.

**Data** **understanding**

Hieronder worden de scope, de gebruikte datasets en de basale kenmerken beschreven.

**Gekozen datasets en scope**

Om te beginnen is hieronder een versimpelde datalog weergegeven. Dit bevat informatie over de datasets in de ruwe vorm (te vinden in ".\Data\_raw")".

D001 is na vooronderzoek gekozen als basis dataset. Hierin staan onder andere (hard)skills, functie type (data analist, -engineer, -scientist). Dit legt een goede basis voor het onderzoek naar data specialisaties binnen HoB, hier worden namelijk dezelfde specialisaties ook aangeboden. Er zijn andere datasets overwogen echter bevatte deze te weinig relevantie functie type (< 200 rijen) of ontbraken hier relevante skills.

D002 is gescrapte data van Indeed. Het rationaal achter het zelfstandig scrapen van Indeed data komt voort uit het tekortschieten van bestaande API’s en afwezigheid van relevante datasets.

**Beschrijving van de ruwe data**

Beginnend bij de beschrijving van D001. Deze dataset is al erg opgeschoond. Een aantal basale kenmerken zijn hieronder programmatische weergegeven. Allereerst is duidelijk dat er 43 kolommen aanwezig zijn. Echter voor het doel van dit onderzoek zijn maar twee kolommen echt interessant. Dit zijn “Skill” en “Job\_Type”. Hierin staan respectievelijk een lijst van (hard)skills en de functie soort per observatie.

Zoals hierboven getoond komt duidelijk naar voren dat het aantal unieke “Skills” bijna gelijk is aan het aantal unieke rijen. Dit betekent echter niet dat er zoveel verschil zit tussen de vacatures. Het is namelijk zo dat [SQL, Python] en [Python, SQL] hetzelfde betekenen maar wordt gezien als twee unieke rijen. Deze informatie wordt gebruikt in latere ETL stappen. Daarnaast is opvallend dat de verdeling tussen data analist (31%), data engineer (24%) en data scientist (45%) niet volledig evenredig is.

Dataset D002 bevat datavan Indeed. Hiervoor is er gezocht naar alle functies waar het woord "data" in voorkomt. D002 bestaat in de ruwe vorm uit meerdere losse bestanden (een voor elke dag dat er een run is geweest van de scraper). Om een indruk te krijgen van wat er in een dergelijk bestand staat kan er gekeken worden naar ".\Data\_raw\D002\2022-03-08 scraped indeed data.csv". Hieronder worden weer de voornaamste kenmerken.

Het is gelijk duidelijk dat de data niet verder terug gaat dan 6 februari 2022. Dit zal effect hebben op de trend analyses in dit onderzoek. Ook wordt duidelijk dat een dataset meer rijen heeft dan unieke URL’s. Dit betekent dat dezelfde URL waarschijnlijke meerdere keren is gescraped. Het derde feit dat duidelijk wordt, euro symbolen (€) worden foutief weergegeven (ï¿½). Dit en andere datakwaliteitsproblemen worden in het ETL process opgelost.

**Data preperation en modeling**

Onderstaande processflow geeft in grote lijnen weer hoe er vanaf de ruwe data wordt gekomen tot een eindproduct.

Het volledige proces is in drie segmenten op te delen. Bovenin betreft D001. Segment twee (midden) betreft D002. Onderin staat segment drie waarin het eindresultaat weergegeven wordt.

Omdat de tijd en scope beperkt is voor dit onderzoek zijn er een aantal aannames gedaan om problemen en complexiteit te voorkomen:

Er zijn 3 type (analist, engineer, scientist). De dataset wordt verrijkt met een van deze type zelfs al wijkt dit af. Zo kan de functie “data entry” als “data analist” worden geclassificeerd.

ETL (midden-rechts)

Salary

Het volgende is aangenomen:

Salaris > 10.000 is jaarsalaris of Salaris < 100 is uur tarief. Dit wordt omgerekend naar maandsalaris (delen door 12 of maal 168 respectievelijk). Alles tussen de 100 en 1000 wordt als scraping fout gezien en omgezet naar NA.

Aannamens tijdens process stappen,

De scripts voor deze stappen zijn te vinden in "./Supporting\_code", specifiek:

Process stap

Script naam

ETL (linksboven)

Generate\_ref\_skill\_list.R

Scraping

indeed\_scraper.R

ETL (midden-rechts)

Scraped\_indeed\_data\_wrangler.R

process stappen en bijbehorende scripts.

De scripts en betreffende functies zorgen voor een dataset die gebruikt kan worden voor dit onderzoek. De scripts zijn voorzien van helpende comments om het gebruik te vergemakkelijken.

Reproduceren van het onderzoek

De mappen: "Data\_cleaned", "Data\_raw", en "Supporting\_code" bevatten alle data en code die nodig is om dit onderzoek te reproduceren. In dit markdown document worden de data en benodigde scripts aangeroepen om tot het eindresultaat te komen. De benodigde R libraries zijn: "tidyverse", "stringr", "lubridate", "tibble", "textTinyR", "XML", "rvest", “tseries”.

**Resultaten**

Dit hoofdstuk beantwoord de deelvragen aan de hand van diverse analyses. Ook wordt besproken welke vragen (nog) niet kunnen worden beantwoord en welke artefacten worden aangeleverd.

**Hoe verhoudt het salaris van HoB zich tot de nederlandse markt?**HoB biedt verschillende manieren om salaris verhoging te halen, dit tezamen met de gesloten HR systemen maakt het lastig om een referentie salaris te bepalen. In dit geval heeft HoB op haar website aangegeven wat een collega kan verdienen in jaar 1, 2 en 3 (€2.763,- , €3.087,- , €3.591,- respectievelijk). Om deze deelvraag te beantwoorden worden deze drie salarissen gebruikt.

De grafiek hieronder geeft de salaris verdeling van data vacatures op Indeed weer. Het eerste wat duidelijk wordt is dat het loon van HoB bij aanvang al boven modaal ligt. Daarnaast is het salaris in jaar 3 hoger dan het mediaan salaris op Indeed. Ofwel, HoB biedt een hoger maandsalaris in jaar 3 dan 50% van de markt (afgaande op Indeed).

Bovenstaande analyse zegt echter niet alles. Hierin wordt het 3 verschillende maandsalarissen van HoB vergelijken met de gehele onderzoekspopulatie. Op Indeed worden diverse ervaringseisen gesteld. Dit gebeurt in de vorm van “1 – 3 jaar ervaring” o.i.d. Tijdens het schonen is altijd de kleinste hoeveelheid opgeslagen (1 – 3 is opgeslagen als 1).

Hieronder is de salaris verdeling te zien gesegmenteerd op gevraagde ervaring. Door de dataset op te delen in negen segmenten (gemiddeld 166 observaties) is de verdeling gevoeliger en minder accuraat. Toch geeft het de indruk dat het salaris geboden door HoB doorgaans beter is dan het salaris van Indeed (jaren 6 en 7 buiten beschouwing gelaten).

Het salaris bij HoB wordt niet alleen bepaalt door werkervaring. Ook het afronden van cursussen en het opdoen van skills wordt beloond. Wat betreft hard skills kan er vanuit worden gegaan dat een consultant 4 tot 6 skills bezit na de eerste 3 jaar bij HoB. Dit zijn data modeling, excel, R, Python, SQL en PowerBI. In onderstaande drie grafieken is het mediaan salaris op Indeed weergegeven verdeeld over het gevraagde aantal skills. De zwarte lijn geeft het bijbehorende HoB salaris aan.

De drie grafieken met elkaar vergeleken laat zien dat HoB vooral vanaf het 3e jaar echt een competatief salaris biedt. Hierbij kan verwacht worden dat een consultant zo’n vier tot zes vaardigheden bezit. Daarmee verslaat het de markt kijkend naar skill,-, loonratio.

Bovenstaande analyse zou nog verdiept kunnen worden in een vervolg onderzoek door het aantal jaren werkervaring mee te nemen. Dit levert nu echter te weinig observaties (11) per categorie op.

**In hoeverre sluit het “curriculum” van HoB zich aan bij de ontwikkelende markt?**

Deze deelvraag behoeft twee analyses. Enerzijds is het van belang om te weten hoe het huidige aanbod van HoB aansluit op de huidige vraag. Daarnaast is het van belang om te weten hoe de vraag naar Skills zich ontwikkelt. Binnen dit onderzoek is het laatste lastig te realiseren. Een goede trend analyse zal ~3 jaar aan data behoeven. Hierin schiet de dataset te kort, dit zal impact hebben op de analyse.

Het eerste waar naar gekeken wordt, de huidige vraag naar skills. Hieronder zijn twee grafieken te vinden. De eerste grafiek geeft de top 20 vaardigheden weer, de tweede grafiek de minst gevraagde 20.

Zoals gesteld zijn de consultants in de eerste 3 jaar voornamelijk bezig met de skills data modeling, excel, R, Python, SQL en PowerBI. Zoals te zien sluit HoB met vier van de skills aan bij de meest gewilde skills. Ook wordt er ingezet op cloud oplossingen binnen HoB wat aansluit op de gevraagde skills “azure” en “aws”. Tableau is een onverwachte vaardigheid, in plaats daarvan was de verwachting namelijk dat PowerBI in de top zou staan hier zit dus een gat in vraag en aanbod. Daarnaast wordt duidelijk dat er op het gebied van data science een gat bestaat tussen HoB en de markt. Zo zijn data science vaardigheden (ai, machine learning, spark, scala, go) niet direct onderdeel van het curriculum.

De minst gevraagde skills zijn voor dit onderzoek minder relevant. Toch is het goed om kaart te hebben welke skills het minst gevraagd worden. Dit zijn voornamelijk skills om te vermijden.

Interessant om nog op te merken dat het voornaamste aandeel van gevraagde skills zich rond de top 5 bevindt. Dit zijn dan ook gebieden waar HoB zich het best op kan focussen om relevant te blijven voor de markt. Wilt HoB uitbreiden richting minder gevraagde maar toch relevante skills? Dan is een tweede stap om ontwikkeling te bieden in de top 20. Dit kan voor programmeertalen en frameworks een cursus zijn, voor concepten zoals “rest” (API’s) zijn dit informatie avonden.

**Ontwikkeling skill behoeften**

De vraag “in hoeverre sluit het “curiculum” van HoB aan op de ontwikkelende markt” is in de ten delen beantwoord. Het is duidelijk wat de huidige vraag is. Met de huidige ontwikkelingen op het gebied van big data en cloud technologie die in de praktijk zichtbaar zijn is het aannemelijk dat deze top 20 skills ook belangrijk zullen zijn om verder in de gaten te gaan houden.

Om een sterke trend analyse te doen is minstens ~3 jaar aan data nodig. Dit is een beperking die verder in het hoofdstuk “discussie” wordt besproken. Wat wel aangeleverd kan worden is de functionaliteit om deze analyse op een later moment te doen.

Onderstaande grafieken, met de verloop per skill, en de bijbehorende code generen de inzichten om een simpele trend analyse te doen. In onderstaande figuren wordt duidelijk dat er momenteel een neerwaartse trend is, de aannamen is dat dit wanneer er meer data beschikbaar is door scrapen de trend een positief verloop laat zien.

De onderstaande grafieken geven de dagelijks vraag naar skills aan en geven de algehele trend weer. Op het moment van schrijven zijn dit enkel negatieve trends. De aannamen is dat, na verloop van tijd, wanneer meer data gescraped wordt deze skills in iedergeval een positieve trend zullen laten zien.

**Welke opdrachtgevers actief op de markt worden nog niet bediend door HoB?**

De laatste deelvraag binnen dit onderzoek richt zich op de mogelijke opdrachtgevers. Hiervan was het oorspronkelijk de bedoeling om een trend analyse te doen bij opdrachtgevers. Echter is dit door het tekort aan data niet te realiseren. In plaats daarvan wordt hier gekeken naar de huidige gap tussen de organisaties die vacatures uit hebben staan op Indeed en de organisaties die HoB bedient.

Allereerst wordt in onderstaande grafiek de top 25 bedrijven getoond op basis van het aantal geplaatste vacatures vanaf de eerste datum in de dataset. Grofweg 50% van deze organisaties betreffen detacheerders (waaronder House of Bèta). Dit is een interessant gegeven, de aannamen hier is namelijk dat de reden dat er zoveel vraag vanuit detacheerders en consultants is komt door de grote vraag naar data professionals in het algemeen.

Buiten de bevestiging van de grote vraag naar data professionals is het vooral interessant om te kijken naar de organisaties met de hoogste vraag die niet vallen onder consultancy of detacheerders. In de tweede grafiek is een subset gemaakt van deze 25 bedrijven zonder consultancies of detacheerders. Uit deze grafiek zijn op dit moment 2 partijen waar HoB geen personeel aan levert: Bertelsmann SE & Co, en Amarant.

Uiteraard is dit een summiere analyse om een daadwerkelijke gap aan te tonen vandaar nog twee andere aanvullende analyses onderbouwd door de twee onderstaande grafieken. De eerste grafiek laat zien dat voornamelijk python, sql, R, ai en (neural) network vaardigheden veel gezocht wordt door de subset van bedrijven. De tweede grafiek laat eenzelfde verdeling zien maar dan vanuit een andere invalshoek. De eerste grafiek beschouwt de meest gevraagde skills vanuit de hele onderzoeks populatie. De tweede grafiek daarentegen laat toont alleen de top skills die daadwerkelijk door deze bedrijven worden gezocht. Uit deze laatste grafiek komen drie inzichten naar voren. De belangrijkste skills blijven ongewijzigd, SAS (een programeertaal waar HoB zich sinds kort op richt) komt naar voren op de 5e plek. De overige 5 vaardigheden hebben veelal te maken met big data toepassingen. Een kanttekening die hierbij moet worden gemaakt is dat Bertelsmann SE & Co. Deze uitslag lijkt te beïnvloeden door hun sterke aanwezigheid (NN heeft dit effect minder).

Om deze deelvraag kort samen te vatten: HoB heeft een kans om bij Bertelsmann SE &Co. en Amarant nieuwe kansen te creëren. Om haar positie bij partijen te versterken is het naast bestaande cursussen nog waardevol om te kijken naar big data en gevorderde data science skills.

**Discusie**

In dit hoofdstuk komen de verbeterpunten en tekortkoming aanbod. Ook wordt er aandacht besteed aan de deployement en wat er daadwerkelijk van waarde wordt opgeleverd.

**Data omvang**

Allereerst de omvang van de dataset. Een groot deel van dit onderzoek is beantwoord met de beschikbare data. Echter zijn verdiepingen op trend analyses achterwegen gelaten alsmede analyses over meerdere dimensies waarvoor de dataset te klein is. Bij aanvang van dit project was de inschatting van de hoeveelheid data die binnengehaald kon worden niet direct duidelijk. Het voornaamste verbeterpunt is om de dataset uit te breiden voor verdere analyses. Een leerpunt is dat analyses over drie dimensies een dataset erg zullen versnipperen, hier moet volgende keer rekening mee gehouden worden bij het beoordelen van het aantal nodige observaties.

**Data kwaliteit**

Datakwaliteit is al kort aangestipt door het project heen. Het ontbreken van een API en de tijdsdruk voor dit onderzoek heeft betekend dat er keuzes zijn gemaakt ik en het scrapen, schonen en vereiken van de data. De belangrijkste om hier te benoemen zijn: functie type en vaardigheden.

Allereerst zijn er fouten aanwezig in de functie type. Er is onderscheid gemaakt tussen data analist, data engineers, en data scientisten. Hoewel er vacatures kunnen zijn voor hele andere data functies is iedere functie geforceerd in een van deze catagorieën.

De tweede kwaliteit issue zit in de vaardigheden. Een goed voorbeeld is de vaardigheid “network”. Dit had waarschijnlijk “neural network” moeten zijn. Wanneer er gezocht wordt op “power bi” zal “powerbi” niet worden gevonden hierdoor mist er data. Ook zijn skills zoals “sql” in veel verschillende varianten te vinden zoals “mysql”. Deze skills zijn apart opgeslagen, dit geeft een vertekende weergaven van de vraag naar de echte skills.

**Data scraping**

Als laatste zijn er twee kanttekeningen bij het data scrapen. Allereerst is alleen Indeed als bron gebruikt. Qua website omvang is dit misschien een goede representatie maar het zou kunnen dat hier nog een beter alternatief voor is. Het gevaar van dubbele vacatures binnenhalen bij het gebruik van meerdere scrapers is dan natuurlijk een gevaar. Wegens tijdsgebrek is hier geen verder onderzoek.

Het tweede punt betreft de query. De scraper is gericht op een zo breed mogelijk profiel. Om deze reden is als zoekopdracht "data" gebruikt. Een voorgestelde verbetering is de scraper om te bouwen naar een generieke scraper. Hierbij kan de query vrij worden ingevuld en opgeslagen in een variabele die ook wordt opgeslagen in de dataset. Voor dit onderzoek zou dit hebben geholpen om accurater te zoeken naar analisten, engineers, en scientisten. Dit sluit aan op de verdere kritiek onder "data kwaliteit".

**Deployment**

Als laatste deployement. Bij dit onderzoek worden een aantal artefacten aangeleverd die gebruikt kunnen worden voor ad-hoc analyses. De scripts en code zijn zonder de markdown te gebruiken en makkelijk te bewerken. De code is voorzien van commentaar en kan relatief makelijk worden uitgebreid met nieuwe functionaliteiten waar nodig.

Artefact

Gebruik

rmd\_Capstone\_DMAE\_MaelvDijk.Rmd

Markdown document met de analyses voor dit onderzoek.

Data\_cleaned / Data\_raw

diverse .CSV bestanden met brondata.

indeed\_scraper.R (Supporting\_code)

webscraper voor de Indeed vacature website. Te gebruiken om nieuwe data binnen te halen.

Scraped\_indeed\_data\_wrangler.R (Supporting\_code)

script om de gescrapte data (opgeslagen in .csv format) op te schonen en verijken met nieuwe datapunten.

Scraped\_indeed\_plotter.R (Supporting\_code)

script om de gescrapte data na schonen en verrijken op een uniforme wijze te plotten.

# Conclusie

Als laatste het antwoord op de vraag: “Hoe positioneert HoB zich ten opzichten van de Nederlandse arbeidsmarkt op het gebied van data professionals kijkend naar loon, vaardigheden en potentiële opdrachtgevers?”

Het is duidelijk geworden data HoB een competatief salaris biedt aan haar medewerkers,In iedergeval vanaf jaar 3 waar zij meer dan 50% van de vacatures in loon voorbijgaan. Voorzowel jaar 1, 2, en 3 verdient een HoB consultant boven modaal. Kijkende naar de gehele onderzoeks populatie.

Op het gebied van vaardigheden en opdrachtgevers doet HoB het erg goed. Zij bedient een groot deel van de bedrijven die de meeste vacatures uit hebben staan. Hierin bestaat voor HoB de kans om met Amarant en Bertelsmann zaken te gaan doen. Wat betrecft de vaardigheden is het belangrijk dat HoB blijft inzetten op de belangrijkste vaardigheden (sql, python en r) maar ook de verdieping aanbiedt op het gebied van data science (ai en neurale netwerken) en big data oplossingen (scala, spark en hive).

Kortom HoB lijkt competitief zowel op loon als aanbod waarbij er kansen bestaan om vooral naar opdrachtgevers toe verder uit te breiden in zowel vaardigheden als partners.