# Identifying and mitigating misclassification:

## A case study of the Machine Learning lifecycle in price indices with web-scraped clothing data

**Authors: William Spackman, Greg DeVilliers, Christian Ritter, Serge Goussev**

*Presented by: Serge Goussev*
*serge.goussev@statcan.gc.ca*

*Presented at the Meeting of the Group of Experts on Consumer Price Indices, 2023-06-08*

Delivering insight through data for a better Canada

# Research objective & problem statement

- Context:
  - NSOs shifting to Alternative Data Sources (ADS), scale leading to adoption of Machine Learning (ML) for classification
- Problem statement: misclassification is generally known to cause measurement error in statistics
  - Classification could impact price statistics if (a) enough product relatives that have a different movement affect the distribution of correctly classified price relatives; or (b) if enough product relatives that should be in a category are absent from it, affecting the distribution of remaining relatives
  - Misclassification may occur at one period, but could also build over time
  - Authors are unaware of a comprehensive discussion on the impacts of misclassification on price indices within the context of applying ML on ADS
- Objective: Study misclassification on key aspects of consideration when applying ML for production
  a) Look at data labelling (or annotation) – as labelled datasets used for ML model training or validation of data in production;
  b) Evaluate how misclassification could impact the elementary indices: the building blocks of the CPI;
  c) Evaluate ML model decay over time and how to mitigate it through model retraining;
  d) Evaluate outlier detection strategies to flag products for manual review in order to improve ML model performance

# Research questions

- RQ1: How can human annotator consistency or inconsistency guide NSOs in designing labelling or validation processes?
  - **Experiment**: 3 annotators independently label each unique product in dataset 1 (next slide). If there is any disagreement, a 4$^{th}$ sees all proposals and arbitrates the correct decision. Evaluate consistency between annotators, subjectivity in the annotation behaviour, and heterogeneity in the categories.
- RQ2: Can misclassification affect an elementary price index?
  - **Experiment 1**: Inject various levels of random misclassification into the data to see if an elementary prices index could be affected in one reporting period;
  - **Experiment 2**: Inject various levels of simulated misclassification (proxy of behaviour of real classifier) to see if a typical elementary index shows movement different than the correct value.
- RQ3: Does performance of ML classifiers decline due to dataset drift?
  - **Experiment**: Evaluate model decay and frequency of retraining appropriate to mitigate it
- RQ4: Which outlier detection methods are useful for NSOs to utilize to maintain classification performance?
  - **Experiment 1**: Evaluate confidence outlier method (likely impactful as it's an application of Active Learning) and compare it against random flagging method;
  - **Experiment 2**: Compare a method for flagging products in small categories, and a certain price range (trial various percentiles) as context for how many records are flagged and the level of F1 reached.

# Data and methods

- Data: One web-scraped dataset obtained from scraping seven Canadian Clothing and recreation retailers:
  - Subset 1: 19,569 unique product/retailer combination in four Clothing retailers were labelled to answer RQ1.
    - Subset spans June 2018 – Dec 2019
  - Subset 2: 155,254 unique product/retailer combinations and approximately 20m price observations from other additional Clothing and Recreation retailers – utilized to answer RQ2-4
    - Subset spans two periods;
      - Initial period of June 2018-Dec 2019 (14,309 annotated, ML model predicted remainder and 100% validated)
      - Second phase of Jan 2020-Dec 2021 (ML model predicted the whole and 100% validated)
- Methods:
  - Misclassification – used for RQ2:
    - Random (unbiased) – depictive of the concept, used on one period and one elementary index (jevons)
    - Non-random (simulated) – designed to scale the misclassification a real ML model contains by setting proportions of mistakes – and as a scale of overall misclassification is varied, the mistakes are assigned to the categories by this proportion
  - Price index method – used for RQ2:
    - GEKS-Jevons utilized as this method is preferred to bilateral methods and is used for unweighted web scrape data
  - Supervised ML model – used for RQ3 and 4:
    - As these research questions required retraining ML models we selected a representative one from the literature (and our experience): Support Vector Machine (SVM) classifier, word tokenization, custom stop word removal, and TF-IDF vectorization

## Results for RQ1 (How can human annotator consistency or inconsistency guide NSOs in designing labelling or validation processes?)

Takeaways:

- Fig. 1: Fleiss Kappa is high at 0.84 (level of agreement attained above the level that could be obtained by arbitrary annotation)
- Some subjectivity present, and some categories quite heterogenous.
- Expertise differed by annotator, expert annotators performed better. An average F1 for all 3 annotators was 0.845 on average, with 0.92 for expert annotators. Non-expert was still consistent for homogeneous categories.
- Fig. 2: Resources could allocated by category if needed.
  - Process could also be designed (on the whole dataset or the challenging categories) to leverage multiple annotators
  - For ex: 1 expert annotator for simple categories, 2 initial + 3rd for review appropriate for harder categories, or 3 initial + 4th.
  - Effort scales with level of robustness

Figure 1

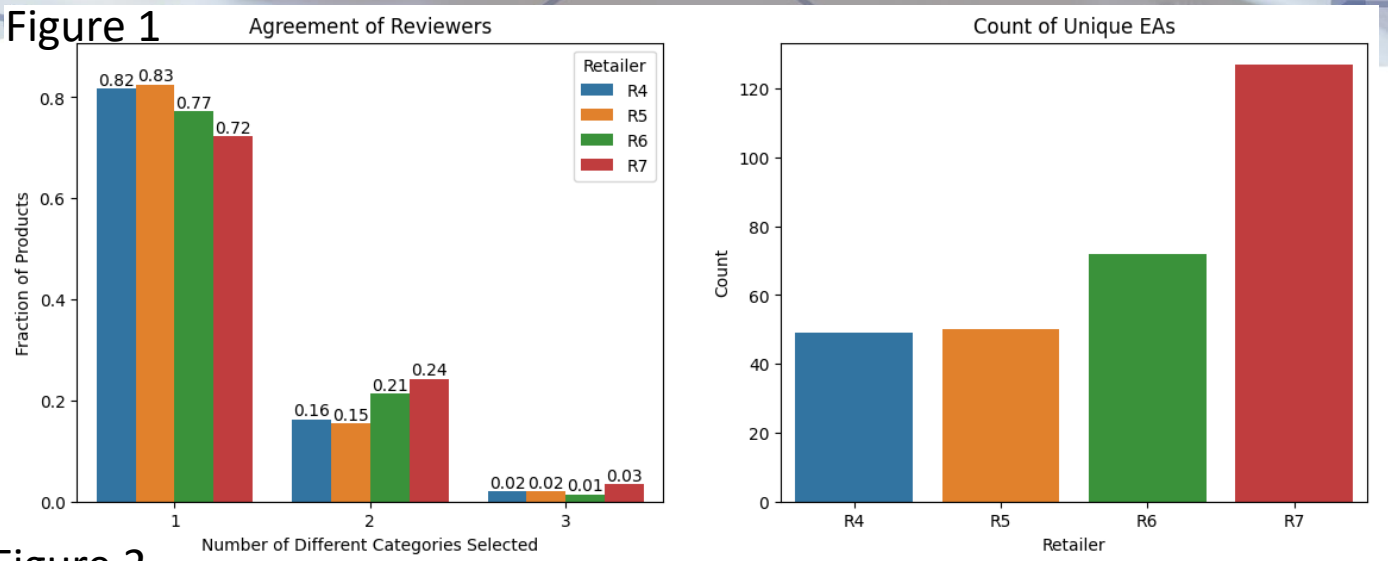
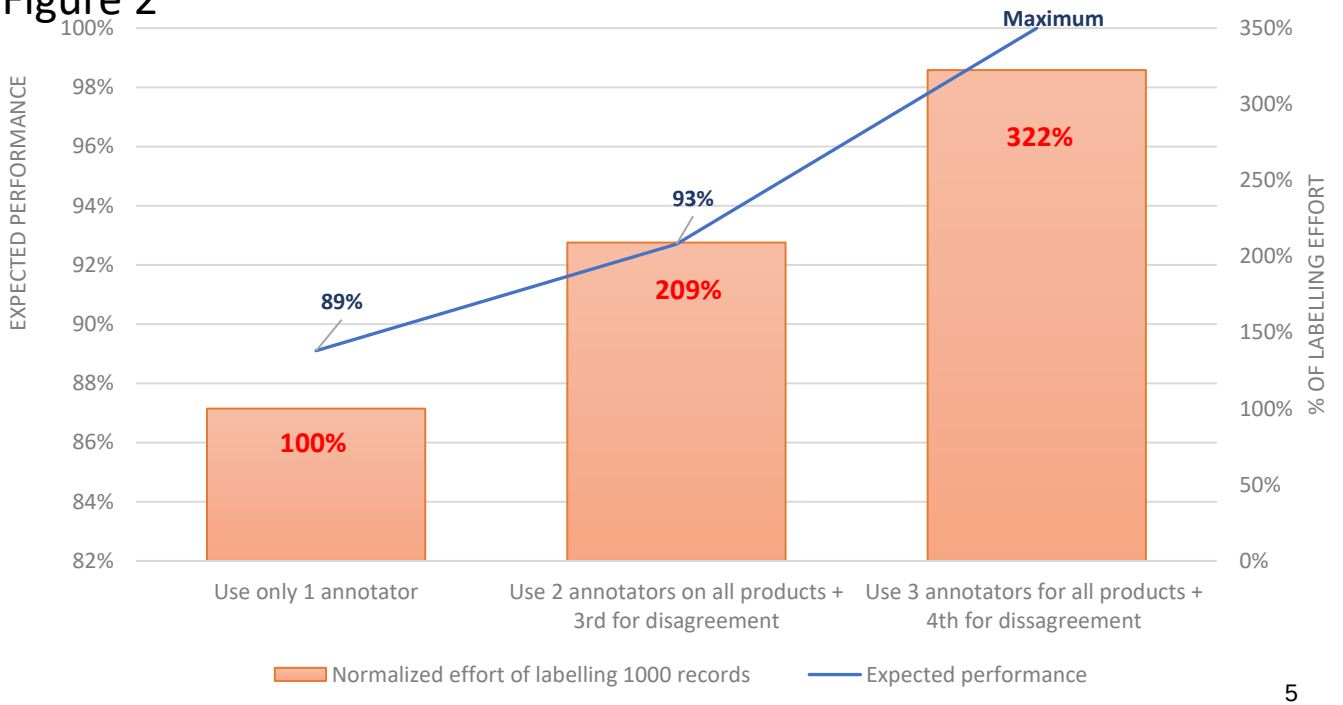



Figure 2



# Results for RQ2: Can misclassification affect an elementary price index? (1/2)

Takeaways:

- Fig. 1: Misclassification can cause bias and variance - various thresholds trialed
- Fig. 2: Fixing precision = 1 and varying recall (level of FNs) increases variance but does not look like its increasing bias
- Fig. 3: Fixing recall = 1 and varying precision (level of FPs) looks like it is increasing bias

Figure 1

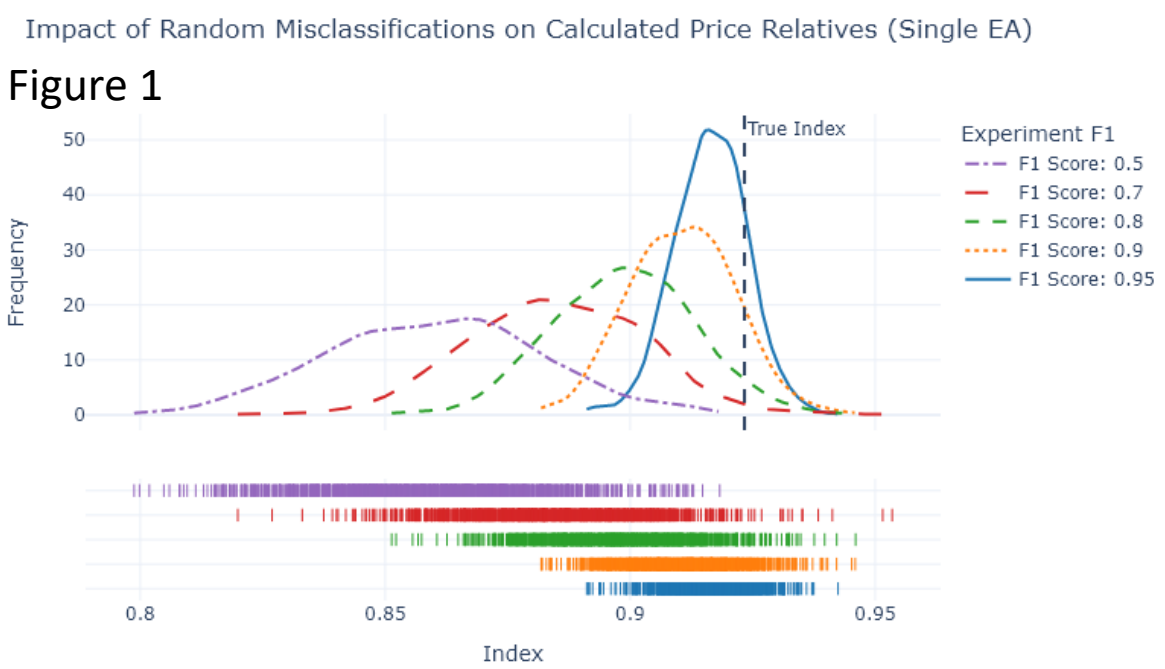



Figure 2

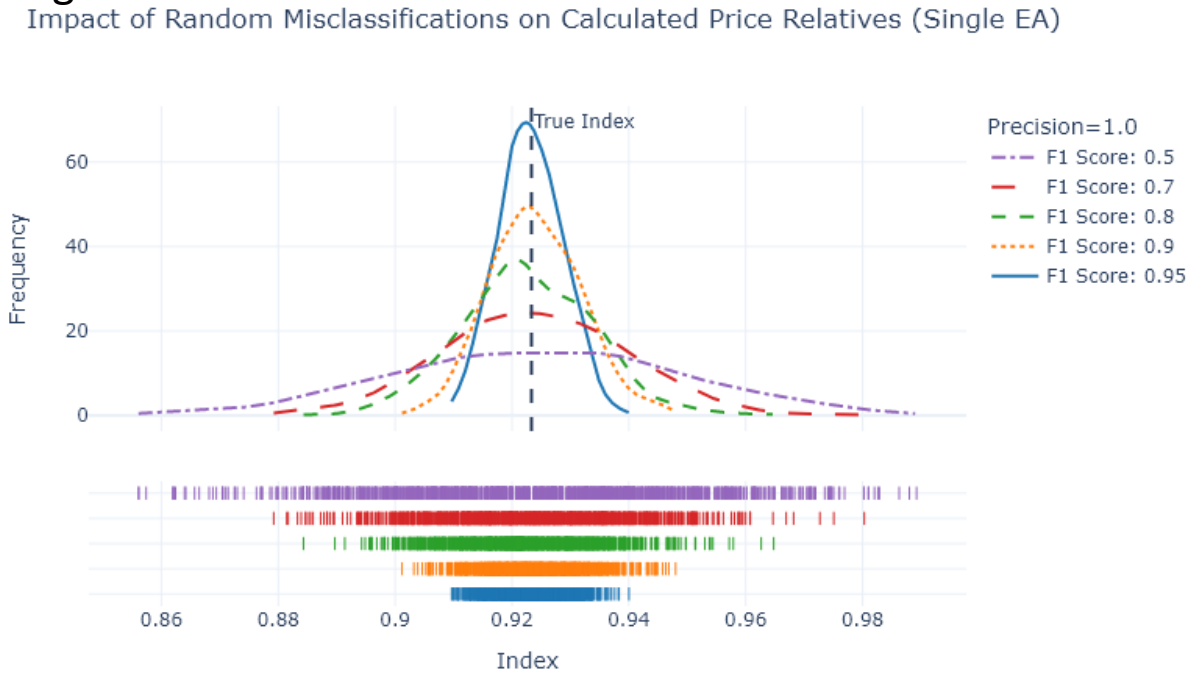
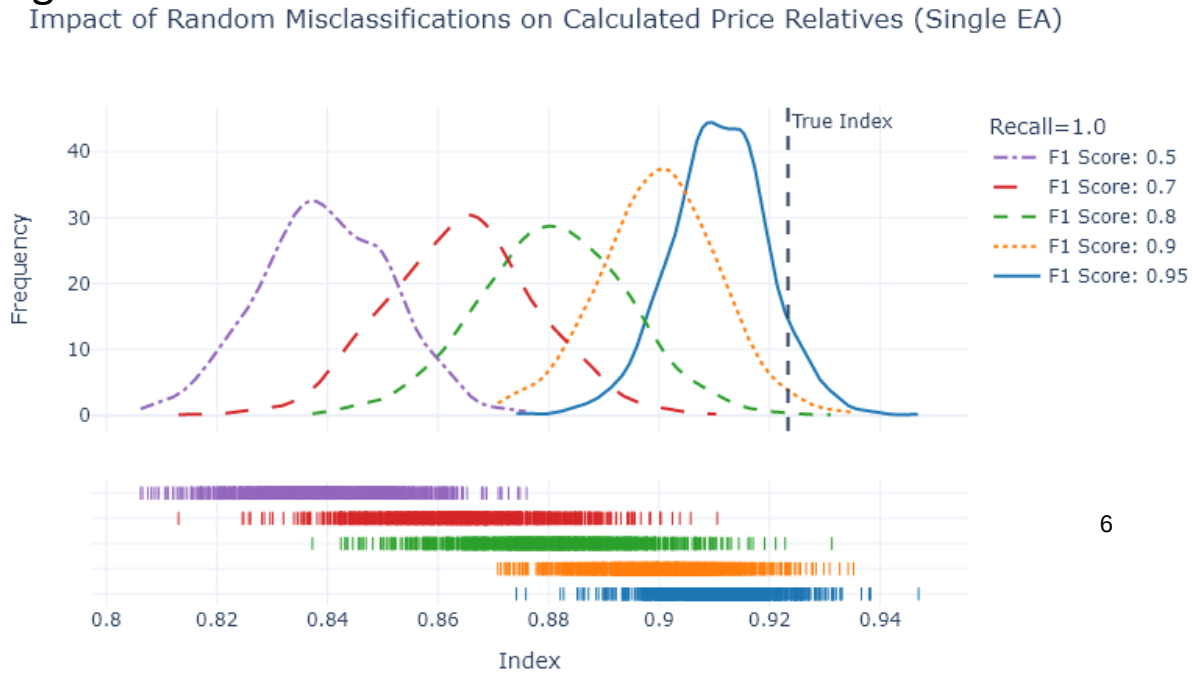



Figure 3



# Results for RQ2: Can misclassification affect an elementary price index? (1/2)

Takeaways:

- Fig. 1&3: GEKS-Jevons index tolerant to some misclassification, it still deviated from the expected (both 13 month window with extension and 25 month without extension)
- Fig. 2: At the same time the level of misclassification built up over time in the category.
- More investigation is needed, both with longer time periods and with other index methods

Figure 1

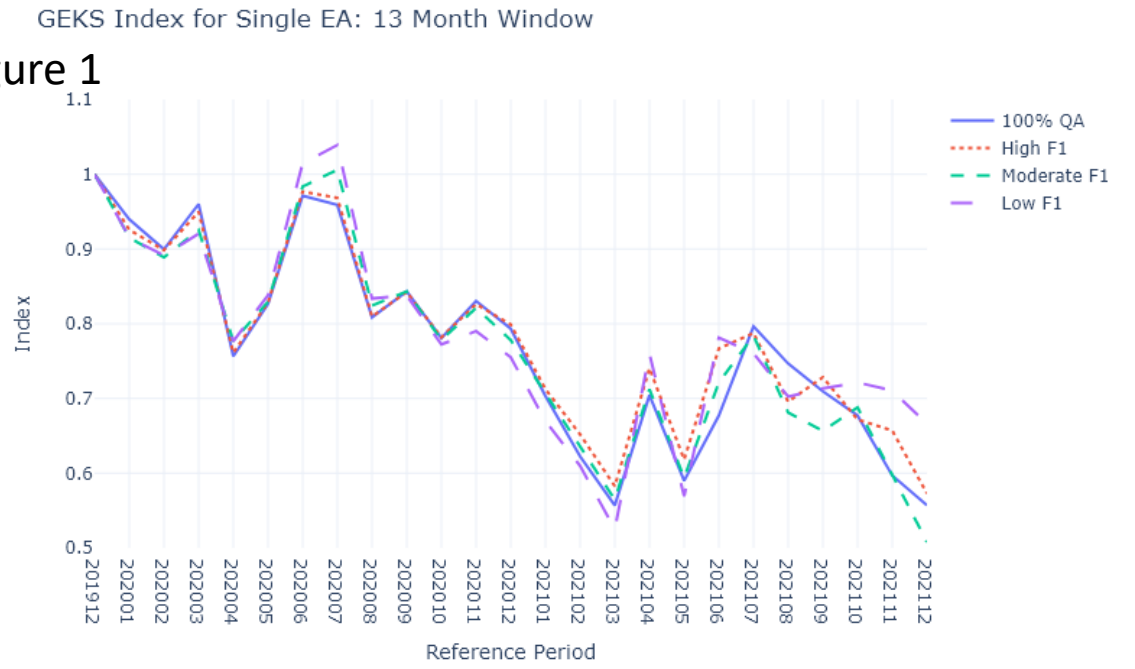



Figure 2

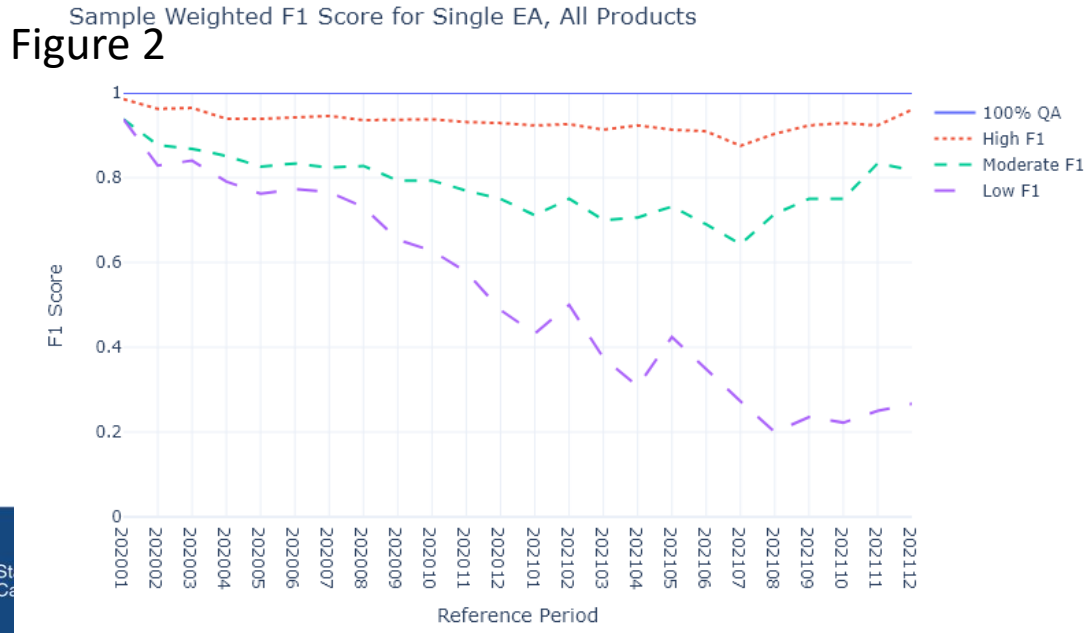
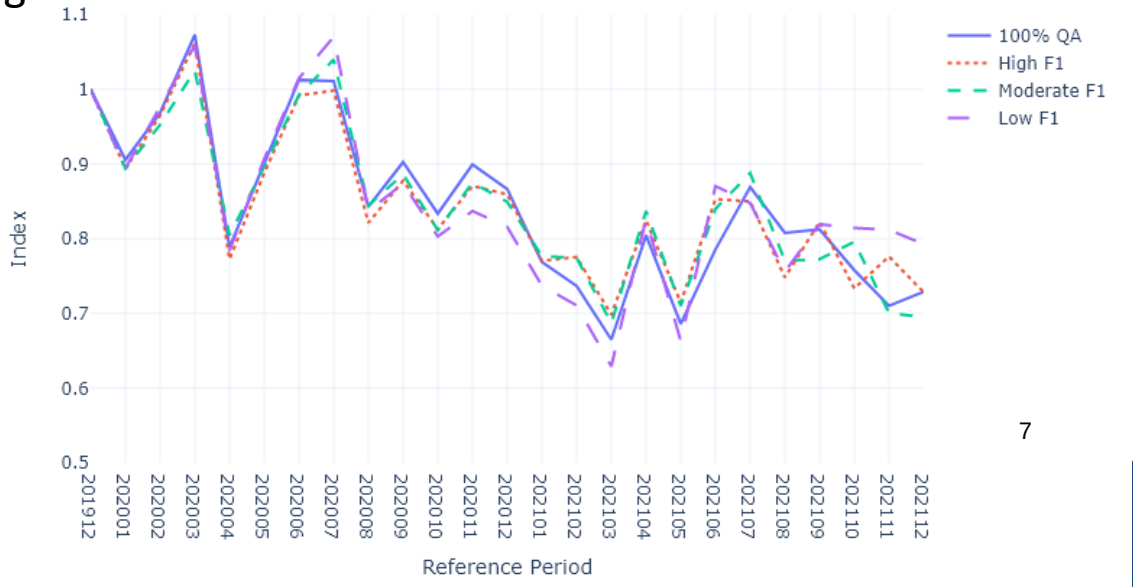



Figure 3



# Results for RQ3: Does performance of ML classifiers decline due to dataset drift?

Takeaways:

- Fig. 2: All 3 retailers showed model decay, although retailer 3 was less sensitive
- Fig. 2: Sudden shifts were seen in all.
- Fig. 3: New products entering the dataset over time showing compounding effects of increasing misclassification in the monthly sample
- Fig. 1 & 4: Retraining mitigated the situation – with a possible finding that retraining every 3 months seemed to be a practical choice

Figure 1

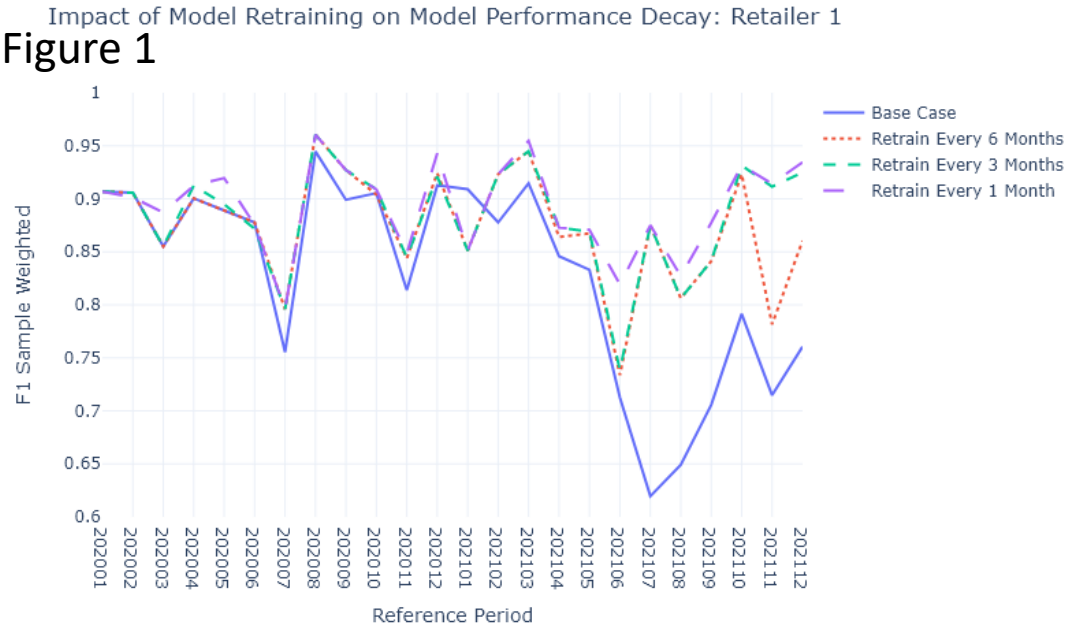



Figure 2

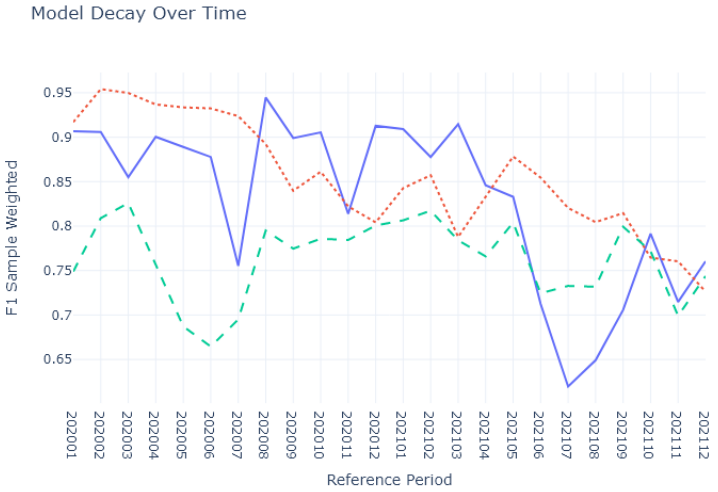



Figure 3

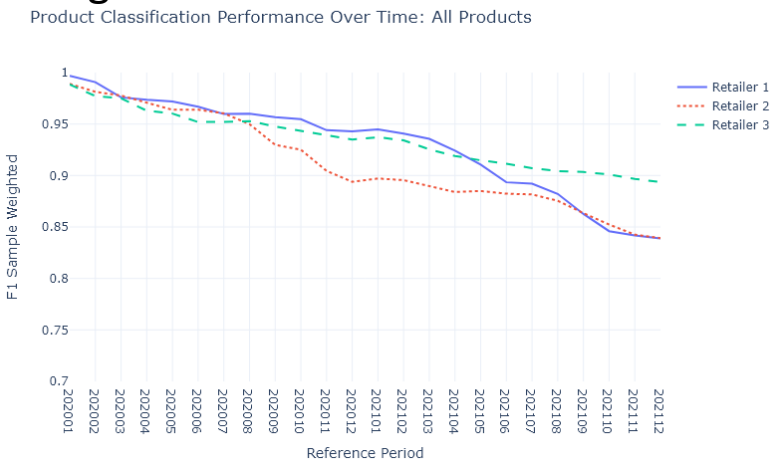
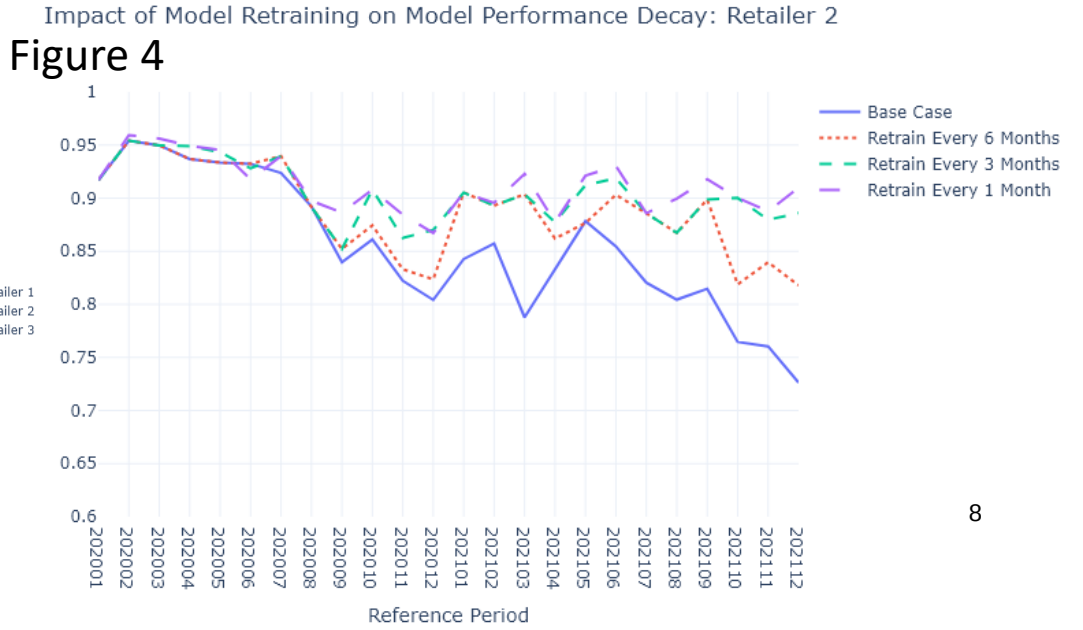



Figure 4



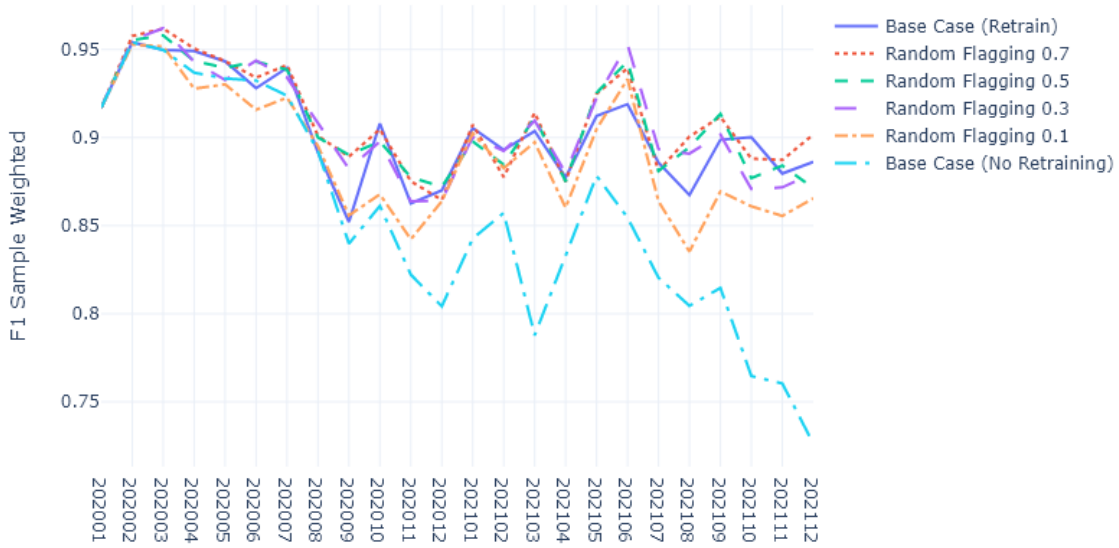
# Results for RQ4: Which outlier detection methods are useful for NSOs to utilize to maintain classification performance? (1/3)

Takeaways:

- Fig. 1: Even a small amount of random flagging (flagging a proportion of products for validation) is effective at bringing up classifier performance with retraining
- Random flagging not efficient at catching misclassified products
- Fig. 2: At same time, considering that there is a natural accumulation of new products that are entering the monthly sample (while some also leave), increasing levels of misclassification will enter the sample. The sample classification accuracy will approach that of the classifier.
  - Random flagging leads to an improvement of the overall sample that feeds the index.
  - The decline is smooth over time compared to the more pronounced monthly performance

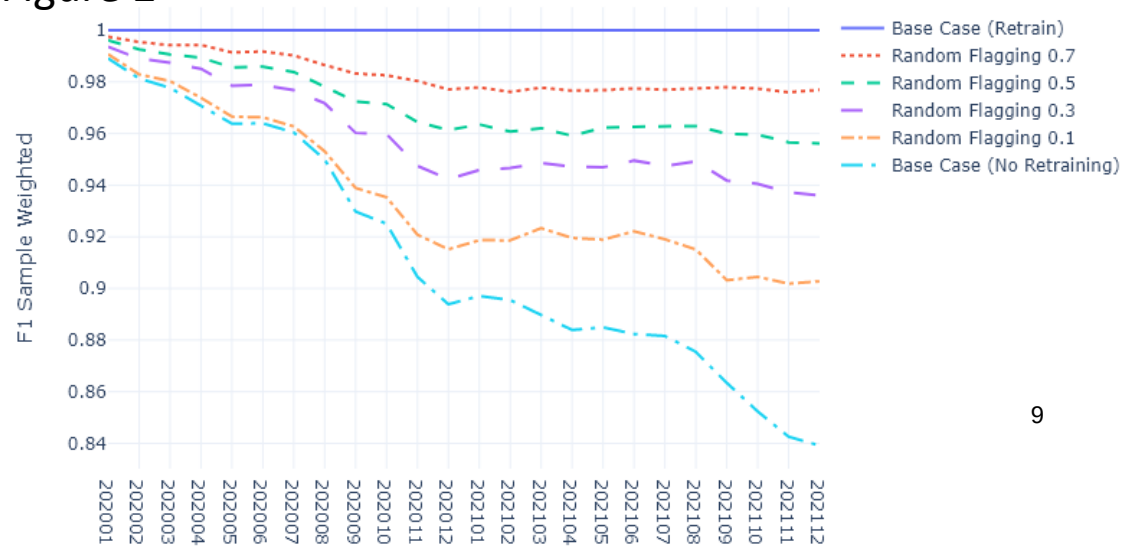
F1 Sample Weighted Comparison: New Products - Random Flagging

Figure 1




F1 Sample Weighted Comparison: All Products - Random Flagging

Figure 2



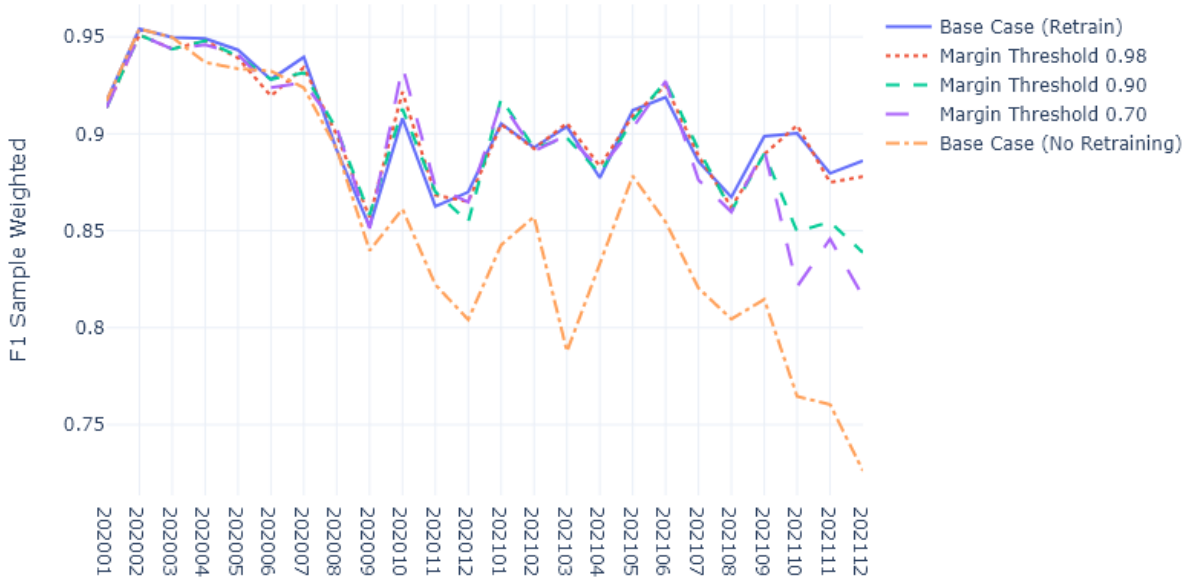
# Results for RQ4: Which outlier detection methods are useful for NSOs to utilize to maintain classification performance? (2/3)

Takeaways:

- Fig. 1: Confidence-based misclassification flagging (based on the margin threshold in the SVM we used) was efficient at creating retraining datasets to improve performance of the model
- Fig. 2: Confidence-based also efficient at having less misclassified products built up in the monthly sample.
- Confidence-based flagging also caught more mistakes compared to random-based.
- Choosing a lower threshold lowers the amount of products that need to be validated. This could be balanced with maintaining classifier performance

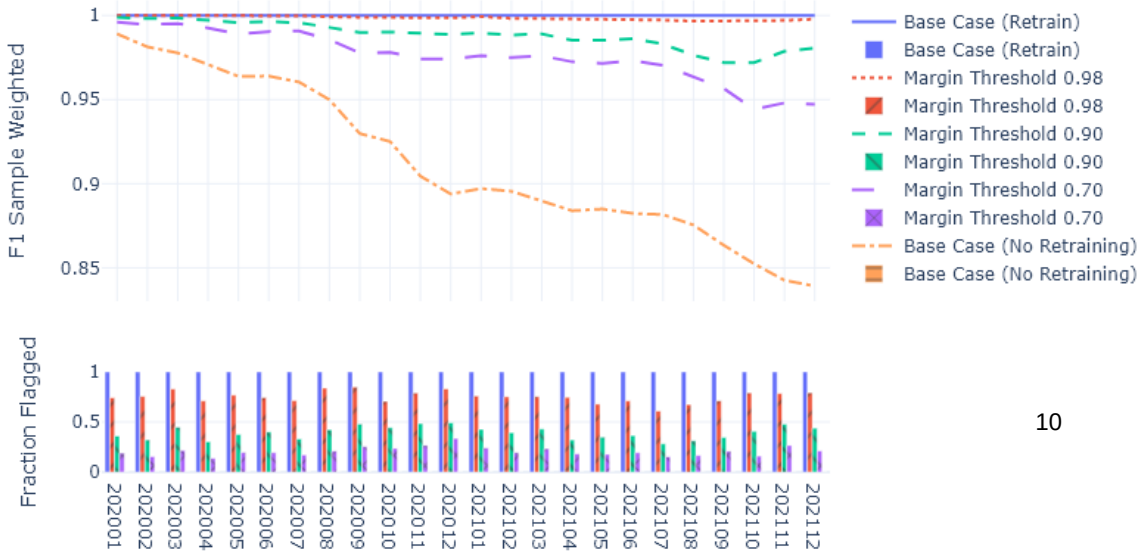
F1 Sample Weighted Comparison: New Products - Uncertainty Flagging

Figure 1




F1 Sample Weighted Comparison: All Products - Uncertainty Flagging

Figure 2



# Results for RQ4: Which outlier detection methods are useful for NSOs to utilize to maintain classification performance? (3/3)

Takeaways:

- Fig. 1: Comparing between random and uncertainty-based – for almost the same proportion of records flagged, uncertainty was more effective at bringing up classifier performance if the dataset was used for retraining models. This aligns with the confidence-based Active Learning method.
- Uncertainty-based flagging would create biased datasets however – and it is not recommended to use unbiased datasets to evaluate model performance. It is recommended to combine confidence-based flagging method with a certain threshold of random (stratified) for unbiased model evaluation
- Fig. 2: Other outlier methods (flagging all products in small categories (minimum number of products) or price outliers) are less likely to be useful for model retraining datasets, but would be necessary to minimize the impact of misclassification on the elementary index.
- Further research needed to design a global optimization process of model retraining and price index based outlier methods most appropriate

Figure 1

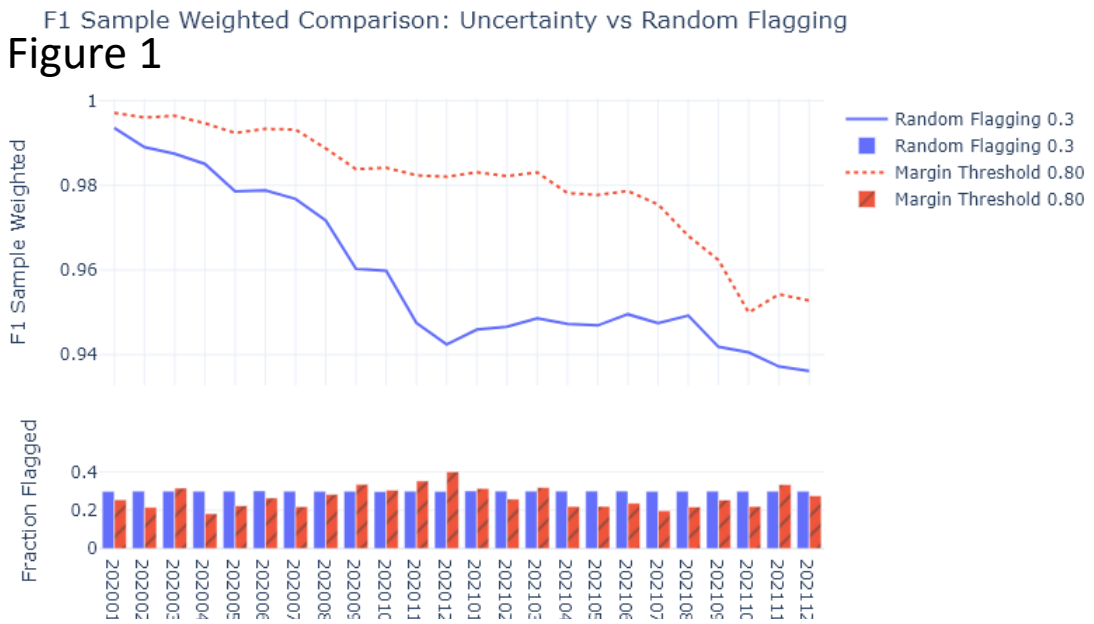



Figure 2

## Discussion & Conclusion

- Our empirical case study showed that misclassification is present in all key steps of the lifecycle of ML in price statistics and how it could be mitigated:
    1. Annotators disagree and robust processes must be designed to mitigate this. A ‘ceiling’ benchmark of ~92% is realistic based on our findings.
    2. Misclassification can affect an elementary aggregate – both bias and variance could enter the index in one representative reporting period. Misclassification could also build over time.
    3. Model decay is present, thus misclassification could grow over time if not addressed. Retraining utilizing the data from a validation process could mitigate decay by bringing performance of the model back up.
    4. Of several outlier methods for retraining available to NSOs – confidence-based method shown to be most useful for retraining models. However as confidence-based flagging results in a biased dataset, random flagging is recommended for evaluating model performance. Other flagging methods should be useful for mitigating impacts on the price index – such as all products in small EA categories or products with large price movements.

# Thank you!

Questions, feedback, ideas?

serge.goussev@statcan.gc.ca