Using Machine Learning to Assign North American Industry Classification System Codes to Establishments Based on Business Description Write-Ins

Joint Statistical Meetings

Denver, CO

July 29, 2019

Brian Dumbacher and Anne Russell

U.S. Census Bureau

United States Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Outline

- Background
- Write-in data
- Machine learning methodology
- Model evaluation
  - 2012 Economic Census
  - 2017 Economic Census
- Future Work

North American Industry Classification System

- U.S. Census Bureau classifies business establishments according to NAICS based on primary economic activities
- NAICS is utilized throughout the survey life cycle
  - Sample selection
  - Data collection
  - Publication
- Hierarchical six-digit coding scheme
  - First two digits of NAICS code represent economic sector (22 – Utilities)
  - Additional non-zero digits add industry detail (221210 – Natural Gas Distribution)

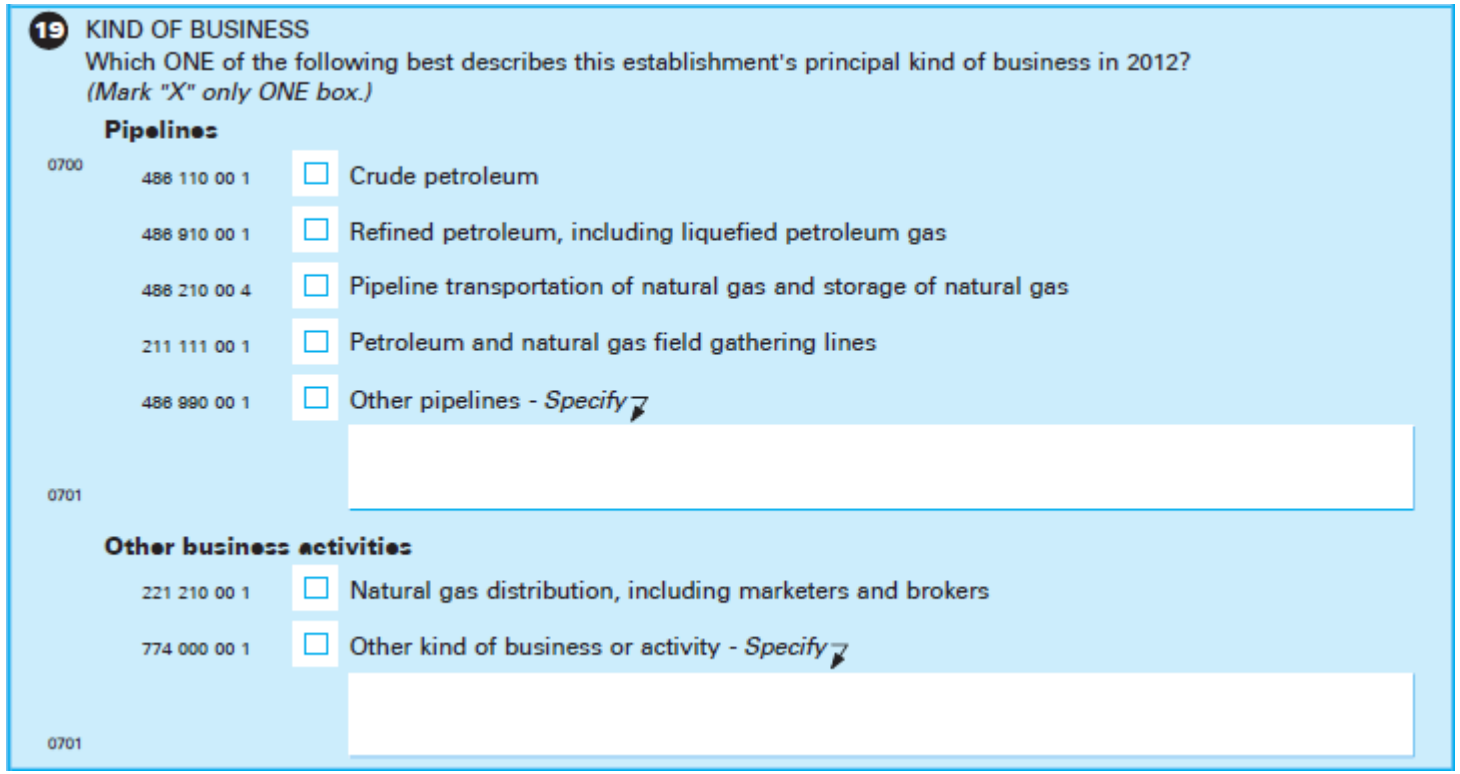
Economic Census

- Conducted every five years for years ending in “2” and “7”
- Extensive survey of approximately eight million establishments that covers most industries and all geographic areas of the U.S.
- Key statistics include
  - Total number of establishments
  - Total number of employees
  - Value of sales, shipments, receipts, and revenue
  - Total annual payroll
- Data products are presented by NAICS and geography

Self-Designated Kind of Business Question from 2012 Economic Census

- Question asks respondents to describe their business
- Respondent has the option to write in a business description if none of the checkbox descriptions is accurate
- Clerical analysis of write-in text and manual NAICS assignment are resource-intensive tasks

Source: 2012 Economic Census



19 KIND OF BUSINESS

Which ONE of the following best describes this establishment's principal kind of business in 2012?
(Mark "X" only ONE box.)

Pipelines

0700

488 110 00 1	☐	Crude petroleum
488 910 00 1	☐	Refined petroleum, including liquefied petroleum gas
488 210 00 4	☐	Pipeline transportation of natural gas and storage of natural gas
211 111 00 1	☐	Petroleum and natural gas field gathering lines
488 990 00 1	☐	Other pipelines - *Specify*

0701

Other business activities

221 210 00 1	☐	Natural gas distribution, including marketers and brokers
774 000 00 1	☐	Other kind of business or activity - *Specify*

0701

NAICS Autocoder for New Establishments

- Developed in collaboration with the Internal Revenue Service (IRS) and Social Security Administration
- Business names and descriptions come from the IRS's SS-4 form, which is used to apply for an Employer Identification Number
- Based on dictionary of words, two-word sequences (bigrams), and complete write-ins that occur frequently and are highly associated with certain NAICS codes
- Logistic regression model with dictionary mapping percentages as predictors

Other Autocoding Efforts

- NAICS autocoding for the 2017 Economic Census
  - Compare write-in text to a look-up list of 5,000 descriptions and their associated NAICS code
  - If exact match, then assign a NAICS code
  - Able to assign NAICS code for 68,897 of 511,251 write-ins
- "Throw-away" write-in identification for the 2017 Economic Census
  - Compare write-in text to a look-up list of text not predictive of NAICS
  - If exact match, then flag to process separately
  - Example text includes "NA" and "business closed"

Write-In Data

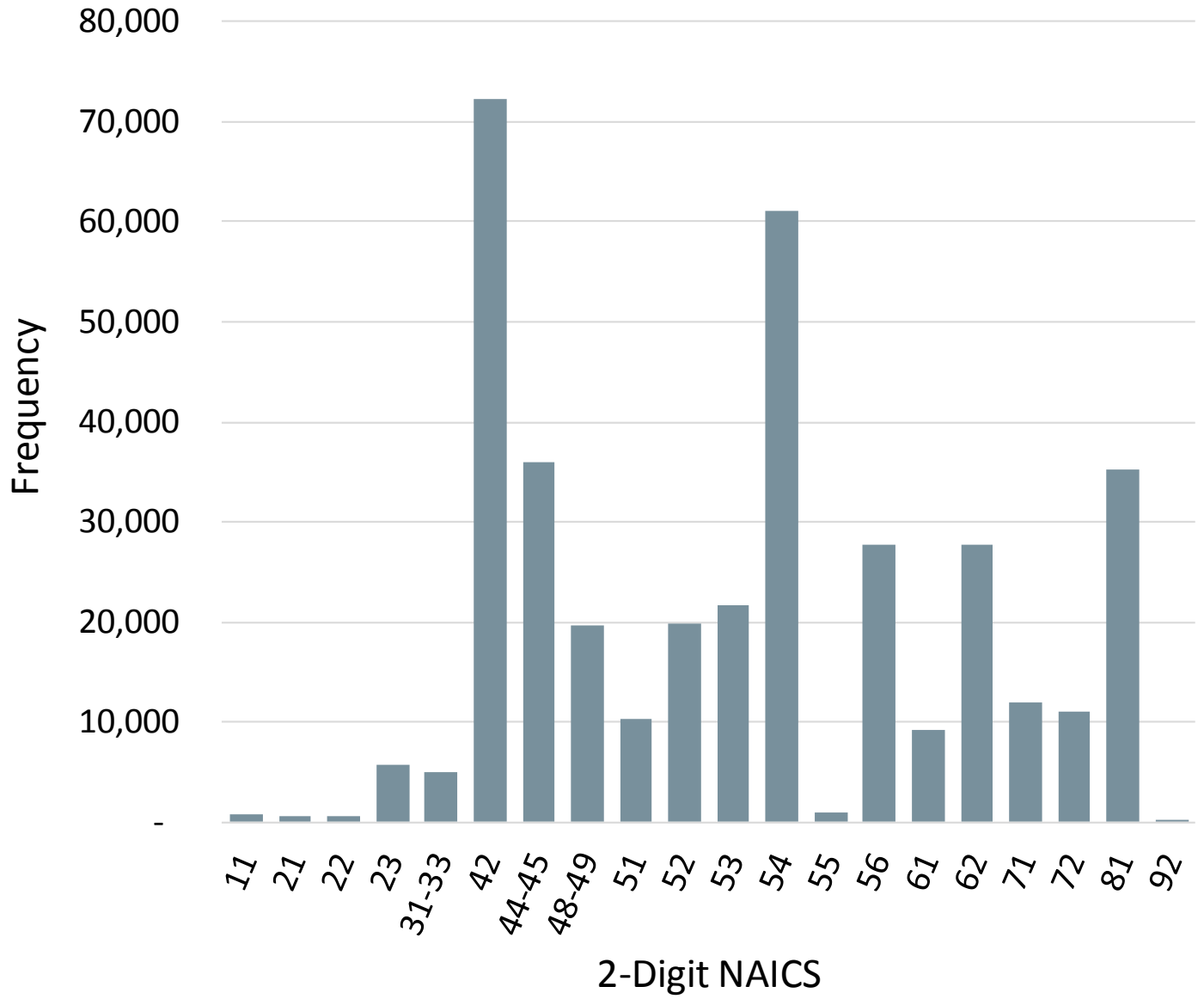
- Self-designated kind of business write-in observations from the 2012 Economic Census
  - Observations with throw-away write-in text are removed
  - Dataset covers all 20 sectors of the economy
  - **377,708** observations
- Other text variables besides write-in text
  - Business name
  - Line label (checkbox description associated with the write-in text box)

Breakdown of 2012 Write-In Dataset by 2-Digit NAICS Code

The four most frequently occurring 2-digit NAICS codes are:

- 42 (wholesale trade)
- 44-45 (retail trade)
- 54 (professional, scientific, and technical services)
- 81 [other services (except public administration)]

Source: 2012 Economic Census



Bag of Words Approach to Text Classification

- Models based on the occurrences of individual words and bigrams in the text variables
- Model predictors, or features, are binary indicators of all the words and bigrams appearing in the dataset
  - Equals 1 if word or bigram is in the text
  - Equals 0 otherwise
- Features created separately for write-in, business name, and line label

Text Variable Standardization

- Convert to lowercase
- Deal with punctuation
- Remove extra whitespace
- Remove common English “stop” words
- Fictional example
  - Original: `Sea-Doo and Jet Ski sales,PARTS & service.`
  - Standardized: `sea doo jet ski sales parts service`
- Determine words and bigrams based on standardized text

Machine Learning Algorithms

- Two commonly used learning algorithms for text classification
  - Naïve Bayes
    - Bernoulli implementation (suited for binary features)
    - Smoothness parameter – α
  - Logistic regression
    - One-versus-rest multiclass classification with L2 penalty
    - Inverse of regularization strength parameter – C
- Stratified 5-fold cross-validation with a grid search to optimize parameter values

Model Evaluation – 2012 Economic Census

- Randomly split **377,708** observations into training and test sets
  - Fit models using training set
  - Apply models to test set
- Stratified simple random sample with strata defined by 2-digit NAICS and sampling fraction equal to 90 percent
- Training set has **339,936** observations
- Test set has **37,772** observations

Cross-Validated Parameter Values and Test Set Accuracies

- Given the same features, logistic regression achieves higher accuracy than naïve Bayes
- Test set accuracies greater than 70 percent are highlighted
- Logistic regression with WI, BN, and LL features is the best

Source: 2012 Economic Census

Learning Algorithm	Text Features	Parameter Value (CV)	Test Set Accuracy
Naïve Bayes	WI	$\alpha = 0.1$	0.6424
Naïve Bayes	WI, BN	$\alpha = 0.1$	0.6593
Naïve Bayes	WI, LL	$\alpha = 0.2$	0.7147
Naïve Bayes	WI, BN, LL	$\alpha = 0.1$	0.7336
Logistic Regression	WI	$C = 1.5$	0.6454
Logistic Regression	WI, BN	$C = 1$	0.6866
Logistic Regression	WI, LL	$C = 1.5$	0.7483
Logistic Regression	WI, BN, LL	$C = 1.5$	0.7697

WI – write in

BN – business name

LL – line label

CV – cross-validation

Confusion Matrix for Best Logistic Model

True 2-Digit NAICS Code \ Predicted 2-Digit NAICS Code	11	21	22	23	31-33	42	44-45	48-49	51	52	53	54	55	56	61	62	71	72	81	92
11	31	0	0	0	0	13	5	3	0	5	6	4	0	4	1	1	4	0	1	0
21	0	25	1	0	2	2	2	3	0	5	1	13	0	5	0	0	0	0	0	0
22	0	0	44	0	0	5	2	4	0	2	5	4	0	3	0	0	0	0	0	0
23	1	4	2	271	3	32	45	8	7	12	52	46	1	65	1	0	5	2	24	0
31-33	0	0	0	7	240	97	36	4	8	3	3	42	0	17	0	4	5	6	29	0
42	8	7	4	28	70	6191	546	33	27	28	21	70	6	34	7	5	9	63	62	0
44-45	8	0	2	35	29	840	2217	18	25	39	32	62	2	45	8	21	18	82	112	0
48-49	3	2	2	7	6	36	8	1723	3	38	23	16	1	46	1	10	11	2	22	0
51	0	0	1	7	7	23	15	1	681	11	4	164	1	31	21	5	34	3	25	0
52	4	16	5	9	1	14	10	14	6	1630	141	60	20	13	0	22	3	4	12	0
53	11	8	1	30	4	34	29	37	4	167	1679	46	7	32	4	13	22	28	21	0
54	5	20	1	54	64	171	44	19	164	169	68	4771	11	221	63	98	39	15	114	1
55	0	0	0	3	0	2	3	0	1	16	6	4	43	6	0	2	2	3	1	0
56	4	3	4	67	16	74	44	55	43	43	38	282	6	1855	16	56	33	38	78	15
61	2	0	0	1	1	8	12	6	8	2	4	43	0	10	719	40	34	2	37	0
62	0	0	0	1	3	7	10	7	5	14	11	69	2	35	42	2459	12	3	98	1
71	10	0	0	4	4	9	22	6	63	4	16	39	5	30	37	17	848	24	52	0
72	2	0	0	2	4	73	58	4	0	8	18	4	3	7	1	6	19	880	16	0
81	1	0	2	32	37	53	91	35	28	25	48	108	2	71	57	106	51	12	2757	1
92	0	0	0	0	0	0	1	0	0	1	0	0	0	5	0	1	0	0	3	10

Source: 2012 Economic Census

Model Evaluation – 2017 Economic Census

- Pulled **226,124** write-in observations from the 2017 Economic Census database
- Distribution of 2-digit NAICS similar to that in 2012 dataset
- Fit logistic regression models using full 2012 data and apply to 2017 data
  - Features: WI, BN, LL – accuracy of 0.4387
  - Features: WI, BN – accuracy of 0.6118
- Differences in line label wording between 2012 and 2017 could explain underperformance of model with WI, BN, and LL features

Future Work

- More advanced machine learning algorithms
- Different text analytics techniques
- Prediction at a more detailed NAICS level
- Combining data from multiple Economic Census years
- Non-text predictors
  - Class of customer information to help distinguish retail from wholesale
  - Other NAICS predictions (for example, naïve Bayes prediction and estimated 2-digit NAICS at time of questionnaire mail-out)

Contact Information

- Brian.Dumbacher@census.gov
- Anne.Sigda.Russell@census.gov