Evaluating Machine Learning Performance on Large Scale Shipment Data for the 2022 CFS

Cecile Murray, Senior Data Engineer, U.S. Census Bureau

Federal Committee on Statistical Methodology Conference

October 24, 2023

United States® Census Bureau

Disclaimers

- Any opinions and conclusions expressed herein are those of the author and do not reflect the views of the U.S. Census Bureau
- The Census Bureau has reviewed this data product to ensure the appropriate access, use, and disclosure avoidance protection of the confidential source data (Disclosure Review Board (DRB) approval number: Project No: P-7504831, DRB Approval Numbers: CBDRB-FY24-ESMD002-003)

United States® Census Bureau

Context

- 2022 Commodity Flow Survey (CFS) collected roughly 16x more data
- This was possible partially because we used machine learning (ML) to categorize shipments by commodity code instead of asking respondents to do so
- This change reduced respondent burden, but it also reduced the amount of human-validated data available for evaluating our ML process

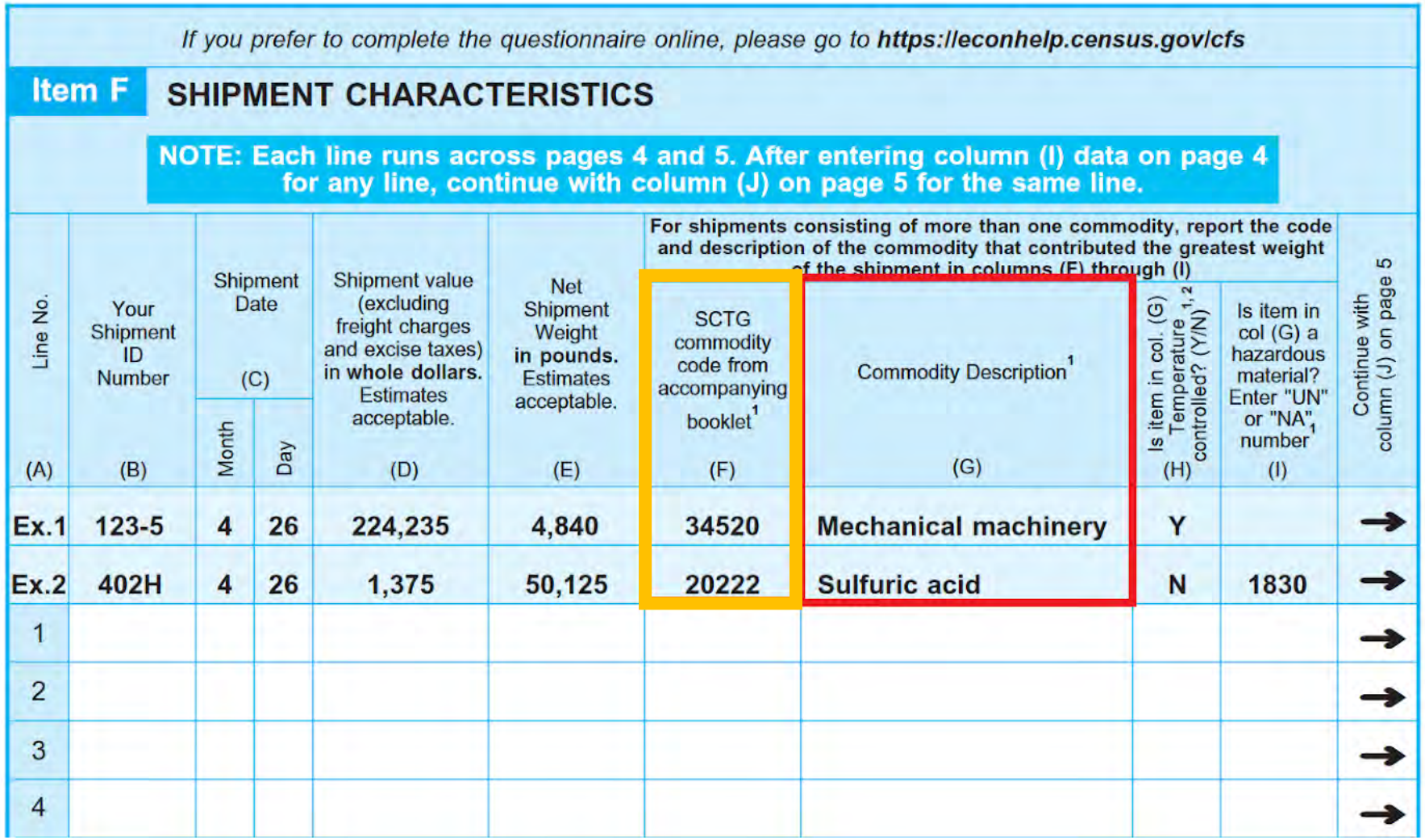
United States® Census Bureau

Overview

- Describe ML problem and our architecture
- How we evaluated model performance
- Impact of ML on data quality

United States® Census Bureau

ML goal: label shipments with SCTG code



If you prefer to complete the questionnaire online, please go to **https://econhelp.census.gov/cfs**

Item F **SHIPMENT CHARACTERISTICS**

NOTE: Each line runs across pages 4 and 5. After entering column (I) data on page 4 for any line, continue with column (J) on page 5 for the same line.

Line No. (A)	Your Shipment ID Number (B)	Shipment Date (C) Month	Shipment Date (C) Day	Shipment value (excluding freight charges and excise taxes) in whole dollars. Estimates acceptable. (D)	Net Shipment Weight in pounds. Estimates acceptable. (E)	For shipments consisting of more than one commodity, report the code and description of the commodity that contributed the greatest weight of the shipment in columns (F) through (I): SCTG commodity code from accompanying booklet[1] (F)	Commodity Description[1] (G)	Is item in col. (G) Temperature controlled? (Y/N)[1,2] (H)	Is item in col (G) a hazardous material? Enter "UN" or "NA"[1] number (I)	Continue with column (J) on page 5
Ex.1	123-5	4	26	224,235	4,840	34520	Mechanical machinery	Y		→
Ex.2	402H	4	26	1,375	50,125	20222	Sulfuric acid	N	1830	→
1										→
2										→
3										→
4										→

Report Online - Do Not Return

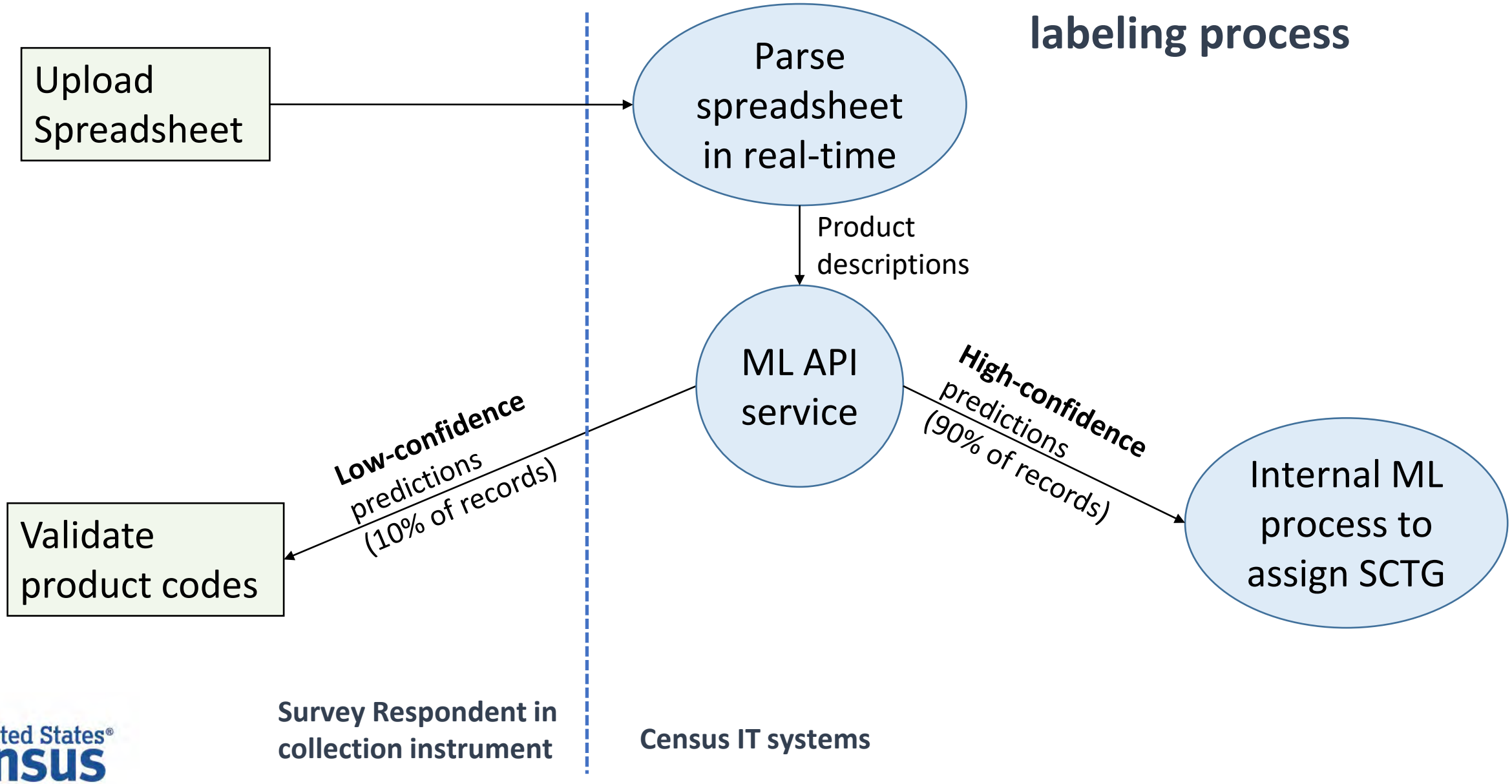
United States® Census Bureau

Brief overview of SCTG codes

- SCTG = Standard Classification of Transported Goods
- Product classification system for transportation analysis
- There are 42 2-digit SCTG major groupings
  - SCTG 24 is Plastics and Rubber
- Within those 42, there are 514 5-digit SCTG commodities
  - SCTG 24221 is “Plastics tubes, pipes, hoses, and fittings, including joints, elbows, and flanges”
  - SCTG 24225 is “Plastics household or toilet articles”

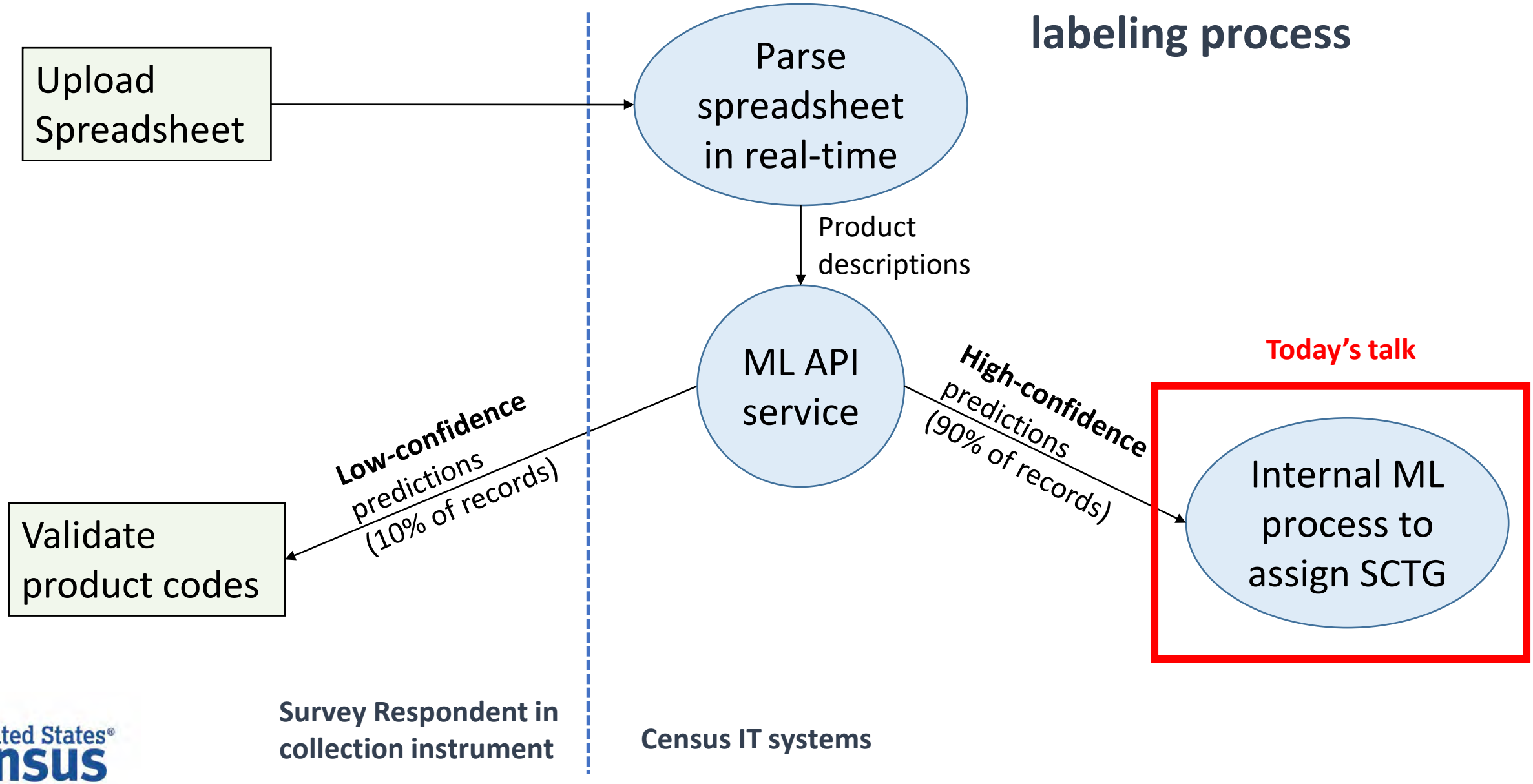
United States® Census Bureau

2022 cycle shipment labeling process



United States® Census Bureau

2022 cycle shipment labeling process



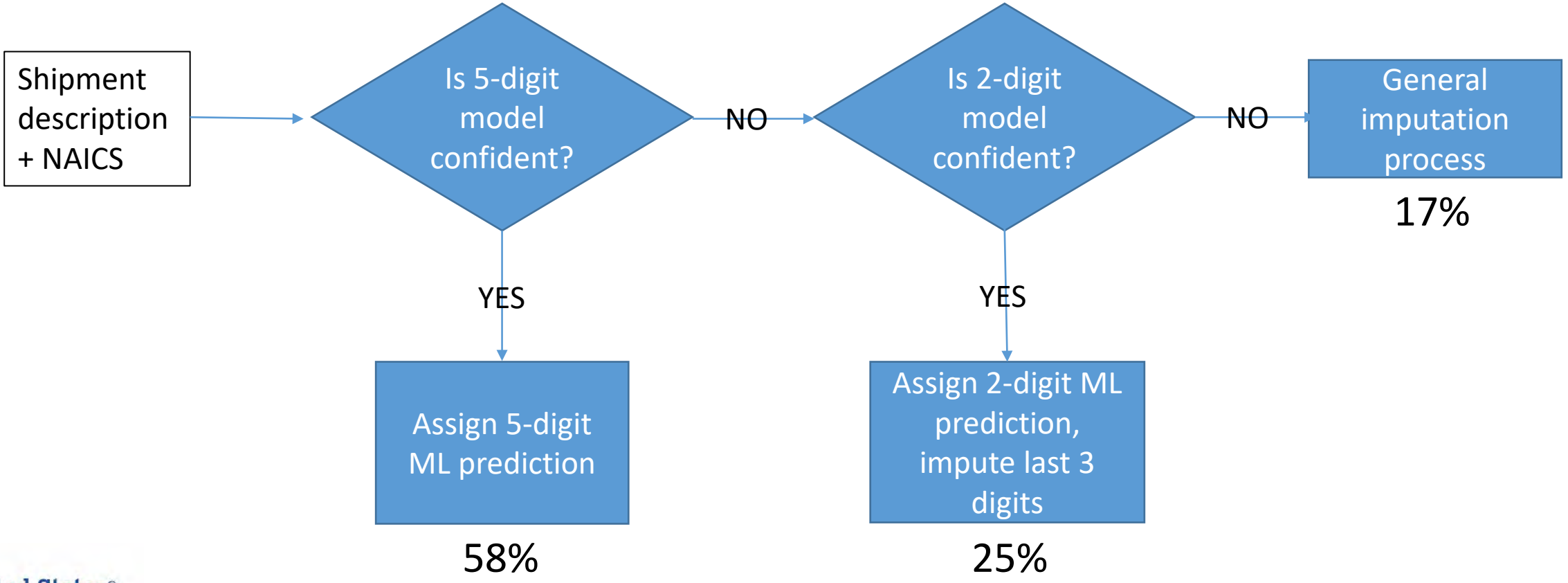
United States® Census Bureau

Internal model details

- Training data:
  - 2017 CFS shipment microdata
  - 2022 CFS shipment microdata where respondents validated description
- Models:
  - Two logistic regressions using industry NAICS code and shipment description as features
  - Models produce an SCTG label, as well as a probability score, which we use as a measure of model confidence
  - One model predicts at the 5-digit level, the other predicts at the 2-digit level

United States® Census Bureau

How we assign SCTG to shipments



United States® Census Bureau

How do we know if the models are “good”?

- This type of model has worked well on similar problems in the past, but data and model drift commonly crop up over time
- Evaluate both model accuracy and overall system performance
- Quantitative approaches:
  - Have humans review model predictions for correctness
  - Cross-validation using training data

United States® Census Bureau

Hand-coding shipments

- Have human SMEs validate SCTG codes for a sample of 882 shipments
  - 42 SCTG groupings * 21 shipments = 882
- Human reviewers provided correct SCTG if the model prediction was incorrect

United States® Census Bureau

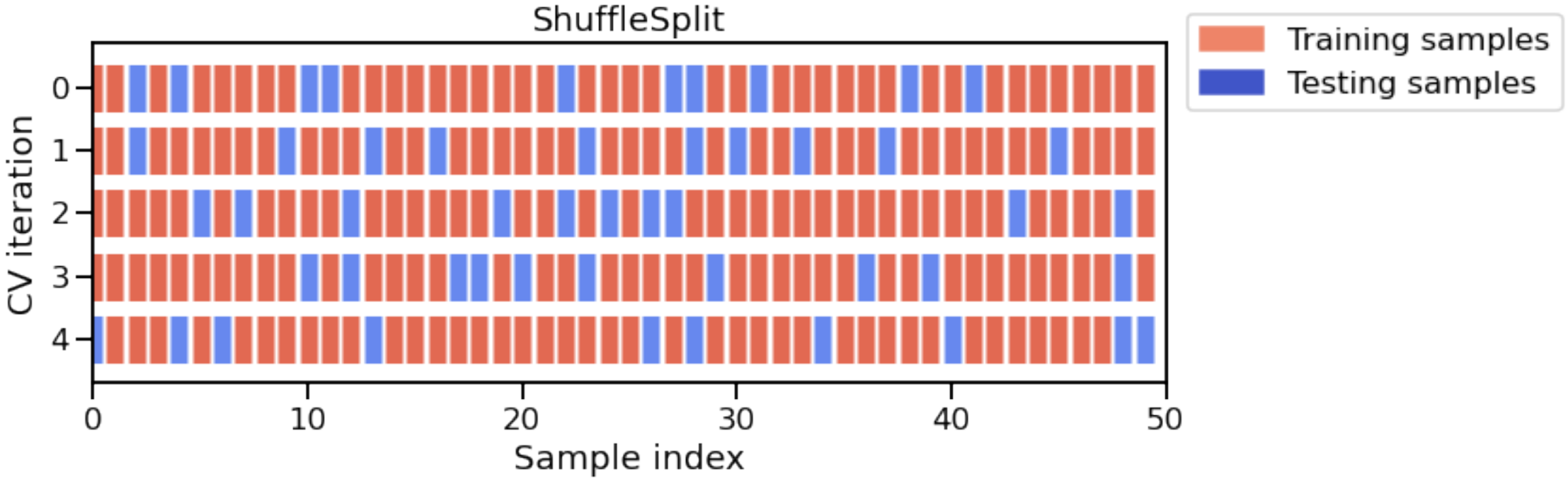
Results of hand coding

- 5-digit model accuracy for SCTG assigned via ML only: **91%**
- 2-digit model accuracy for SCTG assigned via ML and imputation: **68%**
  - Accuracy of first 2 digits of 5-digit prediction on SCTG assigned via ML and imputation: **78%**
- When model confidence is low, so is measured model accuracy => we're not leaving good model predictions on the table

- Caveat: sample sizes are small

United States® Census Bureau

Cross-validation overview



United States® Census Bureau

Cross-validation results

- 5-digit model accuracy on SCTG assigned via ML: **84%**
- 2-digit model accuracy on SCTG assigned via ML and imputation: **72%**
  - Accuracy of first 2 digits of 5-digit model on SCTG assigned via ML and imputation: **66%**
- We see strong model performance across 2-digit SCTG major groupings

- Caveat: “real” data may look different than training data

United States® Census Bureau

Takeaways from model evaluation

- Prediction quality overall is strong
- Strong performance across 2-digit SCTG groupings
- Importance of multiple ways to evaluate models and overall model pipeline

United States® Census Bureau

Automated monitoring

- Automated ML training and prediction process runs roughly weekly
- We deployed the evaluation code along with that process so we always have a current snapshot of performance
- This can help us proactively and systematically identify quality problems with models

United States® Census Bureau

Impact on data quality

- We can use machine learning to code shipments with high quality and with minimal human validation
- Guaranteed consistency in coding shipments
- Ability to quantify relative accuracy of shipment labels and compare within and across survey cycle

United States® Census Bureau

Future work

- Develop process to generate more gold-standard data for training and evaluation
- Continue investigating ways to tune models
- Experiment with different model families, including neural networks

United States® Census Bureau

Thank you!

Email: cecile.m.murray@census.gov

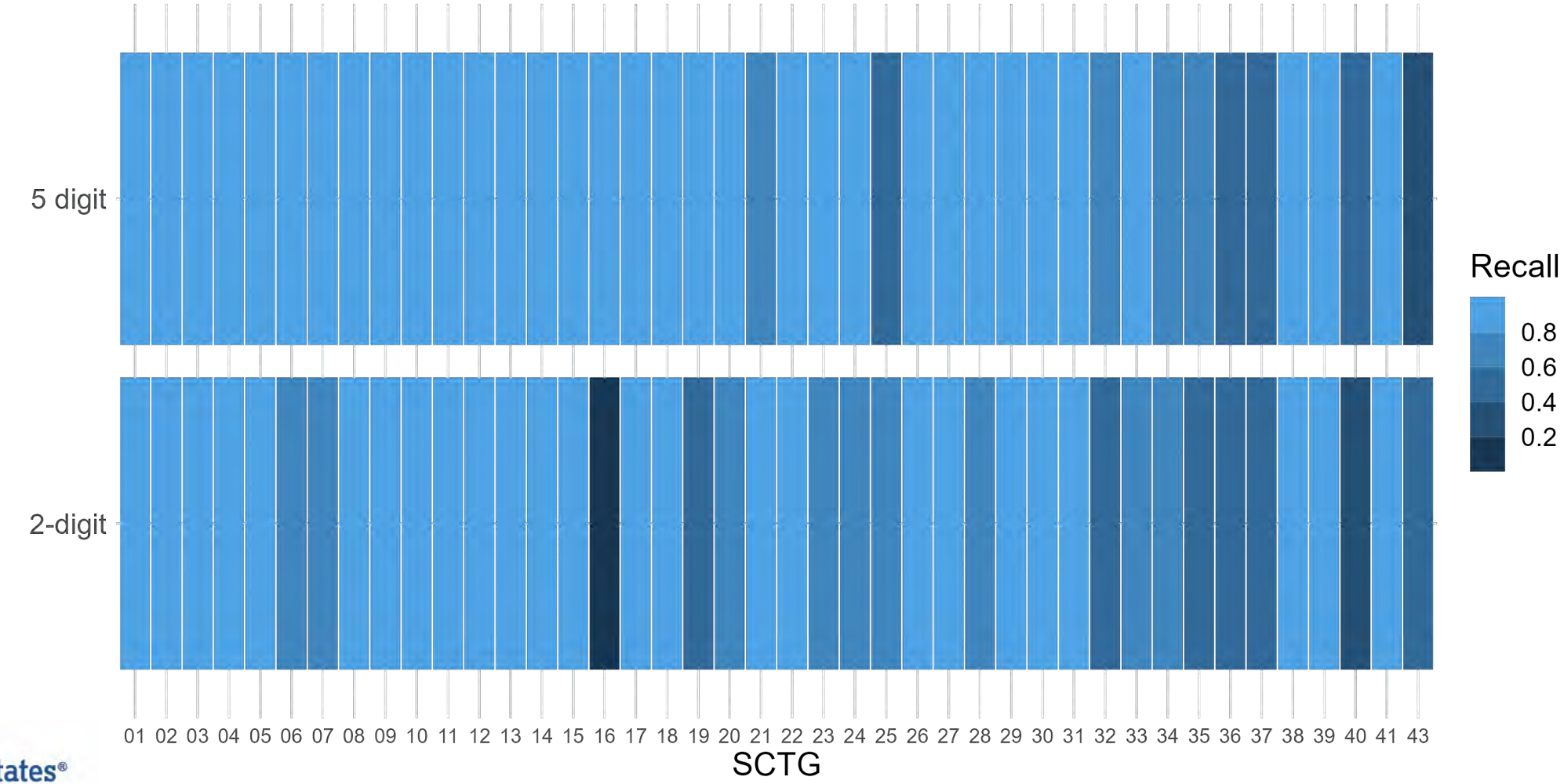
United States® Census Bureau

Results of hand coding

Data segment	# observations	5-digit model accuracy	2-digit model accuracy	2-digit accuracy of 5-digit model
SCTG directly assigned by ML	511	91%	90%	93%
2-digit SCTG assigned by ML, last 3 imputed	210	80%	68%	78%
SCTG assigned by imputation only	80	61%	44%	48%
Invalid/out of scope/too vague	81			

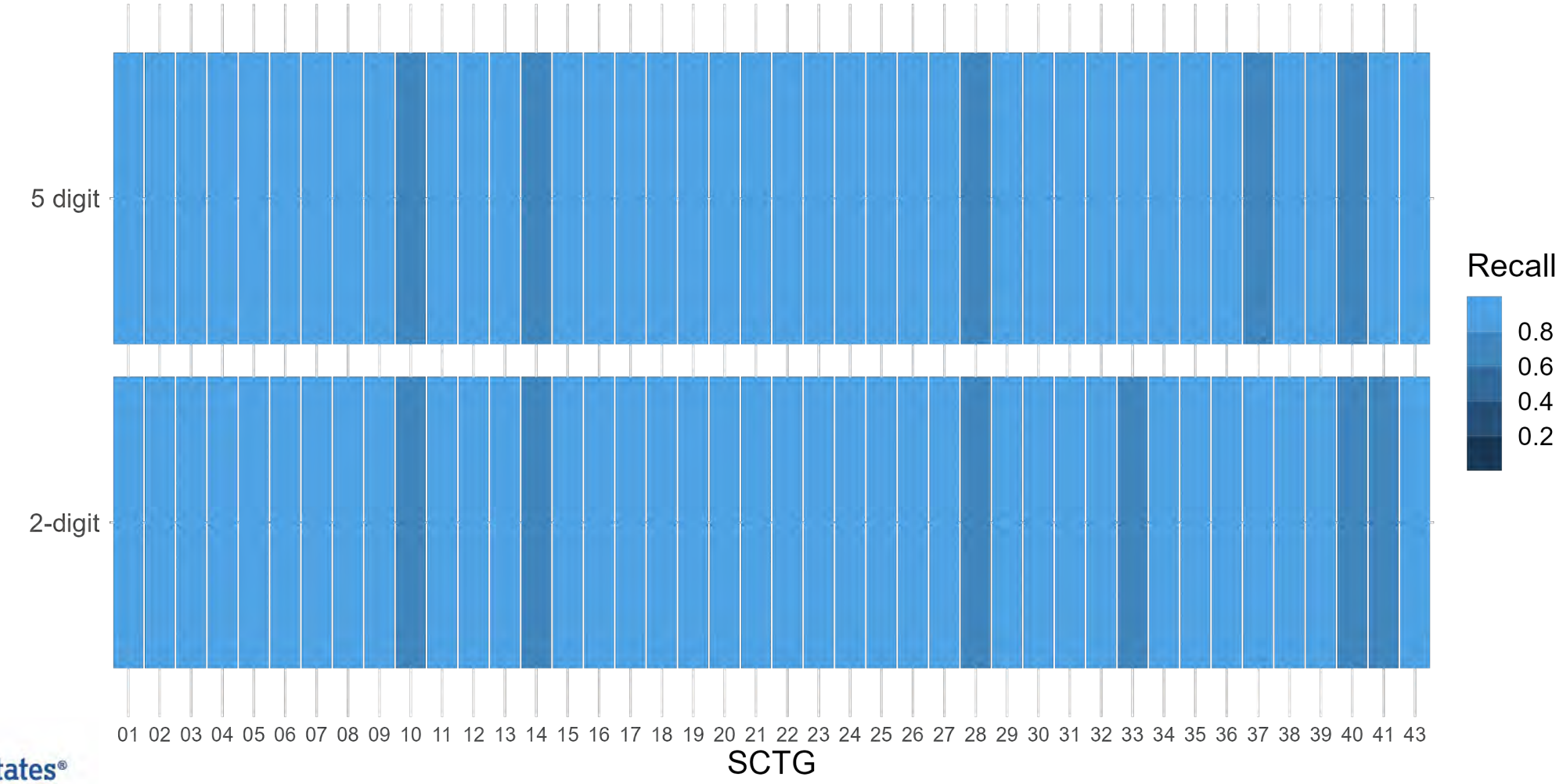
United States® Census Bureau

Hand-coding results by SCTG



United States® Census Bureau

Cross-validation results by SCTG



United States® Census Bureau