NAICS Code Prediction Using Supervised Methods

Christine Oehlert, Evan Schulz, and Anne Parker

U.S. Internal Revenue Service, Research, Applied Analytics, and Statistics, St. Paul, MN

ABSTRACT

When compiling industry statistics or selecting businesses for further study, researchers often rely on North American Industry Classification System (NAICS) codes. However, codes are self-reported on tax forms and reporting incorrect codes or even leaving the code blank has no tax consequences, so they are often unusable. IRSs Statistics of Income (SOI) program validates NAICS codes for businesses in the statistical samples used to produce official tax statistics for various filing populations, including sole proprietorships (those filing Form 1040 Schedule C) and corporations (those filing Forms 1120). In this article we leverage these samples to explore ways to improve NAICS code reporting for all filers in the relevant populations. For sole proprietorships, we overcame several record linkage complications to combine data from SOI samples with other administrative data. Using the SOI-validated NAICS code values as ground truth, we trained classification-tree-based models (randomForest) to predict NAICS industry sector from other tax return data, including text descriptions, for businesses which did or did not initially report a valid NAICS code. For both sole proprietorships and corporations, we were able to improve slightly on the accuracy of valid self-reported industry sector and correctly identify sector for over half of businesses with no informative reported NAICS code.

1. Introduction

The North American Industry Classification System (NAICS), introduced in 1997 to replace Standard Industry Classification (SIC), is a business classification system based on industry production processes. The NAICS was developed by the Office of Management and Budget (OMB) through their Economic Classification Policy Committee (ECPC) in collaboration with the Bureau of Economic Analysis (BEA), Bureau of Labor Statistics (BLS), Census Bureau, and corresponding agencies in Canada and Mexico (OMB 2017). The NAICS categories are updated every five years under a fixed schedule for years ending in a two or seven. In the United States, the NAICS is used for publishing national statistics by industry including Gross Domestic Product (GDP), Gross Output, Employment, and Input-Output Accounts. NAICS codes in tax data are used in developing and executing targeted tax policies to benefit or customize taxes for specific industries. For example, NAICS codes were central in identifying eligible taxpayers when implementing the CARES Act provision that provided small business support to doctors and hospitals.1

The NAICS has a six-digit hierarchical structure. Table 1 summarizes the information contained in the code structure reading the code from left to right. For example, consider Go-Fast Transmission Repair Shop (NAICS: 811113) which breaks down as: other services (Economic Sector: 81), repair and maintenance (Industry Sub-Sector: 811), automotive repair and maintenance (Industry Group: 8111), automotive mechanical and electrical repair and maintenance (NAICS Industry: 81111), and automotive transmission repair (National Industry: 811113). The two-digit Economic Sector categories have not been changed in any NAICS updates.

NAICS codes are self-reported (since 1985) on tax forms so they are subject to error. To assist taxpayers, instructions for these forms include charts of Principal Business/Professional Activity Codes with associated NAICS codes. These charts have varying degrees of granularity depending on the form type.

- **Forms SS-4 Application for Employer Identification Number** have 12 principal activity codes that loosely correspond to a subset of NAICS Economic Sectors (IRS 2017)
- **Forms Schedule C Profit or Loss from Business (Sole Proprietorship)** have principal activity codes that correspond to all applicable NAICS Economic Sectors and 315 specific six-digit codes (IRS 2018a)
- **Forms 1120 U.S. Corporation Income Tax Return Schedule K—Other Information** have principal activity codes that correspond to all applicable NAICS Economic Sectors and 418 specific six-digit codes (IRS 2018b)
- **Forms 1065 U.S. Return of Partnership Income** have principal activity codes that correspond to all applicable NAICS Economic Sectors and the same 418 specific six-digit codes (IRS 2018c)
The goal of this research is to develop effective predictive supervised models that can be used to validate reported NAICS codes in IRS administrative data and fill them in when reported codes are missing or invalid. This initial work is focused on sole proprietorships filing with Form 1040 Schedule C and corporations filing with the various Forms 1120 (1120, 1120S, 1120-L, 1120-PC, 1120-RIC, 1120-REIT, and 1120-P, here collectively referred to as Form 1120) at the two-digit NAICS (Economic Sector) level.

1.1. Related Research

Although NAICS codes are used extensively by both government agencies and industry for statistical and administrative purposes, there is no single government agency that assigns and maintains NAICS codes for business entities. Federal agencies such as the Census Bureau, Bureau of Labor Statistics, Environmental Protection Agency, Occupational Safety and Health Administration, and Small Business Administration all independently collect or assign NAICS codes to meet their respective program needs. The Census Bureau assigns NAICS codes for their surveys using the Economic Census as well as data from other federal agencies together with experienced staff. This process can be manually intensive and costly and as a result the Census Bureau has explored using statistical model-based approaches to automate the NAICS code assignment process (Dumbacher and Russell 2019) and extend earlier work on a joint Census-IRS autocoder application (Kearney and Kornbau 2005) using business description write-ins from the Economic Census focusing on the NAICS Economic Sector (first two digits). Using a bag-of-words approach for text classification, they fit naïve Bayes and regularized logistic regression using 2012 data to train their models and 2017 data to test achieving accuracy rates of 76.9% on training data and 61.2% on test data with their best performing models. Cuffe et al. (2019) extended this work, incorporating public data, including website text and user reviews obtained using public APIs such as Google Places, Google Types, Yelp, and other APIs. The objective of this novel work was to assess the usefulness of these data together with machine learning methods for generating NAICS codes to produce federal statistics. Although the authors reported a modest predictive accuracy of 59% for their best fitting random forest model, they noted substantial variation in predictive accuracy from 5% in sectors with little data to 83% among sectors with large amounts of data; an interesting result given many machine learning methods for generating NAICS codes are missing or invalid. This initial work is focused on sole proprietorships filing with Form 1040 Schedule C and corporations filing with the various Forms 1120 (1120, 1120S, 1120-L, 1120-PC, 1120-RIC, 1120-REIT, and 1120-P, here collectively referred to as Form 1120) at the two-digit NAICS (Economic Sector) level.

1.2. Problem Description

Researchers and IRS business operations often use IRS administrative data by industry when samples cannot provide the required granularity and precision. NAICS codes are important as the only straightforward way to do this. However, NAICS code reporting is prone to error, and because of the lack of tax consequences, these errors are often overlooked even in examined returns. We have no global statistics on the error rate in administrative data. However, Statistics of Income (SOI) draws a sample of returns for computing national statistics, and these returns are validated in detail, including correction of NAICS code if necessary.

Types of coding error include

- Missing or invalid codes entered. We label these as “noninformative.”
- Code 999999 (“other”), while technically valid for Schedule C, is usually misapplied and is functionally the same as a missing code. We also considered these “noninformative”; approximately 1% of Schedule C cases reported as Sector 99 actually were 99.
- Valid code entered but incorrect for the entity.
- Partially correct codes that is, correct for Economic Sector but incorrect for Industry Sub-Sector.

The rate at which SOI corrects the NAICS Economic Sector (first two digits) for the 1040 Schedule C filers in its samples was fairly stable over the time frame for this research. However, the type of SOI corrections trended slowly toward more taxpayers failing to identify a valid NAICS code. The rate at which SOI corrects the NAICS Economic Sector for Form 1120 filers in its sample trended slightly upward over the time frame of this study, while Form 1120 filers failed to provide valid NAICS codes at consistently low rates (less than 2%). See Table 2 for details.

### Table 2. NAICS code error rates in SOI sample for matched noneliminated 1040 Schedule C and Form 1120 Economic Sector.

| Code Category                           | 2012 (%) | 2013 (%) | 2014 (%) | 2015 (%) | 2016 (%) |
|----------------------------------------|----------|----------|----------|----------|----------|
| 1040 Schedule C                        |          |          |          |          |          |
| Overall Error Rate (unweighted)        | 20.4     | 21.1     | 21.4     | 19.5     | 21.0     |
| 1040 Schedule C                        |          |          |          |          |          |
| No Informative Code (unweighted)       | 11.9     | 13.0     | 13.8     | 13.7     | 13.6     |
| 1040 Schedule C                        |          |          |          |          |          |
| Overall Error Rate (weighted)          | 29.7     | 31.6     | 32.5     | 30.9     | 30.8     |
| 1040 Schedule C                        |          |          |          |          |          |
| No Informative Code (weighted)         | 21.9     | 23.5     | 25.2     | 25.7     | 24.6     |
| 1120 Corporate                         |          |          |          |          |          |
| Overall Error Rate (unweighted)        | 18.6     | 19.1     | 19.4     | 19.6     | NA       |
| 1120 Corporate                         |          |          |          |          |          |
| Overall Error Rate (weighted)          | 0.9      | 0.9      | 0.7      | 0.7      | NA       |
| 1120 Corporate                         |          |          |          |          |          |
| No Valid Code (unweighted)             | 16.0     | 16.5     | 16.7     | 16.5     | NA       |
| 1120 Corporate                         |          |          |          |          |          |
| Overall Error Rate (weighted)          | 1.7      | 1.5      | 1.2      | 1.0      | NA       |
| 1120 Corporate                         |          |          |          |          |          |
| No Valid Code (weighted)               |          |          |          |          |          |

Source: Combined IRS/SOI Schedule C data, 2012–2016, and SOIF1120 data, 2012–2015.
SOI Schedule C data have several complications that SOI Form 1120 data do not because a single individual may file multiple Schedule Cs. (IRS records for tax year 2016 indicate that about 11% of Schedule C filers file two or more Schedule Cs, accounting for over 20% of the schedules themselves.) First, the SOI cases are weighted by F1040, not by Schedule C; the weighted values provided are the product of naively extending the weight of each F1040 to its associated Schedule Cs. Second, the accuracy of the reported NAICS code cannot be determined without a successful data link, which reduced the usable sample size; see Methods for details. Third, it is possible for the SOI examiner to determine that there was no business and the Schedule C should not have been filed at all; these cases were dropped.

2. Methods

2.1. Data

2.1.1. Data Linkage

This study involved a combination of statistical data from SOI and IRS tax administrative data. Ground truth NAICS code data from SOI’s validated sample was combined with as-filed tax return administrative data to create models which could be applied to unverified incoming return data. This as-filed data was included with SOI’s Form 1120 data, but to obtain the data for Schedule C we had to link with Individual Income Tax administrative records (IRTF).

The IRTF data and SOI data are structured differently. IRTF data has one record per Schedule C. Individuals with multiple sole proprietorships are instructed to file a Schedule C for each, in descending order of gross receipts, for up to eight Schedule Cs electronically or many more on paper. However, if there are more than three, then the records will be consolidated into only three Schedule Cs before being added to IRTF. SOI data is on a by-taxpayer level, with different fields to record information on up to three Schedule Cs. Therefore, one SOI record may legitimately match up to three IRTF Schedule C records. The Cs are ordered in both datasets. The SOI data has fields for first, second, and third Schedule C; the IRTF Schedule C records have “section numbers” for that purpose. One might therefore, naively match Section 1 to “first SOI C,” and so on, and indeed a naive match for those values is already attached to the SOI data for some line items, called “as filed.”

Taxpayer Identification Number (TIN) is the key identifying field in IRTF data; while all SOI records include a TIN, the SOI unique identifying field for SOI records is record ID. Some data errors occur where multiple record IDs appear in the same filing period with the same TIN. At least one of these TINs must be incorrect, so any Schedule Cs pulled under that TIN matched to those SOI fields would also be incorrect.

Since the SOI “as filed” fields were drawn from IRTF data, they should always match IRTF values unless they have mistakenly not been included at all. In the small number of cases where they are not missing and do not match, the cases are eliminated as probable mismatches (less than 0.5% of initially matched Cs). This also eliminates most of the duplicate cases caused by multiple record IDs with the same TIN or more than three Schedule Cs appearing in IRTF, but a few remain (in all study years except TY2016 less than 0.1% of remaining Cs). These remaining unexplained duplicate cases are set aside. Remaining matched Cs are our sample (Table 3).

While the naïve approach will produce a correct match of predictor data to corrected NAICS codes in most cases, the actual situation is more complicated. SOI’s first, second, and third C fields are for the Cs which should have been in the first, second, and third position. If a taxpayer filed two Schedule Cs with distinct NAICS codes and SOI concluded the Cs were in the wrong order, then the naïve approach would link the Section 1 Schedule C with the NAICS code that belongs to the Section 2 Schedule C and vice versa. If the taxpayer filed two Schedule Cs and SOI deemed the first one to not actually be a business, then the naïve approach would conclude the second one was deemed not to be a business, and the Section 1 Schedule C would again be linked with the NAICS code which should be associated with the Section 2 C. Various other rearrangements are also possible.

The situation is still more complicated because IRTF and SOI use different rules for rolling up multiple filed Cs into the three that will go in the data. SOI tries to combine Schedule Cs in the same industry together; if more than three industries are present, the three largest are kept and all smaller ones are combined with the largest C.1 IRTF data entry tries first to combine Cs with the same proprietor SSN (which may be an issue when the F1040 represents more than one person), and if possible combine Cs with net profit separately from Cs with net loss.2 This leaves some situations ambiguous; for example, if a single taxpayer has four Schedule Cs with net loss, it does not make clear which two should be merged.

Because SOI does not correct the “gross receipts” field, it can usually be matched to the “gross receipts” field in IRTF to confirm a match or identify a rearrangement or simple merge. Using this and a SOI field for count of Cs, over 97% of sample IRTF Schedule Cs could either be matched to a corrected NAICS code or identified as having been deemed not a business. The remainder could not be, and without a good outcome value, had to be excluded from the sample (Table 3).

While the number of cases excluded is relatively small, it is important to note that they are not a random subset. They are concentrated in the sample strata associated with very high positive or negative income.

2.1.2. Text Analysis

The tax return data include: (dollar value) line-items, for example, Gross Receipts; return characteristics, such as date incorpor-

---

1 Correspondence, Mike Strudler, SOI, 2019/10/10
2 Internal Revenue Manual 3.11.3.12.(1-1-2016 revision)
rated or filing status of sole proprietor; and text predictors. There is one text field available for Schedule C and two for Form 1120.

Each Schedule C has a field for “description of business or profession” which may be freely filled in by the taxpayer or left blank. If the taxpayer consults a NAICS table, they will find a short description corresponding to each NAICS code, and some of them copy this description into the field, but many do not. They may also leave the field blank, but this is uncommon. Even among Schedule Cs in our sample with missing, invalid, or noninformative NAICS codes, only 7% had a blank description field; in the full sample, 2% had a blank field.

In the IRTF data the field is truncated to twenty characters. While 62%–70% of descriptions in the sample already had fewer than twenty characters and so are unaffected, many others are cut off mid-word. (The range is due to ambiguity in descriptions whose truncated form is nineteen characters; they may or may not have contained more words after a space.) After tokenizing, we cleaned the data by identifying the most common obviously truncated tokens and creating rules to replace them with their un-truncated form. This also served to combine some synonymous tokens. (For example, any token containing the string “extracti” was turned into “extracting.” This captured the truncations “extracti” and “extractin” but also the synonymous, as far as business descriptions go “extraction.”) A complete list of rules is available upon request. These rules were individually defined and while they covered the most common truncations, many other truncations and spelling errors were not corrected for, and many other synonymous tokens were left distinct.) After removing stop words and applying the data-refining rules, over 25,000 distinct tokens remained. Due to the short length of the field, few Schedule Cs had more than three tokens, and none more than five. The most common tokens are listed in Table 4.

Table 4. Twenty most common tokens, Schedule C.

| Word   | Count |
|--------|-------|
| consulting | 50,757 |
| services  | 40,037 |
| sales    | 33,811 |
| real     | 32,423 |
| estate   | 29,828 |
| oil      | 16,629 |
| management | 15,055 |

Source: Combined IRTF/SOI Schedule C data, 2012–2016.

Although only 253 tokens appeared more than five hundred times, this is still unwieldy as individual predictors and cutting off there would exclude clear sector-linked tokens such as “medicine,” “spa,” or “actor.”

To avoid losing useful data and limit the number of predictors, we took advantage of the background information in the NAICS. Each industry sector (two-digit NAICS code) has a several-paragraph “long” description in the NAICS manual. In addition to this, each NAICS code in use corresponds to a “short” description in words as well as a number for each digit level of specificity. (For example, NAICS code 713950 is “Bowling Centers”; 7139 is “Other Amusement and Recreation Industries”; 713 is “Amusement, Gambling, and Recreation Industries.”) All of the short descriptions falling within each sector were combined with the long sector description to create a comparison “document.” We created such a “document” for each sector, tokenized, removed common stop words, and calculated term frequency-inverse document frequency. (We used R package tidytext.)

We calculated term frequency-inverse document frequency on the data, creating each Schedule C as a document. We then calculated cosine similarity between each Schedule C and each sector according to our comparison data.

The final product for text analysis was a similarity predictor for each industry sector, measuring the similarity of the case’s description to the NAICS definition text for that sector. (We also reran the analysis using more standard bag-of-words TF-IDF analysis of the most common stemmed terms but obtained poorer results.)

For Form 1120 we used a relatively simple term frequency-inverse document frequency approach to create text predictor variables. Form 1120 Schedule K contains two text fields related to industry. These are fields for the filer to describe the “business activity” and “product or service.” Each field is limited to 30 characters. We combined these fields into one text string for processing, removed any punctuation and standard stop words, and “stemmed” the words to remove common word endings in order to combine synonymous terms. (For this text processing we used the R package tm.) This resulted in a corpus of over 16,000 unique word stems. After processing, most returns had two to four terms in their combined text fields, though some had as many as ten and a relatively large proportion (12.4%) had no terms. The most common tokens are listed in Table 5.

In order to limit the number of text variables to work with when modeling, we retained terms that occurred in at least 0.5% of text entries and removed the sparser terms. This left 64 terms, which were made into term frequency-inverse document frequency variables.

2.1.3. Sample Characteristics

We used SOI sample strata when subsampling for cross-validation and creation of holdout samples. The individual strata are based on levels of positive and negative income (deflated to a 1991 baseline to enable comparison) and the presence or absence of special forms or schedules, high business income, and nontaxable status. (IRS 2018d, 205) Corporate strata are based on assets and proceeds. (IRS 2019a, 384–385)

Businesses are not evenly distributed across NAICS sectors in the population or in the sample. Sample breakdown of validated NAICS sector for Schedule C and Form 1120 is reported in Table 6; for comparison, SOI’s population estimates for sole proprietorships and corporations appear in Table A1 in the Appendix A. Sector 52 is overrepresented in both samples, and Sector 21 overrepresented in Schedule Cs.

Table 5. Twenty most common tokens (word stems) on Form 1120.

| Word | Count |
|------|-------|
| service | 79,511 |
| compani | 42,521 |
| estat | 37,018 |
| manufactur | 27,721 |
| construct | 19,133 |
| product | 17,975 |
| equip | 10,812 |

Source: SOI Form 1120 data, 2012–2015.
Two perennial challenges of working with tax return data are (a) they tend to be highly correlated with one another (due to some being, in part, linear combinations of others), and (b) their individual distributions are heavily skewed, with many small or zero values and a few very large values. The text predictors have fewer extreme values, but still exhibit scarcity, and we anticipate many interaction effects.

Random Forest models can handle scarcity, extreme values, correlation, and interactions without difficulty. Random forest is a machine learning technique in which multiple small classification trees are constructed from sub-samples and the output of all trees is combined for the final classification. We present the results of random forest models, using the R package randomForest. (Version 4.6-14). After the model is fitted, the “forest” can be retained to classify new data.

Our final random forest models involved a mix of line item and text predictors. Random forest models are tuned with two parameters: number of trees and predictors tried per tree. We optimized with cross-validation. Variable importance for the models appears in Tables 7 and 8.

### 3. Results

#### 3.1. Baseline Results

To evaluate model success we looked at two measures. First, we computed the simple success rate (also commonly referred to as “accuracy”), or what percentage of the predicted values matched the actual values. Second, we computed the Matthews Correlation Coefficient (MCC). MCC is another way of evaluating classification success which is especially useful on heavily unbalanced datasets (Chicco and Jerman 2020). While it was originally designed to evaluate binary classification, there is a generalization to the multiclass case, also known as the Rk correlation coefficient. It incorporates the actual occurrences and predicted occurrences in each class (Gorodkin 2004).

For the rest of this analysis, we will be looking at models fitted while excluding a one-eighth holdout sample (of the relevant set); model success and MCC will be evaluated on the holdout sample.

For a naïve baseline, we assume all valid reported NAICS codes are correct and assigned cases with noninformative reported NAICS codes to be the most common valid sector for their form type – Sector 54 in both cases (although Sector 52 is most common in the F1120 sample). The results are in Table 9.

Note that for Schedule C, valid reported NAICS codes are correct in over 90% of cases.

#### 3.2. Random Forest Results

Summary results for random forest models are presented in Table 10. In all cases this approach yields improvements over the naïve baseline, especially for Form 1120. Improvement on the Schedule C valid reported set is less noteworthy (or significant).

In addition to a class prediction, random forest models can give an estimated probability for each class, which we may inter-

---

**Table 6. Industry sectors and verified sample counts.**

| Two-Digit NAICS code | Industry Sector | Verified count in Schedule C | Verified count in Form 1120 |
|----------------------|----------------|-------------------------------|----------------------------|
| 11                   | Agriculture, Forestry, Fishing, and Hunting | 8528 | 8485 |
| 21                   | Mining, Quarrying, and Oil/Gas Extraction | 21,117 | 6905 |
| 22                   | Utilities | 571 | 1949 |
| 23                   | Construction | 31,593 | 29,592 |
| 31                   | Manufacturing I | 2448 | 7248 |
| 32                   | Manufacturing II | 2039 | 14,090 |
| 33                   | Manufacturing III | 3697 | 30,035 |
| 42                   | Wholesale Trade | 9521 | 42,723 |
| 44                   | Retail Trade I | 14,701 | 25,010 |
| 45                   | Retail Trade II | 24,480 | 5933 |
| 48                   | Transportation and Warehousing I | 19,280 | 9263 |
| 49                   | Transportation and Warehousing II | 1802 | 1013 |
| 51                   | Information | 8246 | 13,087 |
| 52                   | Finance and Insurance | 31,816 | 86,512 |
| 53                   | Real Estate and Rental and Leasing | 50,744 | 44,339 |
| 54                   | Professional, Scientific, and Technical Services | 103,702 | 36,707 |
| 55                   | Management of Companies and Enterprises | Not valid for Schedule C | 34,566 |
| 56                   | Administrative and Support and Waste Management and Remediation Services | 32,229 | 9685 |
| 61                   | Educational Services | 10,729 | 2006 |
| 62                   | Health Care and Social Assistance | 47,445 | 13,558 |
| 71                   | Arts, Entertainment, and Recreation | 37,746 | 4681 |
| 72                   | Accommodation and Food Services | 15,900 | 9646 |
| 81                   | Other Services (Except Public Administration) | 31,531 | 8818 |
| 92                   | Public Administration | Not valid for Schedule C | Not valid for Form 1120 |
| 99                   | Other/Unclassifiable | 6806 | Not used for Form 1120 |

Source: Combined IRTF/SOI Schedule C data, 2012–2016, and SOI F1120 data, 2012–2015.

---
Table 7. Variable importance for Schedule C random forest models.

| Predictor | Mean decrease GINI | Predictor | Mean decrease GINI |
|-----------|-------------------|-----------|-------------------|
| Reported NAICS code | 7693.5 | Sector 54 Text Similarity | 3804.2 |
| Schedule C Profit/Loss | 4717.0 | Schedule C Profit/Loss | 2460.2 |
| Total Deductions | 4256.4 | Sector 56 Text Similarity | 2264.3 |
| Total Gross Receipts | 4124.3 | Total Gross Receipts | 2248.3 |
| Other Expenses | 2932.9 | Total Deductions | 2082.2 |
| Sector 54 Text Similarity | 2167.3 | Sector 52 Text Similarity | 1788.1 |
| Car and Truck Expenses | 1334.3 | Sector 23 Text Similarity | 1529.9 |
| Office Expense Amount | 1444.7 | Sector 81 Text Similarity | 1485.9 |
| Sector 56 Text Similarity | 1434.7 | Sector 62 Text Similarity | 1446.0 |
| Sector 81 Text Similarity | 1347.3 | Sector 71 Text Similarity | 1381.3 |
| Travel Expenses | 1344.4 | Sector 53 Text Similarity | 1380.3 |
| Meals and Entertainment [Expense] Amount | 1335.0 | Other Expenses | 1338.8 |
| Sector 52 Text Similarity | 1230.8 | Sector 11 Text Similarity | 1327.5 |
| Insurance [Expense] Amount | 1221.2 | Sector 61 Text Similarity | 1320.0 |
| | | Sector 33 Text Similarity | 1182.0 |

Source: Combined IRTF/SOI Schedule C data, 2012–2016.

Table 8. Variable importance for 1120 Corporate random forest model.

| Predictor | Mean decrease GINI |
|-----------|-------------------|
| Reported NAICS code | 80339.1733 |
| Cost of Goods Sold | 14824.4270 |
| Total Assets | 9744.8938 |
| Gross Receipts | 9418.8677 |
| Gross Profits | 9066.6443 |
| Net Depreciation | 8054.8538 |
| Text stem "construct" | 7819.1310 |
| Year incorporated | 6919.7084 |
| Rents [Paid] | 6889.3664 |
| Text stem "manufactur" | 6458.8715 |
| Other Deductions | 6393.8707 |
| Repairs | 6333.5619 |
| Text stem "real" | 6101.2038 |
| Text stem "wholesale" | 6073.8337 |
| Salaries and wages | 5981.8439 |

Source: SOI F1120 data, 2012–2015.

Table 9. Naive baseline results.

| Model type | Accuracy | MCC |
|------------|----------|-----|
| Form 1120 NAICS code prediction | 81.0% | 0.791 |
| (weighted) | 84.5% | 0.740 |
| Schedule C error prediction, valid set | 91.4% | 0.404 |
| (weighted) | 92.6% | 0.402 |
| Schedule C NAICS code prediction, valid set | 92.1% | 0.913 |
| (weighted) | 92.1% | 0.913 |
| Schedule C NAICS code prediction, noninformative set | 65.7% | 0.627 |
| (weighted) | 64.1% | 0.627 |

Source: Combined IRTF/SOI Schedule C data, 2012–2016, and SOI F1120 data, 2012–2015.

Table 10. Results for random forest models.

| Model type | Accuracy | MCC |
|------------|----------|-----|
| Form 1120 NAICS code prediction | 91.4% | 0.906 |
| (weighted) | 91.6% | 0.906 |
| Schedule C error prediction, valid set | 92.4% | 0.404 |
| (weighted) | 92.6% | 0.402 |
| Schedule C NAICS code prediction, valid set | 92.1% | 0.913 |
| (weighted) | 92.1% | 0.913 |
| Schedule C NAICS code prediction, noninformative set | 65.7% | 0.627 |
| (weighted) | 64.1% | 0.627 |

Source: Combined IRTF/SOI Schedule C data, 2012–2016, and SOI F1120 data, 2012–2015.

Table 11. Results for random forest models.

For Schedule C, we have focused on predicting NAICS codes on the noninformative set and predicting reporting error on the set of valid reported NAICS code, with only brief mention of predicting correct NAICS code on the valid reported set. The reason is that NAICS code prediction on the valid reported set has an element of risk that is not present when working on the noninformative data. A prediction on the noninformative set may be wrong, but it is replacing an unusable absence of information. A prediction on the valid reported set, if wrong, is replacing something which might be right—it is potentially a “false correction.” We can define a “false correction rate” as the percentage of cases where the NAICS code was correctly reported but predicted as incorrect. Table 11 displays confusion matrices for the results of different modeling approaches on the valid reported set.

Approaching the problem as prediction of reporting error is more conservative, that is, less likely to identify a correctly reported industry sector as incorrect.

Referring to Tables 9 and 10, we see that the lift over baseline for predicting NAICS code on the valid reported set is under 1%; while the model corrects more than it falsely corrects, the false correction rate is larger than the net gain.

4. Discussion

Random forest models using tax return line-items and text fields provide a viable way of predicting two-digit NAICS codes for both Form 1120 and Form 1040 Schedule C businesses, and can also provide an estimated level of “confidence” in this prediction. In the case of Schedule Cs with valid reported NAICS codes, accepting the reported value performs similarly to the best model and avoids introducing errors. In the case of Form 1120 with valid reported NAICS codes, using a model should reduce the error rate; for cases of both form types with missing, invalid,
Figure 1. Form 1120 random forest probability of predicted NAICS code compared to prediction success; boxplot and conditional density plot.

Figure 2. Schedule C random forest predicted probability of reporting errors compared to actual errors on holdout sample; boxplot and conditional density plot.

Figure 3. Schedule C random forest probability of predicted NAICS code on noninformative set compared to prediction success; boxplot and conditional density plot.

or noninformative NAICS codes, using a model can produce usable results.

Extending this methodology to three- or four-digit NAICS codes would be a more complex problem (and would also
Table 11. Schedule C valid reported set random forest model results confusion matrices.

| Reporting error prediction | Actual error | Actual correct |
|---------------------------|--------------|----------------|
| Predicted error           | 1500         | 916            |
| Predicted correct         | 3338         | 50,211         |
| Falsely identified error rate: | 1.6% | |

| NAICS code prediction | Predicted incorrectly | Predicted correctly |
|-----------------------|-----------------------|---------------------|
| Reported incorrectly  | 3060                  | 1778                |
| Reported correctly    | 1358                  | 49,769              |
| False correction rate: | 2.5% | |

Source: Combined IRTF/SOI Schedule C data, 2012–2016.

require more consideration of trends in NAICS revision) but is an avenue worth exploring.

Appendix A: Population Industry Estimates

Table A1. Population industry estimates.

| Two-Digit NAICS code | Industry Sector                  | SOI 2016 sole proprietorship estimate | SOI 2015 corporate return estimate |
|----------------------|----------------------------------|--------------------------------------|-----------------------------------|
| 11                   | Agriculture, Forestry, Fishing, and Hunting | 297,031                              | 141,534                           |
| 21                   | Mining, Quarrying, and Oil/Gas Extraction | 113,475                              | 35,936                            |
| 22                   | Utilities                          | 14,390                               | 7,123                             |
| 23                   | Construction                       | 2,840,001                            | 711,742                           |
| 31                   | Manufacturing I                    | 82,038                               | 34,379                            |
| 32                   | Manufacturing II                   | 96,812                               | 64,639                            |
| 33                   | Manufacturing III                  | 191,231                              | 139,829                           |
| 42                   | Wholesale Trade                    | 342,687                              | 391,422                           |
| 44                   | Retail Trade I                     | 599,876                              | 411,637                           |
| 45                   | Retail Trade II                    | 1,729,622                            | 173,076                           |
| 49                   | Transportation and Warehousing I   | 1,739,836                            | 249,632*                          |
| 49                   | Transportation and Warehousing II  | 245,155                              | 4,141*                            |
| 51                   | Information                        | 367,766                              | 128,987                           |
| 52                   | Finance and Insurance              | 609,412                              | 245,155                           |
| 53                   | Real Estate and Rental and Leasing | 1,282,230                            | 679,956                           |
| 54                   | Professional, Scientific, and Technical Services | 3,466,550 | 936,361 |
| 55                   | Management of Companies and Enterprises | Not valid for Schedule C | 72,394 |
| 56                   | Administrative and Support and Waste Management and Remediation Services | 2,557,938 | 293,137 |
| 61                   | Educational Services               | 823,153                              | 64,191                            |
| 62                   | Health Care and Social Assistance  | 2,127,345                            | 481,700                           |
| 71                   | Arts, Entertainment, and Recreation | 1,642,700                            | 136,124                           |
| 72                   | Accommodation and Food Services    | 487,183                              | 320,769                           |
| 81                   | Other Services (Except Public Administration) | 3,449,846 | 406,069 |
| 92                   | Public Administration              | Not valid for Schedule C             | Not valid for Corporations         |
| 99                   | Other/Unclassifiable               | 419,638                              | 245,155                           |

Source: Statistics of Income—2016: Sole Proprietorship Returns (IRS 2018a); Statistics of Income—2015: Corporation Income Tax Returns Table I (IRS 2019b).

Acknowledgments

The authors thank Barry Johnson, Director SOI, for the data and support and Mike Strudler, SOI Individual and Tax Exempt Branch for invaluable assistance with SOI treatment of multiple Schedule Cs. The authors also thank Dr. Karl Branting, our panel discussant, at JSM 2020, for helpful feedback. Portions of this paper appeared in the Proceedings of the 2020 Joint Statistical Meetings.

Disclosure Statement

The authors have no personal or financial stakes in the results of this study.

References

Chicco, D., and Jerman, G. (2020), "The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation," BMC Genomics, 21, 6. doi: 10.1186/s12864-019-6143-7. [62]

Cuffe, J., Bhattacharjee, S., Etudo, U., Smith, J. C., Basdeo, N., Burbank, N., and Roberts, S. R. (2019), "Using Public Data to Generate Industrial Classification Codes," in Big Data for 21st Century Economic Statistics, eds. K. G. Abraham, R. S. Jarmin, B. Moyer and M. D. Shapiro, Chicago, IL: University of Chicago Press, Forthcoming-Cited with permission. [59]

Dumbacher, B., and Russell, A. (2019), "Using Machine Learning to Assign North American Industry Classification System Codes to Establishments Based on Business Description Write-Ins," in Proceedings of the American Statistical Association 2019. [59]

Gorodkin, J. (2004), "Comparing Two K-Category Assignments by a K-Category Correlation Coefficient," Computational Biology and Chemistry, 28, 367–374. [62]

Internal Revenue Service (2017), "Instructions for Form SS-4, Application for Employer Identification Number," [online], Available at https://www.irs.gov/pub/irs-prior/i1120--2018.pdf. [58]

Internal Revenue Service (2018a), "Instructions for Schedule C (Form 1040 or Form 1040-SR), Profit or Loss From Business (Sole Proprietorship)," [online], Available at https://www.irs.gov/pub/irs-prior/i1040sc--2018.pdf. [58]

Internal Revenue Service (2018b), "Instructions for Form 1120, U.S. Corporation Income Tax Return," [online], Available at https://www.irs.gov/pub/irs-prior/i1120--2018.pdf. [58]

Internal Revenue Service (2018c), "Instructions for Form 1065, U.S. Return of Partnership Income," [online], Available at https://www.irs.gov/pub/irs-prior/i1065--2018.pdf. [58]

Internal Revenue Service (2018d), "Statistics of Income—2016: Individual Income Tax Returns," [online], Available at https://www.irs.gov/pub/irs-soi/16inalcr.pdf. [61]

Internal Revenue Service (2018e), "Statistics of Income—2016: Sole Proprietorship Returns," [online], Available at https://www.irs.gov/pub/irs-soi/16sp01br.xls. [65]
Kearney, A. T., and Kornbau, M. E. (2005), "An Automated Industry Coding Application for New US Business Establishments," in Proceedings of the American Statistical Association 2005. [59]
Office of Management and Budget (2017), “2017 North American Industry Classification System (NAICS) Manual,” [online], Available at https://www.census.gov/eos/www/naics/2017NAICS/2017_NAICS_Manual.pdf. [58]