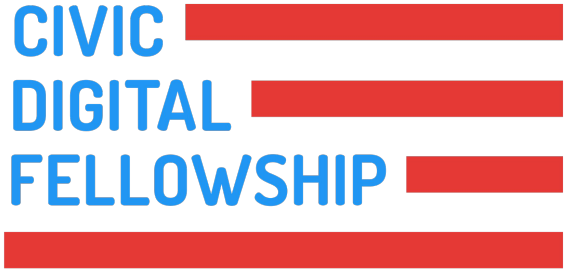


Automating Industry Classification for the Economic Census

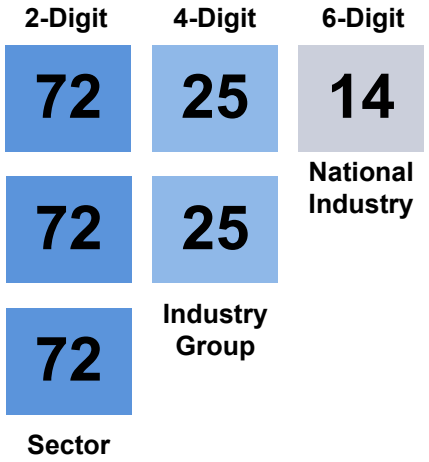
Sudip Bhattacharjee, Javier Miranda, Anne S. Russell, **Rahul Shukla**,
Justin C. Smith, **Lan Zhang**

Economy-Wide Statistics Division (EWD)
Center for Optimization and Data Science (CODS)

August 8, 2019

The Census Bureau has reviewed this data product for unauthorized disclosure of confidential information and has approved the disclosure avoidance practices applied (Approval ID: CBDRB-FY19-EWD-B00008.)

North American Industry Classification System (NAICS)



→ Cafeterias, Grill Buffets, and Buffets

→ Restaurants and Other Eating Places

→ Food and Accomodation Services

Definitions

Economic Census

- Official 5-year measure of American businesses and the economy
- Forms mailed to nearly 4 million businesses
- Collects operational and performance data

Business Register

- The Census Bureau's master database of business establishments
- Most records are updated annually with administrative data (IRS, SSA, BLS)

Motivation

- Industry classification codes are crucial to describe economic activity and its dynamics
- U.S. Census Bureau spends **considerable time and resources** identifying and classifying the industry of establishments
- Burden of collection significant on Census and businesses
- Surveys have **declining response rates**

Implement a machine learning pipeline that uses public data to generate NAICS codes

Roadmap

- Data
- Machine Learning Pipeline
- Hierarchical Modeling Approach and Application
- Applications to Economic Census Surveys
- Applications to Economic Census Operations
- Future Steps

Machine Learning Pipeline Overview

Data Collection

Economic Census or Business Register
- Name
- NAICS Code

⟷

Google API (CODS)
- Google type tags
- Reviews
- Homepage text

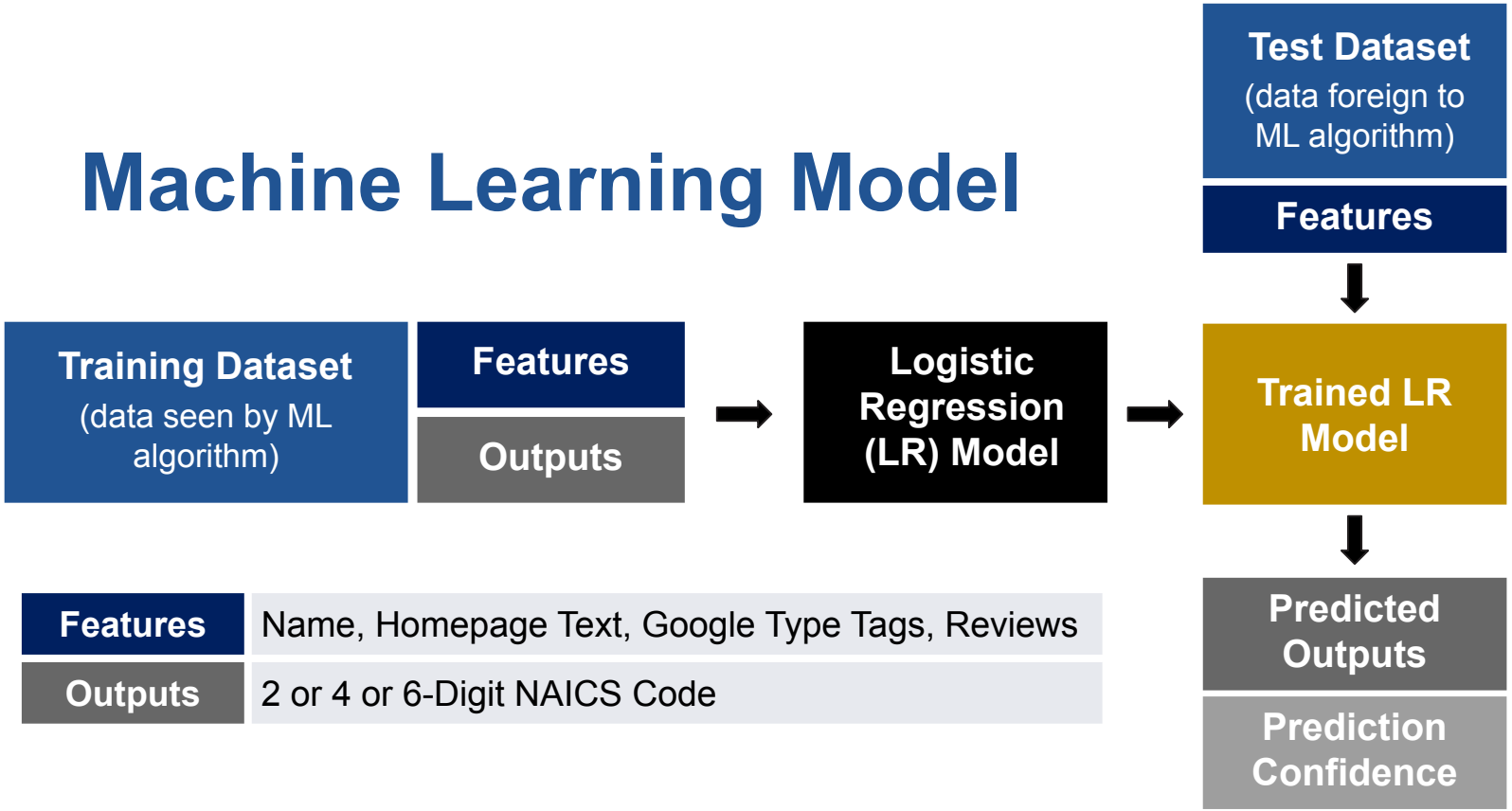
Training Dataset
(data seen by ML algorithm)

Test Dataset
(data foreign to ML algorithm)

Business Name & Address →  → Google Places API

types	bar, restaurant, food, establishment
url	www.business11.com
reviews	“This place had amazing service”, “the coffee is good but the waffles were a little dry…”

Machine Learning Model



Features	Name, Homepage Text, Google Type Tags, Reviews
Outputs	2 or 4 or 6-Digit NAICS Code

Hierarchical Approach

How can we innovate to **build upon the classification accuracy of the current machine learning pipeline** developed by the Center for Optimization and Data Science (CODS)?

Hierarchical Approach

- Pipeline has been using flat classifiers to predict NAICS codes at the 2 (sector), 4 (industry group), and 6-digit (national industry) levels
- Each additional digit provides increasingly specific information about a business establishment
- **How can we use information provided by the ML model when predicting an establishment's lower digit NAICS code to predict codes at higher digit levels?**

Different Combinations of New Probability-Vector Features

- 2-Digit Probabilities
- 4-Digit Probabilities
- 6-Digit Probabilities
- 2-Digit x 4-Digit x 6-Digit

Prediction Probabilities on Original Training Dataset

Original Training Dataset

Training

Trained Hierarchical Logistic Regression Model

6-Digit Classifier

6-Digit Hierarchical Classifier Results

Features	Name, Homepage Text, Google Type Tags, Reviews, **2-Digit, 4-Digit, 6-Digit, 2 x 4 x 6-Digit Probability Vectors**
Outputs	6-Digit NAICS Code

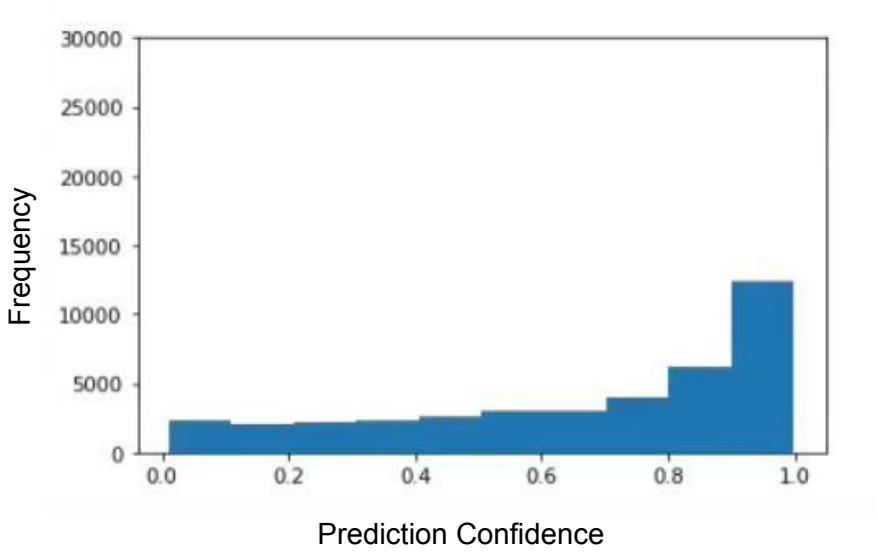
Accuracy of confident predictions over entire test set

6-Digit Flat Classifier	22,990 / 39,980 = **0.5752**
6-Digit Hierarchical Classifier	25,790 / 39,980 = **0.6451**

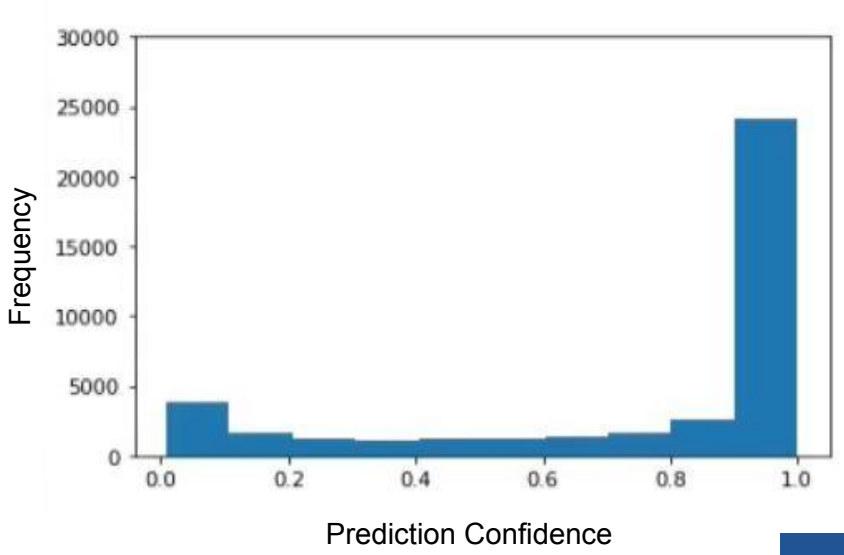
~2,800 more accurate and confident results
Improved accuracy by 7%

Prediction Probability Distributions

Flat 6-Digit Classifier



Hierarchical 6-Digit Classifier

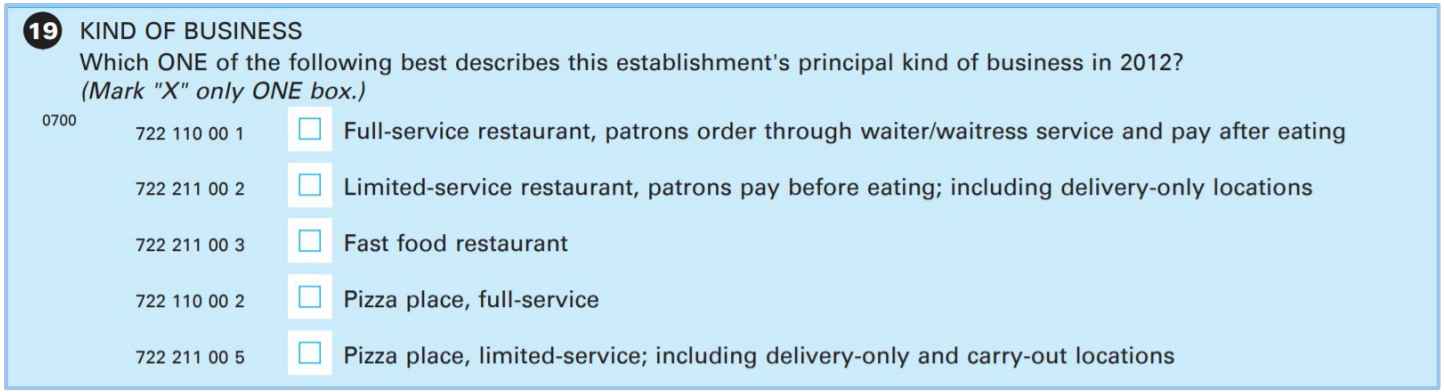


Application to the Economic Census

How does the current machine learning pipeline **perform on Economic Census surveys**, specifically classification cards?

Classification Cards

- Collects **principal kind of business of an establishment**
- Mailed according to 2-digit administrative NAICS codes
- Sector 72 Class Card (Food and Accommodation Services)



19 KIND OF BUSINESS

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

0700

722 110 00 1	☐	Full-service restaurant, patrons order through waiter/waitress service and pay after eating
722 211 00 2	☐	Limited-service restaurant, patrons pay before eating; including delivery-only locations
722 211 00 3	☐	Fast food restaurant
722 110 00 2	☐	Pizza place, full-service
722 211 00 5	☐	Pizza place, limited-service; including delivery-only and carry-out locations

United States® Census 2020

389,300 class cards mailed in 2017…

$1.2 million in mail-out costs

\+ **$300k** in analyst costs

$1.5 million in total costs

Only **37.7%** of mailed class cards do not require additional processing

Shape your future START HERE >

United States® Census 2020

2-Digit Administrative NAICS Accuracy by Sector (Benchmarks)

Sector 11	N/A	Sector 51	66.30%
Sector 21	67.23%	Sector 52	62.85%
Sector 22	N/A	Sector 53	79.94%
Sector 23	86.96%	Sector 54	80.99%
Sector Part 31	56.94%	Sector 55	N/A
Sector Part 32	72.27%	Sector 56	67.69%
Sector Part 33	79.95%	Sector 61	70.40%
Sector 42	N/A	Sector 62	70.43%
Sector Part 44	69.93%	Sector 71	72.08%
Sector Part 45	76.39%	Sector 72	85.96%
Sector Part 48	85.63%	Sector 81	73.32%
Sector Part 49	47.02%	Sector 92	N/A

ML Model Performance with Class Card Data as Test Set

Test & Training Split

- **Training Set:** Business Register (BR) across all sectors that overlaps with Google API data
- **Test Set:** Class card establishments that overlap with Google API data (389,300 → 2,083)
- **Aim:** Gauge 2-digit classification accuracy by ML model when applied to class card data

Train Business Register, Test Class Cards

Sector 11	N/A	Sector 51	0.25
Sector 21	0.0	Sector 52	0.73
Sector 22	N/A	Sector 53	0.83
Sector 23	0.74	Sector 54	0.29
Sector Part 31	0.34	Sector 55	N/A
Sector Part 32	0.17	Sector 56	0.39
Sector Part 33	0.21	Sector 61	0.71
Sector 42	N/A	Sector 62	0.77
Sector Part 44	0.64	Sector 71	0.46
Sector Part 45	0.61	Sector 72	0.89
Sector Part 48	0.58	Sector 81	0.61
Sector Part 49	0.2	Sector 92	N/A

Applications to Census Operations

Using this technology, **what tools can we create that would benefit analysts** at the Census with their daily tasks?

Business Register

1
(high reliability)

9
(low reliability)

For every establishment in the BR, include an ML-predicted code and associated reliability score

NAICS (Admin)	Reliability (Admin)
332813	9
561520	3
488410	2
327320	8

Business Register

1 (high reliability)

9 (low reliability)

For every establishment in the BR, include an ML-predicted code and associated reliability score

NAICS (Admin)	Reliability (Admin)	NAICS (ML)	Reliability (ML)
332813	9	332911	0.7
561520	3	611699	0.2
488410	2	488410	0.8
327320	8	322230	0.9

1 (high reliability)

0 (low reliability)

Classification Cards

- Send class cards based on 2-digit ML predictions with high accuracies
- Accurate 6-Digit NAICS code predictions can eliminate class card mailing altogether

Applications can be extended to cut similar costs in other Economic Census surveys

Future Steps

- **Public Data**
  - Gathering more public data – expand training and test datasets (full access to Google API in acquisition)
  - Targeted Web-Scraping
- **Economic Census**
  - Apply technology to full Economic Census other than class cards – response cases, non-mails, delinquents
- **Feature Engineering**
  - Yelp Data
  - Satellite Data

Thank You!

Shape
your future
START HERE >

United States Census 2020