RSG finance
Datascience in Python

Opdracht

Data-impact met Python!

Voor de portfolio opdracht ga je aan de slag met voor jou interessante data-bron(en). Je begint met het formuleren van een bussinesvraag. Dit geeft richting aan hoe je uiteindelijk je analyse zult gaan presenteren. waar is het bedrijf/de afdeling in geinteresseerd? Vervolgens haal je je data binnen in Python en start je met je analyse. In je notebook blijf je noteren in welke analyse-stap je je bevindt. Dit vergroot de reproduceerbaarheid van je werkzaamheden voor jou (en in de praktijk eventuele collega's). Uit deze eerste grove analyse blijf je over met verschillende uitkomsten en een rommelig script. Hieruit destilleer je wat van vitaal belang is voor jou, reproduceerbare, uitkomst. Ofwel: Kies de belangrijkste metrieken voor jou bussiness-vraag. Voor deze specifieke metrieken bouw je een werkende pipeline. Het grove script, de pipeline, en het rapport of dashboard wat daaruit komt lever je allemaal bij ons in. Naast reproduceerbaarheid is een heldere en inspirerende communicatie van je uitkomsten belangrijk: Zo komt er bedrijfsbreed nieuwe inzicht in de data.

Inleveren: 10 juni. Grove schets, pipeline & evt product.

Advies. Lever gaandeweg, tussen 13 en 27 mei een schets van je huidige proces in voor feedback en ondersteuning. Hierna blijven we natuurlijk beschikbaar voor ondersteuning en vragen. Maar ook naderen we de project-deadline.

Criteria

Data-inzicht 30% Je zorgt voor een goede communicatie over bruikbaarheid v.d. uitkomst. Denk hierbij aan de stakeholders (doelgroep/afdeling) wat voor hen van belang is om te weten, en in welke vorm je de uitkomst het best aanlevert.

Analytics box framework 30%

Je hebt een ruwe maar uitgebreide schets gebruikt om daaruit een pipeline te destilleren. In de scripts van beide, de schets en de pipeline, blijf je goed aangeven op welke plek van het analyseproces je je bevindt. Deze overzichtelijke scripts (schets en pipeline) lever samen met het product je bij ons in.

Reproduceerbaarheid van de uitkomsten. 20% Je Uitkomsten zijn reproduceerbaar en je kunt door middel van de pipeline duidelijk aantonen hoe je tot deze uitkomsten bent gekomen.

Python basics 30% Je toont aan je data goed te kunnen verkennen, door middel van indexeren met behulp van de blokhaken, de (i)loc- functies en boolean arrays en door te groeperen en te aggregeren.

Je kunt Matplotlib en Seaborn gebruiken om een duidelijk beeld te geven van de tendensen in de gekozen data. Dit doe je op een functionele manier op verschillende momenten in je pipeline.

Je voorkomt boilerplating in je pipeline door slim for/while loops, Map/Filter en custom hulp-functies in te zetten.

Beredenering extra technieken 10%:

De laatste tien procent van van de reguliere punten kan je behalen door een kloppende redenering van de waarde van web-scraping en machine learning voor jou bussiness-vraagstuk, zonder deze technieken daadwerkelijk te hebben toegepast. De waarde van deze technieken is per vraagstuk en per moment binnen het project voortkomend uit het vraagstuk variabel. Het wel of niet gebruiken van bovenstaande technieken is goed te onderbouwen in een project-document en levert extra inzicht op over toekomstige mogelijkheden van de huidige data-architectuur.

Extra punten:

Je kunt extra punten verdienen door het daadwerkelijk toe te passen van Web-scraping en/of Machine learning. Het op het juiste moment inzetten van scraping of ML kan je totale score behoorlijk omhoog brengen.

Voor machine learning kan je punten verdienen voor het importeren, instancen en gebruiken (trainen en fitten) van een machine learning model. Meer punten krijg je als je met behulp van parameter-tuning een goede balans vindt tussen de bias & variance van je model. Ook met notatie over de uitkomst zijn extra punten te verdienen, bijvoorbeeld in een juiste redenering van de gedraging van je model.

Wat betrefd web-scraping kan je extra punten verdienen voor: Het vinden van waardevolle web-informatie voor jou bussinessvraag (de beredenering van de waarde). zowel als het technisch uitvoeren van het targeten van de informatie. Een scraper die

Tussen de basis-punten en de extra punten zit overlap wat betrefd python-technieken. Zo kun je door het toepassen van bepaalde genoemde basis-technieken tijdens scrapen/ML voor deze basis-technieken wel punten scoren. Gebruik van basis-technieken tijdens scraping/ML levert geen **extra** punten op voor het gebruik van de Basis-technieken.

Punten voor basis-technieken zullen nooit zwaarder wegen dan 30% van je totale score.