NAICS Code Prediction Using Supervised Methods[1]

Christine Oehlert, Evan Schulz, Anne Parker
US Internal Revenue Service

Abstract

When compiling industry statistics or selecting businesses for further study, researchers often rely on North American Industry Classification System (NAICS) codes. However, NAICS codes are self-reported on tax forms and mistakes have no tax consequences, so they are often unreliable. IRS's Statistics of Income (SOI) program validates NAICS codes for businesses in their samples, including sole proprietorships (those filing Form 1040 Schedule C) and corporations (those filing Form 1120). 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 (rpart and random forest) 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. For both sole proprietorships and corporations, we were able to improve slightly on the accuracy of valid self-reported industry sector and correctly identify NAICS for over half of businesses with no informative reported NAICS.

Key Words: NAICS Codes, Tax Compliance, CART, Random Forests, Data Linkage, Machine Learning

1. Introduction

The North American Industry Classification System (NAICS) codes, introduced in 1997 to replace Standard Industry Classification (SIC) codes, comprise a business classification system based on industry production processes.[2] The NAICS code system 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), and Census Bureau. NAICS codes are updated every five years under a fixed schedule for years ending in a two or seven. In the United States, NAICS codes are used for publishing national statistics by industry including Gross Domestic Product (GDP), Gross Output, Employment, and Input-Output Accounts.

NAICS codes have a six-digit hierarchal structure. The table below summarizes the information contained in the code structure reading the code from left to right. For example, consider *Go-Fast Transmission Repair Shop (NAICS code: 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).

Table 1: NAICS code structure -- number of digits

Economic Sector	1 - 2

[1] The views expressed in this presentation are those of the authors and do not necessarily reflect the views of the U.S. Internal Revenue Service or the Department of the Treasury.
[2] See Forward in North American Industry Classification System manual.

Industry Sub-Sector	3
Industry Group	4
NAICS Industry	5
National Industry	6

NAICS codes are self-reported (since 1985) on tax forms so they are subject to error. To assist taxpayer, 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.[3]

- **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 research is to develop effective predictive supervised models that can be used to validate reported NAICS codes on IRS administrative data and fill them in when reported codes missing or invalid. This initial work is focused on Forms 1040 (Individual Schedule C) and Forms 1120 (Corporate) at the two-digit NAICS code (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.[4] The Census Department 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 Census has explored using statistical model-based approaches to automate the NAICS code assignment process .[Dumbacher and Russell, 2019] 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 code 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

[3] The SS-4 form has 12 principal activity codes that loosely correspond to a subset NAICS Economic Sectors; 1040 Sch C form has 21 principal activity codes that correspond to all NAICS Economic Sectors but have limited sub-sector and industry group detail; 1120 form has 19 principal activity codes but considerable detail down to the NAICS industry.

[4] Examples of federal agencies include: Census, Bureau of Labor Statistics, Environmental Protection Agency, Department of Environmental Protection, Occupational Safety and Health Administration, Small Business Administration all collect or assign NAICS codes to meet their respective program needs.

to 83% among sectors with large amounts of data; an interesting although not unexpected result given many machine learning methods require large amounts of data to achieve good results.

A common theme throughout this work is the focus on automating the NAICS code assignment process due to resource constraints. The IRS Statistics of Income (SOI), one of 13 federal statistical agencies, also uses an extensive manual review process with experienced staff to validate and correct, if necessary, self-reported NAICS codes for their statistical products. Our work focuses on a different, albeit similar problem of validating NAICS codes in our administrative systems to support IRS programs.

1.2 Problem Description

Researchers and IRS business operations both often want to look at IRS administrative data by industry. NAICS codes are important as the only straightforward way to do this. They are also 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 if necessary.

Types of coding error include:

- Missing or invalid codes entered. We label these as "noninformative".
- Code 999999 ("other"), while technically valid, is usually misapplied and is functionally the same as a missing code. We also considered these "noninformative".[5]
- Valid code entered but incorrect for the entity
- Partially correct codes i.e. correct for Economic Sector but incorrect for Industry Sub-Sector.

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

Table 2: NAICS code error rates in SOI sample for 1040 Schedule C and Corporate 1120 Economic Sector[6]

Year	*1040 Schedule C Overall Error Rate (%)*	*1040 Schedule C No Valid NAICS Code (%)*	*1120 Corporate Overall Error Rate (%, unweighted)*	*1120 Corporate No Valid NAICS Code (%, unweighted)*	*1120 Corporate Overall Error Rate (%, weighted)*	*1120 Corporate No Valid NAICS Code (%, weighted)*
2012	21.4	10.4	18.6	0.9	16.0	1.7

[5] In the Schedule C data, approximately 1% of cases reported as 99 were confirmed as 99.

[6] In these and later calculations, both unweighted and weighted values (based on SOI sample weights) appear for Form 1120. Schedule C values do not incorporate sample weights. While the Schedule C data also included weights, they were defined on a *by return* basis, as opposed to the *by*

Year	*1040 Schedule C Overall Error Rate (%)*	*1040 Schedule C No Valid NAICS Code (%)*	*1120 Corporate Overall Error Rate (%, unweighted)*	*1120 Corporate No Valid NAICS Code (%, unweighted)*	*1120 Corporate Overall Error Rate (%, weighted)*	*1120 Corporate No Valid NAICS Code (%, weighted)*
2013	22.0	11.3	19.1	0.9	16.5	1.5
2014	22.5	12.2	19.4	0.7	16.7	1.2
2015	20.5	12.4	19.6	0.7	16.5	1.0
2016	22.0	12.5	NA	NA	NA	NA

Source: SOI Individual Sch C and Form 1120 samples

2. Data Set-up

This study involved a combination of statistical data from SOI and administrative data from IRS's Compliance Data Warehouse (CDW).

2.1 Data Linkage for Forms 1040

Our predictor data is from tax returns, line items and descriptions as given by the taxpayer, and is found in the Compliance Data Warehouse (CDW) Individual Returns Transaction File (IRTF) Schedule C table. In order to associate the cases with 'ground truth' corrected NAICS, we must link them to SOI records.

The IRTF data and SOI data are structured differently: the SOI data has one record per taxpayer, while the IRTF data has one record per Schedule C, for 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 C; the IRTF C records are marked as Section 9, 10, or 11. One might therefore naively match Section 9 to 'first C', and so on, and indeed a naïve match for those values is already attached to the SOI data for some line items, called 'as filed'.

The unique identifying field for SOI records is *record ID* rather than *TIN* as it is for IRTF data. Some data errors occur where multiple record IDs appear 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 CDW 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 too many 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.

While the naïve approach will produce a correct match of predictor data to corrected NAICS 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 and SOI concluded the Cs were

Schedule C of our analysis; each return was associated with up to three Schedule Cs, so distributing the weights appropriately is not straightforward and we did not attempt it at this time.

in the wrong order, then the naïve approach would link the Section 9 Schedule C with the NAICS code that belongs to the Section 10 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 9 Schedule C would again be linked with the NAICS code which should be associated with the Section 10 C. Various other rearrangements are also possible.

The situation is still more complicated because taxpayers may attach up to eight Schedule Cs when filing electronically, and many more than that filing on paper, but these will be combined into at most three Schedule C records for both IRTF and SOI. IRTF and SOI use different rules for combining Cs. 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.[7] IRTF 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. It is not clear if there is a hard guideline on how to deal with the situation beyond that.[8]

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 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.

While the number of cases excluded thusly 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 large profit or loss.

Table 3: Schedule C sample matching outcomes by study year

	2012	*2013*	*2014*	*2015*	*2016*
Correctly matched or eliminated by naïve approach	93.99%	93.72%	93.61%	93.56%	93.55%
Misidentified by naïve approach but matched or eliminated by our approach	3.87%	4.02%	3.91%	3.89%	4.05%
Unidentifiable by our approach	2.13%	2.26%	2.48%	2.55%	2.39%

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

2.2 Text Analysis

2.2.1 Schedule C

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 table of possible NAICS codes, they will find a short description for each NAICS, and some of them copy this description into the field, but many do not. They may also leave the field blank, but this is

[7] Correspondence, Mike Strudler, SOI, 2019/10/10
[8] Internal Revenue Manual 3.11.3.12.(1-1-2016 revision)

uncommon. Even Schedule Cs in our sample with missing, invalid, or noninformative NAICS, 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[9] and so are unaffected, many others are cut off midword. 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.[10] (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.

Table 4: Twenty most common tokens, Schedule C

"consulting"	50,757	"gas"	14,400	"restaurants"	6341
"services"	40,037	"investment"	10,620	"care"	6317
"sales"	33,811	"physicians"	9328	"director"	6269
"real"	32,423	"insurance"	8817	"financial"	6180
"estate"	29,828	"medical"	8592	"attorney"	5566
"oil"	16,629	"construction"	7923	"retail"	5415
"management"	15,055	"development"	6609		

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

Table 5: Number of tokens per Schedule C

No tokens (field blank or all contents removed as stop words)	1.87%
1 token	30.2%
2 tokens	48.9%
3 tokens	18.2%
4 or 5 tokens	0.9%

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

Although only 253 tokens appeared more than five hundred times, this is still enough to be unwieldy as individual predictors and cutting off there would exclude clearly sector-linked tokens such as "medicine", "spa", or "actor".[11]

To avoid losing useful data and limit the number of predictors, we took advantage of the background information in the NAICS definitions. Each industry sector (two-digit NAICS) has a several-paragraph description in the NAICS manual. In addition to this, each NAICS in use has a description in words as well as a number and each digit level of specificity.[12]

[9] 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.

[10] 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.

[11] We reran the analysis using a more standard bag-of-words TF-IDF analysis of the most common stemmed terms but obtained poorer results.

[12] For example, NAICS 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.[13]

We calculated term frequency-inverse document frequency on the data, treating 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. In discussion of results, these text similarity measures are referred to as Txt11, Txt21, and so on, where the numbers refer to the corresponding economic sector (see Table 8 for a list of sectors).

2.2.2 Form 1120

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 NAICS code. 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.[14] 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 a relatively large proportion (12.4%) had no terms.

Table 6: Twenty most common tokens (word stems) on Form 1120

"servic"	79,511	"hold"	50,151	"invest"	43,702
"compani"	42,521	"sale"	42,073	"real"	38,281
"estat"	37,018	"bank"	35,491	"insur"	32,259
"manufactur"	27,721	"properti"	26,815	"casualti"	23,430
"construct"	19,133	"wholesal"	18,381	"rental"	18,023
"product"	17,975	"retail"	16,255	"manag"	14,496
"equip"	10,812	"consult"	10,666		

Source: SOI Form 1120 data, 2012-2015

Table 7: Number of tokens in combined text fields on Form 1120

No tokens (field blank or all contents removed in processing)	12.4%
1 token	0.2%
2 tokens	18.5%
3 tokens	32.0%
4 tokens	25.3%
5 tokens	6.5%
6 tokens	5.0%
7 or more (up to 10)	0.3%

Source: SOI Form 1120 data, 2012-2015

[13] We used R package tidytext.
[14] For this text processing we used the R package tm.

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.

2.3 Outcomes and Predictors

Approaching NAICS codes, there are several possible modeling "targets".

- Is the stated NAICS code correct, or is it a reporting error? How likely is it to be a reporting error? This need only be modeled on the subset of cases where the reported NAICS code was valid; when the code is invalid or missing, it must be an error.
- What is the correct NAICS code, and how confident are we in that? This can be modeled on the entire set, on only those with valid reported codes, or on only those without valid reported codes.

We focused on attempting to predict errors in valid reported NAICS and on predicting NAICS of cases with noninformative reported NAICS, with some notes on results on predicting correct NAICS on the valid reported set.

In addition to the text predictors discussed above, we incorporate line-items as reported on the tax forms F1040 Schedule C and Form 1120, respectively. Most line-items are dollar amounts, although they also include values such as date incorporated (Form 1120) or filing status of the sole proprietor (Schedule C).

Two perennial challenges of working with line-items 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 ones. The text predictors have fewer extreme values, but still exhibit scarcity, and we anticipate many interaction effects.

Classification And Regression Tree (CART) models can handle scarcity, extreme values, correlation, and interactions without difficulty, as can ensemble tree methods such as Random Forests. We present results for both single trees and random forest models.[15]

We used SOI sample strata when subsampling for cross-validation and creation of holdout samples.

Businesses are not evenly distributed across NAICS in the population or in the sample. Sample breakdown of validated NAICS sector for Schedule C and Form 1120 is reported in Table 8. The confusion matrices of reported versus validated NAICS are in the appendix.

Table 8: Industry Sectors and Verified Sample Counts

Two-Digit NAICS	*Industry Sector*	*Verified count in Schedule C sample*	*Verified count in Form 1120 sample*
11	Agriculture, Forestry, Fishing, and Hunting[16]	8,528	8,485

[15] We used the R packages rpart and randomForest respectively.
[16] Strictly agricultural sole proprietorships file Schedule F rather than Schedule C.

Two-Digit NAICS	*Industry Sector*	*Verified count in Schedule C sample*	*Verified count in Form 1120 sample*
21	Mining, Quarrying, and Oil/Gas Extraction	21,117	6,905
22	Utilities	571	1,949
23	Construction	31,593	29,592
31	Manufacturing I[17]	2,448	7,248
32	Manufacturing II	2,039	14,090
33	Manufacturing III	3,697	30,035
42	Wholesale Trade	9,521	42,723
44	Retail Trade I	14,701	25,010
45	Retail Trade II	24,480	5,933
48	Transportation and Warehousing I[18]	19,280	9,263
49	Transportation and Warehousing II	1,802	1,013
51	Information	8,246	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	102,702	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	9,685
61	Educational Services	10,729	2,006
62	Health Care and Social Assistance	47,445	13,558
71	Arts, Entertainment, and Recreation	37,746	4,681
72	Accommodation and Food Services	15,900	9,646
81	Other Services (Except Public Administration)	31,551	8,818
92	Public Administration	Not valid for Schedule C	Not valid for Form 1120
99	Other/Unclassifiable	6,806	Not used for Form 1120

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

3. Results

3.1 First Stage Modeling

To evaluate model success we looked at two measures. First, 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.[19]

[17] Unofficially, Sector 31 is mainly food and clothing, Sector 32 is mainly manufacturing of raw materials to an intermediate stage, and Sector 33 is mainly durable goods.

[18] Unofficially, Sector 48 is all transportation except postal service/couriers, while Sector 49 is warehousing plus postal service/courier transport.

[19] (Chicco 2020)

Going forward, 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 are correct and cases with noninformative reported NAICS to be the most common true sector for their form type – for Schedule C Sector 54, and for Form 1120 Sector 53. The results are in Table 9.

Table 9: Naive baseline results

Model type	*Error prediction success*	*Error prediction MCC*	*NAICS prediction success (valid set)*	*NAICS prediction MCC (valid set)*	*NAICS prediction success (noninf. set)*	*NAICS prediction MCC (noninf. set)*
Schedule C	91.4%	0	91.4%	0.905	16.8%	0
Form 1120 (weighted)	81.5% (85.1%)	0	81.5% (85.1%)	0.797	24.7% (9.6%)	0

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

Note that for Schedule C, assuming all valid reported NAICS to be correct gives good results in terms of identifying NAICS; results are not as good for Form 1120, but they are still correct over 80% of the time with a strong MCC.

3.2 CART Models

Summary results for classification tree models are in Table 10.[20] There is, unsurprisingly, a striking improvement over the lack of information in the 'no informative NAICS' set and over the zero MCC in assuming no reporting errors. The accuracy improvement on the valid sets is not so good, especially for Schedule C, and the Schedule C NAICS prediction on the valid set doesn't even have much of an MCC increase.

Table 10: Results for CART Models

Model type	*Error prediction success*	*Error prediction MCC*	*NAICS prediction success (valid set)*	*NAICS prediction MCC (valid set)*	*NAICS prediction success (noninf. set)*	*NAICS prediction MCC (noninf. set)*
Schedule C	91.6%	0.370	91.9%	0.911	58.9%	0.555
Form 1120 (weighted)	89.6% (88.0%)	0.626	86.0% (86.8%)	0.847	56.3% (41.2%)	0.488

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

In addition to a class prediction, CART models give an estimated probability for each class, which we may interpret as the confidence we can have in a reported or predicted NAICS. Figure 1a shows predicted probability of reporting error versus actual error for Schedule C; there is as clear difference in the predicted probability between cases with and without reporting errors, but it is not as good as it looks due to the much larger number of cases in the no error (0) category. The irregularities of the predicted probabilities are clearer in the conditional density plot in Figure 1b. Figure 2 shows the same plots for Form 1120; they are somewhat better-behaved.

[20] CART models are tuned with a complexity parameter (cp) which we optimized with cross-validation. For Schedule C reporting error prediction only, the optimal cp for accuracy was not the same as the optimal cp for MCC; we optimized MCC.

We can do something similar for NAICS prediction on the noninformative set by comparing the probability given to the predicted NAICS to whether the prediction was a success on the holdout samples. The plots for Schedule C and Form 1120 are in Figure 3 and Figure 4 respectively.

Figure 1: Schedule C CART predicted probability of reporting errors compared to actual errors on holdout sample; boxplot and conditional density plot

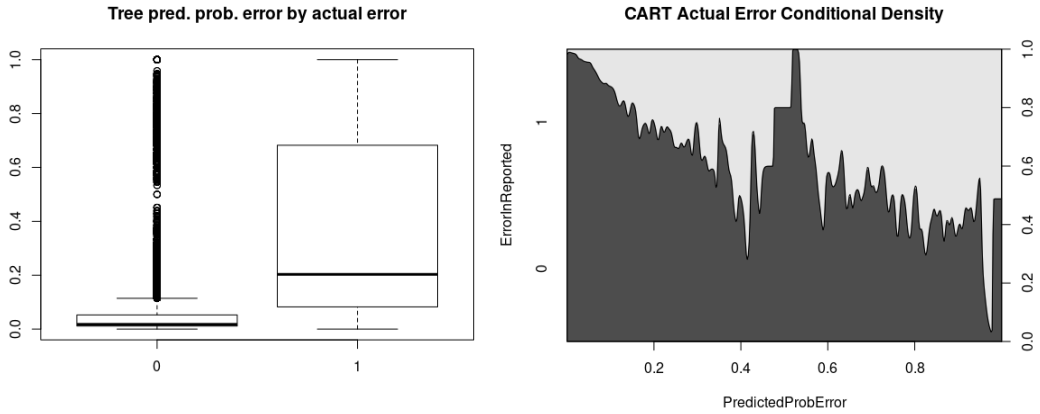


Figure 2: Form 1120 CART predicted probability of reporting errors compared to actual errors on holdout sample; boxplot and conditional density plot

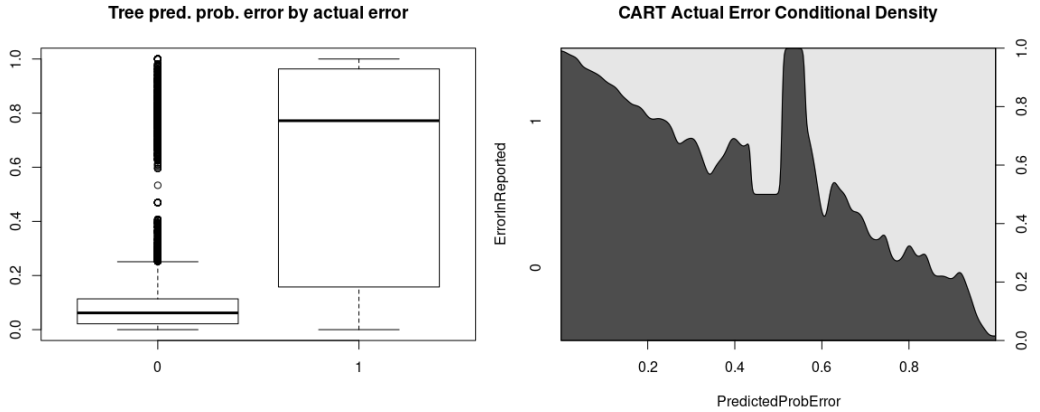


Figure 3: Schedule C CART probability of predicted NAICS on noninformative set compared to prediction success; boxplot and conditional density plot

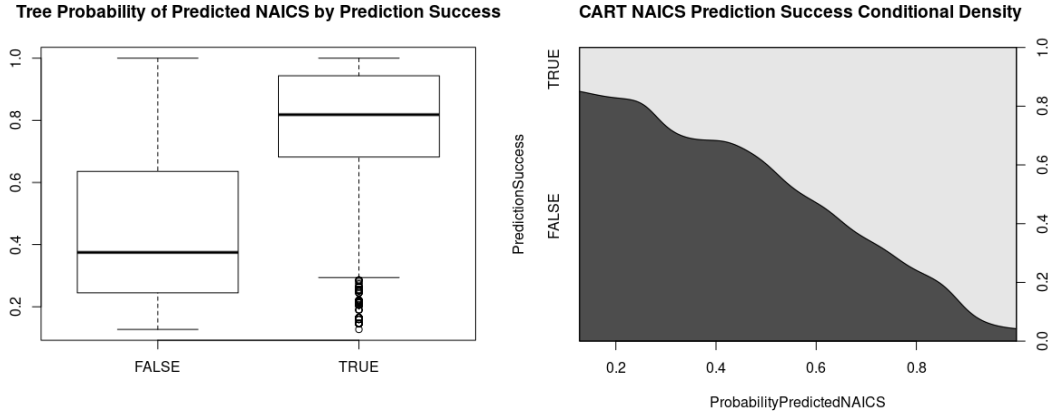
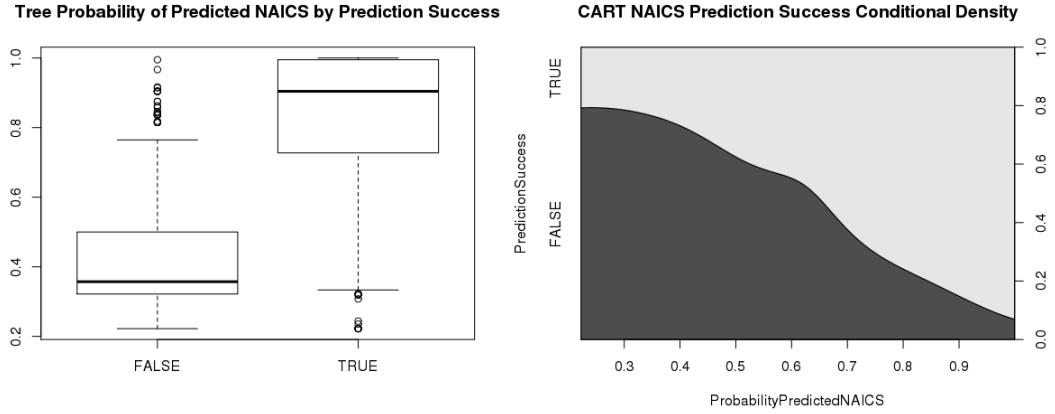


Figure 4: Form 1120 CART probability of predicted NAICS on noninformative set compared to prediction success; boxplot and conditional density plot



The optimized complexity parameters for most of these models produce trees with hundreds of leaves, but we can look at simpler trees to get an idea of how the predictors are behaving. Such trees appear in Figure 5-Figure 8.

Both these simpler results and CART models' built-in variable importance measurements make it clear that when a valid reported NAICS is available, it is the most important predictor of reporting error as well as of true NAICS. For Schedule C it was followed by text similarity measures; for 1120 by a mix of line-items and a few terms, notably "bank".

In the absence of a valid reported NAICS, the Schedule C text similarity measurements dominate and function generally as we would expect; many of the splits assign cases with high similarity to a particular sector to that sector. For the Form 1120 noninformative set, we see a mix of line-items and terms; variable importance measurements generally assign more importance to line-items.

Figure 5: Reporting error prediction classification tree with CP=0.001 (Schedule C)

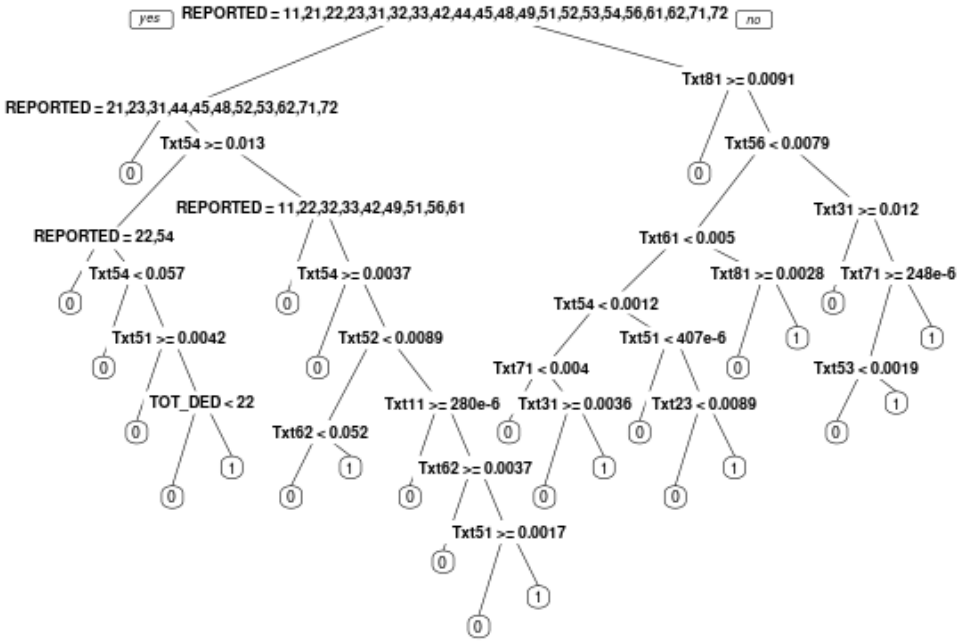


Figure 6: Classification tree predicting NAICS of noninformative group, CP=0.003 (Schedule C)

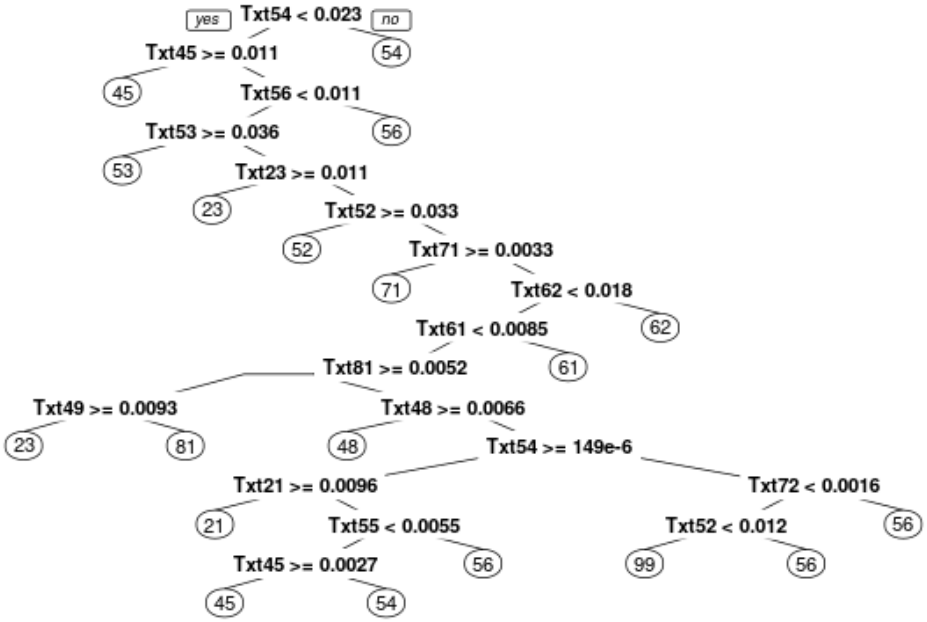


Figure 7: Reporting error prediction classification tree with CP = 0.0013 (Form 1120)

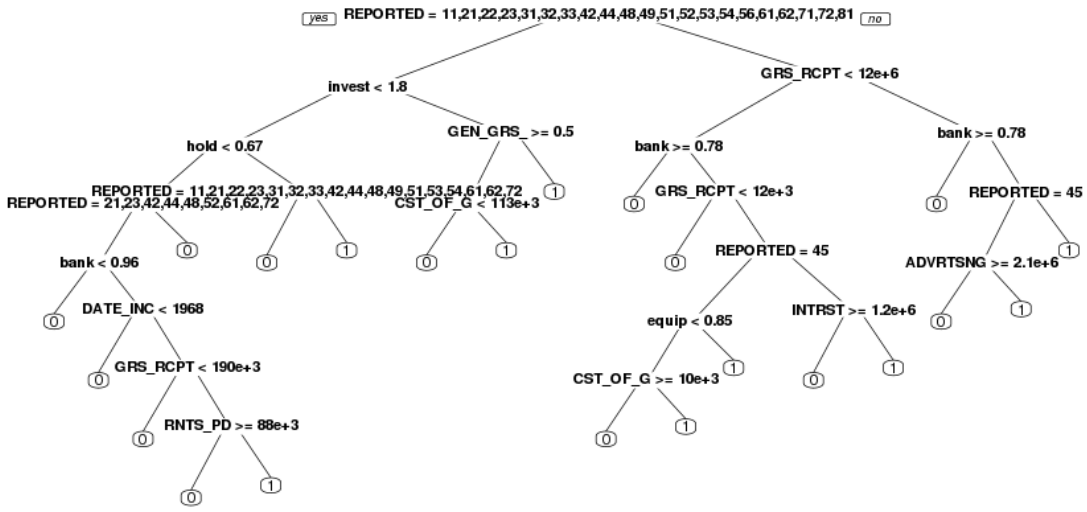
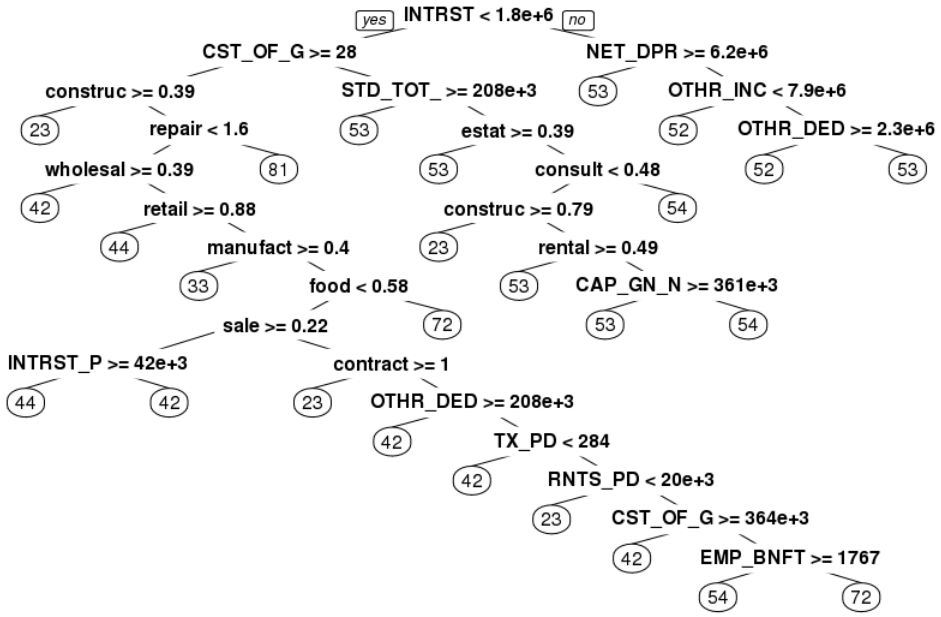


Figure 8: Classification tree predicting NAICS of noninformative group, CP=0.00375 (Form 1120)



3.3 Random Forest Models

Summary results for random forest models are in Table 11.[21] In all cases it improves over single CART models, especially for Form 1120. Improvement on the Schedule C valid reported set is less noticeable.

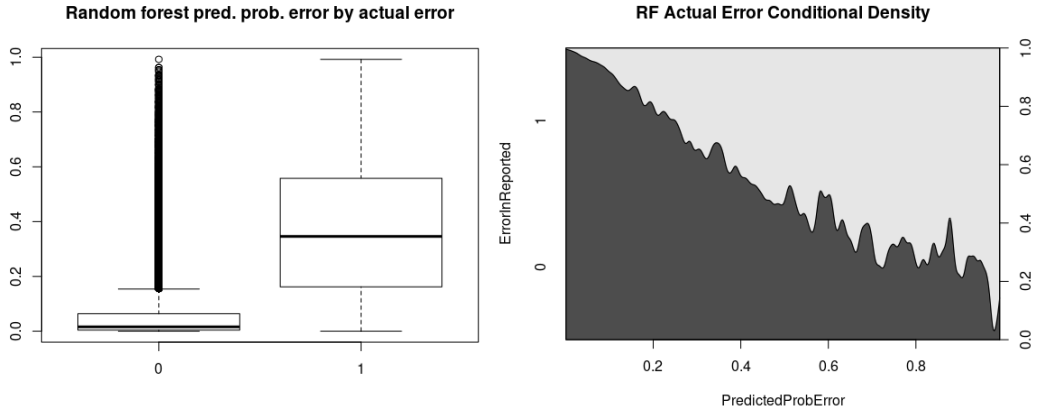
Table 11: Results for Random Forest Models

Model type	*Error prediction success*	*Error prediction MCC*	*NAICS prediction success (valid set)*	*NAICS prediction MCC (valid set)*	*NAICS prediction success (noninf. set)*	*NAICS prediction MCC (noninf. set)*
Schedule C	92.4%	0.404	92.1%	0.913	65.7%	0.627
Form 1120 (weighted)	91.7% (91.3%)	0.705	91.6% (91.7%)	0.908	74.7% (65.2%)	0.706

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

Figure 9-Figure 10 compare predicted probability of reporting error to actual error as we did with CART, and Figure 11-Figure 12 compare probability of predicted NAICS with prediction success on noninformative set. Especially for reporting error predictions, the plots are much cleaner. For Schedule C NAICS prediction we actually have less separation in the boxplots, but this is in part due to random forest producing higher probabilities for its predictions overall.

Figure 9: Schedule C random forest predicted probability of reporting errors compared to actual errors on holdout sample; boxplot and conditional density plot



[21] Random forest models are tuned with two parameters: number of trees and predictors tried per tree. We optimized with cross-validation.

Figure 10: Form 1120 random forest predicted probability of reporting errors compared to actual errors on holdout sample; boxplot and conditional density plot

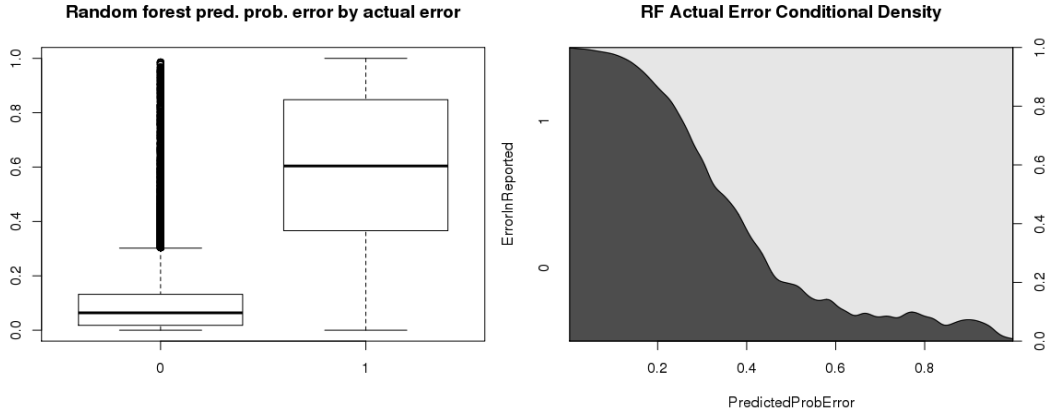


Figure 11: Schedule C random forest probability of predicted NAICS on noninformative set compared to prediction success; boxplot and conditional density plot

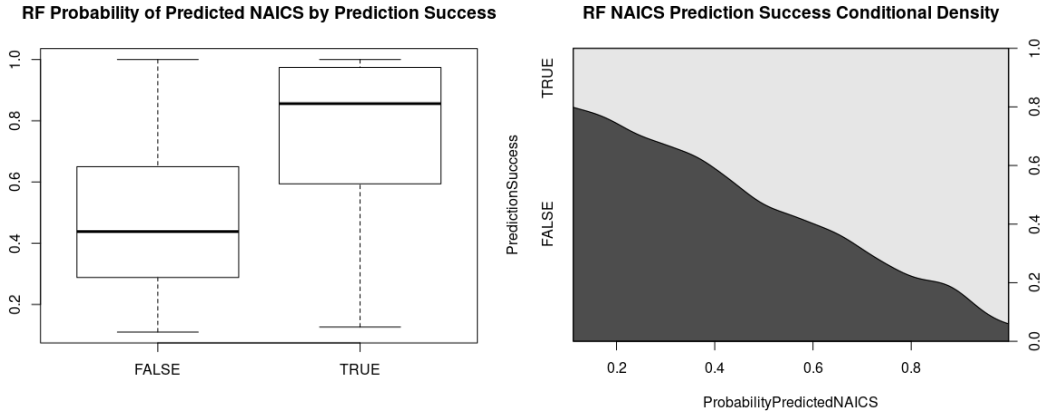
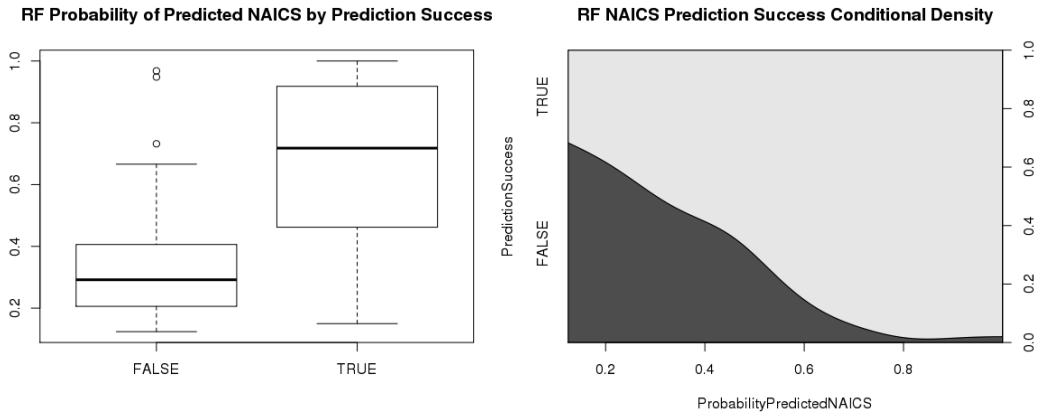


Figure 12: Form 1120 random forest probability of predicted NAICS on noninformative set compared to prediction success; boxplot and conditional density plot



It is not possible to plot a random forest, but we do have variable importance measured by mean decrease in node impurity when a predictor is included in a tree. Table 12 gives the

top fifteen predictors in each of the reporting error models. As in the CART models, when a valid reported NAICS is available it is by far the most important predictor. For Schedule C reporting error prediction there is a clearly defined second tier containing three line-item predictors, followed by a mix of line-items and text similarity measures. For Form 1120, there are only a few term predictors, but one is the most important after reported NAICS. In both cases line-item predictors are more prominent than they were in single trees.

Table 12: Scaled variable importance for random forest error prediction

Schedule C reporting error prediction (Top 15)		*Form 1120 reporting error prediction (Top 15)*	
Predictor	*Mean decrease GINI*	*Predictor*	*Mean decrease GINI*
Reported NAICS	7693.5	Reported NAICS	14243.6
Schedule C Profit/Loss	4717.0	"invest"	5405.9
Total Deductions	4256.4	*Total assets*	4115.3
Total Gross Receipts	4124.3	"hold"	4046.7
Other Expenses	2932.9	*Balance of Gross Receipts or Sales*	4046.3
Txt54	2167.3	Year incorporated	3939.4
Legal and Professional Services [Expense] Amount	1905.1	*Other Deductions*	3745
Car and Truck Expenses	1534.3	*Gross Profit*	3732.6
Office Expense Amount	1444.7	*Taxes and Licenses*	3329.1
Txt56	1434.7	*Cost of Goods Sold*	3253.8
Txt81	1347.3	*Depreciation*	3071.6
Travel Expenses	1344.4	*Rents*	3027.2
Meals and Entertainment [Expense] Amount	1335.0	*Interest* (paid)	2988.0
Txt52	1230.8	"compani"	2944.3
Insurance [Expense] Amount	1221.2	*Salaries and Wages*	2763.7

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

Table 13 gives the top fifteen predictors in each of the models for NAICS prediction on noninformative sets. Rather than being almost exclusively text predictors, the Schedule C model now includes some line items (which are also the most important line items in predicting reporting error); however, text predictors still dominate. In contrast the Form 1120 top fifteen is exclusively line items.[22]

Table 13: Scaled variable importance for random forest NAICS prediction on noninformative set

Schedule C NAICS prediction (Top 15)		*Form 1120 NAICS prediction (Top 15)*	
Predictor	*Mean decrease GINI*	*Predictor*	*Mean decrease GINI*
Txt54	3804.2	*Interest*	204.6
Schedule C Profit/Loss	2460.2	*Total assets*	149.6
Txt56	2264.3	*Other Deductions*	118.7
Total Gross Receipts	2248.3	*Balance of Gross Receipts or Sales*	111.1
Total Deductions	2082.2	Year incorporated	102.0
Txt52	1788.1	*Cost of Goods Sold*	99.8
Txt23	1529.9	*Gross Profit*	99.8

[22] The Form 1120 variable importance values are markedly smaller than those of the other models; this is because only a tiny fraction of Form 1120s do not have a valid NAICS when filed.

Schedule C NAICS prediction (Top 15)		*Form 1120 NAICS prediction (Top 15)*	
Predictor	*Mean decrease GINI*	*Predictor*	*Mean decrease GINI*
Txt81	1485.9	*Special Deductions* (inc. *Net Operating Loss Deduction*)	93.2
Txt62	1446.0	*Taxes and Licenses*	91.6
Txt71	1381.3	*Depreciation*	88.0
Txt53	1380.3	*Repairs and Maintenance*	72.7
Other Expenses	1338.8	*Rents*	72.4
Txt11	1327.5	*Salaries and Wages*	58.3
Txt61	1320.0	*Interest* (paid)	54.1
Txt33	1182.0	*Other Income*	54.0

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

We have focused on predicting NAICS on the noninformative set and predicting reporting error on the set of valid reported NAICS, with only brief mention of predicting NAICS on the valid reported set. The reason is that NAICS 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 an "un-correction". Table 14 holds confusion matrices for the results of different modeling approaches on the valid reported set.

Table 14: Valid reported set random forest model results confusion matrices

	Schedule C			*Form 1120*		
Reporting error prediction		Actual error	Actual correct		Actual error	Actual correct
	Predicted error	1500	916	Predicted error	6,211	549
	Predicted correct	3338	50,211	Predicted correct	4,015	44,508
	Falsely identified error rate:	1.6%		Falsely identified error rate:	1.0%	
NAICS prediction		Predicted incorrectly	Predicted correctly		Predicted incorrectly	Predicted correctly
	Reported incorrectly	3060	1778	Reported incorrectly	3,750	6,476
	Reported correctly	1358	49,769	Reported correctly	920	44,137
	Un-correction rate:	2.5%		Un-correction rate:	1.7%	

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

In both Schedule C and Form 1120, approaching the problem as prediction of reporting error is more conservative, i.e. less likely to identify a correctly reported industry sector as incorrect.

Comparing Table 14 and Table 15 (Schedule C results summary), we see that for Schedule C the un-correction rate is actually higher than the model's lift over assuming all reported

sectors are correct. Form 1120 has both a smaller un-correction rate and (see Table 16, Form 1120 results summary) a better improvement over baseline.

4. Conclusions

Table 15 summarizes the accuracy and MCC results for the different kinds of models and different modeling targets for Schedule C. Table 16 contains the summary for Form 1120. For both form types, in all three cases and in both accuracy and MCC, random forests were the best model.

Table 15: Summary of Schedule C model results

Model type	*Error prediction success*	*Error prediction MCC*	*NAICS prediction success (valid set)*	*NAICS prediction MCC (valid set)*	*NAICS prediction success (noninf. set)*	*NAICS prediction MCC (noninf. set)*
Baseline	91.4%	0	91.4%	0.905	16.8%	0
Classification tree	91.6%	0.370	91.9%	0.911	58.9%	0.555
Random forest	92.4%	0.404	92.1%	0.913	65.7%	0.627

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

Table 16: Summary of Form 1120 Results

Model type	*Error prediction success*	*Error prediction MCC*	*NAICS prediction success (valid set)*	*NAICS prediction MCC (valid set)*	*NAICS prediction success (noninf. set)*	*NAICS prediction MCC (noninf. set)*
Baseline (weighted)	81.5% (85.1%)	0	81.5% (85.1%)	0.797	24.7% (9.6%)	0
Classification tree (weighted)	89.6% (88.0%)	0.626	86.0% (86.8%)	0.847	56.3% (41.2%)	0.488
Random forest (weighted)	91.7% (91.3%)	0.705	91.6% (91.7%)	0.908	74.7% (65.2%)	0.706

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

For both Schedule C and Form 1120, random forest models using line-items and text fields can substantially improve on the lack of information provided by a missing, invalid, or noninformative reported industry sector. The models can also provide a 'confidence' in the prediction which researchers might use in deciding whether to accept a predicted value.

For Schedule C cases where there is a valid reported industry sector, the reporting accuracy rate is high and the small improvement given by the model comes at the cost of overriding correct values with incorrect ones. We feel it is preferable to accept the valid reported sector as filed but use a random forest reporting error prediction model to get some idea of how secure to be in it.

Form 1120 cases with valid reported industry sector have a higher error rate and the model is less likely to introduce mistakes, so researchers may want to replace as-filed values with model predictions.

Extending this methodology to three- or four-digit NAICS codes would be a more complex problem but is an avenue worth exploring.

Acknowledgements

Thank you: Barry Johnson, Director SOI, for the data and support. Mike Strudler, SOI Individual data, for invaluable assistance with SOI treatment of multiple Schedule Cs. Dr. Karl Branting, our panel discussant at JSM 2020, for helpful feedback.

References

Atkinson, Beth, and Terry Therneau. 2018. "Documentation for R package rpart v. 4.1-13." February 22.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 5-32.

Breiman, Leo, J.H. Friedman, R.A. Olsen, and C.J. Stone. 1984. *Classification and Regression Trees.* Monterey, CA: Wadsworth & Brooks/Cole Advanced Books and Software.

Chicco, Davide and Jerman, Guiseppe. 2020. "Advantages of the Matthews Correlation Coefficient." *BMC Genomics.*

Cuffe, John, Sudip Bhattacharjee, Ugochukwu Etudo, Justin C. Smith, Nevada Basdeo, Nathaniel Burbank, and Shawn R. Roberts. Forthcoming-Cited with permission. "Using Public Data to Generate Industrial Classification Codes." In *Big Data for 21st Century Economic Statistics*, by Katharine G. Abraham, Ron S. Jarmin, Brian Moyer and Matthew D. Shapiro. University of Chicago Press.

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

Gweon, H., M. Schonlau, L. Kaczmirek, M. Blohm, and S. Steiner. 2017. "Three methods for occupational coding based on statistical learning." *Journal of Official Statistics.*

Internal Revenue Service. 2018. "Instructions for Form 1065, U.S. Return of Partnership Income."

—. 2018. "Instructions for Form 1120, U.S. Corporation Income Tax Return."

—. 2017. "Instructions for Form SS-4, Application for Employer Identification Number."

—. 2018. "Instructions for Schedule C (Form 1040 or Form 1040-SR), Profit or Loss From Business (Sole Proprietorship) ."

Kearney, A. T., and M. E. Kornbau. 2005. "An automated industry coding application for new US business establishments." *Proceedings of the American Statistical Association.*

Liaw, A. 2018. "Documentation for R package randomForest v. 4.6-14." March 22.

Appendix: Additional Tables

Table 17: Confusion matrix for Schedule C reported vs. actual NAICS, part 1

Reported NAICS	11	21	22	23	31	32	33	42	44	45	48	49
11	6859	13	2	52	19	12	4	63	44	34	99	3
21	9	19845	3	21	0	2	2	5	5	14	25	0
22	4	3	413	14	0	3	0	6	0	5	2	1
23	19	40	8	24791	2	24	26	22	27	56	59	1
31	5	0	0	3	1939	5	7	12	17	16	0	0
32	0	34	0	28	5	1528	14	8	8	18	2	0
33	2	8	1	43	37	53	3006	43	35	38	34	2
42	22	20	3	37	37	30	37	7770	157	346	21	6
44	7	1	0	23	28	5	16	63	12455	381	14	0
45	10	5	2	20	22	13	30	136	316	16382	18	6
48	26	4	0	15	0	3	5	15	23	23	16478	57
49	2	0	0	0	0	1	3	4	8	53	32	1248
51	1	4	2	9	4	1	1	7	6	58	4	2
52	16	39	3	35	2	3	2	15	27	72	29	0
53	36	45	9	1051	1	6	10	18	31	70	108	6
54	82	168	11	197	15	50	58	128	96	483	71	5
56	11	11	0	184	6	8	3	42	27	85	65	7
61	4	1	0	1	1	1	4	2	1	24	1	0
62	10	2	0	12	1	6	8	6	42	31	10	2
71	103	4	0	20	3	18	8	24	13	101	63	3
72	13	2	0	53	53	0	1	12	104	63	12	8
81	159	20	4	540	41	29	22	72	140	508	208	45
Missing, Invalid, or 99	1128	848	110	4444	232	238	430	1048	1119	5619	1925	400

Table 18: Confusion matrix for Schedule C reported vs. actual NAICS, part 2

Reported NAICS	51	52	53	54	56	61	62	71	72	81	99
11	5	19	41	121	68	10	12	146	20	32	7
21	3	38	38	80	23	2	3	4	3	4	0
22	3	0	0	7	12	0	0	3	0	2	3
23	5	27	167	230	294	1	8	42	9	111	5
31	1	1	8	28	6	1	2	7	18	17	0
32	8	2	6	34	34	1	5	19	1	6	0
33	12	10	13	93	28	0	4	8	0	35	1
42	18	33	18	100	38	2	10	44	29	38	3
44	14	12	46	116	43	7	28	42	96	126	1

Reported NAICS	51	52	53	54	56	61	62	71	72	81	99
45	34	54	29	174	76	14	9	85	41	57	1
48	10	34	206	65	82	18	17	15	2	18	5
49	7	1	12	6	8	0	0	3	0	3	0
51	6210	24	37	293	56	10	11	324	1	10	2
52	15	27468	210	1041	543	7	18	51	15	25	7
53	20	226	46471	564	239	11	35	80	85	55	6
54	414	789	462	84412	4458	416	1057	1580	58	387	24
56	41	115	134	1095	15697	24	89	82	30	120	1
61	15	5	8	315	27	6305	141	275	3	13	0
62	11	18	25	427	126	66	41576	98	6	243	1
71	300	25	50	383	95	336	25	27490	77	70	1
72	5	18	43	63	121	6	28	58	14425	43	17
81	64	115	153	1449	2145	805	556	1390	72	25621	10
Missing, Invalid, or 99	1031	2782	2567	11606	8010	2687	3811	5900	909	4515	6711

Table 19: Confusion matrix for Form 1120 reported vs. actual NAICS, part 1

Reported NAICS	11	21	22	23	31	32	33	42	44	45	48	49
11	7921	11	27	22	173	46	8	581	45	11	38	50
21	7	5961	13	70	0	72	32	142	4	4	54	1
22	0	6	1358	68	0	1	4	42	0	7	4	0
23	10	124	30	26044	3	80	205	221	89	15	54	0
31	90	0	0	2	5698	118	75	526	75	8	1	5
32	19	47	12	52	56	10040	604	569	48	17	6	0
33	10	43	19	195	251	1149	24035	1484	67	52	36	6
42	70	27	21	65	291	327	674	31532	550	265	41	17
44	4	5	2	67	32	39	50	1473	21804	221	15	3
45	8	2	4	21	20	52	69	1699	888	4658	30	2
48	1	18	4	33	0	15	17	117	34	3	7797	54
49	3	0	0	2	2	0	6	41	4	0	86	714
51	2	5	1	12	0	37	32	64	32	42	11	3
52	15	53	31	101	17	130	186	218	79	16	38	19
53	87	39	12	1483	21	29	46	166	109	54	90	24
54	34	84	26	212	24	169	464	740	102	114	118	26
55	88	414	358	590	599	1666	3263	2304	693	270	590	74
56	5	11	21	129	17	33	44	144	38	34	38	5
61	0	3	0	2	0	0	0	11	0	12	0	0
62	1	0	0	3	0	9	42	52	48	6	0	0
71	14	0	0	2	0	3	22	27	14	26	27	2
72	4	0	0	6	16	1	0	25	35	5	6	0
81	25	14	4	227	3	23	76	353	110	34	114	1

Reported NAICS	11	21	22	23	31	32	33	42	44	45	48	49
Missing, Invalid, or 99	67	38	6	184	25	51	81	192	142	59	69	7

Table 20: Confusion matrix for Form 1120 reported vs. actual NAICS, part 2

Reported NAICS	49	51	52	53	54	55	56	61	62	71	72	81
11	50	3	27	208	83	66	34	0	5	38	13	0
21	1	0	22	76	59	71	26	0	4	0	4	8
22	0	33	7	7	28	50	58	0	0	0	3	4
23	0	4	10	336	221	83	363	5	7	6	12	26
31	5	0	4	12	10	48	7	0	0	1	16	8
32	0	53	8	33	365	70	53	0	11	0	0	21
33	6	192	13	44	528	108	45	4	29	7	0	131
42	17	120	73	90	205	136	99	4	21	12	31	35
44	3	49	21	163	71	52	14	2	28	20	245	155
45	2	92	114	84	140	18	145	2	4	8	36	71
48	54	5	12	236	57	26	133	7	23	6	8	68
49	714	3	1	22	13	4	11	0	0	0	0	0
51	3	8700	73	53	874	72	89	23	12	50	8	45
52	19	101	79836	1045	382	14017	173	8	66	28	34	32
53	24	47	568	39396	277	1652	157	16	71	107	274	1212
54	26	2183	519	365	30013	222	1029	156	324	129	37	255
55	74	999	4064	856	1634	17484	654	100	459	161	469	177
56	5	90	194	102	664	35	5484	11	92	13	29	118
61	0	13	4	0	25	0	6	1437	77	2	0	6
62	0	18	27	55	222	54	53	12	12104	13	4	35
71	2	248	3	36	76	60	48	119	0	3852	90	40
72	0	5	3	83	20	78	63	0	12	33	8235	18
81	1	59	101	155	436	36	831	79	133	161	14	6227
Missing, Invalid, or 99	7	70	808	882	304	124	110	21	76	34	84	126