Using Machine Learning Approaches to Improve Industry Coding in IRS Master File Data

IRS
Research, Applied Analytics & Statistics
STRATEGY & BUSINESS SOLUTIONS

September 6, 2019

Introduction

- NAICS Codes comprise a business classification system based on industry production processes

- The system was developed by OMB through their Economic Classification Policy Committee (ECPC) in collaboration with Bureau of Economic Analysis (BEA), Bureau of Labor Statistics (BLS), and Census Bureau

- The US implemented NAICS Codes in 1997 replacing the older Standard Industrial Classification (SIC) system

- NAICS Code updates every 5 years (years ending in 2 and 7)

- In the US, NAICS Codes are used to estimate Industry Statistics such as GDP, Gross Output, Employment, and Input-Output Accounts

NAICS Codes at IRS

- NAICS Codes are self-reported (since 1985) on tax forms so they are subject to error
  - Forms SS-4 - Application for Employer Identification Number
  - Forms Schedule C - Profit or Loss from Business (Sole Proprietorship)
  - Forms 1120 - US Corporation Income Tax Return Schedule K – Other Information
  - Forms 1065 - US Return of Partnership Income
- The goal of this project is to develop effective predictive models for NAICS Codes using IRS administrative data. This project uses two parallel approaches:
  - Supervised models – CART, Random Forests, Boosted Trees (XGBoost)
  - Unsupervised models – recommender algorithms
- Initial work is focused on Forms 1040 (individual Sch C) and Forms 1120 (corporate)

NAICS Code structure

- NAICS Codes have a six digit hierarchal structure. The table below summarizes the information contained in the code reading the code from left to right

Economic Sector	1-2
Industry Sub-Sector	3
Industry Group	4
NAICS Industry	5
National Industry	6

NAICS Code Errors

- Types of coding errors include:
  - Missing or invalid codes entered (“noninformative”)
  - While technically valid, code 999999 (“Other”) is usually misapplied and is functionally the same as a missing code
  - Valid code entered but incorrect for entity
  - Codes may be partially correct (Economic Sector correct but Industry subsector incorrect)

- SOI manually validates NAICS Codes on their micro data

- We take SOI validated codes as ground truth for supervised model development and for unsupervised model testing

Data

- The SOI microdata is a stratified probability sample with strata based on presence or absence of a tax form or schedule, and various income factors or other measures of economic size

- 10 years of SOI microdata with NAICS Code corrections:
  - Forms 1040 Tax Years 2007 – 2016
  - Forms 1120 Tax Years 2006 – 2015

- Merged administrative data from IRTF/BRTF tables
  - NAICS Codes as filed by the taxpayer (1040 Sch C filers may have more than one Sch C)
  - Business descriptions

NAICS error rates -- 1040

- The rate at which SOI corrects the economic sector (first two digits) of NAICS codes has been stable over time

SOI Year	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Correction Rate	22.0%	21.8%	22.0%	21.4%	21.0%	21.4%	22.0%	22.5%	20.5%	22.0%

- The type of corrections has evolved, however. Increasingly, taxpayers have not identified a NAICS code on their returns. The table below presents the percent of taxpayers in the SOI dataset that did not identify a valid NAICS code.

SOI Year	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Invalid Rate	9.5%	5.1%	4.7%	4.0%	6.5%	10.4%	11.3%	12.2%	12.4%	12.5%

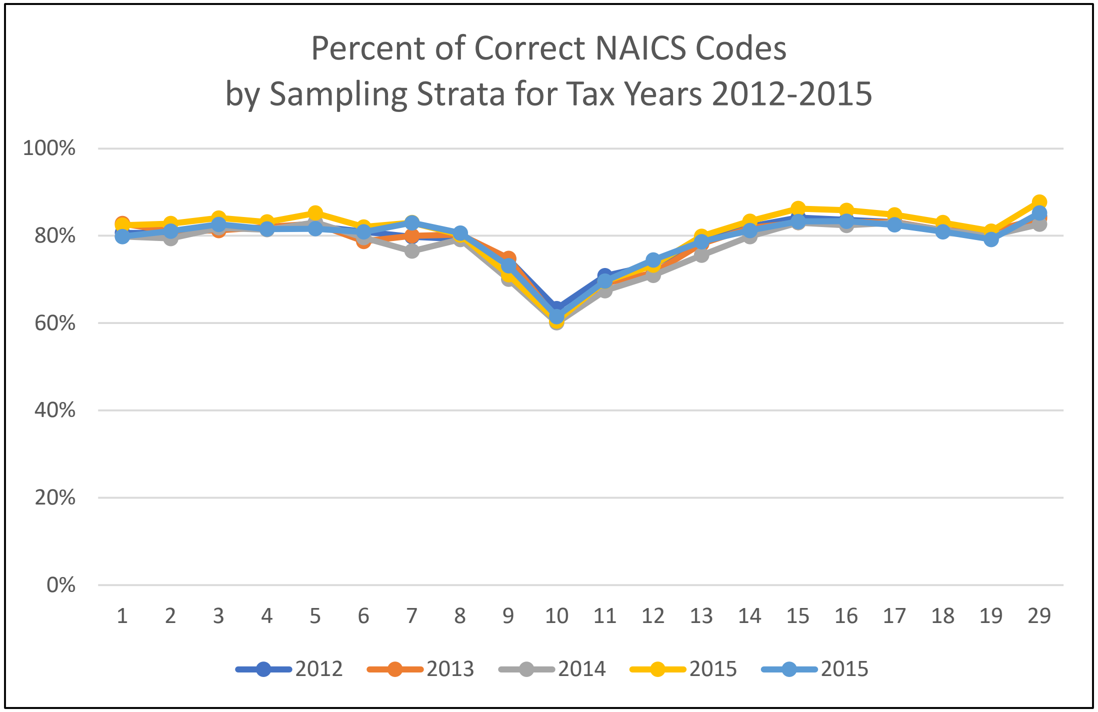
NAICS error rates -- 1120

- The rate at which SOI corrects the economic sector (first two digits) of NAICS codes trended upward

SOI Year	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Correction Rate	15.2%	15.2%	15.4%	16.1%	16.6%	16.8%	16.8%	18.0%	19.1%	19.7%

- As opposed to 1040, Corporate filers tend to identify a valid NAICS code. In all year, there were < 2% of invalid taxpayer-reported NAICS codes.

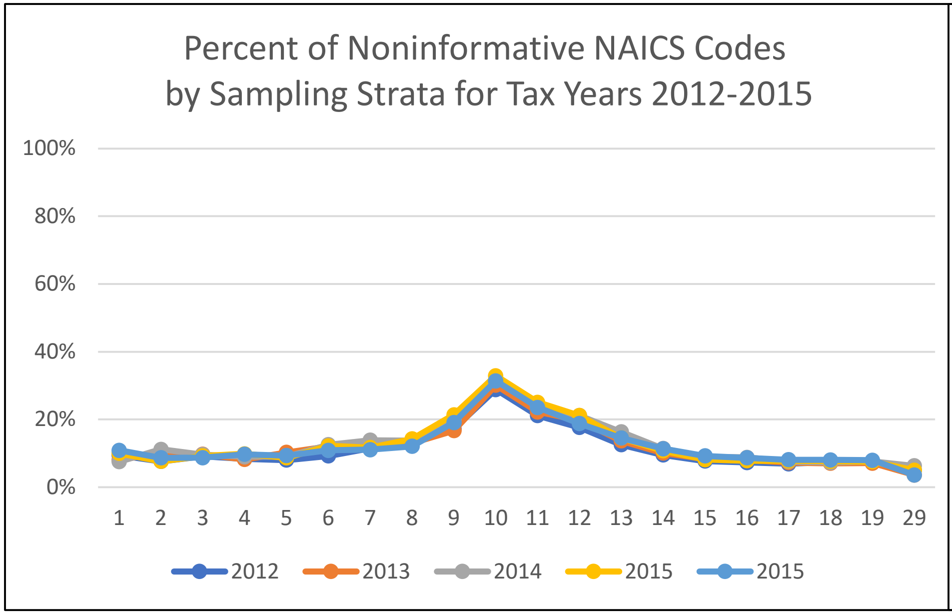
Percent Correct by Sample Strata

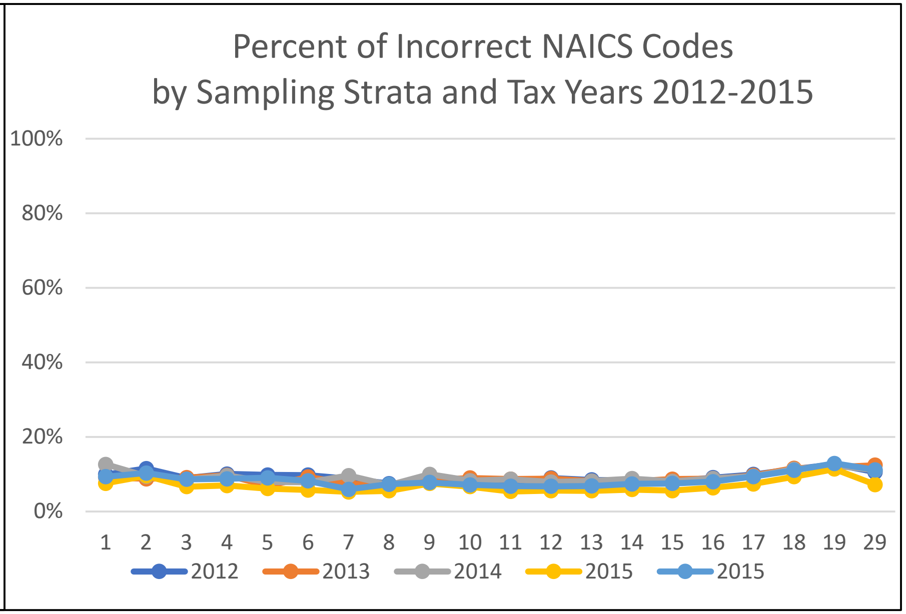


Sampling Strata Categories

Indexed Negative Income		Indexed Positive Income	
1.	$10,000,000 or more	10.	Under $30,000
2.	$5,000,000 under $10,000,000	11.	$30,000 under $60,000
3.	$2,000,000 under $5,000,000	12.	$60,000 under $120,000
4.	$1,000,000 under $2,000,000	13.	$120,000 under $250,000
5.	$500,000 under $1,000,000	14.	$250,000 under $500,000
6.	$250,000 under $500,000	15.	$500,000 under $1,000,000
7.	$120,000 under $250,000	16.	$1,000,000 under $2,000,000
8.	$60,000 under $120,000	17.	$2,000,000 under $5,000,000
9.	Under $60,000	18.	$5,000,000 under $10,000,000
29.	Large Business Receipts	19.	$10,000,000 or more

Percent Noninformative and Incorrect by Sample Strata





Modeling Features

- Textual feature are constructed from business descriptions using a bag-of-words methodology to compare business descriptions from tax returns with NAICS Code descriptions from the NAICS Code Manual
  - Descriptions were tokenized, converted to lower case, then punctuation and stop words were removed
  - Words were quantified by computing the ratio of word frequency (tf) in the description to the inverse document frequency (idf).

    $$\text{idf} = \ln\left(\frac{number\ of\ documents}{number\ of\ documents\ containing\ word}\right)$$

  - For each Sch C description, a cosine similarity score was computed for Sch C description and each NAICS Code description from the handbook (also including sub-headings)
- Numeric features include line items from tax returns

Unsupervised Methods

Predict NAICS economic sectors by calculating the probability that a return belongs to each NAICS sector using recommender algorithms

- Combine line items from 1040 and Schedule C with Term Frequency – Inverse Document Frequency (TF-IDF) of taxpayer-reported business descriptions on Schedule C
- Using only the IRTF, calculate maximum likelihood estimate of mean and covariance for each NAICS category
    - No SOI data is used in calculating model parameters
- For each new return, calculate the NAICS sector probability, assuming a multivariate Gaussian distribution of underlying line-item and text features. The NAICS category with the highest numerical output is the predicted class

Supervised Methods

- CART – Classification and Regression Trees
    - Can mix different data types
    - Easy to interpret and tune model but are prone to accuracy and stability issues
    - Immune to irrelevant variables allowing for variable selection
    - Invariant under monotone transformations of variables
- Random Forests – Ensemble method using CART Trees
    - Combines independently generated trees from bootstrap samples
    - Improves prediction accuracy by reducing overfitting and reducing model variance
- XGBoost – Ensemble method using CART Trees
    - Sequentially improves the fit of previously built trees
    - Prone to overfitting

CART Results

- Data was split into training and test sets. We further split the training data by error type. CART has a single tuning parameter, number of terminal leaves, which is a measure of tree complexity
    - The best model achieved an accuracy rate of 67% on the test data set
    - The best model using only returns with noninformative NAICS given (0, 99, or invalid) achieved an accuracy rate of 58%
    - The best model using only returns with valid NAICS found not to be correct achieved an accuracy rate of 52%

- Text features were by far the strongest predictors

CART Accuracy Results

Full data, rpart, cp=.000006, tested on 12.5% holdout sample

	Predicted by rpart																						
Actual NAICS	11	21	22	23	31	32	33	42	44	45	48	49	51	52	53	54	56	61	62	71	72	81	99	Total actu
11	682	7		32	2			9	14	25	21		3	6	48	31	24	7	14	121	8	34	3	1091
21	6	2084	2	35				6	7	33	9		7	21	67	76	26	2	44	167	15	46		2653
22		7	29	8	1			2	2	5				3	3	6	2		1	5		3		77
23	17	30	4	2906	3	10	12	13	19	51	34	2	2	20	331	108	51	2	45	188	34	158	3	4043
31	7	1		12	105	4	16	13	27	28	1		2	1	7	17	2		7	30	31	8		319
32	4	15		22	2	75	23	9	7	14	1	2	6	3	13	15	10		7	25	4	7		264
33	2	3		36	9	11	156	29	29	39	10	1	1	6	20	39	5		10	35	8	32	1	482
42	27	15	2	35	11	3	17	401	89	273	14	3	7	13	44	91	20	3	20	69	30	43	2	1232
44	11	12	1	39	16	3	10	81	931	345	10		2	13	52	89	18		41	80	64	81	1	1900
45	12	12	1	41	7	9	7	76	182	2069	9	8	18	26	66	139	23	10	39	217	39	116	2	3128
48	16	16		66	1		1	14	6	11	1810	20		12	125	46	36	5	31	178	24	59	1	2478
49	3	2		11				3		8	15	147	1	4	11	2	3		2	11	2	4		229
51	5	1	1	19	3	1	1	9	11	27	4	0	473	9	31	178	14	5	26	208	10	28		1064
52	5	17	2	35	2		4	4	16	48	12	0	3	3061	190	307	89	4	47	140	15	62	4	4067
53	17	26	3	211	3		4	9	29	96	59	4	2	110	4992	229	120	6	137	325	70	106	1	6559
54	20	35	3	163	6	6	14	26	41	144	39	1	79	232	313	9908	472	49	433	734	59	253	10	13040
56	29	15	3	130	3	5	1	11	17	64	37	5	20	92	158	481	2273	8	130	379	53	232	7	4153
61				18	2			1	4	8	15		2	5	22	107	16	798	48	231	7	71		1355
62	6	8		71			3	8	24	25	15	1	1	15	108	283	55	20	4517	526	46	263	8	6003
71	58	11	1	67	4	3		6	24	95	20		104	31	212	294	61	74	203	3207	57	275	5	4812
72	18	5		24	11		3	8	56	49	10	1	1	17	70	49	44	2	65	132	1394	92	1	2052
81	23	16	1	170	3	4	8	15	49	125	42	1	9	32	143	218	168	28	182	682	44	2042	5	4010
99	4	4	1	47	1	1		6	6	23	5	1	5	18	103	90	30	6	22	293	6	172	14	858
Total predicte	972	2342	54	4198	195	135	280	759	1590	3605	2192	197	748	3750	7129	12803	3562	1029	6071	7983	2020	4187	68	

Random Forests Results

- Data was split into training and test sets. We further split the training data by error type. Random Forests have two tuning parameters, number of trees, and the number of features randomly sampled for each tree (mtry)
  - For the full data, mtry of 8 produced the best accuracy in cross-validation, achieving about 72%.
  - For the noninformative cases, mtry of 9 produced the best accuracy in cross-validation, achieving about 62%.
  - For cases with valid NAICS found not to be correct, mtry of 11 produced the best accuracy in cross-validation, achieving about 59%.
- Varying the number of trees (ntrees) produced little variation in accuracy using 8-fold cross-validation.

Random Forests Accuracy Results

Random forest, 500 trees, mtry 8, on a holdout sample of 12.5%

	Predicted																						
Actual	11	21	22	23	31	32	33	42	44	45	48	49	51	52	53	54	56	61	62	71	72	81	99	Total actual
11	773	4		21	3			9	9	8	22		2	3	41	36	22	9	16	82	4	23	4	1091
21	3	2172	2	19	1			9	5	26	11		4	17	55	74	34		42	128	13	33	5	2653
22		5	37	4				1		5				3	5	9	1		2	2		1	2	77
23	11	29	3	3012		6	13	14	17	47	44	1	4	16	284	91	62	1	54	114	31	156	33	4043
31	3	1		5	159	3	11	15	16	18	2		5	3	5	16	1		6	18	20	11	1	319
32	3	12		16	1	125	27	10	3	6	1		5	2	5	14	8		6	13	1	5	1	264
33	2	3		27	4	10	234	19	14	33	11	1	2	1	14	35	3		14	27	1	24	3	482
42	16	14	2	25	14	7	14	598	71	223	10	6	7	12	25	66	16		17	47	15	25	2	1232
44	9	7		37	10	2	6	61	1148	304	8	2	2	12	36	63	19		25	45	32	64	8	1900
45	7	22		29	9	6	9	68	130	2283	13	4	7	27	32	127	22	4	38	157	30	89	15	3128
48	15	8		34			3	7	9	7	1914	22		15	97	41	33	3	42	107	20	85	16	2478
49	2			2				1		11	12	169		3	5	5	3			6	1	7	2	229
51	2	2		10	2	2	1	5	3	28	5	1	560	9	18	169	17	4	20	176	5	19	6	1064
52	4	24		17	2		4	5	9	43	14		2	3204	137	275	101	3	64	91	10	33	25	4067
53	10	22	2	163	1	1	2	10	27	73	59	4	3	123	5108	242	131	3	157	220	68	114	16	6559
54	20	28	5	106	8	6	18	14	25	119	41	3	67	233	204	10369	483	45	415	512	44	216	59	13040
56	15	11		92	2	4		8	10	49	38	8	15	95	116	501	2502	12	143	225	40	215	52	4153
61	1			7				1	3	6	8		2	8	13	95	15	889	51	174	5	60	17	1355
62	4	1		41		1	5	7	10	12	17		2	15	73	302	58	23	4826	279	19	247	61	6003
71	45	22		39	1	4	1	9	12	82	26	1	82	30	132	290	75	71	208	3306	39	267	70	4812
72	12	8	1	11	14		2	5	23	45	13	3	2	20	33	52	31	3	75	74	1559	59	7	2052
81	12	19		124	4	3	8	14	39	111	43	1	8	24	91	211	176	38	205	406	23	2365	85	4010
99	3	8		42		1	2	5	3	27	7	1	2	28	52	99	43	6	51	207	6	133	132	858
Total predicted	972	2422	52	3883	235	181	360	895	1586	3566	2319	227	783	3903	6581	13182	3856	1114	6477	6416	1986	4251	622	

Next Steps

- NLP – Word embeddings
- Address population accuracy estimates
- Compare unsupervised methods to supervised methods
- Other …

Contact Information

- Anne Parker (anne.s.parker@irs.gov)
- Christine Oehlert (christine.b.oehlert@irs.gov)