WEB SCRAPING FOR E-COMMERCE VALIDATION

Economics Indicator Division / New Product R&D

U.S. Census Bureau

Rebecca Weaver — New Products & Support Team Lead

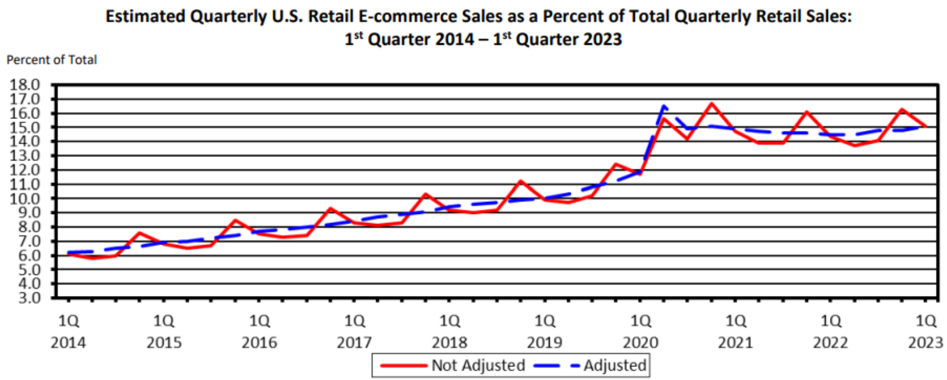
Any opinions and conclusions expressed herein are those of the author(s) and do not reflect the views of the U.S. Census Bureau. The Census Bureau has reviewed this data product to ensure appropriate access, use, and disclosure avoidance protection of the confidential source data (Project No. P-7504206, Disclosure Review Board (DRB) approval number: CBDRB-FY23-ESMD010-031).

CORRINA CALANOC
Georgetown University
Data Science & Analytics

INTRODUCTION & MOTIVATION

- **E-commerce** is a key component of the retail sector and is reported by the EID in the **Quarterly Retail E-Commerce** report

- These numbers are speculated to be **underestimates** due to **missing or inaccurate data**
  - 6,400 retailers unresponsive or report no e-commerce

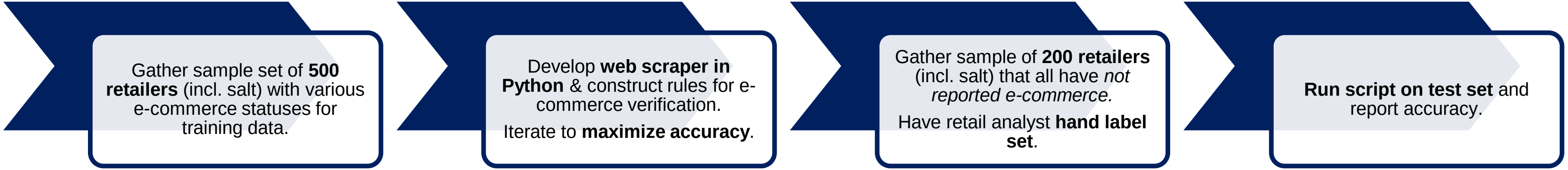


Estimated Quarterly U.S. Retail E-commerce Sales as a Percent of Total Quarterly Retail Sales: 1st Quarter 2014 – 1st Quarter 2023

GOAL: Given a set of retailers and their websites, use web scraping to indicate whether that retailer has e-commerce. This process will then be used to update the retailers’ e-commerce status in the Census database.

Any opinions and conclusions expressed herein are those of the author(s) and do not reflect the views of the U.S. Census Bureau.

PROJECT DETAILS & PROCESS



Gather sample set of **500 retailers** (incl. salt) with various e-commerce statuses for training data.

Develop **web scraper in Python** & construct rules for e-commerce verification.

Iterate to **maximize accuracy**.

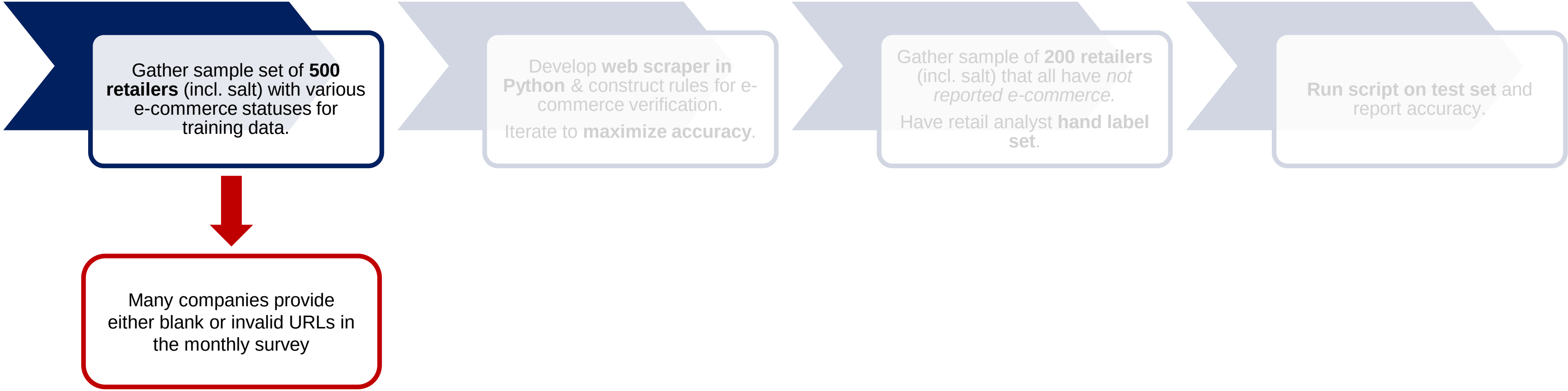
Gather sample of **200 retailers** (incl. salt) that all have *not reported e-commerce.*

Have retail analyst **hand label set**.

Run script on test set and report accuracy.

Any opinions and conclusions expressed herein are those of the author(s) and do not reflect the views of the U.S. Census Bureau.

CHALLENGES



Gather sample set of **500 retailers** (incl. salt) with various e-commerce statuses for training data.

Many companies provide either blank or invalid URLs in the monthly survey

Develop **web scraper in Python** & construct rules for e-commerce verification.

Iterate to **maximize accuracy**.

Gather sample of **200 retailers** (incl. salt) that all have *not reported e-commerce.*

Have retail analyst **hand label set**.

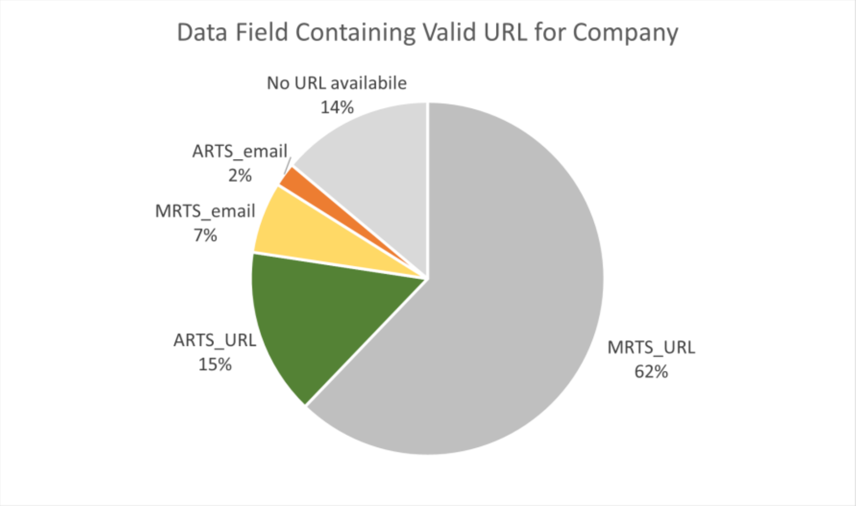
Run script on test set and report accuracy.

Any opinions and conclusions expressed herein are those of the author(s) and do not reflect the views of the U.S. Census Bureau.

RESULTS – URL VALIDATION

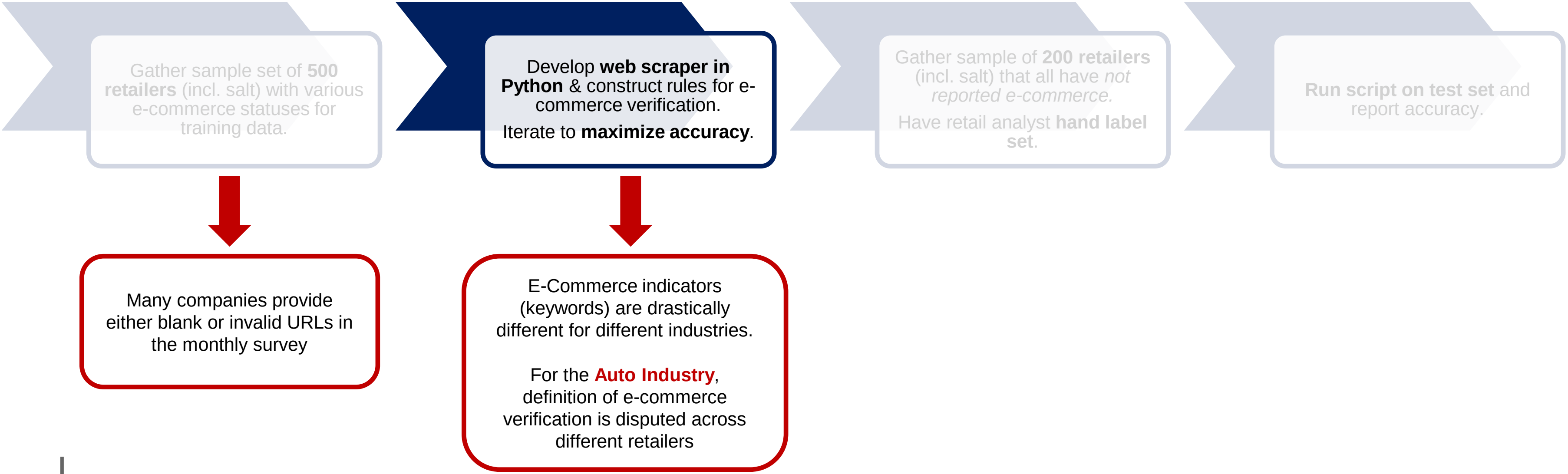
Out of **500 companies** in the Training Set:

- **38% of retailers** had invalid URLs from the monthly survey
- Out of these, the script was able to extract **valid URLs from other sources** for **63% of them**
- URL Validation can be its own standalone process



Any opinions and conclusions expressed herein are those of the author(s) and do not reflect the views of the U.S. Census Bureau.

CHALLENGES



Gather sample set of **500 retailers** (incl. salt) with various e-commerce statuses for training data.

Develop **web scraper in Python** & construct rules for e-commerce verification.

Iterate to **maximize accuracy**.

Gather sample of **200 retailers** (incl. salt) that all have *not reported e-commerce.*

Have retail analyst **hand label set**.

Run script on test set and report accuracy.

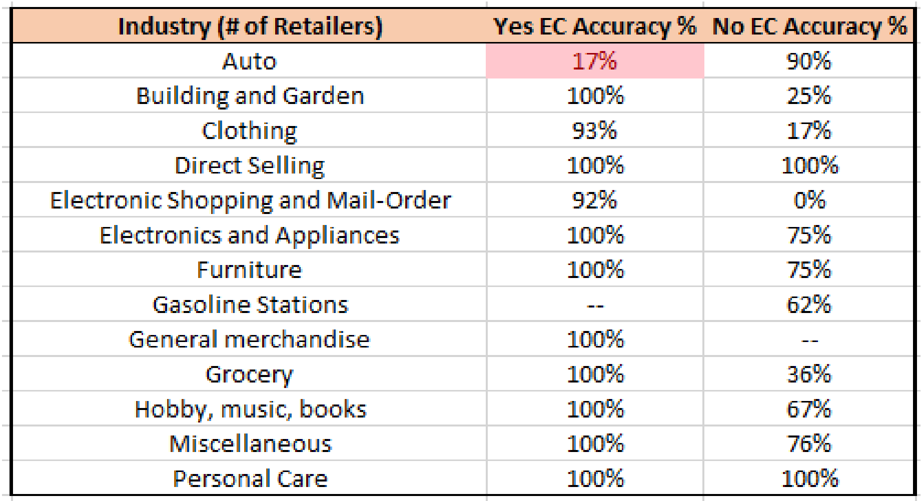
Many companies provide either blank or invalid URLs in the monthly survey

E-Commerce indicators (keywords) are drastically different for different industries.

For the **Auto Industry**, definition of e-commerce verification is disputed across different retailers

Any opinions and conclusions expressed herein are those of the author(s) and do not reflect the views of the U.S. Census Bureau.

RESULTS – E-COMMERCE INDICATOR



Industry (# of Retailers)	Yes EC Accuracy %	No EC Accuracy %
Auto	17%	90%
Building and Garden	100%	25%
Clothing	93%	17%
Direct Selling	100%	100%
Electronic Shopping and Mail-Order	92%	0%
Electronics and Appliances	100%	75%
Furniture	100%	75%
Gasoline Stations	--	62%
General merchandise	100%	--
Grocery	100%	36%
Hobby, music, books	100%	67%
Miscellaneous	100%	76%
Personal Care	100%	100%

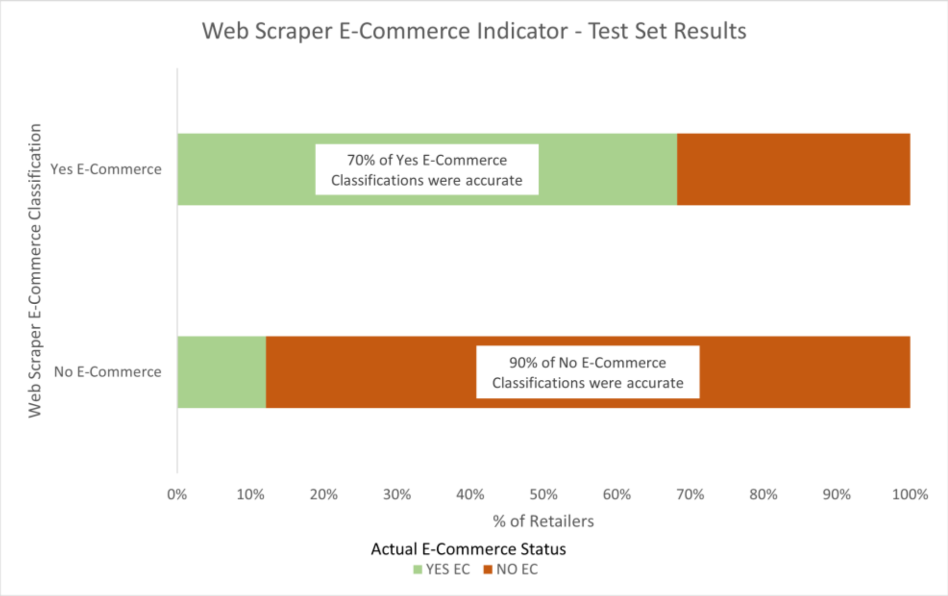
Out of **500 companies** in the Training Set:

- Web scraper had a **74% accuracy**
  - **87%** accuracy for **identifying e-commerce**
  - **63%** accuracy for **identifying No e-commerce**
- Script identified **corrections** to be made
  - **29%** of retailers that **consistently report no e-commerce** were found to have e-commerce
  - **9%** of retailers that are **consistently unresponsive** were found to have e-commerce

Any opinions and conclusions expressed herein are those of the author(s) and do not reflect the views of the U.S. Census Bureau.

RESULTS – E-COMMERCE INDICATOR

Out of **200 companies** in the Test Dataset, the web scraper had a **80% accuracy**



Any opinions and conclusions expressed herein are those of the author(s) and do not reflect the views of the U.S. Census Bureau.

NEXT STEPS & CONCLUSION

Next Steps:

- Run script on set of no e-commerce retailers & update database accordingly
- Periodically will run script in order to maintain e-commerce statuses

Future Enhancements:

- Implement parallel processing for get requests to improve on run time
- Utilize machine learning and NLP techniques for better keyword indicators

Any opinions and conclusions expressed herein are those of the author(s) and do not reflect the views of the U.S. Census Bureau.