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.

Criteria

Data-inzicht *30%*

Je zorgt voor een goede communicatie over bruikbaarheid v.d. uitkomst. Denk hierbij aan de doelgroep/afdeling met behulp van Visualisaties en de 4 levels of listening (U-theory)

Analytics box framework *20%*

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. Je levert overzichtelijke scripts af.

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 en functies in te zetten.

Je weet het pakket Scikit learn te gebruiken en gebruikt op z'n minst één machine learning model om te dataminen. Hierbij pas je parameter-tuning toe om de prestatie van het model te verbeteren. Je voorkomt over- en underfitting. Denk ook aan cross validation.