Project 1 “De gezondheidzorg”

Verslag **Jiyoon Kim 19 maart 2023**

Dit was een bijzonder leerzaam proces om data engineering te ervaren voor het eerst in mijn leven.

De gekozen pipeline voor dit project bestaat uit twee delen. Het eerste deel genereerde de meest geschikte data voor ons project en het tweede paste de lineaire regressie toe voor gebruik in de eindgebruikersinterface. Vanaf het begin zijn wij bewust om ideale data te kiezen die het beste voor de lineaire regressie paste. We verzamelden data via rest API en SQL en als csv bestanden. Daarna hebben we de data transformatie gedaan. Er waren weinig NaN’s, het laten vallen ervan zou geen significant effect hebben voor het eindresultaat. Duplicaten zijn verwijderd en enkele negatieve waarden die ongeldig zijn, werden ook weggelaten. Vreemde tekens omgezet in NaN's en weggelaten. Tijdens EDA (Exploratory Data Analysis) onderzoek was ik bezig om het effect van outliers te ontdekken. Outliers kunnen geknipt worden via IQR methode maar ook Debscan Algoritme kan voor de detectie van Outliers in een dataset gebuikt worden. Het is handig voor het clusteren van datasets en kan Outliers in 2-dementie detecteren. Ik wil dit graag verder onderzoeken. We hebben verschillende sets gedefinieerd om te testen voor de beste resultaten bij het passen van het model. Uiteindelijk hebben wij de meest geschikte data gekozen, die de hoogste score hadden gehaald met R2 coëfficiënt of determination en RMSE (The root-mean-square error).

Ik heb met de gekozen dataset F4, data analysis uitgevoerd. Met deze dataset, lijkt ‘genetic’ het meest gecorreleerd te zijn met ‘lifespan’. ‘Exercise’ heeft positieve correlatie met ‘lifespan’. ‘Smoking’ en alcohol hebben negatieve invloed op de ‘lifespan’. Ik heb niet alleen met lineair regressie maar ook met K-Means clustering geëxperimenteerd. De herkenning van de patronen vind ik erg interessant. ‘Genetic’ en ‘exercise’ hadden duidelijk andere patronen.

Met de gekozen data F4, heb ik een interface gebouwd. Ik heb Pickle gebruikt om de dataset te bewaren en importeren en uiteindelijk de informatie van de patiënten te bewaren. De patiënten kunnen antwoorden krijgen over de levensverwachting, BMI, tot welke categorie de BMI hoort en premieberekening gebaseerd op levensverwachting. De gebruiker kan in de pipeline en git bash runnen om de resultaten te krijgen.

Het was een goede ervaring om het hele proces van data engineering te leren. Het zou interessant zijn ook actuele datasets te analyseren. Bijvoorbeeld kan ik openbare dataset van CBS (Centraal Bureau voor de Statiek) in Python gebruiken. Ik heb deze manier onderzocht en gedeeld met mijn team.

Hartelijke bedankt voor Frank, die erg innovatief en slim is, Hans, die enorme veel ondersteuning en structuur aan me geeft, Stephan, die super enthousiast was om een goede pipeline en interface te genereren. Ik vind onze samenwerking fantastisch!