Assessing the Quality of a Coding Process Generated by a Machine Learning Algorithm

Richard Laroche[1], Pier-Olivier Tremblay[1]
[1]Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, ON K1A 0T6

Abstract
The Retail Commodity Survey (RCS) collects detailed information about retail commodity sales in Canada. The objective is to produce estimates of the sales of various commodities, at the national level, for 12 retail subsectors in Canada. The RCS uses the North American Product Classification System (NAPCS) to classify commodities. Statistics Canada now receives scanner data from some major Canadian retailers. These scanner data files are received on a daily or weekly basis and contain information about products and sales. However, information about the NAPCS is not available on these scanner data files. An automated coding approach was developed using machine learning techniques to assign a NAPCS code to all the product descriptions found on the scanner data files. In order to assess the performance of the automated coding, a quality framework was developed. Different strategies were put in place, going from basic checks when a new scanner data file is received to the manual coding of a sample of products. This will allow the evaluation of the model over time, especially as new products appear. Based on this evaluation, the model will be improved if required.

Key Words: automated coding, machine learning, quality

1. Introduction

Statistics Canada recently began a modernization exercise that is based on five pillars. One of these pillars is 'Leading-edge Methods and Data Integration' (Statistics Canada, 2017). Access to new sources of data and increased use of modeling is therefore strongly encouraged.

Scanner data is among the new sources of data to be exploited. These data are generally provided by companies operating in the retail sector and contain extremely detailed information on the quantity and value of commodities sold. They also have the advantage of reducing the response burden on businesses and also being able to be used by more than one statistical program. On the other hand, the companies producing these files often use their own classification system for commodities which necessitates the coding of these commodities to the standard classification systems used by statistical agencies. In addition, the sheer volume of data contained in these scanner data files demands that automated coding techniques be used.

The purpose of this article is to provide a process for evaluating the quality of the automated coding performed when scanner data is used by the Retail Commodity Survey (RCS). A general description of the survey and the scanner data file will first be presented. The machine learning algorithm used for coding the commodities will then be described.

Finally, the process of assessing the quality of the data obtained by the automated coding will be discussed.

2. The Retail Commodity Survey

The RCS collects detailed information on commodities sold in Canada. The survey is carried out as a supplement to the Monthly Retail Trade Survey (MRTS). The MRTS collects data on total monthly retail sales while the RCS collects a breakdown of these retail sales by commodity. The RCS uses the North American Product Classification System (NAPCS) to collect, process and disseminate product statistics. The NAPCS consists of a 7-level hierarchical system and the RCS collects sales for approximately 140 of these NAPCS products (7-digit codes).

Table 1: NAPCS - Example

NAPCS Code	Description
561	Retail services
56111	Food
561111	Fresh food
5611111	Fresh meat and poultry
5611112	Fresh fish and other fresh seafood
…	…

Recently, a large company in the RCS sample started sharing its scanner data with Statistics Canada. This data is very rich in that the sales are available for thousands of products for each point of sale of the company.

The objective to be achieved in the presence of such scanner data is to assign a 7-digit NAPCS code to each commodity sold so that ultimately the direct collection of data from this company will be stopped. Machine learning techniques are used to assign NAPCS codes to each commodity. Section 3 explains the methodology used.

3. Machine Learning

3.1 Definition of Machine Learning

Machine learning is a way of modeling phenomena in order to make strategic decisions. The idea is to set up an algorithm which will build an "internal representation" in an automated way in order to be able to carry out a certain task (prediction, identification, etc.). This requires a dataset that the algorithm can train and improve on, hence the word learning. This dataset is called "training data".

Machine learning methods can be grouped into two broad categories. In supervised learning, we use annotated data to train the model since this data (i.e. each line or transaction) has already been assigned to a target class. The goal is that the algorithm becomes able to predict this class on new non-annotated dataset once trained. If the data is not annotated, an unsupervised learning algorithm must be used and the aim here is rather to describe associations and patterns between variables (Hastie et al., 2009).

To solve a supervised learning problem, we must separate our initial dataset into three subsets:

- a training dataset, which will allow us to adjust different models and thus train our model and which will be used by the learning algorithm;
- a validation dataset, which will allow us to estimate the prediction error and to select the best model;
- a test dataset, which will allow us to measure the error of the final model on data that the algorithm has never seen (Open Classrooms).

The process that led to the selection of an automated learning algorithm for assigning NAPCS codes to scanner data is described in the following sections.

3.2 Description of the Scanner Data File

The Organization for Economic Co-operation Development (OECD) defines scanner data as detailed data on sales of consumer goods obtained by scanning the bar codes of individual products at points of sale in retail outlets; the data can provide detailed information about quantities, characteristics and values of goods sold as well as their prices (OECD, 2005).

The scanner data file received for the RCS contains around 50 variables. Among these are the UPC code for each product sold, an internal code derived by the retailer, the product name, the brand of the product and the location of the store. A new file is received every week; it contains more than 10 million observations (transactions are grouped by product and store). The following table uses fictitious examples to illustrate the content of the file for some selected variables:

Table 2: Partial layout of the scanner data file

Product description	**Quantity sold**	**Total sales**	**Address**	**City**	**Internal classification code**	**Desc. 1**	**Desc. 2**	**Desc. 3**	**Desc. 4**	**Desc. 5**
Salt and vinegar chips *ABC*	50	$200	123 AAA Street	Guelph	11111	snacks	chips	reg.	reg.	reg.
Soft drink *XYZ*	60	$120	456 ZZ Blvd.	Ajax	22222	drinks	Soft drinks	reg.	reg.	reg.
…	…	…	…	…	…	…	…	…	…	…

Each product sold by the retailer has an internal classification code generated by the retailer itself. A concordance table between this internal code and the NAPCS code has been developed by the RCS analysts (see Table 3). This concordance table made it possible to associate a NAPCS code to each observation in the scanner data file. In the presence of such a concordance table, automated learning techniques are generally not necessary; however, there are three reasons why we chose to use machine learning techniques to derive the NAPCS code:

1) such a concordance table may not exist for retailers whose scanner data will be received in the future;
2) in some cases, more than one NAPCS code is associated with the same internal code;
3) the concordance table would need to be constantly updated every time new internal codes appear if machine learning techniques are not used (Hatko, 2018).

Table 3: Concordance table : Internal classification codes – NAPCS codes

Internal Classification Codes		NAPCS codes	
Code	**Description**	**Code**	**Description**
99991	Apples	5611113	Fresh fruits and vegetables
99992	Bananas	5611113	Fresh fruits and vegetables
99993	Berries/Cherries	5611113	Fresh fruits and vegetables
…	…	…	…

The scanner data file presented in Table 2 was merged with the concordance table presented in Table 3. A manual intervention was required for cases where more than one NAPCS code was associated with the same internal code. In order to have adequate coverage of seasonal products, several months of data were used. The resulting file was then used to develop a machine learning algorithm.

3.3 Machine Learning Algorithm

A group of data scientists at Statistics Canada developed the machine learning algorithm to assign a NAPCS code to each product in the scanner data file. The algorithm uses different variables containing a description of the product; some of these variables are precise (for example, the variable "Product Description" shown in Table 2), others are more general (for example, the variables "Desc. 1", "Desc. 2", etc. also shown in Table 2).

For each observation, these description variables are concatenated to form a "document". The machine learning algorithm will check for the presence or absence of vocabulary terms in the "document". Vocabulary terms can be any combination of consecutive characters. For example, the term "milk" contains three two-character words ("mi", "il" and "lk"), two three-character words ("mil", "ilk") and a four-character word ("milk"). The algorithm is trained to predict the NAPCS code according to the vocabulary terms present.

In order to improve the speed of execution of the algorithm, two consecutive variables containing the same description will not be repeated when the new variable is created. For example, if the variables "Desc1" and "Desc 2" contain the same description then the variable "Desc 2" will be ignored.

The open source library XGBoost for R was used for this exercise. The final result is a linear model that can be written in the form:

$$\hat{Y} = B + XW$$

where

- $\hat{Y}$ is a matrix containing the predictions of the algorithm (the number of rows corresponds to the number of records and the number of columns to the number of NAPCS);
- B is a bias matrix (the y-intercept) (the number of rows corresponds to the number of records and the number of columns corresponds to the number of NAPCS);
- X is a binary matrix containing the input data (the number of rows corresponds to the number of records and the number of columns to the number of vocabulary terms);

- W is a weight matrix (regression coefficient) (the number of rows corresponds to the number of vocabulary terms and the number of columns to the number of NAPCS).

When the algorithm is trained, the matrices W and B are optimized by the algorithm to best fit the training data. Other machine learning algorithms were tested, but none have performed as well as XGBoost in terms of accuracy and speed of execution.

4. Quality Assessment of the Machine Learning Algorithm

This section presents the strategy used to evaluate the quality of the outputs produced by the machine learning algorithm.

4.1 Quality Assessment of the Scanner Data Files Received

It is important to make sure that every scanner data file that is received corresponds to what is expected in terms of format and contents before applying the machine learning model. Basic checks related to the size of the file, the number of observations and the number of variables are done every time a new file is received. More elaborate checks are also made to make sure that the number of products sold, the number of points of sales as well as of total sales (by product and by point of sales) for a given week are consistent with what was observed in the past. When a possible problem is detected, the RCS analysts first need to determine if the issue is due to some erroneous data or if it is simply due to a valid consumer behavior. In case of erroneous data, the Data Acquisition group at Statistics Canada is contacted. Ultimately, the data provider can be asked to submit an updated file.

4.2 Quality Assessment of the Model Used to Derive the NAPCS

To assess the quality of the machine learning model, manual coding of a sample of records is done on a regular basis. Quality indicators are then derived.

4.2.1 Manual coding

A sample of 1,000 product descriptions is selected every four weeks to assess the quality of the algorithm. Prior to the sample selection, all descriptions went through a pre-processing step where the duplicates were removed and the sales generated by each unique product description were calculated. For each description selected, a NAPCS code is manually assigned. The NAPCS codes obtained through manual coding are compared to the NAPCS codes obtained by the model. Note that the result of manual coding is considered to be the “true” value.

In order to ensure a good quality assessment of the model, different elements must be considered when selecting the sample:

- we must select a sample of new product descriptions since we need to make sure that the model is able to assign the right NAPCS code to descriptions that have never seen before;
- for each prediction made, the machine learning algorithm produces a confidence score, which is a number between 0 and 1 (where in general, the higher the score, the greater the confidence in the prediction); the sample should contain products with high and low scores;

- it will be necessary to ensure that the sample covers the most important NAPCS codes since some NAPCS codes generate more sales than others.

In order to select a representative sample, we first stratify the product descriptions into two groups: old and new. The ‘new’ products are the ones that did not exist when the model was built thus these products were not part of the initial training, validation or test sets.

The old and new products are further stratified according to the confidence score associated with each prediction (score greater than or equal to 0.9, score less than 0.9). In each stratum, the products are sorted by NAPCS and sales, and a systematic sample is selected in each stratum.

4.2.2 Quality Indicators

Several indicators are proposed to measure the reliability of the model: accuracy, sensitivity, precision, F1 score as well as the Matthews correlation coefficient.

The accuracy is defined as the ratio of the number of descriptions correctly coded by the model over the total number of descriptions in sample. This ratio is weighted by the total sales of the descriptions in order to obtain a more relevant measure of accuracy in the context of the RCS.

To calculate the other quality indicators mentioned above, a confusion matrix for each NAPCS code is necessary:

Table 4 : Confusion matrix for NAPCS *X*

		Real NAPCS (manual coding)	
		NAPCS *X*	Other NAPCS
NAPCS predicted by the algorithm	NAPCS *X*	True Positive (**TP**)	False Positive (**FP**)
	Other NAPCS	False Negative (**FN**)	True Negative (**TN**)

The following indicators are then derived for each NAPCS:

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{F1-Score} = \frac{2TP}{2TP+FP+FN}$$

$$\text{Matthews correlation coefficient} = \frac{TP*TN-F\quad *FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

In words, **sensitivity** represents the rate of true positive, which is, the ratio of the number of observations predicted in the right class to the number of observations that actually belong to that class. In the context of the RCS, a low sensitivity for NAPCS *X* means that there is an underestimation of the sales for NAPCS *X* (too few descriptions were coded to NAPCS *X* by the algorithm). **Precision** is the ratio of the number of observations correctly predicted to the number of observations predicted in this class. A low precision for NAPCS *X* means that there is an overestimation of the sales for NAPCS *X* (too many descriptions were coded to NAPCS *X* by the algorithm). The **F1-score** is a harmonic mean of sensitivity and precision. It reaches its best value at 1. As for the **Matthews correlation coefficient**, it is a measure of the correlation between the value predicted by the algorithm and the value coded manually. It will take a value of 1 if all the observations have been predicted in the right class (TP and TN) and conversely a value of -1 if they have all been predicted in the wrong class (TN and FP). A value of 0 indicates no correlation with manual coding. In other words, a value close to zero indicates that the model does not perform better than assigning a class at random.

4.2.3 Results

Since the strategy described above was implemented, eight samples of 1,000 descriptions were selected. Table 5 shows the average overall accuracy rate over all eight samples and Table 6 shows the distribution of the Matthews correlation coefficients obtained using the most recent sample.

Table 5: Accuracy rates (average of 8 samples)

		Weighted accuracy	**Economically weighted accuracy**
Product description	Recurring	84%	93%
	New	80%	85%
Confidence score	≥ 0.9	87%	95%
	< 0.9	63%	75%
Overall		84%	93%

In the table above, the weighted accuracy uses the design weight while the economically weighted accuracy uses both the design weight and the sales generated by each product.

Overall, the algorithm is able to correctly code 84% of the products. When the sales generated by each product are taken into account (economically weighted), the accuracy rate goes up to 93%. In other words, 93% of the sales of our retailer are well classified. This accuracy rate will probably get even higher once the model is re-trained.

Table 5 also shows that the algorithm does not do a good job when the confidence score is lower than 0.9. That tells us that we might need to simply manually code any product with a low confidence score. Approximately 10,000 descriptions (over more than 100,000) have a confidence score lower than 0.9, and from these 4,000 have a confidence score lower than 0.8.

Table 6: Matthews correlation coefficient (most recent sample)

Matthews correlation coefficient	**Number of NAPCS**[1]
[0.95 , 1]	34
[0.9 , 0.95[	5
[0.8 , 0.9[	5
[0.7 , 0.8[	4
[0.6 , 0.7[	3
[0.5 , 0.6[	2
[0 , 0.5[	14
[-1 , 0[	4

1. From the 140 different NAPCS codes covered by the survey, roughly 70

NAPCS codes could be found on the scanner data file.

From the most recent sample that was selected for which manual coding was performed, we observe that a majority of the NAPCS have a Matthews correlation coefficient equal or higher to 0.9 which indicates a very high correlation between the predicted NAPCS and the real NAPCS code.

Different options are available if some of the indicators show a lower level of quality for a given NAPCS. Using an exception file that allows to systematically overwrite the results of the automated coding for some specific descriptions is the easiest thing to do in the short term. A systematic manual coding for some NAPCS is also possible. Retraining the model is also an option to consider.

5. Conclusion

The RCS was one of the first surveys at Statistics Canada to use machine learning in its production process and the outcome was extremely positive: the retailer from which we received weekly scanner data saw its response burden considerably reduced as collection stopped for that business and a lot of knowledge on machine learning techniques was gained throughout the project.

Using scanner data also has benefits for other programs at Statistics Canada. The Consumer Price Index uses it to get prices for a selected list of products. There are also benefits for Statistics Canada's Business Register maintenance, as we can almost get in real-time the list of stores that opened or closed. Some unexpected benefits were also seen when a paper was published by some colleagues from the Consumer Price Index about the panic buying at the beginning of the COVID-19 pandemic. The results were largely based on scanner data from a few retailers (Statistics Canada, 2020).

But most of all, we showed that automated coding can be reliable. The accuracy rate (economically weighted) is 93% and will most likely get higher once the descriptions that were manually coded are added to the training set and the model is re-trained. We are now ready to expand the scope of the project by adding scanner data from other retailers in our process. This means that the RCS will rely more and more on these new data sources. It is therefore important to have rigorous quality assessment processes in place to verify the content of the files received and to ensure the validity of the machine learning model used. It should also be noted that since this is a fairly new approach, the proposed process will need to be adapted and improved over time.

Acknowledgements

The authors thank Kenneth Chu, Marie-Claude Duval, Sevgui Erman, Susie Fortier, Stan Hatko, Christiane Laperrière, Robert Philips, Jean-Sébastien Provençal, Saman Rudgar, Simon Sheldrick, Claude Turmelle and Anthony Yeung for their thoughtful comments and assistance in the development of this project.

References

Hastie, T., Tibshirani, R. and Friedman, J. (2009), The Elements of Statistical Learning, Second Edition, Springer, 2009.

Hatko, Stan (2018), Retailer Scanner Data, Internal document.

OECD (2005), Glossary of statistical terms, https://stats.oecd.org/glossary [consulted on July 3, 2020]

Open Classrooms, Initiez-vous au machine learning, https://openclassrooms.com/fr/courses/4011851-initiez-vous-au-machine-learning [consulted on September 23, 2020]

Statistics Canada (2017), The Vision: A Data-driven Society and Economy, Internal document.

Statistics Canada (2020), Canadian Consumers Adapt to COVID-19: A Look at Canadian Grocery Sales up to April 11, Catalogue No. 62F0014M, May 2020.