The use of ML Text Classification for The Retail Commodity Survey

Survey improvements from the use of ML in the commodity classification process

Statistics Canada Statistique Canada

Canada

Overview

- Background of Retail Commodity Survey (RCS)
- Process of Model Creation
- ML Pipeline
- Benefits and challenges of integration
- Lessons learned

Background of RCS

Collects national-level retail sales data, by commodity.

Commodities classified according to North American Product Classification System (NAPCS).

Conventional approach; survey data are collected from respondents.

NAPCS Code	NAPCS Description
5611111	Fresh meat and poultry, at retail
5611112	Fresh fish and other fresh seafood, at retail
5611113	Fresh fruit and vegetables, at retail
5611114	Eggs and dairy products (except frozen desserts), at retail
5611115	Baked goods (except frozen products, cookies and crackers), at retail
5611116	Perishable prepared foods (including fresh sliced deli meats, prepared entrees and fresh pasta), at retail

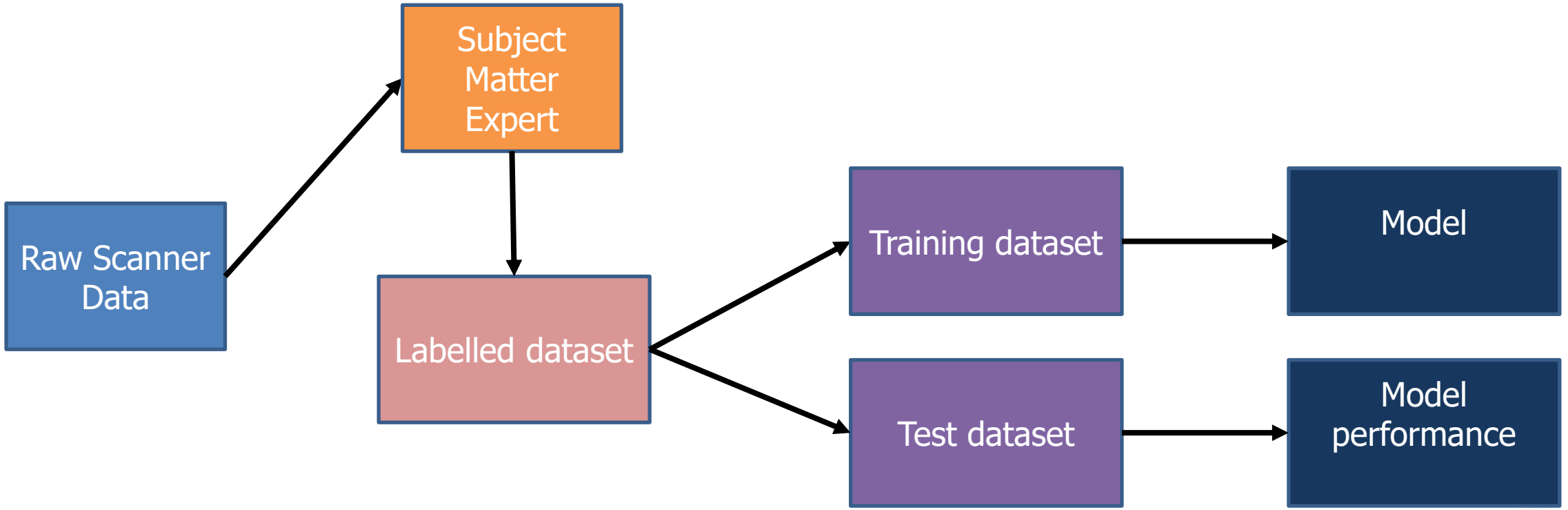
Scanner data and RCS

As of 2016, Statistics Canada has acquired scanner data from select large Canadian retailers on a regular basis.

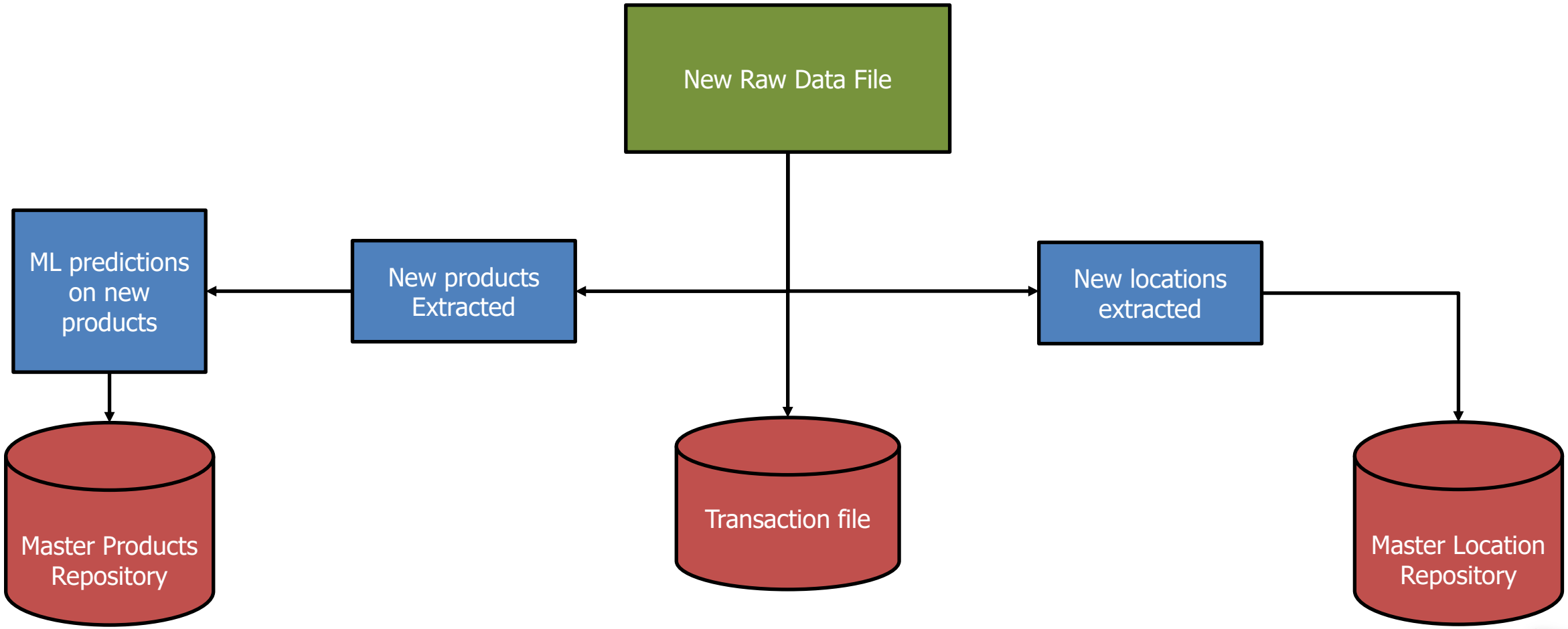
Scanner data used as a replacement for respondent data.

Process of Model Creation

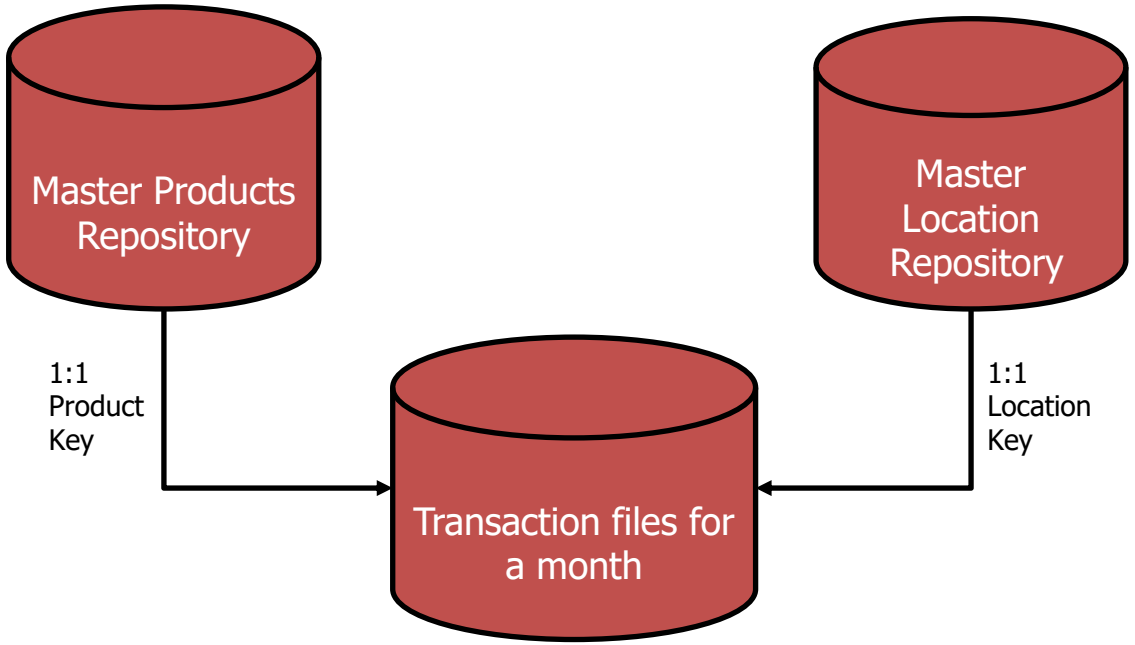
Our machine learning approach was creating a supervised model - meaning a labeled dataset is provided to train the ML model.



Current Pipeline



Current Pipeline cont.



Benefits

Reduced response burden

Data received and processed in a timelier manner

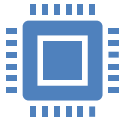
More granular than conventional respondent questionnaire

Challenges

Set up costs

Creating initial labelled dataset

Hardware constraints



Access to open-source software

Model maintenance

Quality assurance

Classification concordance

Lessons Learned

- Importance of exploratory data analysis
- Having a pipeline that can be scaled-up
- Keeping up with the literature
- Bridging the gap between developer and subject matter