HLG-MOS ML Project Pilot Study

NAICS and NOC Models and Journey to Production

Delivering insight through data for a better Canada

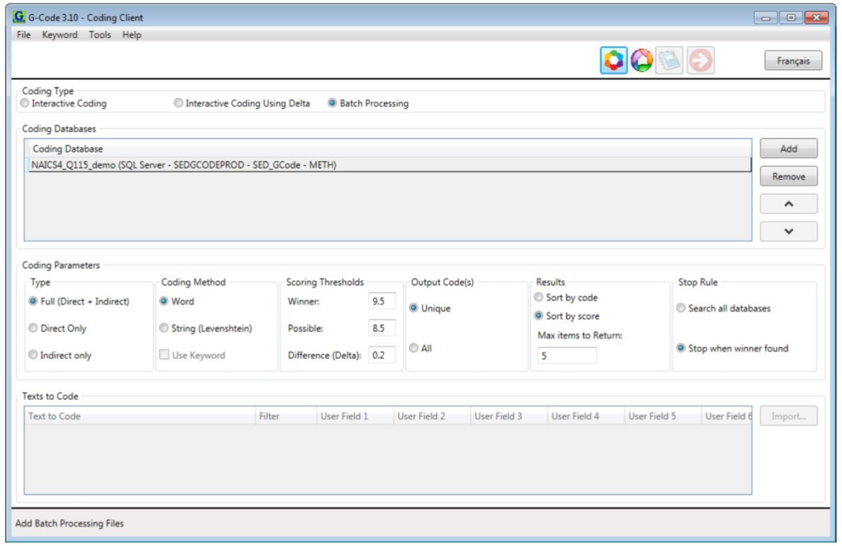
Pilot Study Overview

- Objective: Develop machine learning models to code the North American Industry Classification System (NAICS) and National Occupational Classification (NOC).

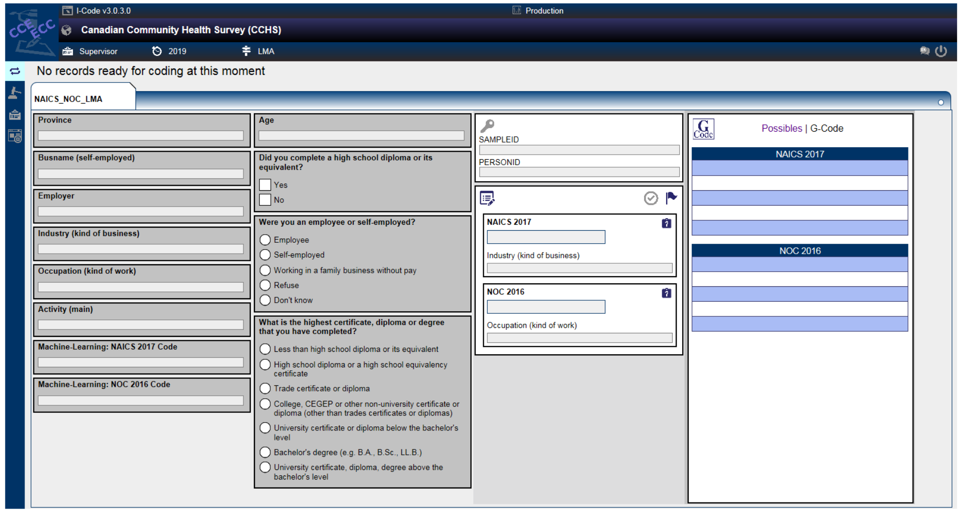
- CCHS: Implementation of NAICS and NOC models in production
  - Models: FastText (approved for use)
  - Requirement: Error rate below 5 % (≥ human coders)
  - Quality Control: Confidence level, by class

System Terminology

G-Code: Generalized coding tool, includes word-matching and ML (FastText, XGBoost)



Coding and Corrections Environment (CCE): Coding platform, which integrates automated and manual coding



Transition of Coding Platforms

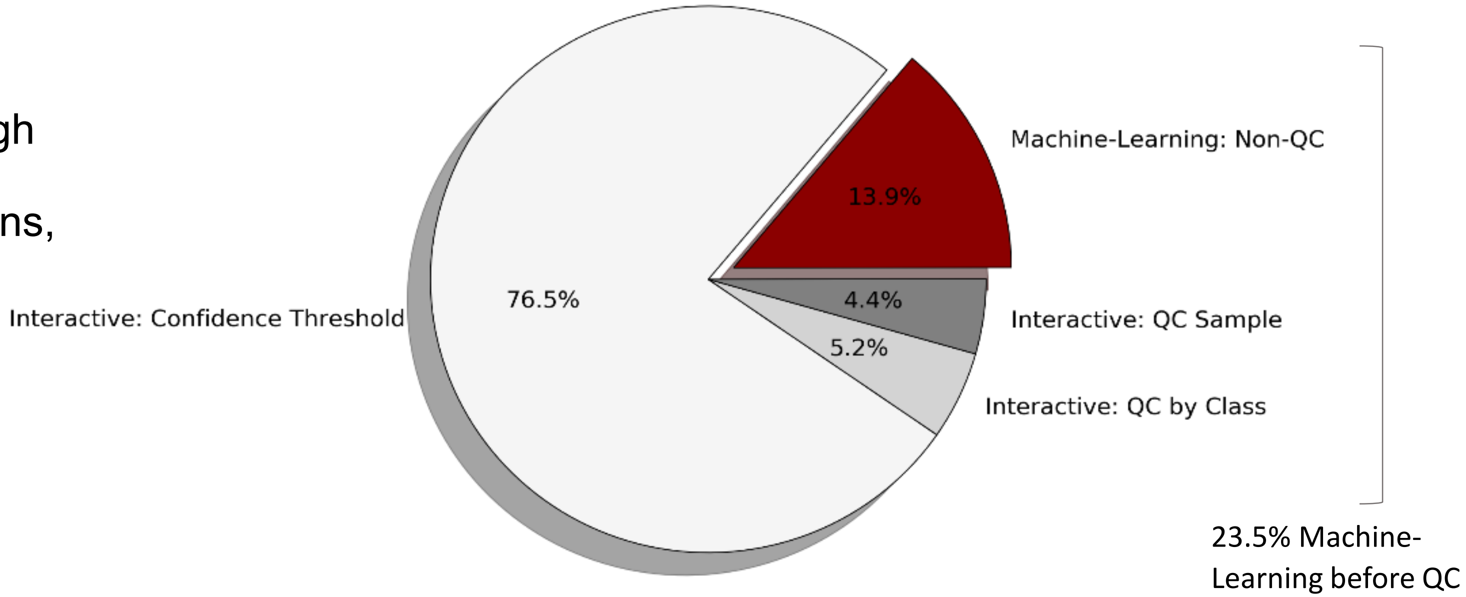
2018: Previous System	Data	→			IOCS	→	Manual
2019-Q1: Manual	Data	→	G-Code	→	CCE	→	Manual
2019-Q2: Word-Matching	Data	→	G-Code	→	CCE	→	Manual
2019-Q3,4: ML (temporary)	Data	→	G-Code	→	CCE	→	Manual
2020: ML (upcoming)	Data	→	G-Code	→	CCE	→	Manual

Pipeline Flow of Records

- Records which did not obtain a high enough confidence score, to ensure to ensure an overall 95% accuracy for both classifications, were sent to be interactively coded

- After applying the threshold models were able to predict a NAICS code in 46.5% of records and a NOC code in 34.2% of records.

- Records with received both a NAICS and NOC code (23.5%) were sent for Methodology QC

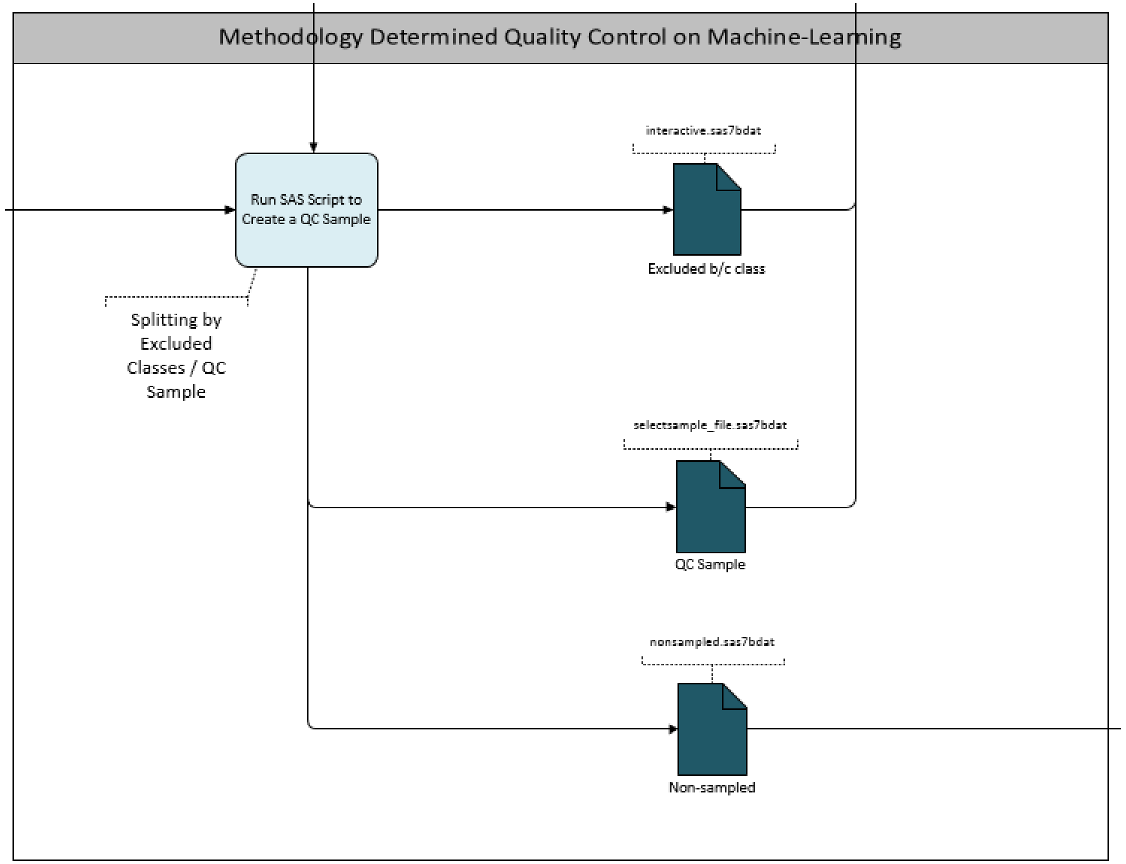


Methodology Quality Control (QC) Framework

- Models run on a testing datasets of 157,527 NOC records and 134,911 NAICS records from multiple surveys including: LFS, JVWS, CCHS, CHMS, CHSCY, SFS.
- QC by Class: Classes which had an error rate above 10% for NAICS and 15% for NOC, when run on the testing dataset, were flagged for interactive coding.
- QC Sample: Given a margin of error of 0.02, where p̂ is 0.05, we calculated the quality control sample size as follows:

$$Margin\ of\ Error = \sqrt{\frac{Z_{0.05}^{2}(P(1-\hat{p}))}{n}}$$

- Overall Error in production
  - Excluded by class: 10.5%
  - QC Sample: 4.2%



Journey to Implementation

Phase 1: Starting Point ~1.5 years ago

- Had management buy-in.
- Were using Word Matching for some auto-coding activities.
- Early CCE version could consume word-matching outputs, in Production as of Jan 2019.
- G-Code being used for auto-coding using Word Matching and integrated FastText as a prototype in Jan 2019.

Journey to Implementation

Phase 2: Evolution of the Pilot / Unit

- Developed technical capacity to develop quality models (learning / consultation)
    - Training Data, Pre-Processing, Feature Selection, Parameters, Analysis, etc.

- Collaborated with Methodology on development of a QC Sampling Strategy

- G-Code with FastText moved in to Production.

- Developed good working relationships with ML partners
    - Data Science Accelerator / Data Science Division
    - G-Code Methodology Team – Statistical Integration Methods Division
    - ML Communities of Practice
    - Subject matter areas

- NAICS and NOC models used in Production for CCHS and CHMS.

Journey to Implementation

Phase 3: Going Forward

- May be able to use models for other small surveys, such as the General Social Survey (GSS).
- Further improvement of models: boosting minority classes, additional data sources.
- Work towards development of models tailored to use for Job Vacancy and Wage Survey (JVWS) and Labor Force Survey (LFS).
- Revisit Methodology QC sampling plan to account for upcoming CCE development.
- Adapt upcoming Quality Validation Framework(s).

Conclusions / Lessons Learned

- **Time Investment / Interdependencies**
  - Technical capacity
  - IT systems development / Testing
  - Methodology consultation
  - Client approval
  - Feasibility / Suitability - ROI
- **Engagement / Buy-In**
  - Subject matter / Management
  - Explainability / ML black box

Thank you!

Questions?

Model Creation

- Algorithm: FastText (approved for use in G-Code)

- Training Data Sources
  - CCHS: 88,782 historical records
  - Labour Force Survey (LFS): 443,464 historical records
  - Standards Classification: 80,000 NOC and 40,000 NAICS index entries

Model Creation

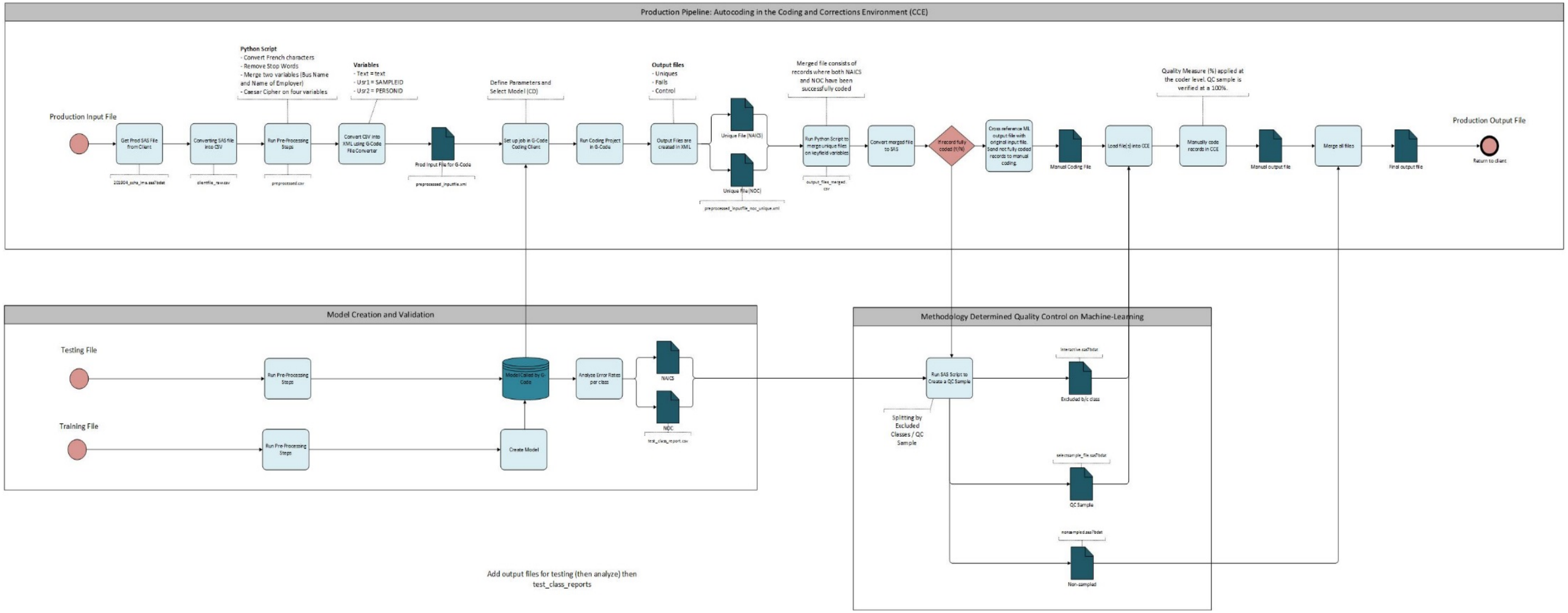
Preprocessing Steps

- Exploratory: Multiple Bags of Words, Up-Sampling of Minority Class, Separation of English and French Models, Stemming and Lemmatization, and Pre-Trained FastText Embeddings.

- Production:
  1. Removal of Stop Words
  2. Lowercasing character conversion
  3. Merging of the variables ‘Business Name’ and ‘Name of Employer’
  4. Application of a Caesar Cipher to differentiate text from “Company”, “Industry”, “Job Title” and “Job Description” when concatenated into a single field

Documented on UNECE, Working Documents: FastText_Techniques [FastText Techniques]

Temporary production pipeline with the CCE



Available in GC Docs: https://gcdocs.gc.ca/statcan/llisapi.dll/Overview/7285192

CCHS Autocoding

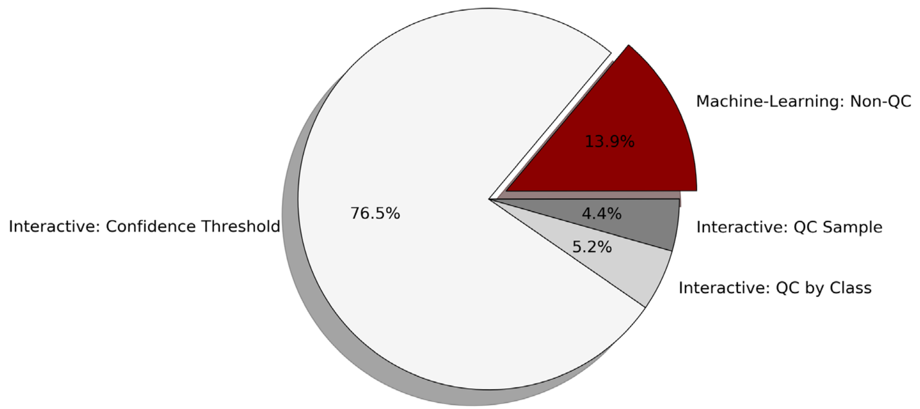
As the Coding and Corrections Environment (CCE) cannot currently consume results from G-Code ML models, we developed a temporary production pipeline to code CCHS

- Validation Pipeline
  - CCHS output files (2019 Q2) used as historical data.
  - Records = 7917.

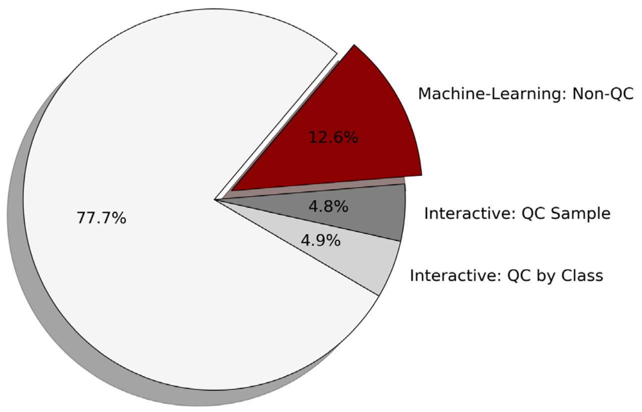
- Production Pipeline
  - CCHS (2019 Q3, Q4) collection period.
  - Records = 7430, 7404

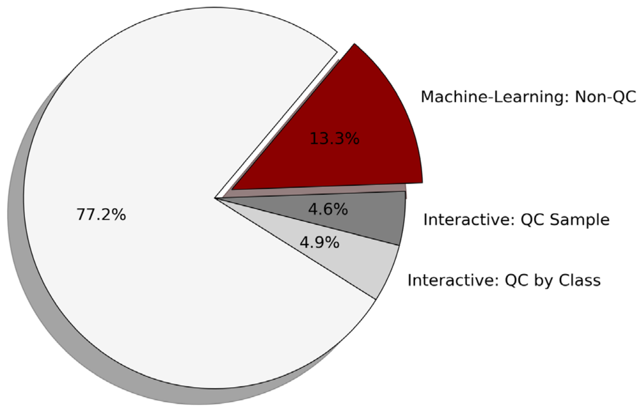
Comparison of pipelines

Validation Pipeline



Production Pipeline





Collection Period:	Q2	Q3	Q4
Records:	7917	7430	7404

Production pipeline error rate

Table 1. Error rate of NAICS and NOC in CCHS production pipeline. Record flow follows the same path as described in Figure 1. Record number = 7430 (Q3), 7404 (Q4). (*) indicates the error rate of manual coders in production before verification is applied.

Collection Period

Q3

Classification	Interactive: Confidence Threshold	Interactive: QC by Class	Interactive: QC Sample	Machine-Learning: Non-QC
NAICS 2017	3.0*	5.5	2.2	N/A
NOC 2017	4.7*	6.3	2.5	N/A
Both	3.9*	10.6	4.2	N/A

Q4

Classification	Interactive: Confidence Threshold	Interactive: QC by Class	Interactive: QC Sample	Machine-Learning: Non-QC
NAICS 2017	1.1*	3.3	0.0	N/A
NOC 2017	1.6*	5.5	1.8	N/A
Both	1.3*	7.1	1.8	N/A

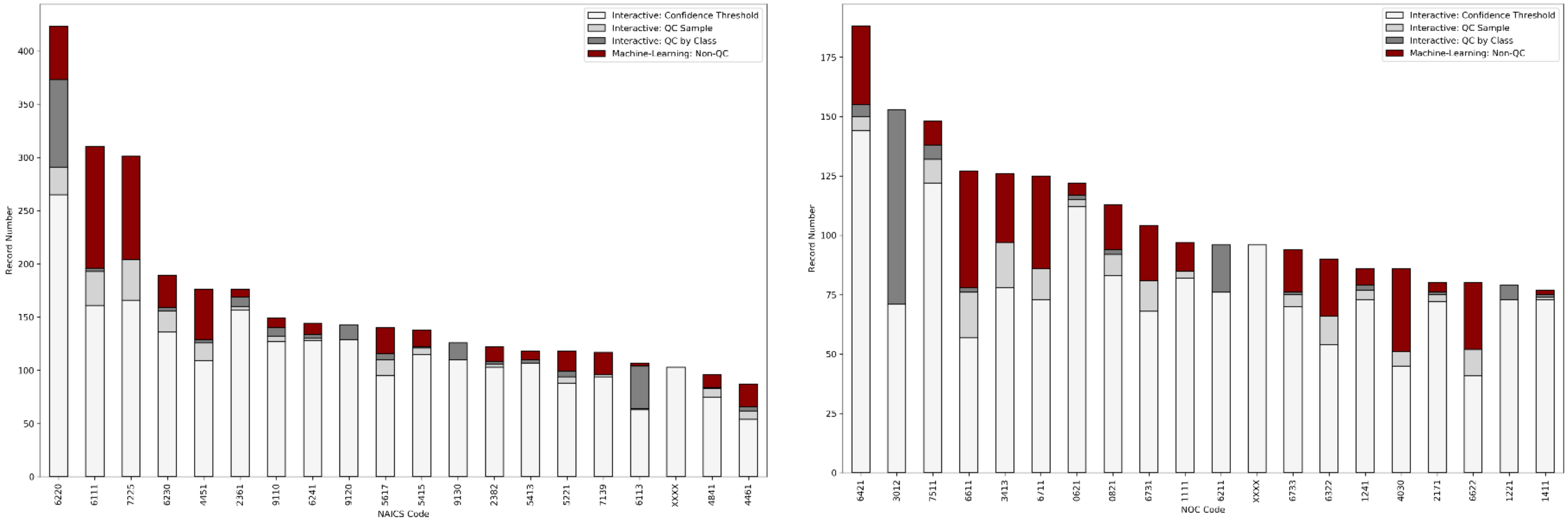


Figure 2. Per class analysis of NAICS and NOC. The 20 most frequently coded classes in each classification are displayed. Each class includes the stage (color coded in the legend) at which the record was coded in.

Table 1. NAICS and NOC model metrics tested on multiple surveys. Overall Accuracy, F1, Precision, and Recall were calculated on the entire training dataset. After a confidence threshold was applied, the error rate and coding rate were calculated on the remaining records. (*) A subset of the NOC training dataset, using only CCHS data, was evaluated using a different confidence threshold.

Measure	NAICS	NOC	NOC – CCHS*
Record #	64,249	157,527	7,088
Overall Accuracy (%)	80.5	64.4	70.8
Weighted Average F1-Score	80.4	51.5	70.6
Weighted Average Precision	81	53.3	71.9
Weighted Average Recall	80.5	51.8	71.0
Confidence Threshold (%)	96.0	99.9	99.0
Error Rate (%)	4.5	5.9	4.8
Coding Rate (%)	61.25	10.9	36.2

Table 4. NAICS and NOC model metrics for the CCHS production pipeline. Overall Accuracy, F1, Precision, and Recall were calculated on the 'Interactive: QC Sample'. Record number = 343.

Measure	NAICS 2017	NOC 2016
Error Rate (%)	2.2	2.5
Weighted Average F1-Score	97.5	96.9
Weighted Average Precision	97.4	96.7
Weighted Average Recall	97.8	97.5