

Data Domain / Owner Assignment Prediction

Problem situation

The assignment of a Critical Data Element to the correct Domain can be a highly subjective, complex process. This knowledge requires SME knowledge of the Business Data Model, cross-Divisional dependencies and Group standards and guidance principles.

Solution

Solution to provide a ML driven prediction for CDE assignment to a Domain and associated Data Owner based on the description in Collibra. Leveraging pre-existing approved CDE/Domain mappings to train clustering model.

The prototype maps text with labelled data into vector space through a standard language model. The domain is predicted through using and training a standard classifier (e.g. RF/XGB) on the data.

This model is refined through applying lines of questioning applied by the BDM team to improve the accuracy of clustering techniques. This is refined further through logic detailed in Group standards.

We can further tune this algorithm to only make suggestions above certain confidence threshold (e.g. >90%) and this will increase accuracy of the predictions.

