Data science minor Portfolio herkansing

Dit is mijn code voor de herkansing. Hierbij heb ik een code toegevoegd om the mean accuracy te laten zien en de individuele score van de training set en de test set.

Maar eerst, de scores die ik probeer te krijgen voor mijn model zijn de accuracy, precision, en recall. Deze scores zijn allemaal metingen van de prestaties van het model en kunnen helpen om de kwaliteit van de voorspellingen van het model te beoordelen. Dit kan mij laten zien hoe goed het model presteert en hoe het presteert op verschillende aspecten van de voorspellingen.
Een hoge accuracy betekent dat het model de meeste voorbeelden correct kan classificeren, maar een lage precision kan betekenen dat het model veel fout-positieven heeft, wat kan leiden tot ongewenste resultaten. Als het model een lage recal heeft betekent dit dat het model veel fout-negatieven heeft wat kan leiden tot gemiste kansen.

Het model model heeft met 63% accuracy een vrij lage score wat betekent dat het model veel niet correct kan classificeren en dit zal ik in eerste instantie willen proberen te verbeteren.

Ik focus me nu vooral op de Decision Tree Classifier, omdat alle 3 de classifiers redelijk dezelfde resultaten geven. Alleen is de Decision Tree Classifier voorlopig heel iets beter en hiervan wil ik dus met name de accuracy gaan verbeteren.

Ik had mijn dataset al gesplits in een training en testset, maar nu wil ik ook kijken of mijn model overfit. Als de accuracy van de trainingsset significant hoger is dan op de testset, dan is er waarschijnlijk sprake van overfitting. Ik train het model op de trainingsset en evalueer de prestaties op de testset. Overfitting treedt op wanneer het model te complex wordt en zichzelf aanpast aan de trainingsgegevens, zodat het niet in staat is om nieuwe gegevens te generaliseren.

Om te bepalen of mijn model overfit heb ik als eerste cross-validation gebruikt, waarbij het model gesplit wordt in 5 subsets. Waarbij de testset steeds 1 subset gebruikt totdat elke subset gebruikt is op de testset. Hier komt uiteindelijk een score van 55% uit.
Elke keer wordt de accuracy score berekend en hiervan de mean wordt uitgeprint. Als de mean accuracy score lager is dan de training accuracy score, dan kan er sprake zijn van overfitting. In mijn model is de training accuracy score 90% en de mean accuracy score 55%, dus is er sprake van overfitting.

Ook heb ik gecontroleerd wat de prestaties zijn op de trainingsset en testset. Wanneer de prestaties van het model op de testset lager zijn dan op de trainingsset, dan is er sprake van overfitting. In mijn model is de training set score 90% en de test set score 51% en dus is er sprake van overfitting.

Ook zie ik dat in de confusion matrix, de False Positives en False Negatives vrij hoog is. Idealiter zou dit omlaag moeten gaan om een hogere accuracy score te geven.

Met een mean accuracy score van mijn model van 55% betekent dat het 55% van de tijd een goede uitkomst voorspelt. Dit is maar iets beter dan met een rood/zwart voorspelling van 50%, alhoewel de meeste tags minder vaak voorkomen dan 50%, maar dit betekent dat het model niet goede resultaten geeft. Het model maakt verkeerde voorspellingen op een groot deel van de dataset.
Ik ga in de Decision Tree Classifier proberen om de accuracy score te verbeteren en overfitting te verminderen. Dit kan door middel van het toepassen van regularisatie, het verminderen van features (tags), het geven van een max-depth aan het model, en eventuele andere oplossingen.

Nu heb ik een max_depth, min_samples_split, min_samples_leaf toegepast als regularisatie middel. Normaal gesproken wanneer deze parameters omhoog gezet worden, wordt het model simpeler en zal overfitting minder worden. Hierbij is het effect nog niet heel geweldig.

	Max_depth = 3	Max_depth = 12
Accuracy	56	67,5
Precision	60	73
Recall	80	72
Mean accuracy score	56	59
Training set score	65	86
Test set score	53	54

Meer data kan een optie zijn om de accuracy te verhogen en dit heb ik dan ook gedaan.
De accuracy score gaat richting 78% en de mean accuracy score naar 75%. Precision 81% en recall 90%. Alleen met een max_depth van 3 is de accurcay score 78%, want met een max depth van 12 gaat het weer richting de 60%. Het model scoort nog niet heel geweldig, maar het is nu niet meer hetzelfde als met een dobbelsteen gooien.
Mean accuracy score 75%, training test score 78%, test set score, 72%. Er is nu veel minder overfitting dan in mijn eerste inleverpoging.

Met de L1 (lasso regularization). Dit probeert ervoor te zorgen om veel parameters naar 0 te halen. Door hiermee te spelen gaat de accuracy score richting 82%.

Uiteindelijk door nog iets meer te spelen met de hoeveelheid features, en de regularisatie, heb ik een accuracy gekregen van 84%, mean accuracy score 84%, training set score 84%, test set score 88% met de decision tree.