Ongoing work - Classification of scanner data into COICOP : a machine learning approach.

Measuring, understanding

A. Montbroussous and M. Monziols (Consummer Price Index Division)

Keywords — Classification, COICOP, Machine Learning

I. Data description and goal

i. Context

- Study on 1 hard discounter scanner data in 2023.
- 265 672 distinct EAN (European Article Number = barcode)
- 42 775 products (16%) representing almost 63% of the expenditure can be classified into COICOP using our current process for scanner data :

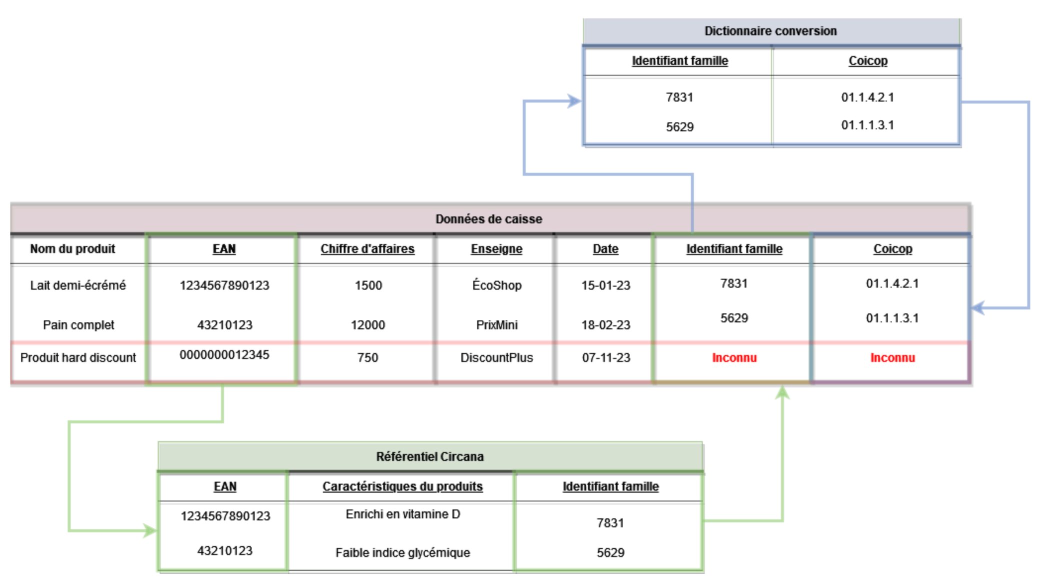

Figure 1: Process to classify scanner data with the dictionary from our external provider

- 11 072 products among the 42 775 are in reality assigned a custom item "99.9.9.9.9 : unfollowed", they represent 13.9% of the total expenditure
- We will try to classify the unclassified data into COICOP using its label.

ii. Data cleaning process

- EAN cleaning: convert to 8 or 13 digit numbers by adding or removing digits. If a label is shared by several EAN, we regroup them.
- Label cleaning: convert to ASCII, remove stopwords (le, la ..), lemmatization and construction of indicators (500g -> #WEIGHT).

II. Model and methodology

i. FastText

- Short training time
- Designed to handle noisy texts, including spelling errors
- Interesting performance compared to other state-of-art methods
- Gives a list of possible classification with a probability for each one :

$$p(C_k \mid x) = \frac{p(x \mid C_k)p(C_k)}{\sum_{i=1}^{K} p(x \mid C_i)p(C_i)} = \frac{e^{a_k(x)}}{\sum_{i=1}^{K} e^{a_i(x)}} \quad (1)$$

with :

$$a_k(x) = log(p(x \mid C_k)p(C_k)) = log(p(C_k \mid x)), \forall k \in \{1, .., K\} \quad (2)$$

ii. Model Training

- Training and test sample following a 80%/20% random partition
- The model was trained to predict a 6 digit COICOP and only at this level.

iii. Unlabeled data description and sampling

- 220 000 unlabeled products are too much to classify manually.
- Our strategy was to stratify according to the following two variables in order to minimize variance in each stratum:
  - the amount of expenditure the product represents
  - an indicator of the **confidence of the model in its prediction (the difference of the two best prediction probabilities)**.

The expenditure share of scanner data correctly classified by the model is the variable of interest:

$$R = \frac{\sum_{k \in U} CA_k \times z_k}{\sum_{k \in U} CA_k} \quad (3)$$

where:

- k is an article.
- U represents the sampling universe of the articles (represented by their EAN) sold during the year 2023.
- CA_k is the cumulative expenditure of the article k in our scanner data in 2023.
- $z_k \in \{0, 1\}$ whether the EAN is classified into the right COICOP item (level to be defined) or not.

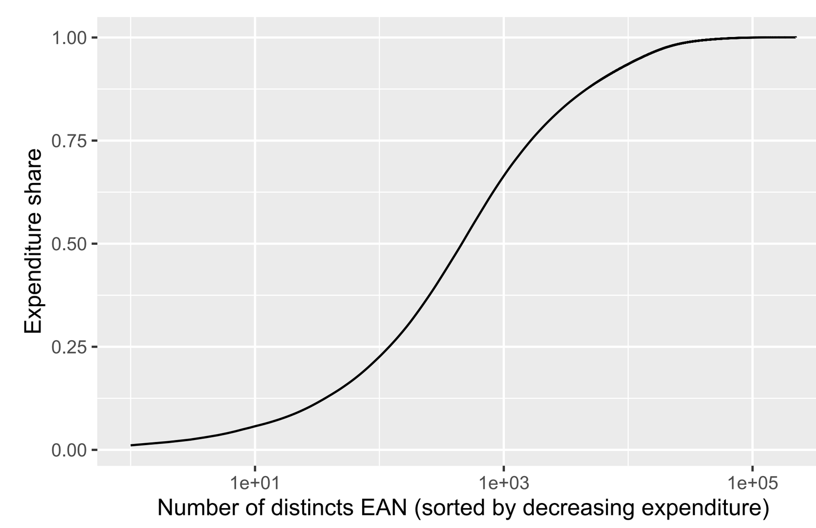

Figure 2: Expenditure share according to the number of observation.

With the given samples for each stratum S_h, we defined our estimator of R for the total sample S as

$$\hat{R} = \sum_h W_h \hat{R}_h \quad (4)$$

with :

- $W_h = \frac{\sum_{k \in U_h} CA_k}{\sum_{k \in U} CA_k}$ the share of expenditure of the stratum h
- $\widehat{R}_h = \frac{\sum_{k \in S_h} CA_k z_k}{\sum_{k \in S_h} CA_k}$ the estimator of well predicted expenditure share in the stratum h

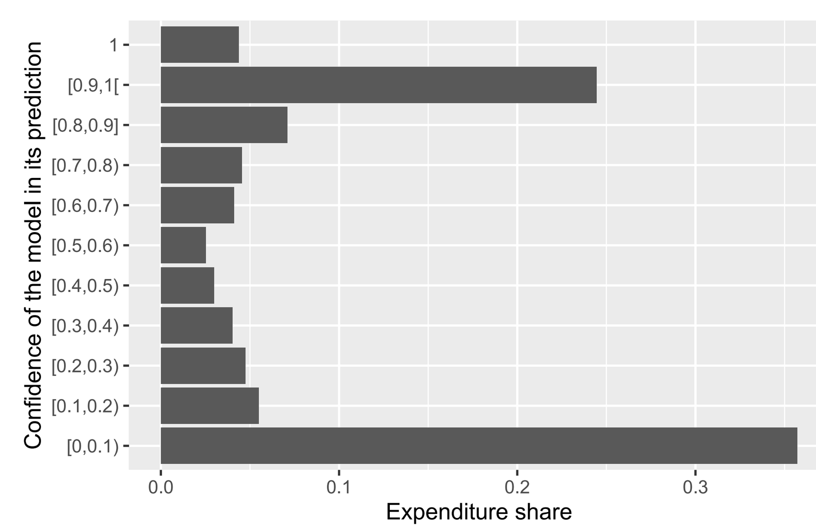

Figure 3: Distribution of expenditure according to the model confidence

Our resources allowed us to classify 3 000 products:.

expenditure	Confidence of the model prediction	Sample size	Number of EAN in the stratum	Sampling ratio
[0,5e+04)	[0,0.1)	177	71429	0.25 %
[0,5e+04)	[0.1,0.9)	116	75665	0.15 %
[0,5e+04)	[0.9,1]	104	61097	0.17 %
[5e+04,2e+06)	[0,0.1)	844	4494	18.78 %
[5e+04,2e+06)	[0.1,0.9)	651	2489	26.16 %
[5e+04,2e+06)	[0.9,1]	394	1556	25.32 %
[2e+06,7.32e+07]	[0,0.1)	209	209	100 %
[2e+06,7.32e+07]	[0.1,0.9)	266	266	100 %
[2e+06,7.32e+07]	[0.9,1]	240	240	100 %

Table 1: Sample Distribution after the allocation of the 3 000 products to manualy label according to each stratum expenditure

iv. Manual classification in practice

- using **LabelStudio**, see Figure 4
- 10 classifiers (200-400 product each)
- Only 1 classifier per product
- Manually assign a COICOP 6 digit item

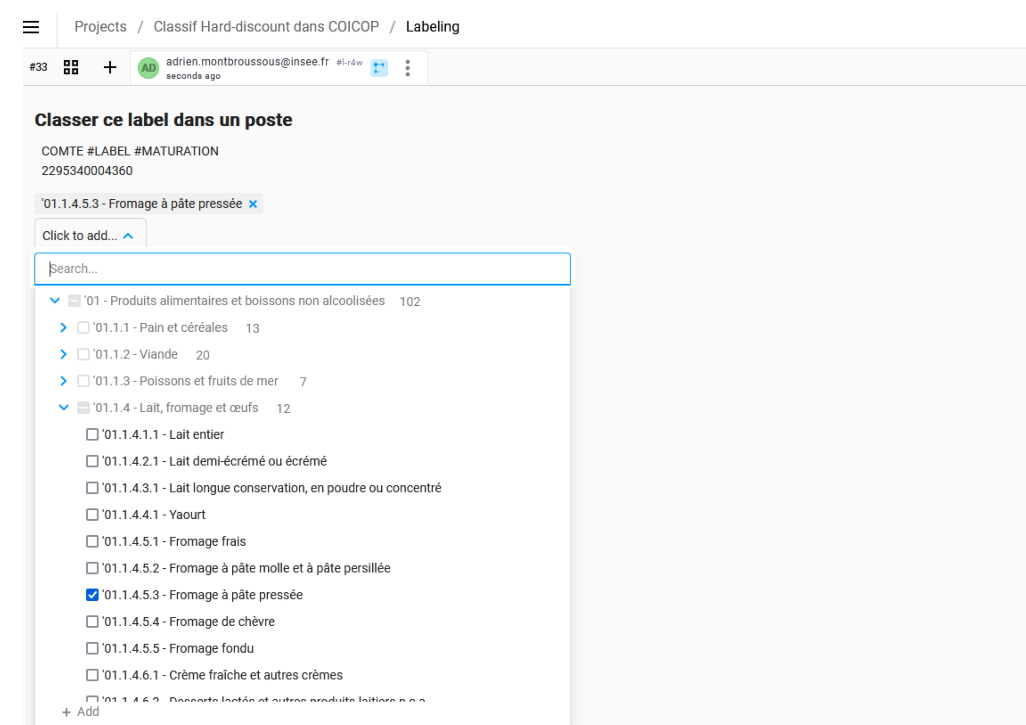

Figure 4: Label Studio screenshot with the use of a taxonomy for COICOP

III. Results

i. Test sample

Confidence of the model prediction	Share of expenditure well classified	Share of observation well classified
[0,0.1)	36.6 %	47.5 %
[0.1,0.9)	87.52 %	81.89 %
[0.9,1]	98.84 %	98.92 %
TOTAL	97.37	96.58 %

Table 2: Share of expenditure well predicted for the test sample at the COICOP 6 digit level, including the classified into 99.9.9.9.9

ii. Unlabeled data

Confidence of the model prediction	Share of expenditure	COICOP 6 Digit level	COICOP 5 Digit level	COICOP 4 Digit level	COICOP 2 Digit level
[0,0.1)	34.36 %	15.6 ± 2.35 %	30.69 ± 2.41 %	56.48 ± 2.48 %	62.64 ± 3.39 %
[0.1,0.9)	38.74 %	50.05 ± 2.79 %	73.91 ± 3.18 %	91.6 ± 3.1 %	93.56 ± 1.05 %
[0.9,1]	26.9 %	72.36 ± 2.49 %	85.11 ± 3.99 %	97.75 ± 4.18 %	98.73 ± 0.79 %
TOTAL	100%	41.44 ± 1.51 %	61.7 ± 1.83 %	80.83 ± 1.86 %	82.39 ± 1.42 %

Table 3: Share of expenditure well predicted for the unlabeled data (without observation classified into '99.9.9.9.9 - unfollowed') according to the COICOP level checked at the confidence interval of 95%

The gaps between the share of expenditure well predicted at 4, 5 and 6 digit level are quite important. The highest the confidence of the model in its prediction, the better the expenditure is classified.

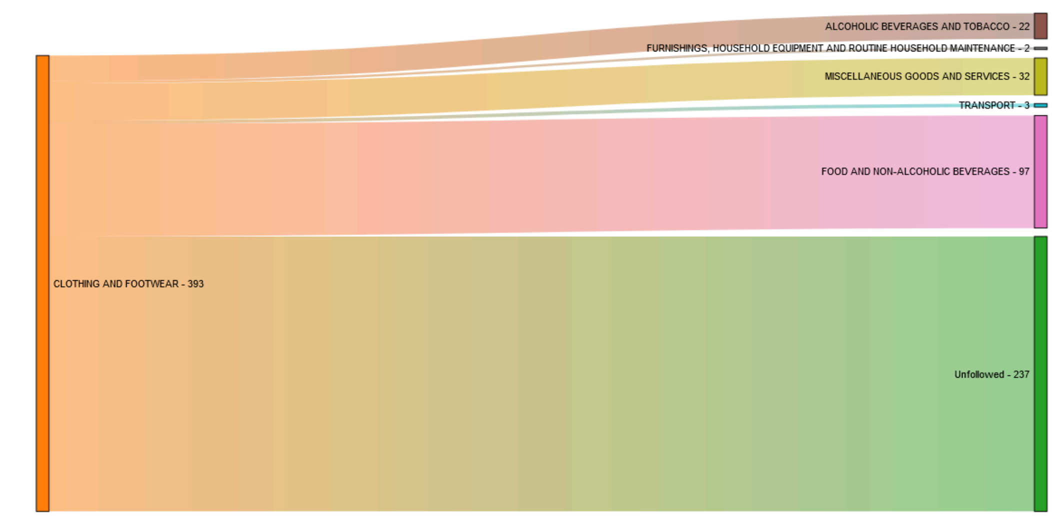

Figure 5: Classification of '05 - Clothing and footwear'

- Clothing is not followed in our current scanner data
- An important number of articles manually classified into clothing aren't classified into "unfollowed" by the model

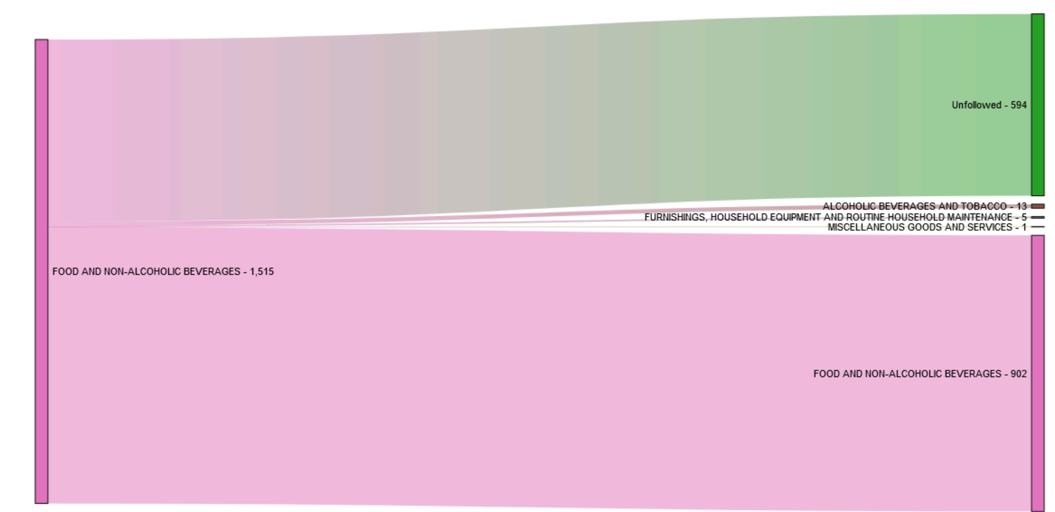

Figure 6: Classification of '01 - Food and non alchoolic beverages'

- " 01 - FOOD AND NON-ALCOHOLIC BEVERAGES " products are mostly classified into the right division or unfollowed.(Figure 6)
- Some products are wrongly classified into the division " 01 - FOOD AND NON-ALCOHOLIC BEVERAGES " (Figure 5)

IV. CONCLUSION

i. About the results

- Global results at our target level are not satisfying at this stage: only a bit more than 40% of expenditure is well-classified.
- The gaps between the performance of the classification at 4, 5 and 6 digit level are quite important
- Hard to expect that effect on indexes is not huge given the misclassifying rate.

ii. About our prediction strategy

- A finer definition of the classifying rules based on the product dictionary could lead to better prediction
- Removing impossible to classify products (like products labeled "non food") at the beginning could help the model.
- Certain label cleaning steps are counterproductive : the gender written in some clothing products is replaced by a "#gender" tag which does not allow us to classify in the right 6 digit COICOP code.
- Manual classification before training the model could be useful

iii. About labelisation strategy

- Developing knowledge of the nomenclature is necessary to be efficient and precise in the manual verification
- Double annotation could be useful to identify easy to classify products and hard ones.
- Issues on specifics articles have to be analyzed (fish, meat, wine...)