OPTIMIZING THE COMMODITY FLOW SURVEY, v2

Economic Reimbursables Division / Commodity Flow Branch

James Hinckley – Branch Chief
OVERVIEW

- Commodity Flow Survey
- Commissioned by BTS
- Conducted every 5 years (2017, 2022)
- Respondents provide sampling of shipments from each quarter
OVERVIEW

- Commodity Flow Survey
- Commissioned by BTS
- Conducted every 5 years (2017, 2022)
- Respondents provide sampling of shipments from each quarter
<table>
<thead>
<tr>
<th>SCTG commodity code from accompanying booklet</th>
<th>Commodity Description</th>
<th>Is item in col (G) a hazardous material? Enter &quot;UN&quot; or &quot;NA&quot; number</th>
</tr>
</thead>
<tbody>
<tr>
<td>(F) 34520</td>
<td>Mechanical machinery</td>
<td>Y</td>
</tr>
<tr>
<td>20222</td>
<td>Sulfuric acid</td>
<td>N 1830</td>
</tr>
</tbody>
</table>
ITEM G - Other Clarifying Information

"Pulling this information was a huge spend of time and resources."

"Just glad this is over!!"
OBJECTIVE

Using Machine Learning, can we automate the assignment of SCTG codes to shipments?
OBJECTIVE

(spoiler alert)
OBJECTIVE

(spiloter alert)

YES!
METHOD

- Training Data: **6.4 million** labelled shipment records from 2017 CFS
- Bag-of-Words, Logistic Regression (logit) model
- Use shipment description and NAICS Code
  - "fasteners"
- Clean records, de-duplicate, disambiguate: ~**400,000 unique training records**
- Initial results: **50% accuracy** (on 40,000 unseen test set records)
- Investigate poorly performing codes

- **40994**: Sewing and knitting needles (includes for machines) crochet hooks, hook and eye *fasteners*, safety pins, straight pins, buttons, buckles and clasps, tubular and bifurcated rivets, *snap-fasteners*, zippers, and similar notions.
METHOD

- **33310**: Nails, screws, bolts, nuts, washers, staples except in strips, and similar **fastening** articles
METHOD

- Manually validating, about 50% of items labelled 40994 by respondents were miscoded.

- However, the model was getting it right!
RESULTS

- **Proof-of-concept**: ran model on 170,000 unlabelled records
- 70,000 with confidence score above predefined threshold
- Manually validate a sample of 350 unique records
- **89% accurate!**
- Batch-edits have saved \(~500 \text{ hours}\) of manual editing time, $35,000
RESULTS

- 2022 CFS: save respondents 50,000+ hours, $2.1 million in respondent lookup costs

- This does not include Census' editing costs, cleaning up invalid / messy data

- And, we get more, more accurate, and real-time shipment data!

- Cleaner data (via Amazon's Mechanical Turk) will improve these savings
  - Better model == more savings
ACKNOWLEDGMENTS

- James Hinckley, CFS Branch Chief
- Jessica Young/Berin Linfors, CFS Section Chiefs
- Ja Lee, Ben Schultz, CFS Analysts
- Julie Parker, Joy Sharp, Mehdi Hashemipour, BTS
- Cavan Capps, Census Big Data Lead
- Andrew Cukierwar, 2017 Civic Digital Fellow
THANK YOU!

Twitter: @c_moscardi
Github: cmoscardi
Email: christian.moscardi@gmail.com