Catalogue no. 11-522-x
ISSN: 1709-8211

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society's Emerging Needs

Machine Learning Classifier Evaluation Criteria: application to price statistics

by Serge Goussev and William Spackman

Release date: November 5, 2021

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society's Emerging Needs

Machine Learning Classifier Evaluation Criteria: application to price statistics

Serge Goussev, William Spackman[1]

Abstract

The modernization of price statistics by National Statistical Offices (NSO) such as Statistics Canada focuses on the adoption of alternative data sources that include the near-universe of all products sold in the country, a scale that requires machine learning classification of the data. The process of evaluating classifiers to select appropriate ones for production, as well as monitoring classifiers once in production, needs to be based on robust metrics to measure misclassification. As commonly utilized metrics, such as the Fβ-score may not take into account key aspects applicable to prices statistics in all cases, such as unequal importance of categories, a careful consideration of the metric space is necessary to select appropriate methods to evaluate classifiers. This working paper provides insight on the metric space applicable to price statistics and proposes an operational framework to evaluate and monitor classifiers, focusing specifically on the needs of the Canadian Consumer Prices Index and demonstrating discussed metrics using a publicly available dataset.

Key Words: Consumer price index; supervised classification; evaluation metrics; taxonomy

1. Introduction

Alternative data sources have become key for National Statistical Offices (NSO), helping augment traditional sources such as field collected data. To be implemented within official price statistics, such as the Consumer Price Index (CPI), high volume alternative sources such as transactional (scanner or point of sale technology) and online data (through scraping retailer websites) need to first be accurately classified to correct commodity classes using Machine Learning classifiers.[2] When developing classifiers, NSOs need to assess and select an appropriate classifier for production uses, as well as monitor its performance on a monthly basis (Eurostat 2017; Martindale, et al. 2020, 19). A question that NSOs face is what metrics are applicable to the evaluation and monitoring of classifiers in the price statistics space, as the Fβ-score and other commonly utilized metrics do not take into account several key aspects applicable to price statistics in all cases, most notably the hierarchical nature of the taxonomy to which all records need to be classified, and the unequal importance of some categories within the taxonomy over others. To help provide insight on the metric space available and contributing to the previous theoretical considerations on metric selection in price statistics (UK Statistics Authority 2019), this working paper investigates hierarchical evaluation metrics as well as other criteria applicable to price statistics to support NSOs in evaluating, selecting, and monitoring classifiers. The specific focus of the paper is on metrics applicable to multiclassification, as most NSO effort is directed towards classifying datasets that contain multiple consumption segments.

This paper is organized as follows: the first section provides an overview of business and machine learning objectives and applicable metrics to evaluate models. Next, an operational framework is proposed that combines applicable metrics for effective analysis and monitoring of deployed classifiers. Finally, an empirical demonstration is outlined, utilizing an open dataset to demonstrate the application of discussed metrics on a realistic use case.

[1]Serge Goussev, Statistics Canada, 150 Tunney's Pasture Driveway, Canada, K1A 0T6 (serge.goussev@statcan.gc.ca); William Spackman, Statistics Canada, 150 Tunney's Pasture Driveway, Canada, K1A 0T6 (william.spackman@statcan.gc.ca);

[2] See chapter "10: Scanner data" of the *Consumer Price Index Manual: Concepts and Methods* (International Labour Office 2020), for more information on the process of integrating alternative data such as scanner data into the Consumer Price Index as an example.

2. Objectives and proposed metrics

2.1 Classification quality metrics

Multiclassification of records in price statistics involves assigning each record of a dataset to the lowest or leaf level of a hierarchical tree-based taxonomy.[3] Within the CPI, classification is usually at the level of the Elementary Product (EP) category, or a set of relatively homogeneous goods or services with similar uses and price movements, the lowest level for which expenditure weights are available (International Labour Office 2020). Classification categories could also be lower depending on NSO needs and data, as some EP categories are known to be not-perfectly homogeneous and several alternatives are being explored (Office for National Statistics 2020). Once all records are classified, price indices are then calculated utilizing different price index formula.

Two types of classification approaches may be applicable to this type of problem, ones that include the consideration of the taxonomy and ones that do not (Silla and Freitas 2011). If misclassification needs to be penalized equally, a case where every EP category is assumed to have a price movement independent of other categories, the simplest case is applicable. In this case, researchers can ignore the class hierarchy and predict at the lowest or leaf node level, an approach that utilizes flat classification consistent with classic supervised machine learning methods (Silla and Freitas 2011; Costa, et al. 2007). Traditional evaluation metrics—such as precision, recall, and Fβ-score (with a selection of an appropriate β)—are applicable to this case and have been widely adopted by NSOs (Office for National Statistics 2021; Office for National Statistics 2020; Martindale, et al. 2020).[4]

Several approaches to Fβ-scores are important to consider in the application of the metric to the case of price statistics. Firstly, a consideration of each classifier's per-category Fβ-score is key to help differentiate classifiers that may perform well overall but tend to predict a number of categories badly. In price statistics this is further important as some categories are more important than others due to their higher relative importance in consumer expenditure. Secondly, per-category Fβ-scores can be aggregated into a one per-model Fβ-score to provide a general view of model performance. Three aggregation techniques are applicable: sample-weighted, equally-weighted, and expenditure-weighted. A sample-weighted Fβ-score is an average of the per-class Fβ-score, weighted by the number of records present in each class during evaluation, which inherently biases the metric towards a higher overall result with models that perform well on larger classes. Alternatively, a macro or equally-weighted Fβ-score can be utilized, which applies a simple arithmetic average of all per-category Fβ-scores without weighting. This guarantees equality among all classes, promoting models that perform well in all cases, even in classes with few data points. Finally, an expenditure-weighted Fβ-score can be utilized, based on a weighted average of all per-category Fβ-scores with weights taken from the CPI expenditure basket weights representing the relative importance of each category. This approach promotes models that perform well on classes that are relatively more important for the final use of the classified data, a case that is highly applicable to price statistics such as the CPI.

2.2 Hierarchical classification quality metrics

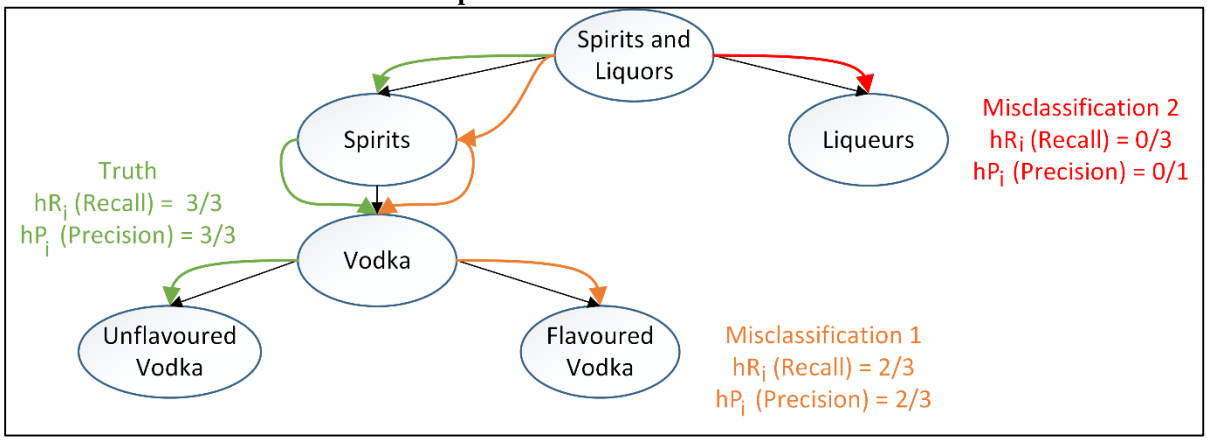
Inherent in standard performance measures is an assumption of independence between categories and an equal penalization of misclassification. This approach is not fully applicable to classification in the context of price statistics. Wrongly classifying a record to a similar category may be less problematic than classifying it to an unrelated one, as in this way the classifier more closely replicates human effort in making more mistakes on closely related categories. This mirrors real life application, as NSO efforts in manually labelling data have shown that maintaining consistency of classification is non-trivial for closely related categories, especially where a product lies on the boundary of two possible classes (Office for National Statistics 2020, 7). The hierarchical approach is also applicable if NSO researchers utilize custom non-flat, more complex taxonomy specific approaches to classification, such as the local classifier approach, or a global (big-bang) classifier approach. For local classifier approaches, a pipeline of sequential classifiers supports a decision tree of choices, where every record is classified through the set to get to the final lowest level; whereas for a global approach, a single and relatively complex classification model is trained on data taking the

[3] Common taxonomies for price statistics include international classifications of goods and services such as Classification of individual consumption by purpose (COICOP), North American Product Classification System (NAPCS), or specific NSO taxonomies such as CPICLS used to calculate the CPI at Statistics Canada.

[4] See (UK Statistics Authority 2019) for a detailed overview of these metrics.

whole class hierarchy into account (Silla and Freitas 2011). NSOs may find that these approaches are performant and applicable for cases where one alternative data provider covers a multitude of disparate consumption segments.[5]
Within hierarchical classification, a key consideration is the 'closeness' of categories in the penalization of misclassification (Sun and Lim 2001), with mistakes in closely related categories tracked as less of an issue than with distant categories. Within price statistics, taxonomies are designed in most cases to categorize similar categories under the same parent, hence misclassification of 'close' records will usually put them only one child edge away from each other. Depth measures, or focusing on the number of edges correctly predicted from the root node to the final leaf node for each record, are an applicable and interpretable way to quantify closeness for price statistics. A possible limitation of depth measures is on classification that can be stopped at the parent node without continuing to the lowest child node, a case that does not apply to price statistics as all records need to be classified to the lowest leaf level.
Similar to traditional classification metrics, three hierarchical metrics can be calculated: hP_i representing hierarchical precision, or the ratio of correctly predicted taxonomy edges to the total edges precited; hR_i representing hierarchical recall, or the ratio of correctly predicted taxonomy edges to the actual number of edges; and $hF\beta$ representing the hierarchical Fβ-score. Figure 1 demonstrates a visual application of these metrics, with three examples. Specifically, classifying a record correctly would yield recall and precision of 1, whereas misclassification penalty depends on the distance from the truth. Classifying Unflavoured Vodka to Flavoured Vodka utilizes two correct edges travelled in the hierarchy, hence resulting in a recall of 2/3 and a precision of 2/3. Classifying the record to Liqueurs results in a precision and recall of 0. Once a per-category Fβ-score is calculated using these hierarchical metrics, macro, sample, and basket weighted aggregated model specific Fβ-scores can also be determined as with classic classification metrics.

Figure 1
Hierarchical classification metrics example with one correct classification and two misclassifications



2.3 Business and process considerations

To support the initial selection of a performant classifier for production, other resource and business constraints need to be taken into account. Firstly, while most price statistics are calculated on a monthly or quarterly basis, data is received regularly and is classified, quality-assured, and aggregated on a strict schedule. While unlikely, models that incur a long prediction delay may limit time for other NSO tasks. Akin to this, if IT infrastructure is constrained, model storage size may also be a consideration, as some machine learning models take up considerable space on disk. Finally, model complexity and explainability may be considered, as well as ability to support the regular quality assurance process if one is utilized prior to the predicted data being incorporated into a published index.

[5] For instance many clothing classification tasks will deal with retailer data where the company sells products that would fall in both COICOP 03 (Clothing and Footwear) and COICOP 09 (Recreation and culture), specifically 09.3.2 (Equipment for sport, camping and open-air recreation).

3. Proposed Evaluation Criteria

Putting the above metrics together, a three-step approach is proposed when evaluating a classification model, as well as in monitoring a model once deployed.

3.1 Step 1 - Aggregate metric analysis

Firstly, the β in the Fβ-score should be selected. Secondly, applicable traditional classification quality metrics ($F\beta_{mw}$ as macro or equal-weighted, $F\beta_{sw}$ as sample weighted, and $F\beta_{bw}$ as the basket weighted Fβ-scores) and hierarchical classification quality metrics ($hF\beta_{mw}$ as macro or equal weighted, $hF\beta_{sw}$ as sample weighted, and $hF\beta_{bw}$ as the basket weighted hierarchical Fβ-scores) from sections 2.1 and 2.2 can be compared in isolate for each model or combined together into an average to generate an overall score per model for easier interpretability (1). Research should be conducted on metrics to be the included in the total score and if some can be excluded, as well as the type of average to be utilized. The final score should be interpretable, which an average achieves by bounding the score to between 0 and 1, and sensitive to outliers, resulting in a lower total score for the model if a specific type of Fβ-score is low. Formula 1 utilizes an arithmetic mean as an example final model score.

$$Model\ score\ = \frac{1}{6}[F\beta_{mw}\ + F\beta_{sw} + \ F\beta_{bw} + hF\beta_{mw} + hF\beta_{sw} + hF\beta_{bw}] \qquad (1)$$

3.2 Step 2 - Per-class analysis

In parallel to step 1 above (section 3.1), an Fβ-score for each lowest level class is calculated for each classification model, allowing the comparison at a lower level. A threshold contextual to the price statistic being calculated and the business processes should be utilized to eliminate models that do not perform on too many categories. The selection of models should be considered in context of the effort required for quality assurance post classification, and the importance of the categories with low performance, such as their importance in the final CPI basket.

3.3 Step 3 - Business and process considerations

If applicable to NSO requirements, specific business and process considerations should be taken into account.

3.4 Putting it all together

The above steps combine to make an operational framework that can be utilized by NSOs to apply to both initial evaluation of selecting a performant classifier for production, as well as helping monitor classifier performance once deployed. For an initial evaluation, all three steps should be completed, with the first two done in parallel on robustly labeled data—selecting models that perform well on multiple categories, and a ranking of resulting models with the goal of selecting an optimal model. Business and process considerations can be taken into account as applicable to evaluate whether the top rated model is ideal for a production process, and whether deployment of the runner-up model is instead preferred. Once a model is deployed to support a production process, monitoring the performance of the deployed model is required, including to assess when a model needs to be retrained (UK Statistics Authority 2019, 9; Eurostat 2017, 23; Martindale, et al. 2020, 19). The proposed framework, specifically steps 1 and 2, can be applied to a sample of new products every month that underwent quality assurance, or validation that the predicted categories were correct. If the NSO is deploying a new model to replace an existing one, the new model should be A/B tested or deployed in parallel with the existing model to assess that it is performing as effectively out-of-sample as expected.

4. Empirical Test

4.1 Data and preprocessing

To demonstrate the performance of the method, we chose a public dataset made available by the State of Iowa on Liquor Sales Data (State of Iowa 2021). The data contain structured variables and resemble scanner data often utilized by NSOs. Specifically, the data is electronic transaction data by unique product sold at a specific date in a specific Class “E” liquor licensee. Furthermore, the dataset contains sale and volume variables necessary for price index

calculation ("Bottles Sold", "Volume Sold (Liters)", "Sale (Dollars)"), categorical variables ("Category" and "Category Name") that can be used to map the dataset products to a hierarchical taxonomy, and product definition variables ("Item Description", "Vendor"), that can be utilized for prediction features of a classification model.
Paralleling common approaches to scanner data implementation (Eurostat 2017, 24), we map the dataset categories to 14 lowest level codes in a custom hierarchical taxonomy, paralleling the lowest taxonomy codes used in price statistics such as Elementary Products taxonomy codes. For specialty categories in the dataset we instead manually map the individual products directly to our taxonomy codes, as these are interpreted to be catch-all categories for heterogeneous types of products that should belong with similar products. Once mapped, we treat the mapped code as a true label of each product, and follow the rest of our demonstration ignoring the dataset category variable. An artificial hierarchy at four levels was created to evaluate hierarchical metrics, with a single root splitting into three types of sprits (Liqueurs, Cocktails/RTD, and Spirits), Spirits further split into to 8 types (Brandy, Gin, Mezcal, Rum, Tequila, Vodka, Whiskey and Other Spirits), with Vodka and Whiskey further split to two and four children respectively (flavored and unflavoured Vodka; and Bourbon, Irish, Scotch, and Other whiskeys).

4.2 Classification and tuning

We utilize two years of the dataset for the empirical test – 2019 data is used to perform the initial training, tuning and evaluation of the model; 2020 data is used to demonstrate the monthly monitoring process. 2019 data contains 3208 unique products; the 2020 contains 941 new products that are not observed in 2019.

The classification pipeline consists of two steps: Pre-Processing and Classification. For preprocessing, common Natural Language Processing (NLP) pre-processing was done, including: special character removal, tokenization to word unigrams, stop word removal, and Term Frequency-Inverse Document Frequency (TF-IDF) vectorization. TF-IDF is commonly used, simple to implement, relatively computationally inexpensive, and it is more robust than a standard simple count vectorizer. For the classification step, five widely-adopted traditional classification models were tested ignoring the hierarchy and utilizing the flat classification approach (see Table 1). The 2019 data was divided into a training (80% of the 2019 data) and validation set (20%), using stratified random sampling. For each of the candidate classification pipelines, model hyper parameters were selected using 3-fold cross validation of the training dataset. Each pipeline was then trained, using the selected hyperparameters, on the full training dataset. Evaluation criteria metrics were then calculated on the validation dataset, with the results summarized in Table 1. The final selected model is then refit using all 2019 data (training plus validation) to be deployed to production. The production model is used to predict new products on a quarterly basis, for 2020. Model performance on new products is evaluated each quarter using the same evaluation criteria; model score on a quarterly basis for 2020 is presented in Table 3.
A β of 1 for all Fβ-scores was selected, as both false negatives and false positives were observed in training models. For the combined score, we included all applicable F1-scores for demonstration purposes and adopted an unweighted arithmetic mean to make the final score sensitive to outliers. A threshold of 0.8 was chosen based on findings that high performing classifiers in excess of 0.8 leads to superior price indices (Office for National Statistics 2020).

4.3. Results

The demonstration dataset shows how the Support Vector Machine (SVM) and Shallow Neural Network (NN) generally performed the best across all evaluation metrics and had the top two aggregate model scores (Table 1). The threshold test demonstrated that of the top 3 models, the SVM model had one low level category (Irish Whiskies) with an F1 score of less than 0.8, the NN had two, Adaboost had 3 (Table 2). Given its high performance, the SVM was selected as the model for selection on 2020 data, with the quarterly results of all metrics summarized in Table 3.

Table 1.
Training and test statistics for 2019 data based on various FB-scores outlined in section 3.1.

Model	$F1_{sw}$	$F1_{mw}$	$F1_{bw}$	$hF1_{sw}$	$hF1_{mw}$	$hF1_{bw}$	Model score
SVM	0.8795	0.857371	0.876771	0.916849	0.908075	0.918435	0.892834
Shallow NN	0.871776	0.86666	0.860297	0.919695	0.918763	0.916391	0.892264
Naïve Bayes	0.841622	0.832533	0.818937	0.901617	0.900732	0.891479	0.864487
AdaBoost	0.844433	0.837334	0.846729	0.88011	0.874676	0.886112	0.861566
Random Forest	0.839274	0.829134	0.843871	0.877714	0.871871	0.884811	0.857779

Table 2.
Training and test statistics for top 3 models based on 2019 data and various Fβ-scores outlined in section 3.1.

Model	Elementary Aggregate (EP)	FB-Score
SVM	Irish Whiskies	0.556
Shallow NN	Other Spirits	0.714
	Irish Whiskies	0.762
AdaBoost	Other Spirits	0.632
	Irish Whiskies	0.632
	Liqueurs	0.746

Table 3.
Model score when applied per quarter on 2020 data

Quarter	$F1_{sw}$	$F1_{mw}$	$F1_{bw}$	$hF1_{sw}$	$hF1_{mw}$	$hF1_{bw}$	$Model\ score$
2020Q1	0.791704	0.761984	0.779637	0.861476	0.85468	0.862852	0.818722
2020Q2	0.81019	0.788765	0.813265	0.854453	0.826787	0.859456	0.825486
2020Q3	0.800554	0.700787	0.780164	0.886857	0.827672	0.877761	0.812299
2020Q4	0.782575	0.784638	0.763883	0.869312	0.869071	0.849384	0.819811

5. Conclusion

This research has outlined and demonstrated key metrics for evaluating and selecting classifiers for production and monitoring their performance for use in the calculation of price statistics. While the approach has introduced the set of metrics that can be considered for an evaluation model performance, the selection of the most appropriate metrics from within this set is a topic that requires further research. Specifically, further research should first explicitly model the impact of misclassification on a price index, and consider this risk jointly with methods to flag records and quality assure them prior to price index calculation to minimize any misclassification bias on the final price index. Secondly, future research should assess empirically the impact of various metrics proposed in section 3.1 on tracking misclassification performance, as well as the ideal weighting of various metrics to achieve the most optimal total score. Finally, future research should consider whether the choice of price index method applied, as well as whether weight data is available, would impact the metrics and their weighing in the final evaluation score.

Acknowledgement

The authors would like to thank Junxiao Ma for support on an earlier version of this paper.

References

Costa, E. P., A. C. Lorena, A.C.P.L.F. Carvalho, and A. A. Freitas (2007), "A review of performance evaluation measures for hierarchical classifiers", *Evaluation methods for machine learning II: Papers from the AAAI-2007 workshop.* pp. 1-6.

Eurostat (2017), "Practical Guide for Processing Supermarket Scanner Data." Harmonised Index of Consumer Prices. International Labour Office. 2020. *Consumer price index manual: Theory and practice.* https://www.ilo.org/global/statistics-and-databases/WCMS_331153/lang--en/index.htm.

Martindale, H., E. Rowland, T. Flower, and G. Clews (2020), "Semi-supervised machine learning with word embedding for classification in price statistics." *Data & Policy* 2.

Office for National Statistics (2020), "Automated classification of web-scraped clothing data in consumer price statistics." Office for National Statistics. https://www.ons.gov.uk/economy/inflationandpriceindices/articles/automatedclassificationofwebscrapedclothingdatainconsumerpricestatistics/2020-09-01.

Office for National Statistics (2021), "Classification of new data in UK consumer price statistics." Office for National Statistiscs.https://www.ons.gov.uk/economy/inflationandpriceindices/articles/classificationofnewdatainukconsumerpricestatistics/2021-04-06.

Silla, C. N., and A. A. Freitas (2011), "A survey of hierarchical classification across different application domains." *Data Mining and Knowledge Discovery 22.1* pp. 31-72.

State of Iowa (2021), "2019 Iowa Liquor Sales." *data.iowa.gov*. 10 01. https://data.iowa.gov/Sales-Distribution/2019-Iowa-Liquor-Sales/38x4-vs5h.

Sun, Aixin, and Ee-Peng Lim (2001), "Hierarchical text classification and evaluation." *Proceedings 2001 IEEE International Conference on Data Mining.* pp. 521-528.

UK Statistics Authority (2019), "Guidelines for selecting metrics to evaluate classification in price statistics production pipelines." Advisory Panel on Consumer Prices – Technical, UK Statistics Authority. https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2019/08/APCP-T1910-Classification-metrics-guidelines.pdf.