Ensemble Modeling Techniques for NAICS Classification in the Economic Census

Federal Committee on Statistical Methodology
Machine Learning Applications
October 24, 2023

Daniel Whitehead
Brian Dumbacher
Economic Statistical Methods Division
U.S. Census Bureau

Disclaimer

Any opinions and conclusions expressed herein are those of the authors 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 used to produce this product [Data Management System (DMS) number: P-7504847, subproject P-7514952; Disclosure Review Board (DRB) approval number: CBDRB-FY23-ESMD002-034].

Outline

- Background: Slides 4 - 10
- BEACON Methodology: Slides 11 - 13
- Model Stacking: Slides 14 - 18
- Results: Slides 19 - 25
- Conclusions/Contacts: Slides 26 - 27

Background: North American Industry Classification System (NAICS)

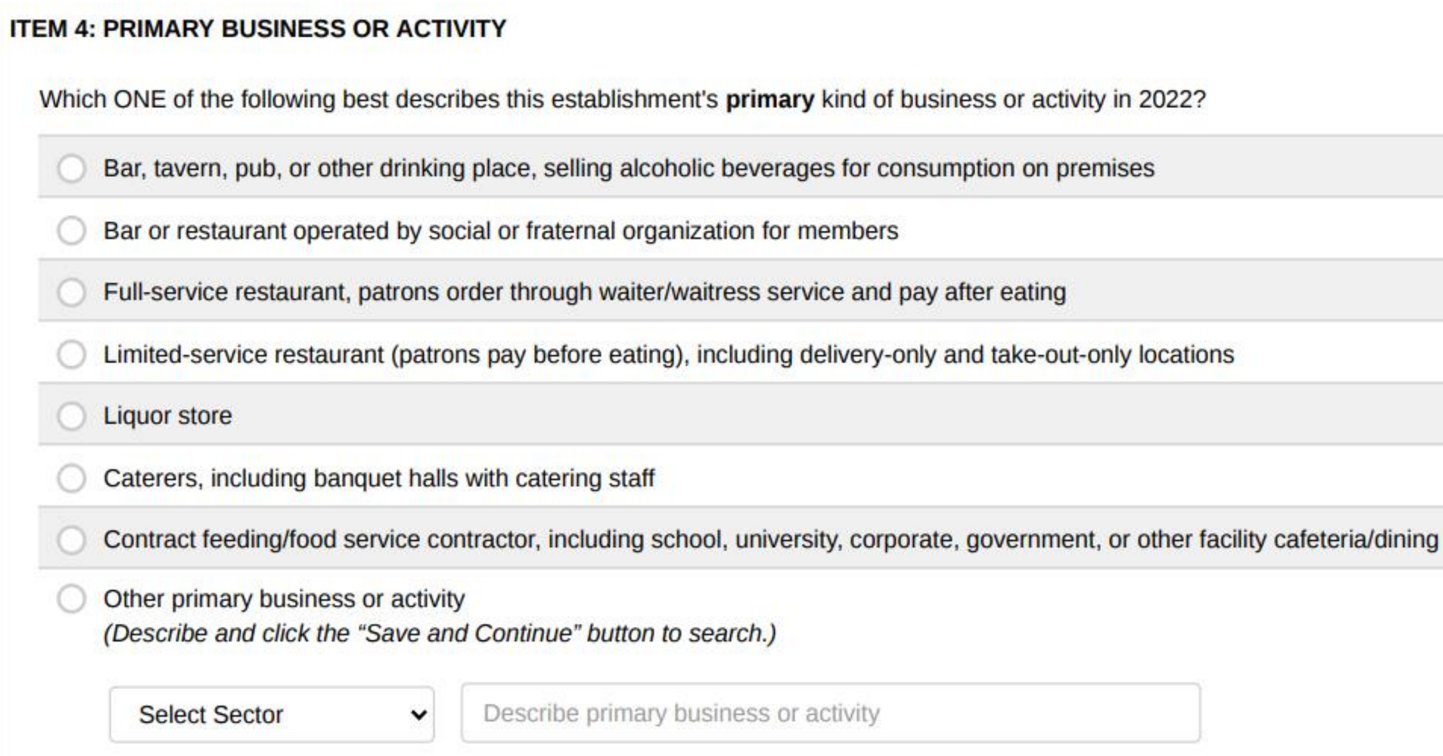
- U.S. Census Bureau classifies business establishments by NAICS code based on primary business activity

- NAICS is utilized throughout the survey life cycle
  - Sample selection
  - Data collection
  - Analytical review
  - Publication

- Hierarchical 6-digit coding structure
  - First two digits of NAICS code represent economic sector (**22** – Utilities)
  - Additional non-zero digits add industry detail (**221210** – Natural Gas Distribution)

Background: Primary Business or Activity Question from the Economic Census

- Question asks respondents to describe their business
- There are prelisted descriptions, but the respondent also has the option of writing in a business description
- Manual coding of write-in text is resource-intensive



ITEM 4: PRIMARY BUSINESS OR ACTIVITY

Which ONE of the following best describes this establishment's **primary** kind of business or activity in 2022?

○ Bar, tavern, pub, or other drinking place, selling alcoholic beverages for consumption on premises

○ Bar or restaurant operated by social or fraternal organization for members

○ Full-service restaurant, patrons order through waiter/waitress service and pay after eating

○ Limited-service restaurant (patrons pay before eating), including delivery-only and take-out-only locations

○ Liquor store

○ Caterers, including banquet halls with catering staff

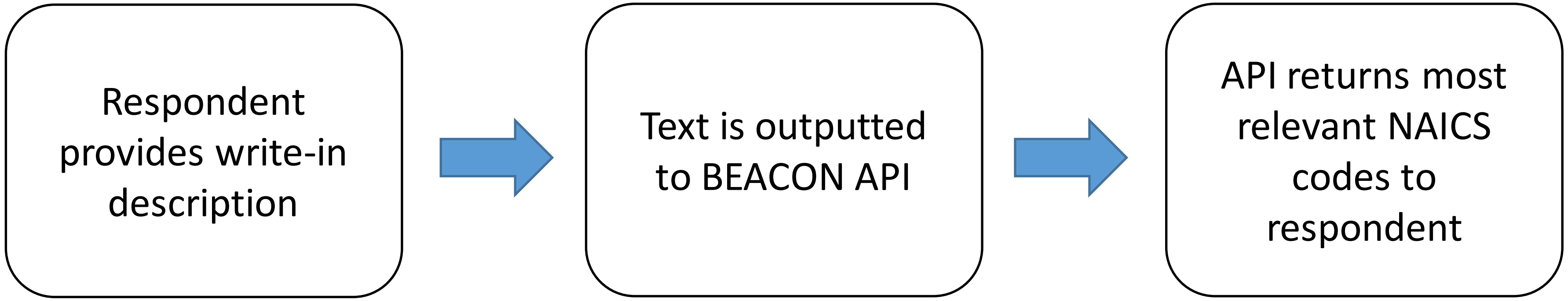
○ Contract feeding/food service contractor, including school, university, corporate, government, or other facility cafeteria/dining

○ Other primary business or activity
(Describe and click the "Save and Continue" button to search.)

Source: 2022 Economic Census

Background: What is BEACON?

- Business Establishment Automated Classification of NAICS
- A machine learning tool developed by the Economic Statistical Methods Division (U.S. Census Bureau) to classify NAICS for establishments based on a write-in business description



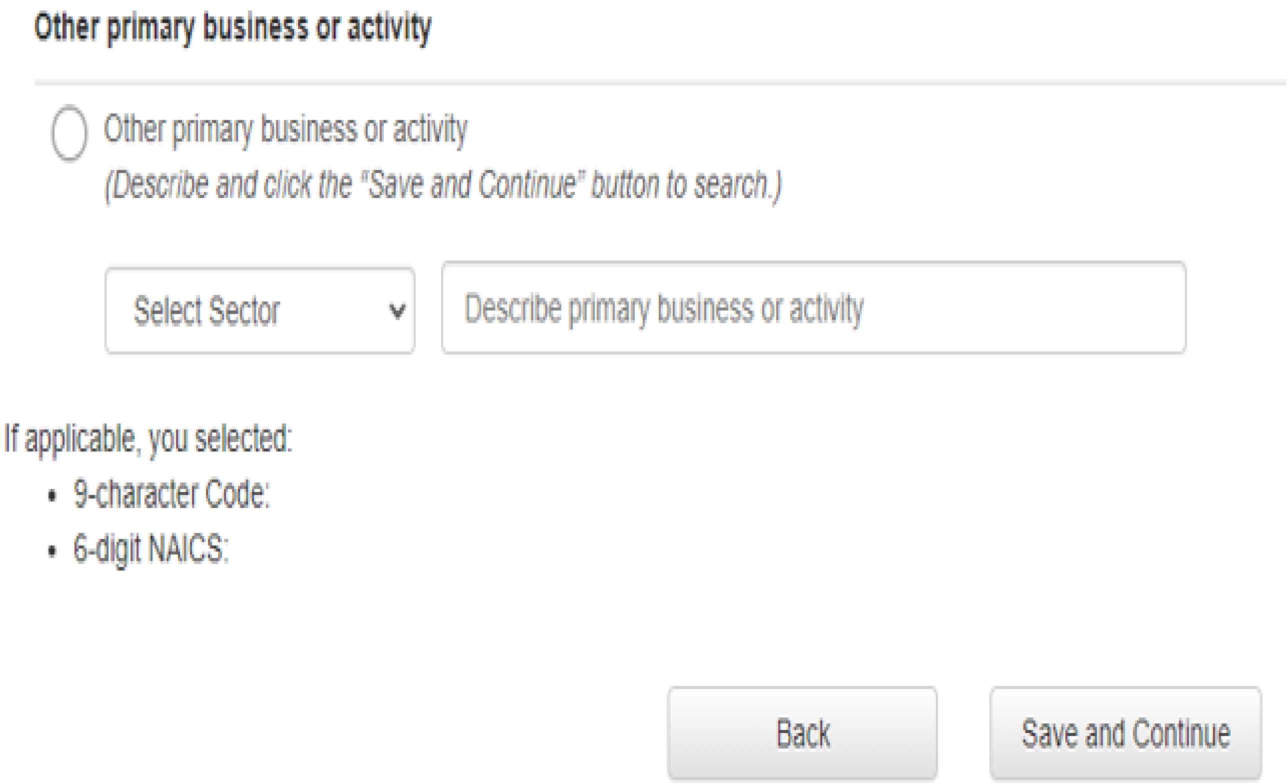
Respondent provides write-in description

Text is outputted to BEACON API

API returns most relevant NAICS codes to respondent

Background: Goals of BEACON

- Assist respondents in self-designating their NAICS codes
- Improve accuracy of self-designated NAICS codes
- Reduce manual coding of write-ins



Other primary business or activity

○ Other primary business or activity

(Describe and click the "Save and Continue" button to search.)

If applicable, you selected:

- 9-character Code:
- 6-digit NAICS:

Background: Training Data

- Historic write-in responses to the Economic Census (EC)
- Frequent write-in text that was autocoded during 2017 EC
- Business descriptions from IRS SS-4 forms*
- Classification Analytical Processing System (CAPS) items
- Harmonized System commodity descriptions
- Variables
  - Business description text
  - Corresponding NAICS code

Business Description Text	NAICS
This is a car dealership.	441110
R&D lab – medical/health	541715
we mainly repair furniture, some sales	811420

*IRS data used for internal statistical purposes only, in accordance with Title 26.

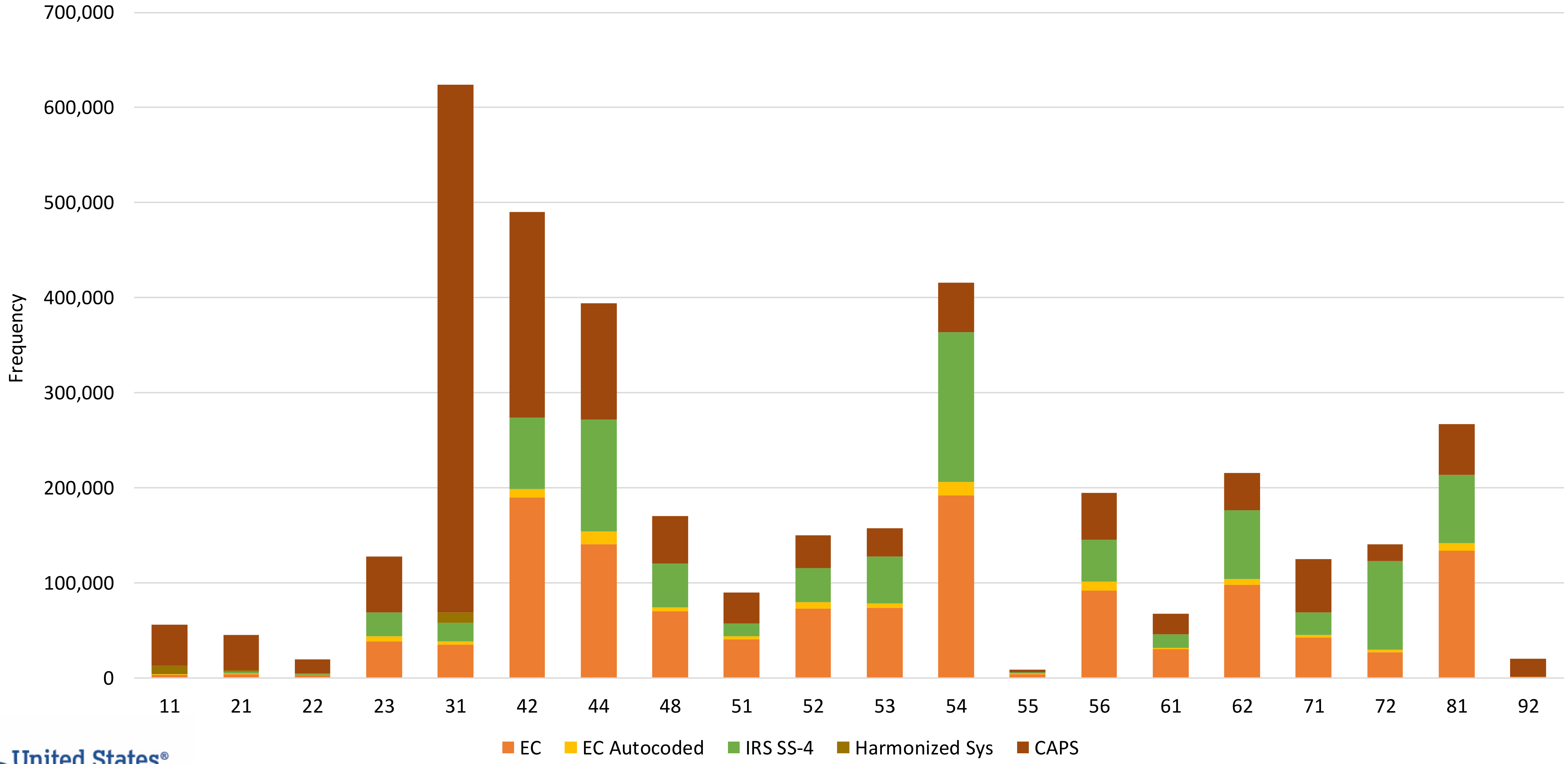
Background: Training Data

Data Source	Number of Observations	Advantages	Disadvantages
EC	~ 1,611,000 (single-unit)	• Represents target population • Reflects natural language	• Descriptions not perfectly classified • Descriptions contain misspellings
EC Autocoded	~ 98,000 *	• Improves consistency with autocoding during 2017 EC	• Relatively small data source
IRS SS-4	~ 865,000 (single-unit)	• Provides timely data • Reflects natural language	• Descriptions not perfectly classified • Descriptions contain misspellings
CAPS	~ 1,508,000 *	• Provides a rich vocabulary • Descriptions are classified correctly	• Does not always reflect natural language
Harmonized System	~21,000	• Provides examples of industry-specific abbreviations/terminology	• Relatively small data source • Does not always reflect natural language

*Includes duplicates and variations of original observations.

United States® Census Bureau

BEACON Training Data Breakdown by Sector and Source



Data Sources: Economic Census (2002–2022), 2021 Industry Classification Report, Internal Revenue Service SS-4 (2002–2016), Classification Analytical Processing System, Harmonized System

Methodology: Text Cleaning

- Removal of extraneous words/symbols
  - Remove extra white space and common “stop” words (“the”, “and”, “or”, etc.)
  - Account for numbers and punctuation
- Correct common misspellings
  - Map stems of misspelled words to stems of correctly spelled words
  - For example, “manifactur” → “manufactur”
- Stem
  - Apply prefix/suffix stripping rules to reduce number of word variations
  - For example, “manufacturing” → “manufactur”, “cars” → “car”
- Lemmatize
  - Map synonyms and abbreviations to a common concept
  - For example, “mfg” → “manufactur”, “auto” → “car”

Methodology: Dictionary

- Words and word combinations that BEACON recognizes

- All model features are based on the data dictionary

- Associations between words and NAICS codes in the training data influence predictions
  - “tutor” is highly associated with NAICS 611691 – Exam Preparation and Tutoring
  - “store” occurs in many NAICS codes and is therefore less predictive

Methodology: Model Ensemble

- Model ensemble
  - Information retrieval models look at how words, combinations, and entire descriptions are distributed across NAICS codes
  - Individual predictions are averaged, yielding relevance scores

- Relevance scores
  - Range in value between 0 and 100
  - Reflect how confident BEACON is that the NAICS code is correct

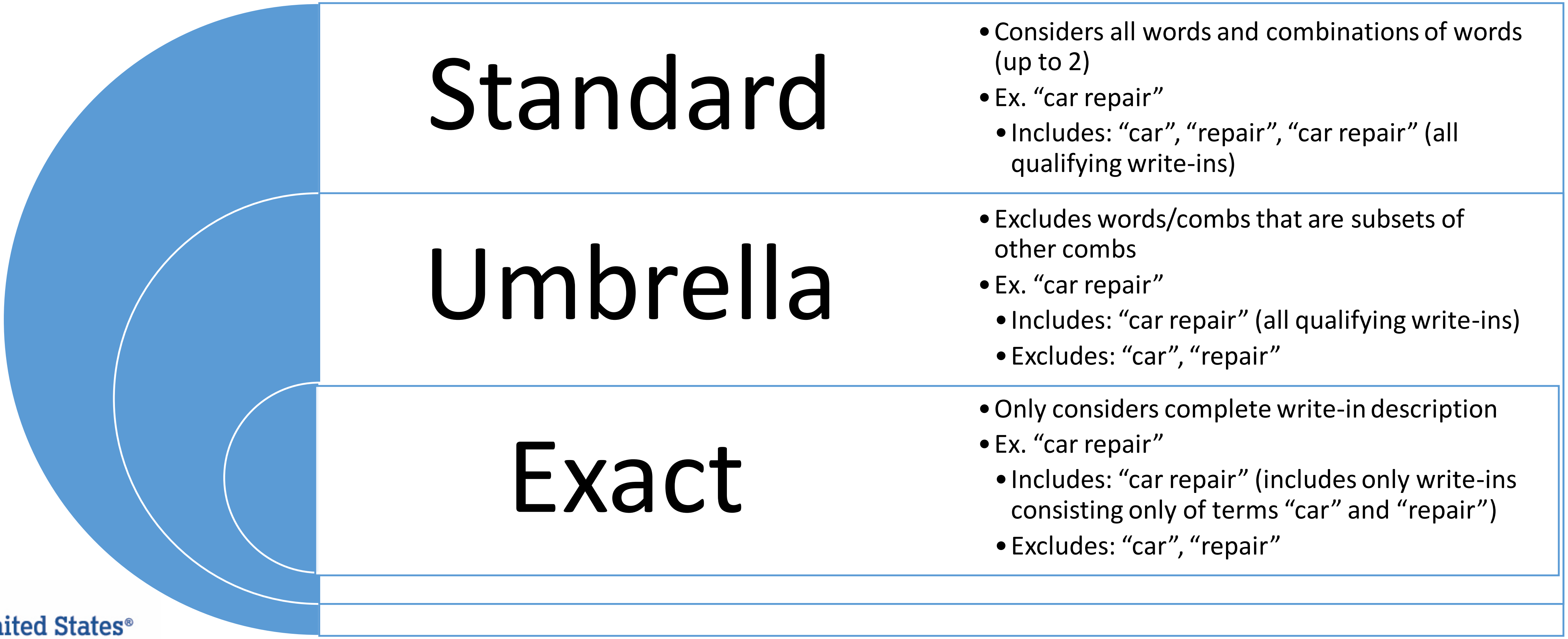
Model Stacking: Overview

- Separate prediction problem into two parts
  - Create multiple models to generate initial predictions
  - Use these predictions as inputs to meta-model

- Proposal:
  - Generate predictions from component models within BEACON
  - Refine initial predictions with predictions from meta-models

- *Sources:*
  - Todorovski, L. and Džeroski, S. (2003). Combining Classifiers with Meta Decision Trees. *Machine Learning*, *50*, 223–249. Netherlands: Kluwer Academic Publishers.
  - Merz, C. (1999). Using Correspondence Analysis to Combine Classifiers. *Machine Learning*, *36*, 33–58.

Model Stacking: Within BEACON



Standard

- Considers all words and combinations of words (up to 2)
- Ex. “car repair”
  - Includes: “car”, “repair”, “car repair” (all qualifying write-ins)

Umbrella

- Excludes words/combs that are subsets of other combs
- Ex. “car repair”
  - Includes: “car repair” (all qualifying write-ins)
  - Excludes: “car”, “repair”

Exact

- Only considers complete write-in description
- Ex. “car repair”
  - Includes: “car repair” (includes only write-ins consisting only of terms “car” and “repair”)
  - Excludes: “car”, “repair”

Model Stacking: Within BEACON

- “Standard” and “Umbrella” models
  - The NAICS distributions of the words/stems and word/stem combinations are averaged using “purity weights” that give more weight to the NAICS distributions of words that are more predictive.
  - The purity weight is a function of the maximum proportion.
- Final scores
  - The scores from the “Standard”, “Umbrella”, and “Exact” models are averaged
  - Three model weight parameters $w_{standard}$, w_{umb}, and w_{exact} (= 1 - w_{all} - w_{umb})

Model Stacking: Proposal

- Generate predictions from component models within BEACON
  - “Standard”
  - “Umbrella”
  - “Exact”

- Refine initial predictions with predictions from meta-models
  - Logistic Regression
  - Decision Tree
  - Random Forest

United States® Census Bureau

Model Stacking: Evaluation Metrics

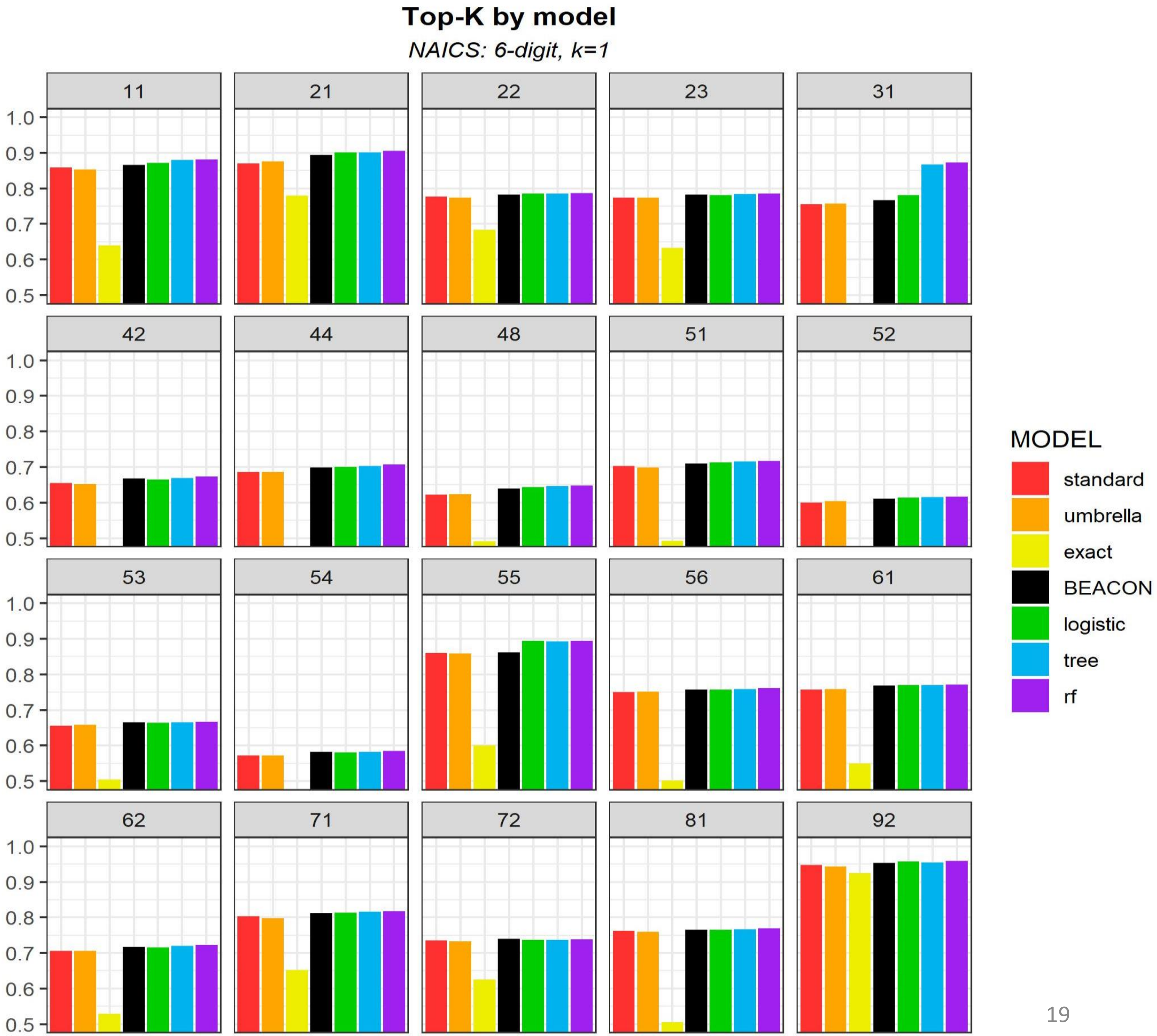
- Top-*k*
  - Measures success rate where success is % of times that true NAICS code is found in top-*k* predictions within sector
  - Evaluated for *k* = 1, 2, 3, 4, 5

- F_1 score
  - Harmonic mean of precision and recall

Results

Exact model is least accurate of individual component models.

BEACON model tends to produce more accurate predictions than any of its component models.



Data Sources: Economic Census (2002–2022), 2021 Industry Classification Report, Internal Revenue Service SS-4 (2002–2016), Classification Analytical Processing System, Harmonized System

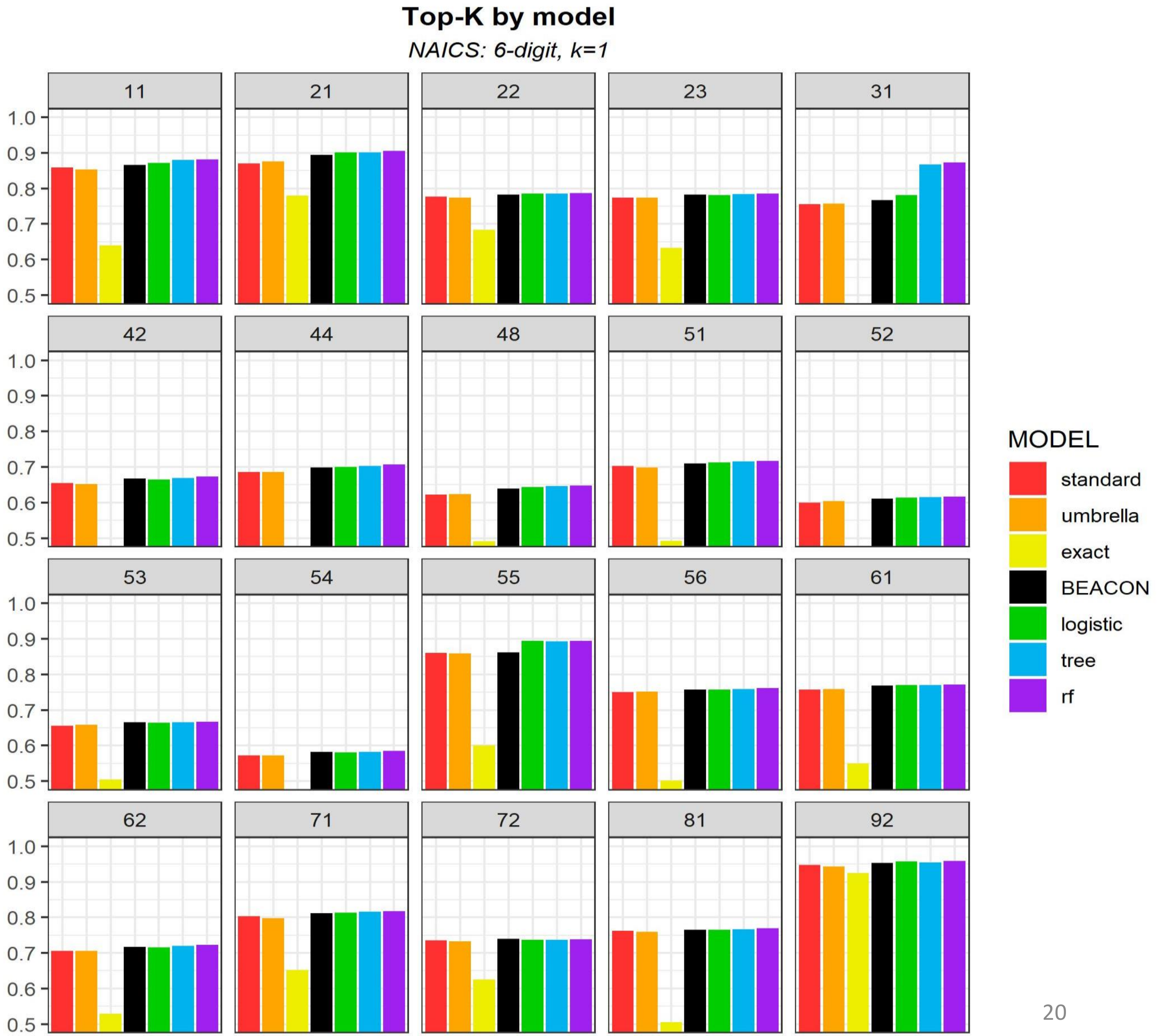
Results

Meta-models offer similar potential for incremental improvement.

Random forest and decision trees methods appear most promising.

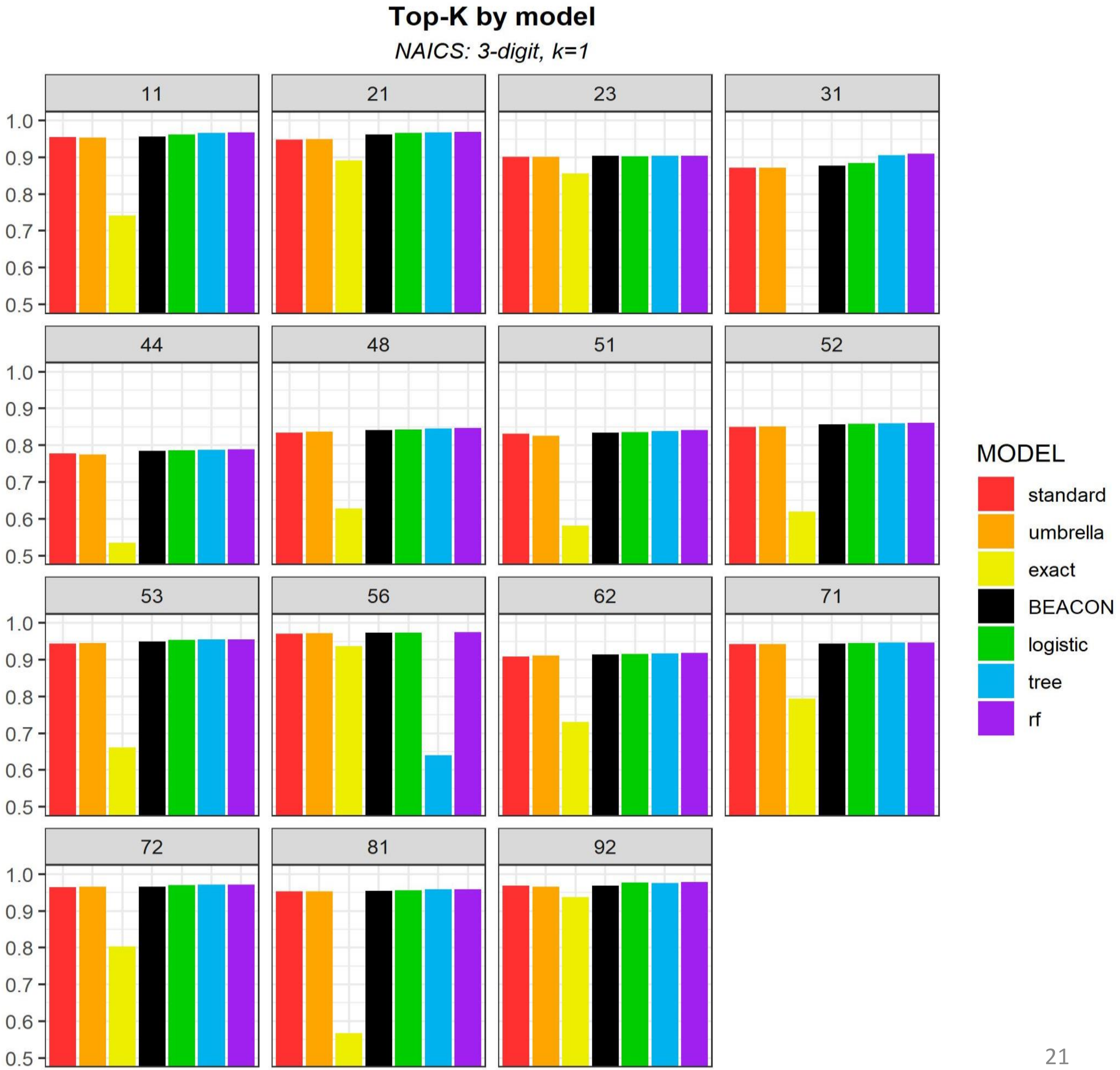
Potential improvement in manufacturing sector (31) is encouraging.

Data Sources: Economic Census (2002–2022), 2021 Industry Classification Report, Internal Revenue Service SS-4 (2002–2016), Classification Analytical Processing System, Harmonized System



Results

Similar results were found at other NAICS levels.



Data Sources: Economic Census (2002–2022), 2021 Industry Classification Report, Internal Revenue Service SS-4 (2002–2016), Classification Analytical Processing System, Harmonized System

Results

Meta-models improved performance in sectors where BEACON was already doing well.

F_1 score of BEACON in NAICS codes where meta-model outperformed BEACON is higher than that of NAICS codes where BEACON outperformed meta-model.

Median BEACON F_1 score: 3-digit NAICS codes

	RF	Tree	LR
NAICS codes where F_1 score of meta-model > F_1 score of BEACON	0.93	0.93	0.95
NAICS codes where F_1 score of BEACON > F_1 score of meta-model	0.85	0.85	0.89

Data Sources: Economic Census (2002–2022), 2021 Industry Classification Report, Internal Revenue Service SS-4 (2002–2016), Classification Analytical Processing System, Harmonized System

United States® Census Bureau

Results

Meta-models improved performance in sectors where BEACON was already doing well.

F_1 score of BEACON in NAICS codes where meta-model outperformed BEACON is higher than that of NAICS codes where BEACON outperformed meta-model.

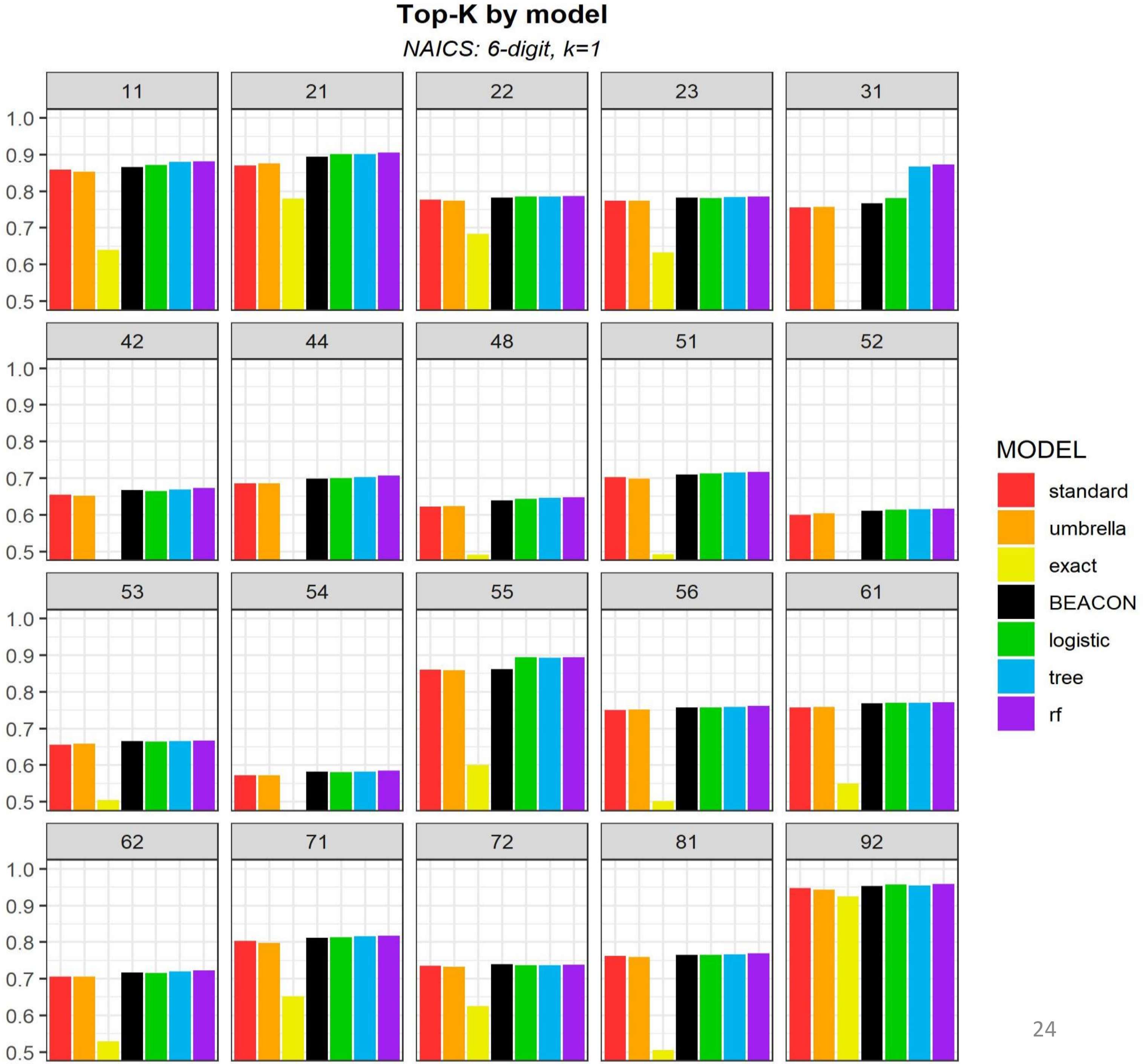
Median BEACON F_1 score: 6-digit NAICS codes

	RF	**Tree**	**LR**
NAICS codes where F_1 score of meta-model > F_1 score of BEACON	0.89	0.89	0.89
NAICS codes where F_1 score of BEACON > F_1 score of meta-model	0.85	0.86	0.85

Data Sources: Economic Census (2002–2022), 2021 Industry Classification Report, Internal Revenue Service SS-4 (2002–2016), Classification Analytical Processing System, Harmonized System

Results

Meta-models may offer more potential for providing a single predicted NAICS than for providing multiple NAICS codes to a respondent.



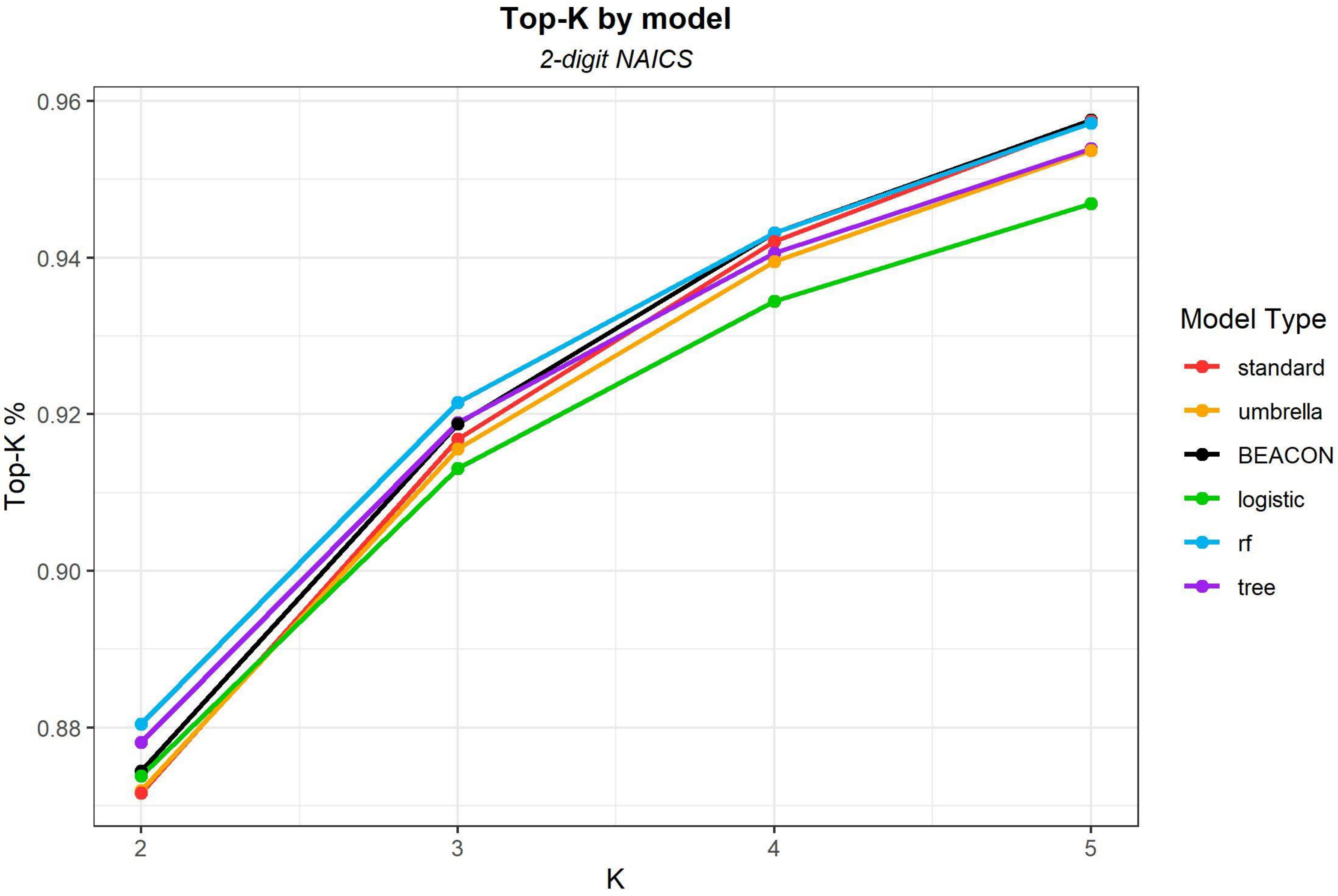
Data Sources: Economic Census (2002–2022), 2021 Industry Classification Report, Internal Revenue Service SS-4 (2002–2016), Classification Analytical Processing System, Harmonized System

Results

BEACON performed as well or better than meta-models as top-*k* goal was relaxed.

BEACON is well-equipped at providing multiple NAICS codes to respondent.

Data Sources: Economic Census (2002–2022), 2021 Industry Classification Report, Internal Revenue Service SS-4 (2002–2016), Classification Analytical Processing System, Harmonized System



Conclusions

- Meta-models performed best, compared to BEACON, at predicting single best NAICS code.

- BEACON performed as well or better than meta-models when goal was to include best NAICS code as one of several potential NAICS codes.

- Meta-models may be helpful in manufacturing sector, which has more NAICS codes than any other sector.

Contacts

- **Email:** Daniel.Whitehead@Census.gov
- **Email:** Brian.Dumbacher@Census.gov