## Use of Big Data and Machine Learning at Statistics Canada

### Christie Sambell, Stan Hatko

Data Science Division, Statistics Canada

### Introduction

Statistics Canada is the national statistical office of Canada and is responsible for collecting statistics about all aspects of Canadian life: commercial, industrial, economic and social. Traditionally the majority of these statistics have been obtained through surveys. As part of modernization efforts, Statistics Canada has been exploring ways to expand the use of alternative data sources to supplement or even replace entire surveys. The Data Science Division is a team within Statistics Canada who tries to solve concrete business cases using big data and Machine Learning. Here are three examples of projects where the Data Science Division at Statistics Canada has used machine learning to incorporate big data sources into its statistical operations.

### Retail Scanner

**Background:** Several years ago, Statistics Canada began receiving scanner data—point of sales data from a major Canadian retailer. The product descriptions from each of the 13 million records per week needed to be classified to match the product classifications on the Retail Commodity Survey. An initial attempt was made using traditional matching methods but the time needed to process the weekly data was over a week.

**Methods:** Used MapReduce to chunk the weekly file up and XGBoost to train a model that could be run in a few hours.

**Results:** 99% accuracy in training and 92% weighted accuracy in Production.

**Status:** This model has now been used in Production for over a year and that retailer is no longer being sent the Retail Commodity Survey.

**Next Steps:** Will be scaling up for two other major retailers.

### Freight Trucking

**Background:** The Freight Trucking Statistics Program has been receiving shipping manifests instead of questionnaires from four major trucking companies for several years. They want to increase the number of trucking companies submitting shipping manifests instead of questionnaires. The current system to classify the product descriptions is outdated and requires a significant amount of manual corrections which is not sustainable. Like the scanner data project, the product descriptions needed to be classified to a standard classification but this project proved to be much more challenging as there were 582 classes and the product descriptions were neither clean nor systematically generated.

**Method:** An XG-Boost model for the first company and active learning to scale the model up to include the other three companies.

**Results:** For the first company the results went from 76% to 91% accuracy with the client labelling an additional 31,000 product descriptions.

**Status:** Currently waiting for the rest of the processing system to be redesigned so that ML can be incorporated.

**Next Steps:** Determine if there are any additional trucking companies that would be good candidates for ML.

### Greenhouse Detection

**Background:** One area at Statistics Canada that really saw the potential of alternative data was the Agriculture Statistics Program. One of their projects was an experiment to see if greenhouses could be identified from images in order to remove questions from the Census of Agriculture and to improve the survey frame for the Annual Greenhouse Survey. There have been two phases of this project so far, both with promising results. The first phase used RapidEye satellite images (at the 5 meter level) while the second uses higher resolution aerial images provided by municipalities (at the 7-15 cm level).

**Methods:** First phase on satellite images used a simple convolution neural network. The second phase on the aerial images used U-Net with a retrained residual network.

**Results:** The first phase produced decent results for two test areas (F1 0.76-0.94) but poor results for the third (F1 0.56 – 0.89). The second phase produced good results across all test areas (F1 0.83 – 0.95) and can still be improved with further hyper-parameter tuning.

**Status:** Planning for the next phase.

**Next Steps:** Subsequent project to explore greenhouse cover type, active status type and vegetation type, in order to further reduce the questions on the Greenhouse Survey. Also a future project to retrain the model on images of solar panels.

### Lessons Learned

While the Data Science Division at Statistics Canada is starting to see success with big data and machine learning, we have learnt many thing along the journey:

**Start small:** Break the project down, start with a small proof of concept and iterate from there.

**Data Science is a team sport:** It is important to keep in regular contact with the client as they are the experts of their domain. Make sure they set aside time to support the project. Also it is important to bring in IT early especially if a machine learning model needs to be integrated into an existing system.

**Machine learning is not magic:** machine learning is not the right solution for all problems so really work with the clients to make sure that the project is really a machine learning problem before starting. Also a machine learning model that was trained for one data source can’t just be reused on a different data source without some work.

**Data quality:** The machine learning model is only as good as the data going in. The messier the data the more difficult it is to achieve a good accuracy.

Infrastructure is half the battle. Having the correct infrastructure can make the difference in whether a project is a success or not. It is also important to have it in place before starting or considerable time will be lost.