ADEP WORKING PAPER SERIES

Ensemble Modeling Techniques for NAICS Classification in the Economic Census

Daniel Whitehead
U.S. Census Bureau

Brian Dumbacher
U.S. Census Bureau

Working Paper 2024-03
June 2024

Associate Directorate for Economic Programs
U.S. Census Bureau
Washington DC 20233

Ensemble Modeling Techniques for NAICS Classification in the Economic Census

Daniel Whitehead, U.S. Census Bureau, daniel.whitehead@census.gov
Brian Dumbacher, U.S. Census Bureau, brian.dumbacher@census.gov

ADEP Working Paper 2024-03
June 2024

Abstract

The Business Establishment Automated Classification of NAICS (BEACON) is a machine learning tool developed by the U.S. Census Bureau to help Economic Census respondents select their establishment's North American Industry Classification System (NAICS) code. BEACON uses the respondent-provided text, in real time, to predict the respondent's most likely NAICS code. BEACON utilizes past Economic Census responses in conjunction with other data sources such as NAICS manual descriptions and Internal Revenue Service data to create a data dictionary for training and testing purposes. Through an ensemble method, BEACON hierarchically predicts a respondent's NAICS code, first at the 2-digit level and then at the 6-digit level. As a potential means of improving BEACON's current prediction method, we are exploring the use of model stacking to incorporate predictions from alternative models. This research paper details the ensemble modeling behind BEACON and explores this application of model stacking to improve predictions.

Keywords:

BEACON, Economic Census, ensemble method, hierarchical modeling, information retrieval, machine learning, NAICS, text classification, model stacking

JEL Classification Codes:

C. Mathematical and Quantitative Methods

- C1. Econometric and Statistical Methods and Methodology: General
    - C18. Methodological Issues: General

Ensemble Modeling Techniques for NAICS Classification in the Economic Census

Daniel Whitehead, Brian Dumbacher
Daniel.Whitehead@Census.gov, Brian.Dumbacher@Census.gov
U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233

Abstract: The Business Establishment Automated Classification of NAICS (BEACON) is a machine learning tool developed by the U.S. Census Bureau to help Economic Census respondents select their establishment's North American Industry Classification System (NAICS) code. BEACON uses the respondent-provided text, in real time, to predict the respondent's most likely NAICS code. BEACON utilizes past Economic Census responses in conjunction with other data sources such as NAICS manual descriptions and Internal Revenue Service data[1] to create a data dictionary for training and testing purposes. Through an ensemble method, BEACON hierarchically predicts a respondent's NAICS code, first at the 2-digit level and then at the 6-digit level. As a potential means of improving BEACON's current prediction method, we are exploring the use of model stacking to incorporate predictions from alternative models. This research paper details the ensemble modeling behind BEACON and explores this application of model stacking to improve predictions.

Key words: BEACON, Economic Census, ensemble method, hierarchical modeling, information retrieval, machine learning, NAICS, text classification, model stacking

Disclaimer: *Any opinions and conclusions expressed herein are those of the authors and do not reflect the views of the U.S. Census Bureau. The Census Bureau has reviewed this data product to ensure appropriate access, use, and disclosure avoidance protection of the confidential source data used to produce this product [Data Management System (DMS) number: P-7504847, subproject P-7514952; Disclosure Review Board (DRB) approval number: CBDRB-FY23-ESMD001-034].*

Organization

Section 1 contains background concerning NAICS, the Economic Census, and BEACON. Section 2 discusses model stacking, its current use in BEACON, and proposed methods. Section 3 details the evaluation methodology. Section 4 contains the results, and Section 5 presents conclusions.

[1] IRS data used for internal statistical purposes only, in accordance with Title 26.

1. Background

1.1 NAICS/Economic Census Discussion

Established in 1997, the North American Industry Classification System (NAICS) is used by the U.S. Census Bureau and other statistical agencies to classify businesses according to their primary business activity (U.S. Census Bureau, 2023d). NAICS plays an important role throughout the economic survey life cycle, including sample design, data analysis, and publication (Dumbacher and Whitehead, 2022). NAICS is hierarchical and allows one to decipher quickly what activities a business is engaged in at different levels of detail. The first two digits of a NAICS code identify the broad sector of economic activity whereas the full six digits denote the specific industry the business operates in. For more on the history of NAICS as well as a complete catalog of NAICS codes, see https://www.census.gov/naics/ (U.S. Census Bureau, 2023d).

For years ending in "2" and "7", the Census Bureau conducts the Economic Census, a large survey that covers over eight million business establishments with paid employees (U.S. Census Bureau, 2023a). About half of the establishments are sent an electronic questionnaire whereas the other half is accounted for through administrative data (Dumbacher and Whitehead, 2022). Most industries and all geographic areas of the U.S., including island territories, are in scope (U.S. Census Bureau, 2023a). The Economic Census provides a wealth of detailed information about U.S. economic activity (Dumbacher and Whitehead, 2022). Some of the crucial statistics provided by it include total revenue, total number of employees, and total annual payroll (Dumbacher and Whitehead, 2022). Data products are broken down by geography and industry, as classified by NAICS (Dumbacher and Whitehead, 2022). For details about the design and methodology of the Economic Census, see https://www.census.gov/programs-surveys/economic-census/technical-documentation.html (U.S. Census Bureau, 2023c). The 1997 Economic Census was the first major survey to use NAICS (U.S. Census Bureau, 2023b). For further information regarding the adoption of NAICS by the Economic Census, see Wiley and Whitehead (2022).

1.2 Motivating Problem

The Primary Business or Activity (PBA) question on the Economic Census questionnaire asks respondents to describe their business (Dumbacher and Whitehead, 2022). Answers to this question help keep NAICS code assignments up to date on the Business Register, the Census Bureau's master list of businesses (Dumbacher and Whitehead, 2022). The respondent is presented with prelisted descriptions but can also provide a short, open-ended response (Dumbacher and Whitehead, 2022). To illustrate, Figure 1 is a screenshot of the PBA question from the drinking places questionnaire. Every Economic Census, there are hundreds of thousands of these PBA "write-in" responses. For the most part, clerks process and assign NAICS codes manually, which is resource intensive. Using automated methods can improve efficiency. To this end, the Census Bureau developed a model called BEACON (Business Establishment Automated Classification of NAICS) to help respondents self-designate their 6-digit NAICS code in real time (Dumbacher and Whitehead, 2022). BEACON takes the respondent-provided write-in as input and returns a ranked list of candidate 6-digit NAICS codes for the respondent to choose from (Dumbacher and Whitehead, 2022).

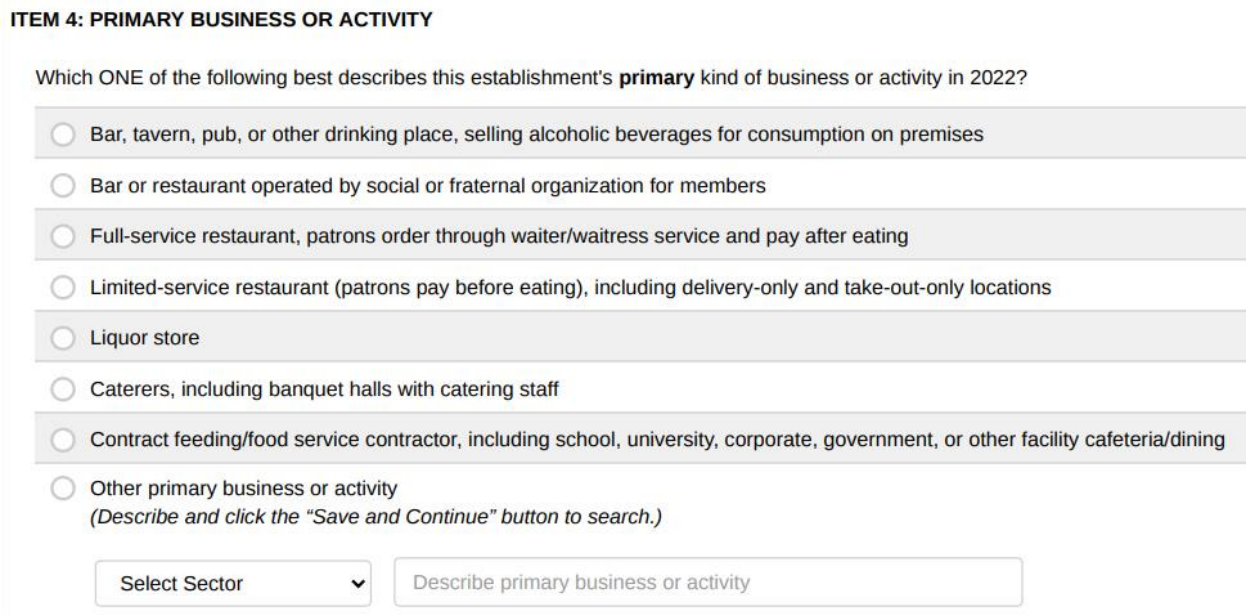

Figure 1. Primary Business or Activity (PBA) question from the 2022 Economic Census drinking places questionnaire (AF-72240). Example write-ins include "liquor distribution" and "brewpub". Source: https://bhs.econ.census.gov/ombpdfs2022/export/2022_AF-72240_su.pdf

1.3 Overview of BEACON's Methodology

BEACON's text classification methodology is based on information retrieval, a field that includes internet and database search (Dumbacher and Whitehead, 2022). Early in the modeling research, we explored various techniques including logistic regression, naïve Bayes, and random forests. Ultimately, we determined that the information retrieval framework provided the best means of recognizing a large vocabulary and generating useful NAICS predictions while being computationally feasible. BEACON has been a success. In both internal testing and real-world rollouts, it has achieved its goals of taking a respondent's short business description and returning relevant NAICS codes for the respondent to choose from.

The idea behind the information retrieval approach is to assign relevance scores to all 6-digit NAICS codes for ranking purposes. The higher the score, the more confident BEACON is that the NAICS code is correct. BEACON is based on a collection, or ensemble, of three information retrieval sub-models. The sub-models use different sets of text features to inform score assignment. The text features that are more highly associated with a particular NAICS code have more influence. The final scores equal a weighted average of the scores from the three sub-models. A cross-validation grid search was used to derive the three ensemble weights used in this weighted average (Dumbacher and Whitehead, 2022). The model ensemble is not applied directly at the 6-digit NAICS level. Instead, a two-stage process is used that takes advantage of the NAICS hierarchy. First, the ensemble assigns scores at the 2-digit sector level. Next, the ensemble assigns sector-conditional scores at the 6-digit level within each sector. Using the conditional probability formula, BEACON calculates the unconditional scores at the 6-digit level.

2. Model Stacking

2.1 Discussion of Model Stacking

Model stacking is the idea of using statistical models to assess, or model, the output from other models. The idea is that because no model is perfect, it is better to approach a problem with multiple models and borrow from each (Merz, 1999, p.33). Thus, modeling is divided into two stages: (1) generate multiple initial models to make predictions, and then (2) apply a so-called "meta-model" that uses the initial models' output as input to make final predictions (Todorovski and Džeroski, 2003, p. 223; Merz, 1999, p. 33). For the meta-model, "the features are the predictions of the base-level classifiers and the class is the correct class of the example at hand." (Džeroski and Ženko, 2004, p. 257). According to Merz (1999, p. 33), "it is important to generate a set of models that are diverse in the sense that they make errors in different ways". Ideally, the meta-model would always be able to lean on correct predictions in each domain from one or more of the input classifiers, even if the classifiers as a whole perform poorly in certain domains.

Examples of meta-models that will be considered later in our study include logistic regression, decision trees, and random forests. "Logistic regression is the standard way to model binary outcomes (that is, data y_i that takes on the values of 0 or 1)" (Gelman and Hall, 2007, p. 79). For example, in the context of NAICS classification, a value of 1 indicates that the correct NAICS code was provided by the model, whereas a value of 0 indicates failure. We will also make use of decision trees, which "partition the feature space into a set of rectangles, and then fit a simple model (like a constant) in each one" (Hastie et al., 2017, p. 305). As each partition is a based on a single feature, when a decision tree is used as a meta-model, each partition is derived from a single classifier. However, the complete tree may take advantage of each classifier as the data dictates. "Bagging, …, is a technique for reducing the variance of an estimated prediction function" (Hastie et al., 2017, p. 587). It has been found to be very useful when applied to decision trees (Hastie et al., 2017, p. 587). This principle may have inspired the use of random forests, which will be our third meta-model. Random forests operate "according to the simple but effective bagging principle: sample fractions of the data, grow a predictor (a decision tree in the case of forests) on each small piece, and then paste the results together" (Biau, 2019, p. 348). In other words, random forests themselves are a form of model stacking where many models (trees) are fit to the training data and their predictions are synthesized into a single prediction (Merz, 1999).

2.2 Current Use of Model Stacking in BEACON

The score averaging currently used by BEACON can be considered a simple version of model stacking as "the individual predictions are combined … to classify new examples" (Džeroski and Ženko, 2004, p. 255). BEACON currently fits three separate information retrieval sub-models using three different sets of text features: "standard" (all combinations of words up to three in length), "umbrella" (combinations of words up to three in length excluding lower-order combinations), and "exact" (the complete and entire respondent write-in). Each sub-model is capable of assigning relevance scores to all 6-digit NAICS codes. The final scores equal a weighted average of the scores from the three sub-models. As discussed in Dumbacher and Whitehead (2022), a cross-validation grid search was used to derive the three ensemble weights used in this weighted average. The three weights sum to one, so there were two degrees of freedom and two parameters to optimize. Table 1 summarizes BEACON's model ensemble.

Table 1. Summary of BEACON's model ensemble

Sub-model	Features	Comment	Ensemble weight
"standard"	All combinations of words up to three in length	Allows all words and word combinations to contribute to the scores	0.1
"umbrella"	Combinations of words up to three in length excluding lower-order combinations	Focuses on the detail of the respondent's write-in	0.6
"exact"	The complete and entire respondent write-in	Feature is unique but may not occur frequently enough to have predictive value	0.3

The three sub-models consider different aspects of the respondent's write-in. The "standard" sub-model takes into account all words and word combinations and allows all of them to contribute to the score calculation. In this regard, the "standard" sub-model helps protect against model overconfidence. The "umbrella" sub-model, which receives the most weight in the ensemble, also considers words and word combinations but focuses on the detail of the text. Individually, the "exact" sub-model is the weakest of the three as often the entire write-in occurs too infrequently to have predictive value. Yet, its uniqueness strengthens the overall prediction; its biases are different from those of the "standard" and "umbrella" models in such a way that it can complement and alleviate the biases of the other models (Merz, 1999, p. 35).

2.3 Proposed Methods

The ensemble weights allow BEACON's prediction to be informed by all three sets of features, but they do not allow for any variation in how the three sub-models inform BEACON's prediction. Regardless of sector or any other context, the relevance scores from the three sub-models are simply averaged using the same ensemble weights (i.e., 0.1, 0.6, and 0.3) to calculate BEACON's final scores, which are used for prediction. In this paper, we propose a more robust version of model stacking for the NAICS classification problem. We will compute relevance scores using BEACON's three information retrieval sub-models but then use this output as input to a group of "meta-models". In a sense, the meta-models will evaluate the best values for the parameters used to combine information from the three sub-models. We will consider logistic regression, decision tree, and random forest as meta-models. To emphasize, these meta-models will use the sub-models' relevance scores as input. The meta-models will not use textual features as input, which, as mentioned in Section 1.3, were found to present computational limitations early in the development of BEACON.

3. Model Stacking Evaluation Methodology

3.1 Evaluation Overview

To evaluate our proposed model stacking method, we compare the accuracy of the predictions of BEACON's current information retrieval sub-models, with some slight adjustments, to the accuracy of meta-models trained on the sub-models' scores. Accuracy is defined in two ways: (1) the percentage of times the true NAICS code is included among the top-k (scikit-learn, 2023d) results, with k ranging from 1 to 5, and (2) as measured by the F_1 score, which "can be interpreted as a harmonic mean of the precision and recall" (scikit-learn, 2023b). To do so, we divide BEACON's training data within each sector into five separate partitions of data such that roughly one-fifth of the data within each sector is omitted from each partition. For each partition, the data within it is treated as the training data, while the missing fold is used as test data for predictions by the models. Just as in BEACON's production model, the text from each respondent write-in is first cleaned and organized into the building blocks for BEACON's information retrieval sub-models: word combinations of varying length and the entire text description, post cleaning. Unlike with the version of BEACON used in production, restrictions on the minimum frequency of words by data source are removed to maximize the size of the training data within each partition. This simplifies the data generation process and allows richer training sets to be developed within each partition of the data.

Within each sector, we fit the BEACON information retrieval models to the training or partition data to compute predictions at the 3-digit and 6-digit NAICS levels. Exploratory work was also done at the 4-digit and 5-digit levels. "In [model] stacking, a learning algorithm is used to learn how to combine the predictions of the base-level classifiers." (Todorovski, L. and Džeroski, S., 2003, p. 223). As described in Section 2.2, BEACON combines the scores from the individual component learning models using fixed ensemble weights derived through cross-validation (Dumbacher and Whitehead, 2022). For our current research, we take this process a step further and feed these scores as inputs to the following meta-models: logistic regression, decision tree, and random forest. We fit these meta-models separately for each NAICS code within the sector to predict whether the write-in corresponds to the given NAICS code. We then calculate the meta-models' performance on the held-out fold from each partition as the test data and compare the performance to that of BEACON, using both the top-k% and the F_1 score of the performance on the held-out folds. Performance metrics are averaged across folds; nearly identical results were obtained by simply summing across the folds before computing the performance metrics. Comparisons are made both at the level of the individual NAICS code and across all of the NAICS codes of an individual sector.

3.2 Evaluation Limitations

Some limitations of this approach are as follows. Unlike production BEACON, to best incorporate and compare the use of the meta-models, we do not apply the hierarchical framework of BEACON. Instead, we only compare training/test data either per individual sector or across all sectors, in the case of the 2-digit NAICS codes. To apply the meta-models, we limit ourselves to using only the scores from the three information retrieval sub-models as predictor variables. However, it would be possible to incorporate the length of the write-in, number of words, or other such features as inputs.

Also, for computational ease, we initially did not deviate from the default settings used by the Python module sklearn (Pedregosa et al., 2011). This meant for the decision tree and random forest models that the minimum number of samples split and minimum number of samples per leaf were originally kept at the default values. We later re-ran the decision tree and random forest models, auto-adjusting parameters such as the number of estimators and maximum depth of each tree for each NAICS code as each partition was run. Results for the decision tree and random forest meta-models are based on these adjustments. By adjusting the parameters of the decision tree based on the model's performance on the training data, its performance increased relative to that of the other meta-models and of BEACON. We saw little change though in the performance of the random forest meta-model from parameter tuning, as its default settings are rather robust (scikit-learn, 2023a). However, for the logistic regression model, results are based on the default solver (LBFGS), penalty term (L2), and inverse regularization parameter (C=1) as attempts to adjust the logistic regression parameters actually lowered performance (scikit-learn, 2023c). As the number of features is limited to the predictions from the three individual component models, we do not believe these choices greatly affected the final results, but we acknowledge these limitations.

4. Results

We compare both the performance of the information retrieval sub-models to BEACON as well as that of the meta-models. For each sector, we analyze performance in predicting the correct 3-digit and 6-digit NAICS code[2]. For sectors with a large number of observations, we select a random stratified sub-sample of observations[3] to partition into 5 folds. This is done to reduce the computational burden in processing the various models across the sectors and NAICS code level. Finally, a sample of approximately 250,000 observations, across all sectors, is selected to assess performance at the 2-digit level.

When looking at the performance of the individual component models by the top-k% we see that the internal model stacking within BEACON tends to improve upon the performance of the individual models. At the 3-digit level, BEACON outperforms all of its component models in all sectors studied, for k = 1 (see Figure 2). As k increases, the differences understandably narrow, as many sectors have a small number of 3-digit NAICS codes (for *k* = 3, see Figure 3). Despite this limitation, for *k* = 5, BEACON is only edged in 2 of 15 sectors by the "standard" model and outperforms the "standard" model in 3 sectors, "umbrella" in 5, and the "exact" model in 5, tying in all other sectors. Although the "exact" sub-model is the least powerful model individually, its inclusion with the "standard" and "umbrella" models may be the difference that allows the "combined" (BEACON) model to edge out each of the component models. As BEACON is a weighted average of the three sub-models, each sub-model influences the final prediction.

[2] At the 3-digit level, the following sectors were excluded: Utilities (22), Wholesale Trade (42), Professional, Scientific, and Technical Services (54), Management of Companies and Enterprises (55), and Educational Services (61).

[3] For manufacturing, a sample of approximately 100,000 observations was chosen. For all other sectors that required a reduced subsample, approximately 250,000 observations were sampled. For the 2-digit sample, a stratified sample of approximately 250,000 observations was selected across all sectors.

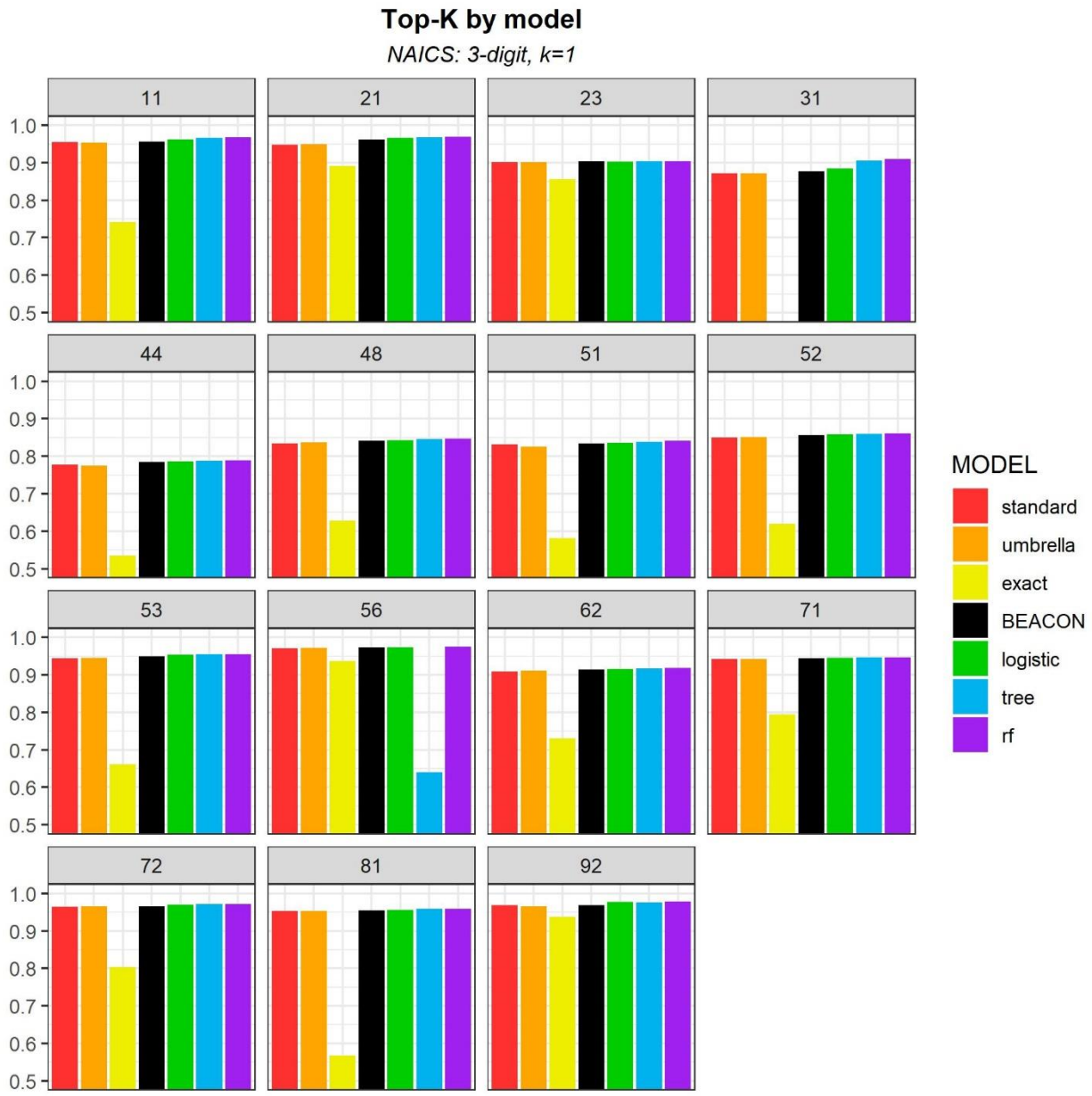

Figure 2: Percentage of test observations by sector where the true 3-digit NAICS code was the prediction with the highest score. Results are averaged across folds.

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

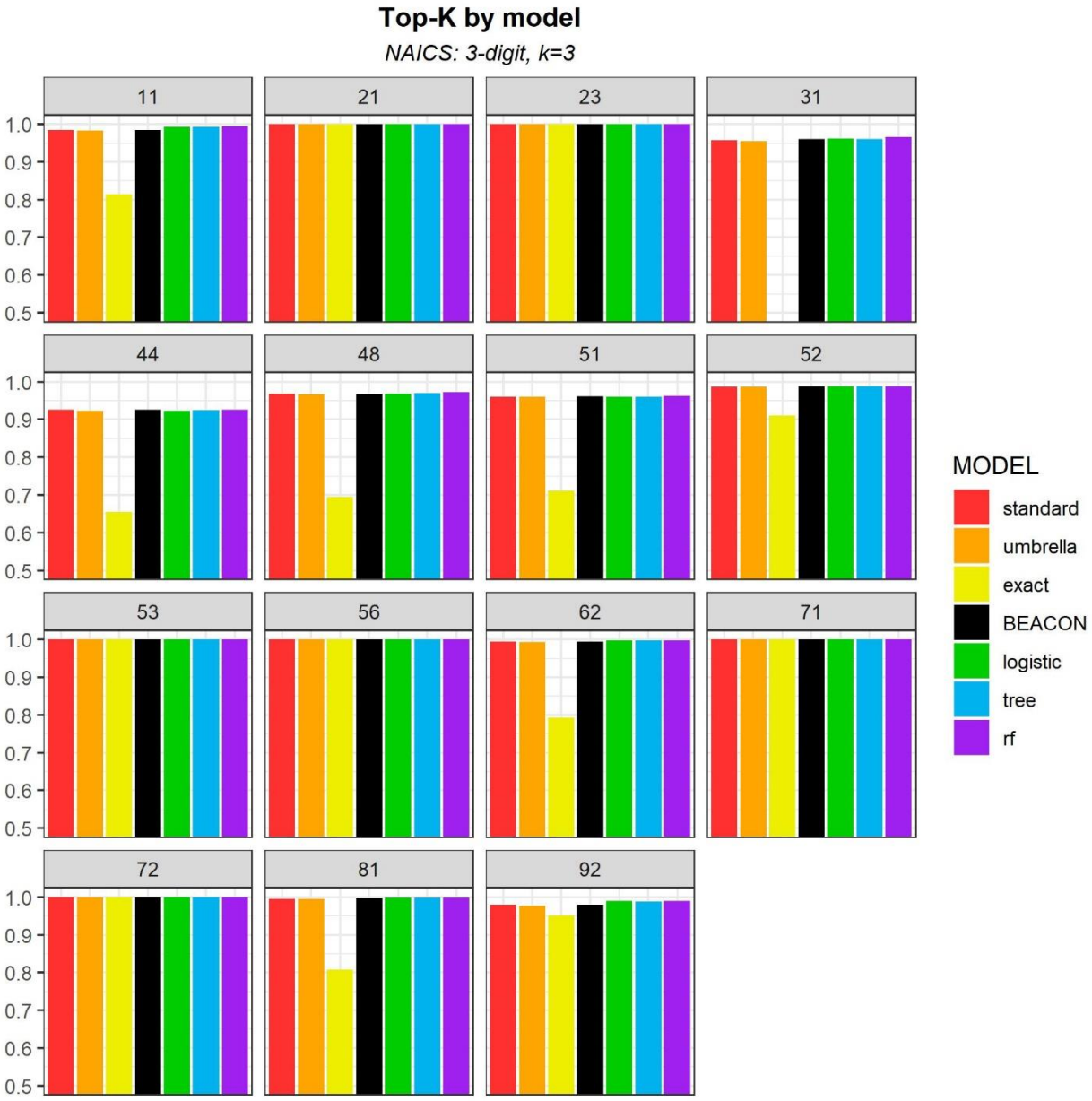

Figure 3: Percentage of test observations by sector where the true 3-digit NAICS code was in the top 3 codes predicted. Results are averaged across folds.
Data Sources: Economic Census (2002–2022), 2021 Industry Classification Report, Internal Revenue Service SS-4 (2002–2016), Classification Analytical Processing System, Harmonized System

Similarly, applying the meta-models to the output from the information retrieval sub-models within BEACON offers potential for slight improvement as measured by the top-k %. At the 6-digit level, for 19 of the 20 sectors studied, the random forest meta-models slightly outperform both the other meta-models and BEACON in including the true NAICS code as its top prediction (see Figure 4). Once again, as *k* increases, this difference narrows (for *k* = 3, see Figure 5). When k is increased to 5, the random forest method outperforms BEACON in only 10 of 20 sectors, while BEACON outperforms it in 9 of 20 sectors.

Similarly at the 3-digit level, when k is only 1, random forest slightly outperforms BEACON in all 15[4] sectors studied. However, when k is increased to 5, the random forest method only outperforms BEACON in 4 out of 15 sectors, tying in 10 sectors. Interestingly, the meta-models offer the most potential improvement in the manufacturing sector (31), where the number of NAICS codes to choose from is highest.

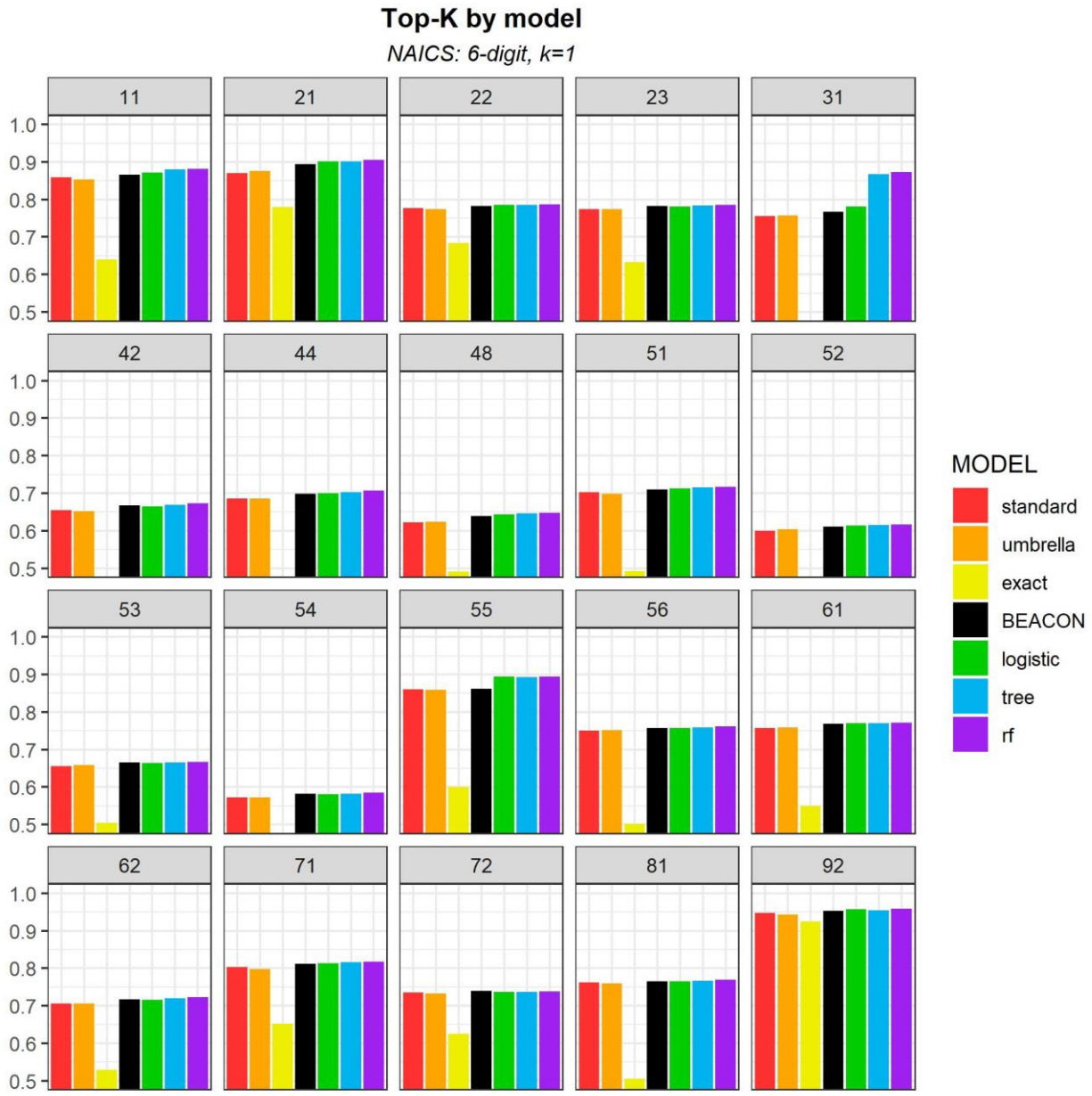

Figure 4: Percentage of test observations by sector where the true 6-digit NAICS code was the prediction with the highest score. Results are averaged across folds.
Data Sources: Economic Census (2002–2022), 2021 Industry Classification Report, Internal Revenue Service SS-4 (2002–2016), Classification Analytical Processing System, Harmonized System

[4] At the 3-digit level, only 15 sectors that contain more than one 3-digit NAICS code were studied.

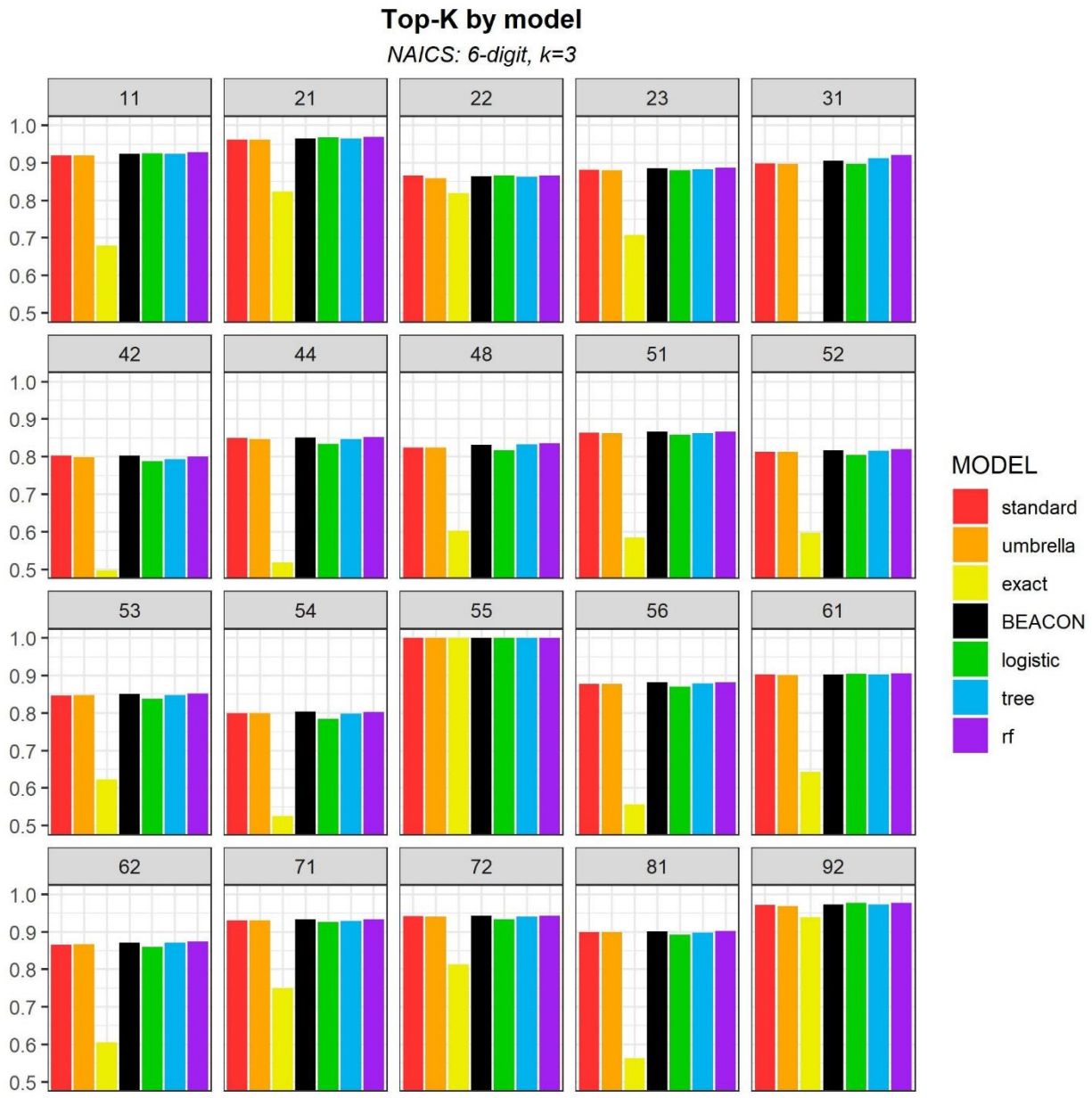

Figure 5: Percentage of test observations by sector where the true 6-digit NAICS code was in the top 3 codes predicted. Results are averaged across folds.

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

Model stacking also offers potential improvement at the 2-digit, or sector level. Here, BEACON slightly improves upon the sub-models. While the "standard" and "umbrella" sub-models correctly predict the true NAICS code as the top choice for 76.6% and 76.9%, respectively, of the test observations, averaged across the 5 folds, BEACON does so at a rate of 77.2%. Similarly, two of the meta-models improve upon BEACON's performance. The random forest meta-model correctly predicts the true NAICS code as its top choice for 78.1% of the testing observations while the logistic regression meta-model does so for 77.5%

of observations. Yet for k = 4 and k = 5, BEACON performs best, correctly predicting the true NAICS code within the top-k observations at a higher rate than all other models, though the random forest and standard models succeed at nearly an identical rate (see Figure 6). This makes sense as BEACON is not optimized to provide a single prediction; rather, its goal is to provide multiple, reasonable NAICS codes to the respondents who then use subject-matter knowledge of their own business to make a selection.

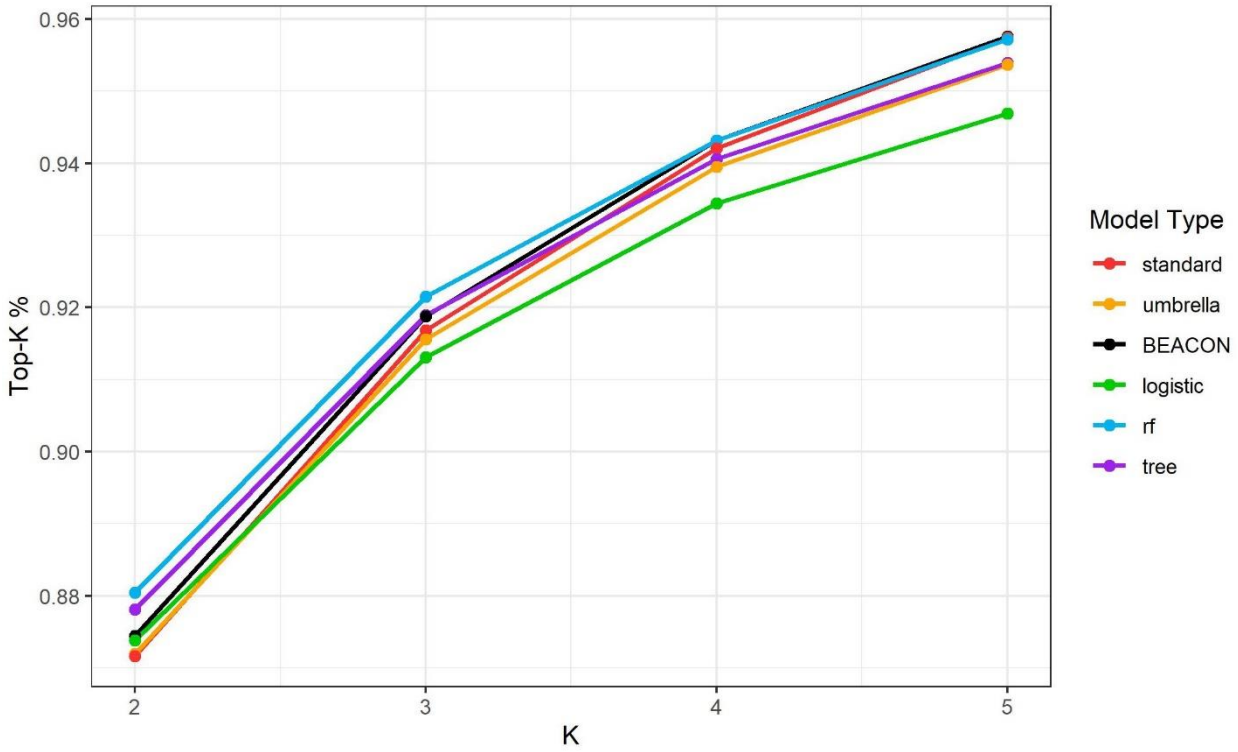

Figure 6: Percentage of test observations where true 2-digit NAICS code was the top choice predicted. Results are averaged across folds. The "exact" sub-model is omitted from this plot.
Data Sources: Economic Census (2002–2022), 2021 Industry Classification Report, Internal Revenue Service SS-4 (2002–2016), Classification Analytical Processing System, Harmonized System

Evaluating the models using the F_1 score provides similar insights. At the 6-digit level, the random forest meta-model outperforms BEACON in 73% of the 6-digit codes studied, while it underperforms BEACON in only 16%. Similarly, the logistic regression and decision tree meta-models outperform BEACON in 67% and 66%, respectively, of the 6-digit NAICS codes. However, the mean difference in F_1 score between the random forest model and BEACON is only 0.024. Likewise, the mean difference is only 0.007 for the logistic regression model and just 0.022 for the decision tree. Thus, any potential improvement offered by the meta-models may be limited without the use of additional information.

Ideally, the meta-models would support BEACON in areas where BEACON does not perform as well. However, if anything, the meta models may perform slightly better in sectors where BEACON already performs well. At the 6-digit level, BEACON has a median F_1 score of 0.88; in sectors where random forest outperforms BEACON, BEACON's median F_1 score is 0.89, compared to 0.85 in sectors where BEACON outperforms the random forest method. Similarly, in sectors 21 (Mining, Quarrying, and Oil and Gas Extraction) and 92 [Other Services (except Public Administration)], for example, we see slight

improvement from the logistic regression and random forest models over BEACON. However, BEACON is already predicting the correct NAICS code as the first choice for around 89% of the test observations in sector 21 and 95% of the test observations in sector 92. Any improvement by the meta-models in these sectors is not as useful as the potential improvement offered by the meta-models in sectors such as 52 (Finance and Insurance) or 54 (Professional, Scientific, and Technical Services) where BEACON does not perform as well. On the other hand, the median F_1 score achieved by random forest and decision tree across NAICS codes in sector 31 (Manufacturing) is 0.95 compared to 0.89 for BEACON. As manufacturing is such a vital sector for the Economic Census, this result is encouraging for the potential of model stacking to complement predictions from BEACON's component models.

5. Conclusion

In this work, we have demonstrated both the current use of model stacking within BEACON as well as the potential to incorporate the scores from BEACON's sub-models as features into a meta-model. Through comparisons based on both the top-k% of true predictions as well as the F_1 score, we saw the use of meta-models to bolster predictions may be worthwhile in sectors for which it is challenging to predict the true NAICS code. However, any potential improvement provided by the meta-models is only an incremental improvement as BEACON's current model-stacking performs very well as is.

When comparing the meta-models, the random forest model tended to slightly outperform the decision tree model. Because the number of input parameters in our experiment is limited to the three internal predictions from the BEACON sub-models, we are not surprised that the "model-stacking" provided by the random forest outperforms a single decision tree model as the random forest is essentially an averaging of many simple decision trees. Also, the random forest and decision tree methods both slightly outperformed the logistic regression model. As noted, we attempted to tune the logistic regression parameters, but doing so only decreased the model's performance on the test data.

One interesting result is that the meta-models may offer more potential for providing a single predicted NAICS than for providing a menu of reasonable NAICS codes to a respondent. Whereas BEACON strives to predict the true NAICS code with the highest probability, it is not necessarily optimized to do so. Rather, if the true NAICS code is included within a reasonable number of alternatives for a single query, that is considered a successful result for BEACON. When comparing the performance of BEACON to the meta-models, we see via the top-k measure that the meta models sometimes outperform BEACON in predicting a single NAICS code, but that they are less likely to do so when the top-k is expanded to include the top 3, 4, or 5 predictions. Such a feature may be useful if BEACON is used to auto-code responses instead of assisting the respondent. For such an application, success would depend solely on the single prediction provided by the model rather than the entire suite of predictions.

Topics for future research include further tuning of the meta models, combining the meta-models into a single meta-model, and incorporating additional parameters into the meta-models such as the length of the write-in or the number of words it contains. Ideally, the meta-models would provide predictions very similar to BEACON for sectors where BEACON already performs outstanding while complementing BEACON in those sectors and NAICS codes that are difficult to predict correctly. The meta-models' performance in the manufacturing sector (31), with its wide assortment of NAICS codes, offers potential in the application of model stacking to supplement the component models within BEACON.

References

Biau, G., Scornet, E., and Welbl, J. (2019). Neural Random Forests. *Sankhyā: The Indian Journal of Statistics, Series A*, *81*(2), 347–386.

Dumbacher, B. and Whitehead, D. (2022). Industry Self-Classification in the Economic Census*. 2022 Proceedings of the American Statistical Association, Section on Statistical Learning and Data Science.* Alexandria, VA: American Statistical Association.

Džeroski, S. and Ženko, B. (2004). Is Combining Classifiers with Stacking Better than Selecting the Best One? *Machine Learning, 54*, 255–273.

Gelman, A. and Hall, J. (2007). Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge, United Kingdom: Cambridge University Press.

Hastie, T., Tibshirani, R., and Friedman, J. (2017). The Elements of Statistical Learning. (2nd ed.) Springer Series in Statistics. New York City: Springer Press.

Merz, C. (1999). Using Correspondence Analysis to Combine Classifiers. *Machine Learning*, *36*, 33–58.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Scikit-learn documentation (2023a). scikit-learn 1.3.0: sklearn.ensemble.RandomForestClassifier. Retrieved July 18, 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier

Scikit-learn documentation (2023b). scikit-learn 1.3.0: sklearn.metrics.f1_score. Retrieved August 15, 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score

Scikit-learn documentation (2023c). scikit-learn 1.3.0: sklearn.metrics.linear_model.LogisticRegression. Retrieved October 17, 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression

Scikit-learn documentation (2023d). scikit-learn 1.3.2: sklearn.metrics.top_k_accuracy_score. Retrieved December 20, 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.top_k_accuracy_score.html

Todorovski, L. and Džeroski, S. (2003). Combining Classifiers with Meta Decision Trees. *Machine Learning*, *50*, 223–249. Netherlands: Kluwer Academic Publishers.

U.S. Census Bureau. (2023a). About the Economic Census. Retrieved October 19, 2023, from https://www.census.gov/programs-surveys/economic-census/year/2022/about.html

U.S. Census Bureau. (2023b). Economic Census. Retrieved May 17, 2023, from https://www.census.gov/history/www/programs/economic/economic_census.html

U.S. Census Bureau. (2023c). Economic Census Technical Documentation. Retrieved December 20, 2023, from https://www.census.gov/programs-surveys/economic-census/technical-documentation.html

U.S. Census Bureau. (2023d). Introduction to NAICS. Retrieved August 15, 2023, from https://www.census.gov/naics

Wiley, E. and Whitehead, D. (2022). Development of Interactive Classification Tools. Federal Committee on Statistical Methodology 2022 Research and Policy Conference.