The Survey Statistician, 2020, Vol. 82, 11-13.

Ask the Experts

Why machine learning and what is its role in the production of official statistics?

Sevgui Erman

Statistics Canada

e-mail: sevgui.erman@canada.ca

Abstract

To remain competitive, statistical organizations need to move quickly to adopt and take advantage of machine learning and new digital data sources. Machine learning is not fundamentally new, and statistical agencies have been using modelling techniques for a very long time. Why do National Statistical Organizations require machine learning in their toolbox and what is its role in the production of official statistics? These are some of the questions discussed in this paper, along with examples of machine learning use in official statistics.

Keywords: machine learning, official statistics, artificial intelligence, open source

What is Machine Learning?

"Machine learning is the science of getting computers to automatically learn from experience instead of relying on explicitly programmed rules, and generalize the acquired knowledge to new settings." [1]

In essence, Machine Learning automates the analytical model building through optimisation algorithms and parameters that can be modified and fine-tuned.

1 Why do National Statistical Organizations require machine learning in their toolbox?

National Statistical Organizations (NSOs) are data-driven organizations, and data are at the centre of today's digital revolution. Data and technology are transforming our society and the way we consume information. The vast amount of digital data available is also transforming the role of NSOs as the premier information providers for evidence-based decision making.

New alternative data sources are already showing many benefits, including: providing faster and timelier products, reducing response burden on households and businesses, producing more

accurate results and lowering costs. This is fundamentally changing the way statistical agencies operate. Many of these new opportunities require the use of machine learning methods. In fact, machine learning is the main computation tool for big data processing.

2 Is machine learning new?

Machine learning, and artificial intelligence, are not fundamentally new [4]. Statistical agencies have been using modelling techniques and data analytics for a very long time. Examples include modelling for stratification, imputation, and estimation purposes. [2] and [3] are excellent references in this context.

What makes today's machine learning methods different than the ones used five or ten years ago is their evolution within the big data processing space. This evolution has been enabled by:

- better computational capacity,
- along with developments in the algorithmic space and applications to unstructured data (text, images, video, censor, etc.)
- more efficient data ingestion
- increased access to structured and unstructured data
- more capabilities offered by big data processing platforms to efficiently manage RAM and CPU, and, when required, GPUs, both in the cloud and on-premises [5]

Another major factor driving this shift in methods is collaboration, especially in the open source community. Using R and Python for machine learning and having an open source first approach are accepted standards today. While previously development of data processing systems has been done independently by organizations, today users can benefit from open source code that results from years of effort, and has been tested at a scale that was not previously feasible. The implementation of open source tools can accelerate development, reduce project costs and result in faster turnaround times, allowing projects to move from development mode to production mode more quickly.

3 Machine learning use in official statistics: examples and benefits

3.1 Machine learning applied to retail scanner data

Statistics Canada receives point of sale data from large retailers. This provides a complete census data for volume and price statistics from the participating businesses. In the short term, the agency reduces reporting burden by eliminating survey collection for the participating businesses, which also reduces collection efforts. Statistics Canada is providing participating businesses with custom user-defined statistics based on their data. In the long term, as more businesses provide scanner data, the agency will be in a position to release local-level data (city and postal code), along with commodity data at a far more granular level. Whereas previously data was produced on a few hundred commodities, based on the North American Product Classification Standard (NAPCS), now it will be possible to potentially release data at the Universal Product Code level, i.e., thousands of different commodities. Another potential output is weekly publications on the value, amount, and average price of each NAPCS product sold at retail by detailed geographic area. A machine learning classifier, XGBoost, with linear base learners using character n-grams and bag of words based approach, is used to associate the presence of substrings in the data with certain NAPCS codes.

3.2 Satellite images use in agriculture

Currently, Statistics Canada has three machine learning projects in the space of agriculture that use satellite images. The in-season crop identification project, for instance, aims at predicting crop type proportions within an image. Landsat-8 satellite images of two census agricultural regions within Alberta are used. The labelled data are derived from crop insurance data. Using this dataset, a state-of- the-art deep learning model is built. This new model is expected to produce real time data and reduce the cost of crop production data collection. Other examples of machine learning use include the estimation of the area of land covered by greenhouses from satellite images, as well as the area covered by solar panels.

3.3 Automation

A broad range of tasks exist where analysts can extract information from unstructured data sources, such as the extraction of financial variables from annual financial reports; financial statements; company information forms; legal reports; news releases; acquisition and merger of assets of publicly traded companies; and financial statements received from federal, provincial and municipal organizations. Many of these tasks can be automated using machine learning, resulting in much more efficient processes.

Closing. Challenges and opportunities

The machine learning context is highly dynamic—which can be both an advantage and a challenge. This type of environment requires an ever-learning mindset. To remain competitive within this transformed data modelling space, statistical organizations need to move quickly to adopt and take advantage of machine learning and new digital data sources. Survey statisticians offer advanced expertise in statistical methods and data quality, and are well positioned to contribute to and benefit from the larger machine learning community. Survey statisticians will play a key role in the algorithmic space by identifying the standards of rigor, ensuring statistically sound methods are used, promoting quality and valid inference when it is needed, and abiding by ethical science practices when deriving insights from data [6]. While new technologies are creating amazing opportunities, these opportunities come with responsibilities. New algorithms and model assessment guidelines will have to be developed, and their monitoring and maintenance in production will pose new type of challenges.

References

1. United Nations Economic Commission for Europe's Machine Learning Team (2018 report). The use of machine learning in official statistics.
2. Carl-Erik Sarndal, Bengt Swensson, Jan Wretman (1992). *Model assisted survey sampling*. Springer-Verlag.
3. J.N.K. Rao, Isabel Molina (2015). *Small area estimation*. Wiley Series in Survey Methodology.
4. Sean Zinsmeister, Andrew Yeung, Ryan Garrett (2019). *AI-Driven analytics – how artificial intelligence is creating a new era of analytics for everyone*. O'Reilly Media. https://www.oreilly.com/library/view/ai-driven-analytics/9781492055785/.
5. Ashish Thusoo, Joydeep Sen Sarma (2017). *Creating a data-driven enterprise with DataOps. Insights from Facebook, Uber, LinkedIn, twitter, and eBay*. O'Reilly Media. https://www.oreilly.com/library/view/creating-a-data-driven/9781492049227/.
6. Michael I. Jordan (2019). *Artificial Intelligence – the revolution hasn't happened yet*. University of California. Harvard Data Science Review Issue 1. Berkley. https://hdsr.mitpress.mit.edu/pub/wot7mkc1.