**Technology Category Recommendation**

**Business Objective:**

Within a technology category, give a rank order list of all the companies in the database who have not used any technology within the subcategory but might start using it in the near future.

**ML Objective:**

Find the propensity for a company to start using any technology within a subcategory based on its previous technology purchase pattern.

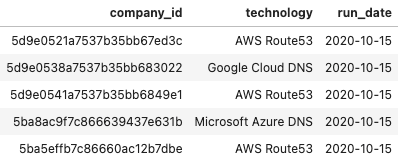
Based on this propensity score we can rank order the companies within a technology subcategory.

**Data:**

There are scheduled processes that run at a regular interval to detect the technologies used by the companies.

We had taken the data from 08/06/2020 to 10/02/2021.

Sample Table:

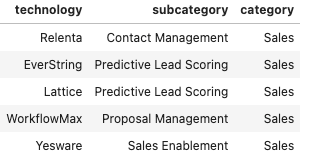


Total Companies: 12410952

Total Technologies: 28375

There is a separate table that keeps a mapping of technology to its category and subcategory.

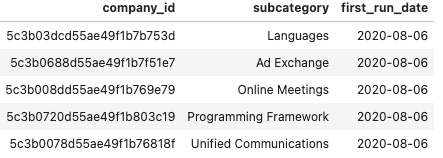
Sample:



By joining the above 2 tables we create a new table that maps a company to a subcategory.

And the min of run-date of the company-subcategory pair gives us the first date on which the company-subcategory pair was detected

Sample:



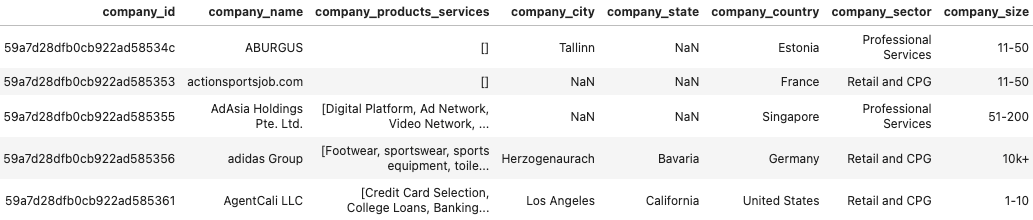
Total Subcategories: 806

We are assuming that the first run date of the company-subcategory pair tells us the first time a company started using any technology in that subcategory.

The above table gives us the training data that we can use to start training the model.

We had another table that tells us the company's intrinsic features like location, sector, and size information that can be used in training.

Sample:



**Data Cleaning:**

There were a lot of subcategories that meant the same but the spellings were different.

We used edit distance to get similarity scores between each pair of subcategory and combined the subcategories which have high similarity score and has the same meaning.

Sample of top subcategory pairs with high similarity scores:



**Data Filters:**

We have removed any companies that have used less than 3 subcategories as it is not enough history to learn meaningful features about the company. (Removes 20% of companies)

We have removed the subcategories which have not been used by at least 100 companies for similar reasons. (Removes 8% of subcategories)

Note: The threshold that has been used are hyperparameters, we have selected the parameters in a way that does not reduce the dataset size significantly.

Hyperparameter: Something that has to be tuned to get better model performance.

**Training:**

Now that we have our data ready we can go on with the training process.

But there is one major problem we have to resolve before we start the training.

From the dataset, we know if the company started using any technology in a subcategory (positive set), but we don't know if a company did-not/will-not use a subcategory.

In the dataset, we observed that out of all the connections which are possible between subcategories and companies only 1.5% of connections are actually present.

So if we pick a random pair of companies and subcategories, there is a 98.5% chance it is not connected.

That is what we are doing for the negative dataset, we are randomly picking a pair of companies and subcategories and assuming that it is not connected for more than 95% of the time.

Since we are using the data from October 2020 till Feb 2021.

We took data till January 2021 for training and Feb 2021 for validation.

And removed any company in validation data that are not part of train data.

We took 2 times more negative data than train using the logic mentioned above.

Using the above data we built a collaborative-based deep learning model to come up with the predictions.

Please refer to the below articles to get more details

<https://medium.com/quantyca/deep-learning-for-collaborative-filtering-using-fastai-b28e197ccd59>

<https://towardsdatascience.com/neural-collaborative-filtering-96cef1009401>