CLASSIFYING ALTERNATIVE DATA FOR CONSUMER PRICES STATISTICS: METHODS AND BEST PRACTICES

SESSION 1

STARTING OUT WITH CLASSIFICATION: FROM THE BUSINESS PROCESS TO COMMON INITIAL METHODS

Workshop on Scanner Data (2023 June, 8) within the Meeting of the Group of Experts on Consumer Price Indices
Presented by the
UN Task Team on Scanner Data

Topics covered in Session 1

- Overview of classification purpose
- Overview of methods NSOs tend to use
- Pre-conditions to classification
- Deciding on appropriate classification method to use
- Discussion of initial methods:
  - Method 0 (Manual labelling or validation of predicted labels)
  - Method 1 (Attribute based classification method
  - Method 2 (Pattern matching classification method
  - Method 3 (Recommendation / Machine-assisted classification)

Overview of classification purpose

- The goal of classification is to assign each unique product and its prices to a taxonomy category utilised by the NSO for aggregation, thus preparing the data source for price index compilation
- This classification category utilized by the NSO as a stratification variable in the aggregation step
- There are two types of classifications NSOs need to handle:
  - Initial - when preparing to integrate a new data source into the CPI or to support the research process – need to classify a large set of products
  - Recurrent - once the dataset is in production – need to classify all new products
- We found classification can be outsourced, but NSOs tend to do them in-house

Overview of 5 common methods NSOs tend to use

Focus for session 1

- Method 0: Manual labelling or validation of predicted labels
- Method 1: Attribute based classification method
- Method 2: Pattern matching classification method
- Method 3: Recommendation / Machine-assisted classification
- Method 4: Machine Learning classification method

Focus for session 2

- Blending classification methods

Typical variables utilized

Variable	Definition	Importance for classification
Product Name	Title of the product being sold	High
Retailer category / categories	The retailer specific category or set of hierarchical categories	High
Product identifier	The unique identifier for the product, ideally the GTIN/UPC, the first 2 codes of which are country, then 3-7 are manufacturer/producer specific.	High
Product description	Text description of the product	Medium
Product brand	Text description of the product	Medium
Product characteristics	Materials or other aspects about the product.	Low (excluding for electronics)

Pre-conditions to classification: Scaling the task

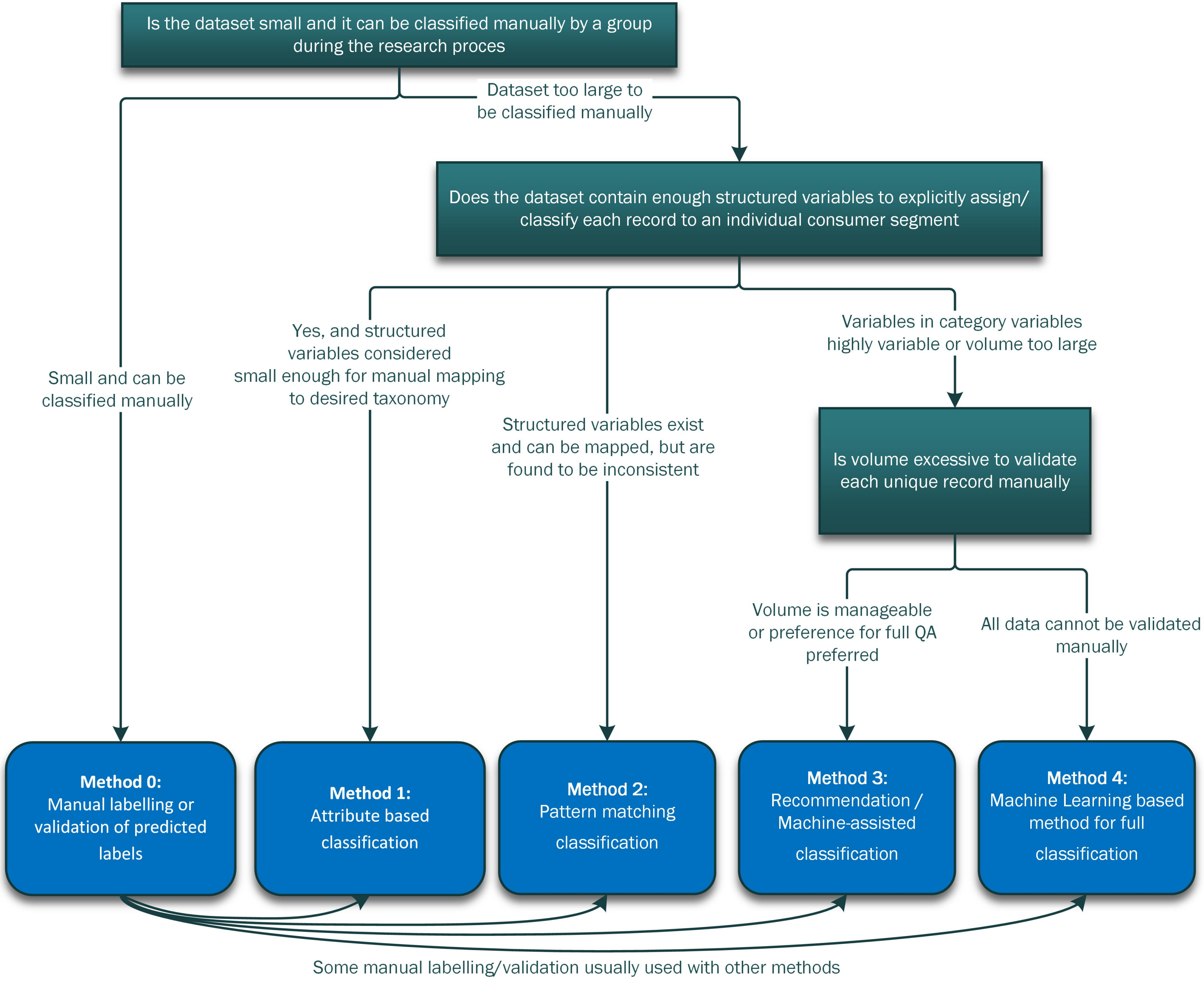
- Only unique products need to be classified, all prices (unit or offer) can be assigned to the classified product
- This may be large for initial classification, but could be manageable in production (on a recurrent basis)

	Store: Electronics	Store: Supermarket	Store: Clothing
# products stocked	300	40,000	20,000
Product churn rate	20%	2%	30%
# labels needed (initial month)	300	40,000	20,000
# labels needed (each month after)	60	800	6,000

Pre-conditions: Initial vs recurrent classification

	Initial classification	Recurrent classification
Time to complete task	• Months typically available • As process typically iterative, different levels of accuracy reached at different times	• Tightly bounded to production schedule of the CPI • Automatization a high priority for NSOs
Scale of task	• Scale depends on the size of dataset that needs to be labelled. Commonly, a year is used, however if multilateral methods with a 25 month window are being researched, a larger dataset may need to be labelled.	• Scale depends on the overall size of the retailer (i.e. how many products are sold in total) and the churn rate. As only new unique products need classification, task is much smaller.
The process usually applied to classify unique records	• Typically a representative time period necessary to study a price index is identified (12, or 13/25 months). • Typically an iterative research process to trial and select method.	• Production classification method was developed beforehand, applied on all new unique products. • It is typically combined with quality assessment and quality control methods to evaluate how well the method works.
How quality is maintained/checked	• Level of accuracy reached iteratively, effort continues until a desired (or achievable) quality level is reached. • Quality is usually tested with manual method (method 0) on a sample of the data.	• Key to check and maintain quality in production – a large investment is needed. • Method 0 (manual validation) often used to evaluate method performance
Supporting processes needed	NA	• Manual validation usually used to retrain the model • MLOps investments made to improve maturity of adoption of advanced (methods 3 and 4)

Pre-conditions: Deciding on appropriate classification method to use



Is the dataset small and it can be classified manually by a group during the research proces

Dataset too large to be classified manually

Does the dataset contain enough structured variables to explicitly assign/ classify each record to an individual consumer segment

Small and can be classified manually

Yes, and structured variables considered small enough for manual mapping to desired taxonomy

Structured variables exist and can be mapped, but are found to be inconsistent

Variables in category variables highly variable or volume too large

Is volume excessive to validate each unique record manually

Volume is manageable or preference for full QA preferred

All data cannot be validated manually

Method 0: Manual labelling or validation of predicted labels

Method 1: Attribute based classification

Method 2: Pattern matching classification

Method 3: Recommendation / Machine-assisted classification

Method 4: Machine Learning based method for full classification

Some manual labelling/validation usually used with other methods

Method 0: Manual labelling or validation of predicted labels

- Why use it:
  - Sample size is small or medium size tasks where simple rule based or keyword based methods (methods 1 and 2) would not work;
  - Utilized as part of manual annotation to support advanced methods (method 3 or method 4), or utilized as part of the recurrent validation process. The latter is usually combined with multiple outlier methods;
- How to apply it:
  - Preparation is key – need to create a clear set of instructions to make sure annotators are consistent;
  - It may be feasible to test annotator consistency by having more than one individual label the same record and test for what categories annotators agree at a high level (and where one annotator may be sufficient) and where they disagree (where multiple may be needed)
- Best practices and quality considerations:
  - Small tasks may be performed with spreadsheets, however longer and more consistent efforts could benefit from a ‘labelling environment’ (especially applicable if used in production to validate records).
  - Maintain quality by designing unambiguous and homogenous classes, having clear instructions, designing quality control processes (where escalation to a more experienced annotator when dealing with uncertain products), etc

Method 1: Attribute based classification method

- Why use it:
  - When data is highly structured, category variables in the retailer dataset is stable over time, and category variables not too numerous – categories can be assigned cleanly to one taxonomy category utilised by the NSO for aggregation (1:N or 1:1 mapping possible);
- How to apply it:
  - Create a mapping dataset (or concordance) between retailer category codes present in the data and taxonomy categories utilized by the NSO;
- Best practices and quality considerations:
  - Evaluation of scope of retailer categories key to validate that all products within retailer category could map cleanly to NSO taxonomy codes.
  - Maintenance of mapping dataset (concordance) is important once method used in production. Alerts should be set up when datasets received contain a new category (or a set) and to periodically validate that the scope of retailer categories remains consistent.

Method 2: Pattern matching classification method

- Why use it:
  - When Method 1 criteria is satisfied (data is highly structured, category variables in the retailer dataset is stable over time, and category variables not too numerous), however scope of retailer categories does not cleanly map to taxonomy categories utilized by the NSO;
  - Scope of retailer categories should be stable over time and can be cleanly and consistently stratified using specific keywords;
- How to apply it:
  - Similar to Method 1, first create a mapping dataset (or concordance) between retailer category codes present in the data and taxonomy categories utilized by the NSO, and then evaluate how to map products in appropriate retailer categories to multiple NSO categories:
    - Exclusively, by first evaluating all retailer categories, and then by deciding on decision boundaries within retailer categories and appropriate keywords to stratify products;
    - Inclusively, by manually labelling (Method 0) of a representative sample of products (of either all retailer products, or products in a specific category) to find the keywords that are appropriate to use of stratifying the retailer category.
  - Ex using SAS code:
    - `if prodgroup_retailer='45678'and index(upcase(product_descr), 'VEGAN') then coicop6=011464';`
    - `if prodgroup_retailer ='45678' and index(upcase(product_descr), 'ALPRO) then coicop6=011464';`
- Best practices and quality considerations:
  - In addition to best practices mentioned as part of Method 1, keywords might need frequent assessment and updates if the product environment changes;

Method 3: Recommendation / Machine-assisted classification

- Difference from previous methods:
  - Every product is still manually scrutinised (thus it is an extension of method 0), however the recommendation system supports the identification of most appropriate class by identifying a short list of options instead of focusing on one (hence it is separate from method 2, which should be confident enough in the choice to select a class);
- Why use it:
  - Most beneficial when the classification task is large enough that investing in machine-assistance techniques provides reasonable efficiency gains, but small enough that the associated manual work is still viable;
  - Could be used as an earlier phase when on the path to utilize method 4, as operating method 3 could create cost-effective way to create larger training datasets method 4 requires.
- How to apply it:
  - **Keyword-based recommender:** Utilize a product name, product description, and hierarchy to identify potential classes for an NSO officer to select from. Simple methodologically but resource intensive to maintain keyword-to-class; recommendation lists;
  - **Hierarchy mapping:** Utilize the hierarchy (similar to method 2) within the retailer categories to recommend multiple classes to an NSO officer to select from. Requires substantial number of recommendation classes
  - **Machine Learning:** Use a supervised Machine Learning Model (similar to method 4) but (1) maintain full manual validation, and (2) propose the top N recommendations from the model for the NSO officer to select from.
- Best practices and quality considerations:
  - As an extension of method 0, designing unambiguous and homogenous classes, having clear instructions, designing quality control processes valuable;
  - Depend on the type of recommendation method utilized.

Questions?

We hope to see you in Session 2, we will discuss:

ADVANCED CLASSIFICATION: APPLICATION OF MACHINE LEARNING

CLASSIFYING ALTERNATIVE DATA FOR CONSUMER PRICES STATISTICS: METHODS AND BEST PRACTICES

SESSION 2

ADVANCED CLASSIFICATION: APPLICATION OF MACHINE LEARNING

Workshop on Scanner Data (2023 June, 8) within the Meeting of the Group of Experts on Consumer Price Indices
Presented by the
UN Task Team on Scanner Data

Overview of classification purpose

- The goal of classification is to assign each unique product and its prices to a taxonomy category utilised by the NSO for aggregation, thus preparing the data source for price index compilation
- This classification category utilized by the NSO as a stratification variable in the aggregation step
- There are two types of classifications NSOs need to handle:
  - Initial - when preparing to integrate a new data source into the CPI or to support the research process – need to classify a large set of products
  - Recurrent - once the dataset is in production – need to classify all new products
- We found classification can be outsourced, but NSOs tend to do them in-house

Overview of 5 common methods NSOs tend to use

Focus for session 1

- Method 0: Manual labelling or validation of predicted labels
- Method 1: Attribute based classification method
- Method 2: Pattern matching classification method
- Method 3: Recommendation / Machine-assisted classification
- Method 4: Machine Learning classification method
- Blending classification methods

Focus for session 2

Topics covered in Session 2

- Situations where Machine Learning is typically chosen
- Overview of common steps taken when applying ML
- Approaches to handle class imbalance and evaluating model performance
- Recommendations on best practices and current topics of research/investigation by NSOs applying ML
  - Model retraining – how often to retrain
  - How to mitigate misclassification
  - How to automate the process (MLOps generally)

Situations where Machine Learning is typically chosen

- The vast amount of products can no longer be classified to COICOP or breakdowns thereof manually but only automatically.
- The classification might come from the data owner, at least to some extent.
  - Supermarkets, for example, have their own classification for scanner data which might be useful to this end.
  - The same holds true for web shops, where the products might be presented in a structured way.
- However, should this information not be available or sufficiently detailed for the purpose, one has to rely on supervised machine learning techniques.
- Yet, this requires the construction of a small labelled data set in order to train the algorithm.

Overview of common steps taken when applying ML

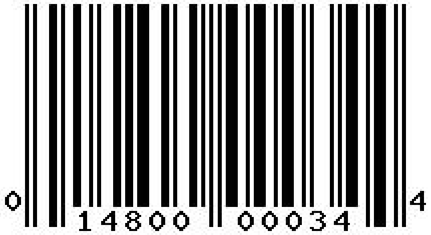
- In addition to information from the data owner, typically
  - product codes (such as GTINs),
  - descriptions (i.e., text), and
  - other metadata (e.g., size)

  are available.
- A major challenge in this respect is feature engineering.
- In most cases, product descriptions are not natural text but use specific vocabularies and rely on different kinds of shorthand.
  - This prevents the use of normalization techniques such as stemming or lemmatization. For example, trigrams can be exploited.
- Product codes, in general, follow some kind of a structure.
  - These can be treated as text strings and decomposed in prefixes.

Overview of common steps taken when applying ML

Example from the Dominick’s Finer Foods (DFF) data set
See Mehrhoff (2018), https://github.com/eurostat/dff

- DFF category: **bottled juice**
- UPC number: (*Universal Product Code*)
  - 0 is the number system digit
  - 14800 is the manufacturer code
  - 00034 is the product code
  - 4 is the check digit (not in the DFF data set)
- Product name: **Mott’s® 100% Original Apple Juice**
  - DFF description: **MOTTS REGULAR APPLE**
- Product size: **64 oz.** (≈ 1.89 l)



Overview of common steps taken when applying ML

Alternative: Term frequency of e.g., “juice” and variants thereof (“juic”, “jui”, “jce”, “ju”, “jc”, “j”)

- Bottled juice category: 118 in 511 products → 23 percent
  - Near misses: “LJ” (lemon juice), “juicy” (*Juicy Juice* also a brand)
- Frozen juices category: 52 in 175 products → 30 percent
  - Near misses: “OJ” (orange juice), “raspberryjc” (raspberry juice)
- 14 other categories: 19 in 6 896 products → 0.3 percent
  - Canned tuna: “clam juice” (in clam juice)
  - Toothbrushes: “J & J” (Johnson & Johnson)
  - Cheese: “Monterey J” (Monterey Jack)

Approaches to handle class imbalance and evaluating model performance

	Condition positive	Condition negative
Predicted condition positive	True positive	False positive
Predicted condition negative	False negative	True negative

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

$$\text{Accuracy} = \frac{\text{true positives} + \text{true negatives}}{\text{Number of observations}}$$

Approaches to handle class imbalance and evaluating model performance

	Class 1	**Class 2**	**Class 3**	***Total***
Predicted class 1	Hit 1s	Missed	Missed	*Predicted 1s*
Predicted class 2	Missed	Hit 2s	Missed	*Predicted 2s*
Predicted class 3	Missed	Missed	Hit 3s	*Predicted 3s*
Total	*True 1s*	*True 2s*	*True 3s*	*Number of observations*

$$\text{Precision}_k = \frac{\text{hit } k\text{s}}{\text{predicted } k\text{s}} \qquad \text{Recall}_k = \frac{\text{hit } k\text{s}}{\text{true } k\text{s}} \qquad F1_k = 2 \cdot \frac{\text{precision}_k \cdot \text{recall}_k}{\text{precision}_k + \text{recall}_k}$$

Approaches to handle class imbalance and evaluating model performance

	Class 1	Class 2	*Total*
Predicted class 1	9 644	252	*9 896*
Predicted class 2	23	81	*104*
Total	*9 667*	*333*	*10 000*

	Precision	Recall	F1 score	*Prevalence*
Class 1	0.9745	0.9976	0.9859	*0.9667*
Class 2	0.7788	0.2432	0.3707	*0.0333*
Prevalence-weighted average (unweighted)	*0.9680* *(0.8767)*	*0.9725* *(0.6204)*	*0.9655* *(0.6783)*	*Accuracy = 0.9725*

Approaches to handle class imbalance and evaluating model performance

- Since 96.67 percent of observations are in class 1, a simple but useless classifier that always predicts class 1, regardless of the features, will result in an accuracy rate of 96.67 percent. In other words, the trivial null classifier will achieve an accuracy rate that is only a bit lower.
- However, of the 333 observations in class 2, only 81 (or 24.32 percent) were hit. So, while the overall accuracy rate is high, the accuracy rate in class 2 is very low. Class-specific performance is important, and the terms precision and recall characterize the performance of a classifier.
  - Precision is the fraction of *predicted* class 2 observations that are correctly identified, i.e., 81 in 104, or 0.7788.
  - Recall is the fraction of *population* class 2 observations that are classified correctly, i.e., 81 in 333, or 0.2432.
  - F_1 score is the harmonic mean of precision and recall.

Recommendations on best practices and current topics of research/investigation by NSOs

- Model retraining – how often to retrain
  - Every month new products will appear and need to be classified as well. Already classified products should not be re-classified in this exercise to avoid revisions.
  - Model decay: The decreasing performance of the baseline model on new products is affecting the performance. The validated data from the initial time period is increasingly being diluted by the wrongly classified new products. This resulting error of all products directly impacts the CPI. This emphasizes the potential effect in the absence of a quality control process, including retrainin (Spackman et al 2023, Figure 10)
  - Model retraining: To address the observed model performance decay and mitigate the impact of low model performance on the elementary prices index, the model can be periodically retrained. With periodic refitting, the classification performance can be stabilized. More frequent retraining shows to be beneficial with less benefit on a short time horizon. This justifies a balancing of the costs of retraining with the improved performance over that horizon span (Spackman et al 2023, Figure 12)

Recommendations on best practices and current topics of research/investigation by NSOs

- How to mitigate misclassification
  - What matters most is how and which features are generated rather than the particular algorithm (as the old quip says “garbage in, garbage out”).
    - Generally, calculate the impact of different classifications on the price indices.
  - Automatic classification eventually needs to be assisted by human beings. As a rule of thumb, manually double-check the most important products, where importance is a combination of expenditure share (weight) and price change (volatility).
    - For example, investigate “near hits”, i.e., a subset of misses where the ground truth is in the top three, say, guessed classes (in terms of highest probabilities).
  - Exclude “near misses” from the performance measures, i.e., hits where the probability is below a certain threshold.
    - And, focus on classes with low performance, e.g., F_1 score.
  - Finally, be aware of the trade-off between the construction and maintenance of the algorithm and features and the performance of supervised machine learning.

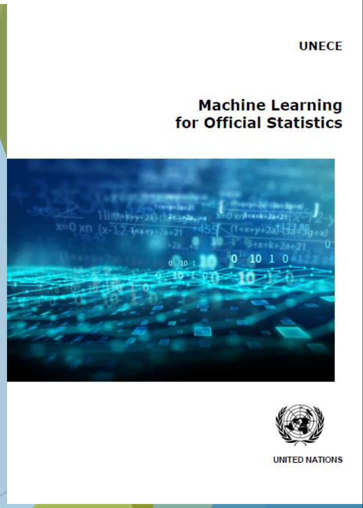
Recommendations on best practices and current topics of research/investigation by NSOs

- How to automate the process
  - Implementing a computer-assisted classification, i.e., automate the process but manually quality-control the model performance for important and/or uncertain products and classes.
    - Machine learning can give reasonable suggestions for the classification, but one must not trust the results blindly; it is no panacea. (Rhetoric questions: “How large should the training data set be?” & “Is 90 percent accuracy sufficient?”)
  - Assessing the quality of the classification over time, i.e., retrain the model.
    - If only the results of the algorithm itself are used in retraining, it reinforces itself over time – a feedback loop is created. The only way to bring “news” into the model is to also manually classify new products proportionally.
  - Focusing on building ML products, not only on developing and building ML models.
    - MLOps is a collection of principles and components that implement those principles (See From theory to practice: detecting model decay (or a journey to better understand MLOps), 2021)

UNECE literature on ML and MLOps from the HLG MOS

- General ML reference materials:
  - (2022) Recent publication from group: Machine Learning for Official Statistics
- MLOps related:
  - Choi, InKyung, Andrea del Monaco, Eleanor Law, Shaun Davies, Joni Karanka, Alison Baily, Riitta Piela, et al. 2022. "ML Model Monitoring and Re-training." ML 2022 Model Re-training Theme Group, UNECE.
  - Piela, Riitta. 2021. "From Theory to Practice: Detecting Model Decay (or a journey to better understanding of MLOps)." ONS-UNECE Machine Learning Group 2021 webinar.
  - Piela, Riitta. 2022. Work Stream 4 - Model Retraining. HLG MOS, Machine Learning Group 2021.
  - Piela, Riitta et al. 2021. Maintainging the Data Quality in ML development. Statistics Finland.
  - Del Monaco, Andrea. 2022. Model Retraining Theme Group. HLG MOS, Machine Learning Group 2022.
  - ML in Statistical Production process III: IT Infrastructure group. HLG MOS, Report from IT Infrastructure Group.

modernstats by HLG - MOS


UNECE

Machine Learning
for Official Statistics

UNITED NATIONS

Questions?

We welcome your thoughts and ideas as we finialize guidance and code on classification!

Feel free to reach out to:

Serge Goussev (Classification Workstream lead, UN Task Team),
serge.goussev@statcan.gc.ca