Machine Learning Classification of Product Categories in Scanner Data used in the United States Consumer Price Index

Brendan Williams and Zach Whitford
Bureau of Labor Statistics
2024 FCSM Research and Policy Conference
24 October 2024

This presentation reports on the results of ongoing research and analysis undertaken by Bureau of Labor Statistics staff. It has undergone more limited review than official publications.

Overview

- Background on production model used for department store scanner data
- Research methods for model updates and maintenance
  - ML Classifiers
  - Confusion matrix analysis
  - Active learning
- Extend technique to grocery data

Background

- Began production use in March 2019 with a department store supplying scanner data
- Bag-of-words approach with a logistic regression classifier
- Many National Statistical Offices have begun using ML classifiers when working with “alternative data”
  - Adapted from classifier used in BLS’s Survey of Occupational Injuries and Illnesses
  - Increasingly common among price index programs working with scanner data

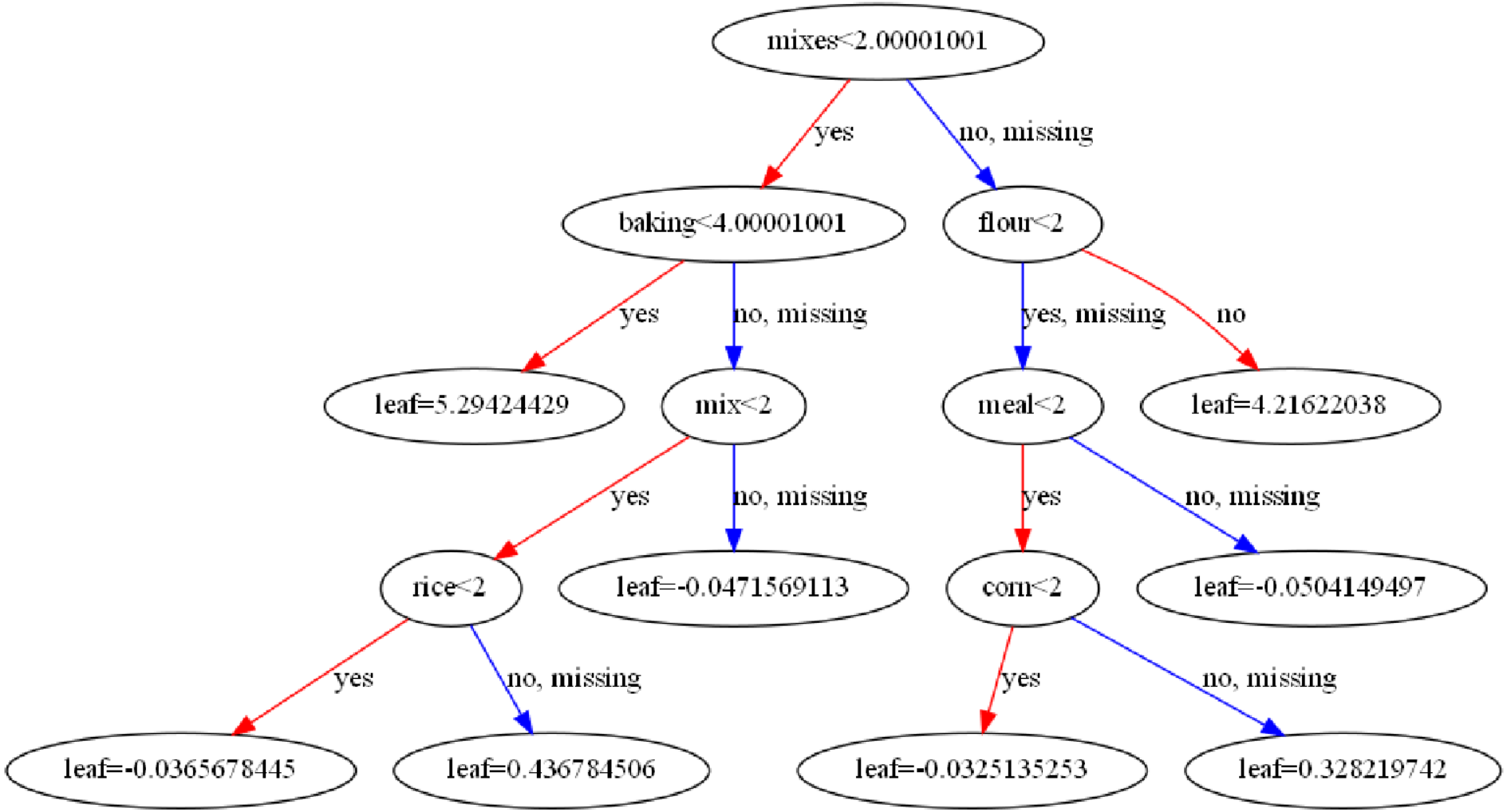
Words Associated with Men's Accessories

Model Updates

Machine Learning Classifiers

- Logistic Regression
- Neural Network
  - Extends logistic regression to multiple layers to capture interactions
- SVM
  - Classifies items based on hyper-dimensional boundaries
- XGBoost
  - Gradient-boosted decision trees

Example Decision Tree from GROCER XGBoost

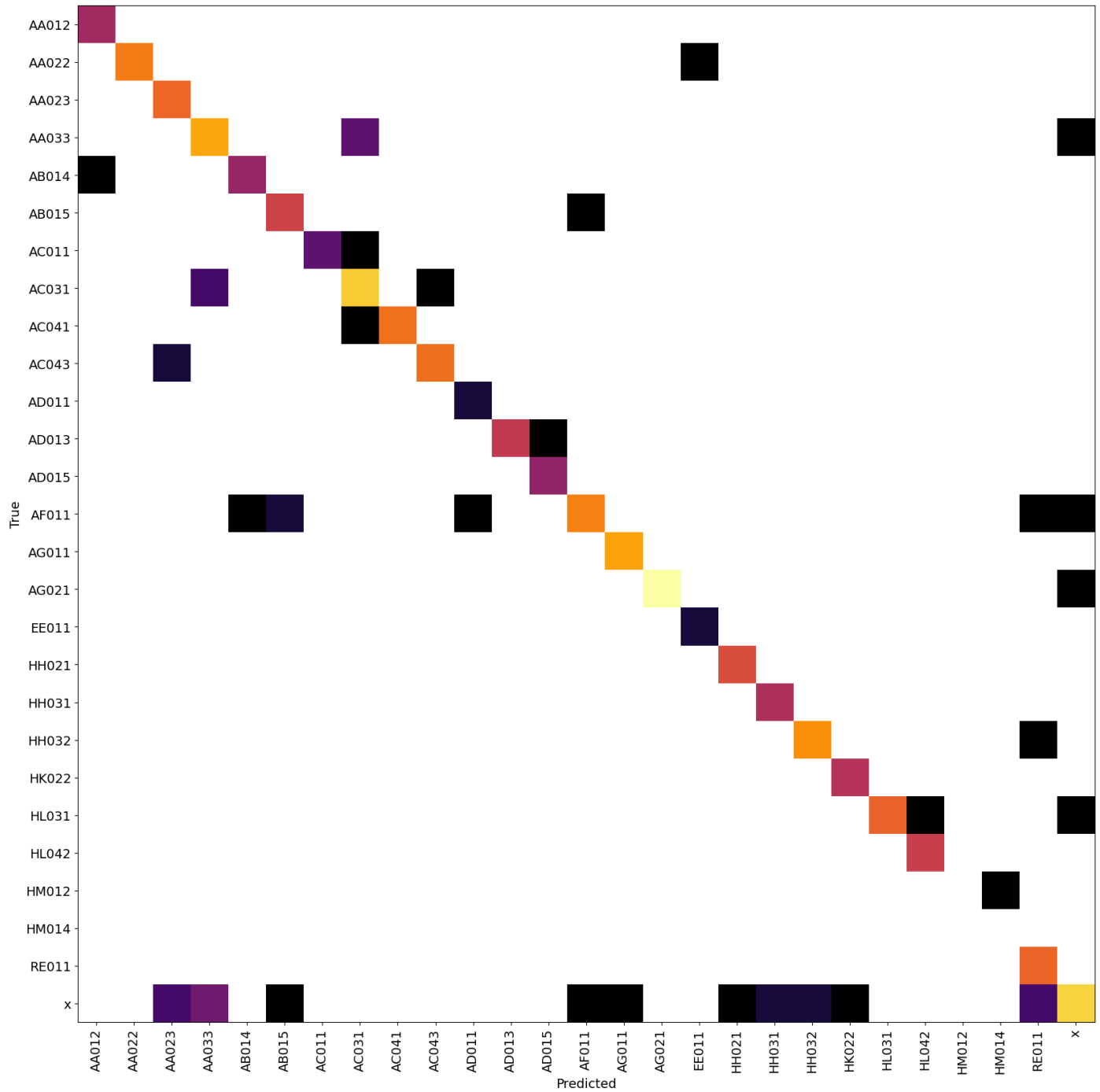


F1-Scores Department Store

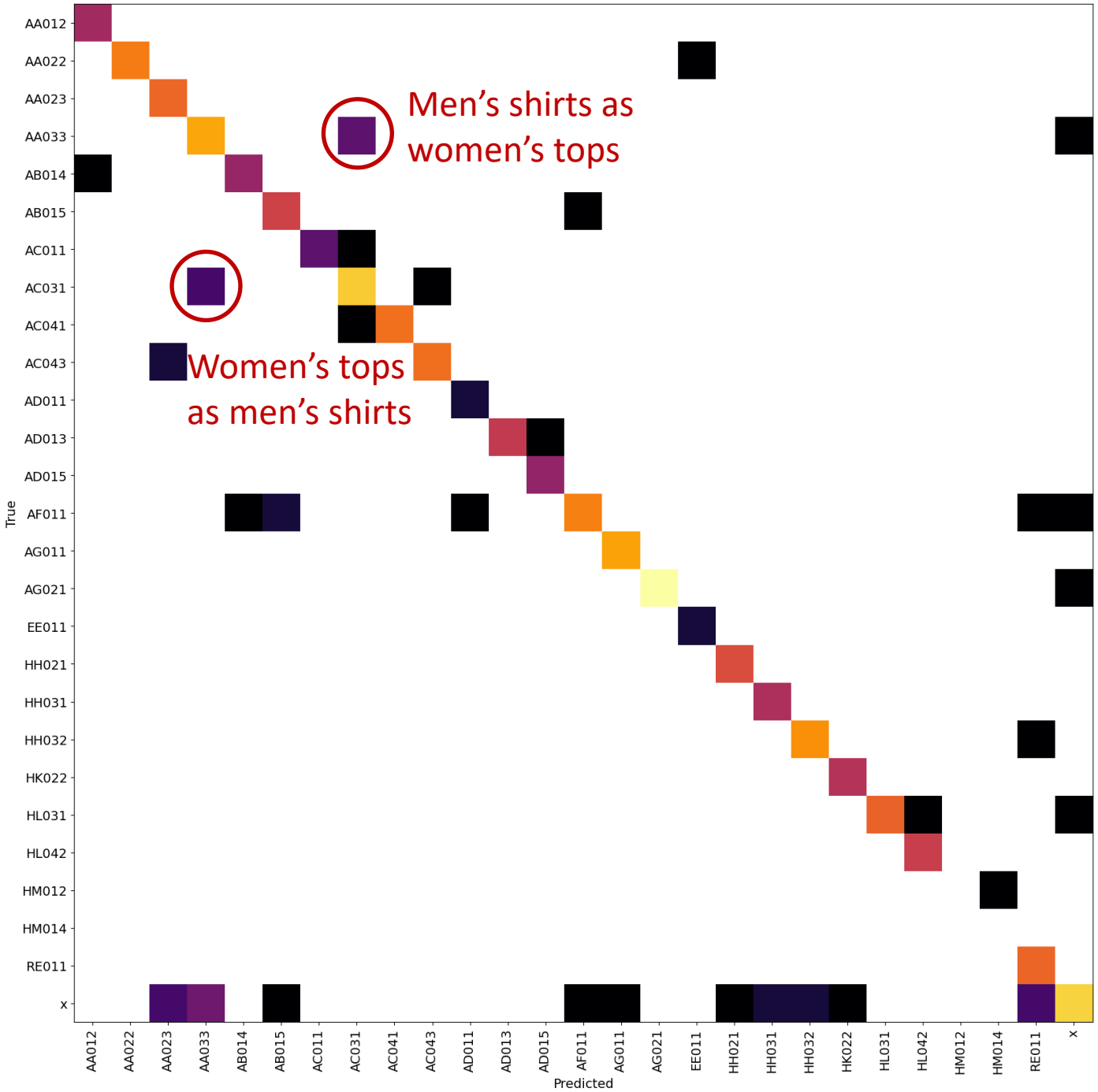
	DEPT (BoW)
Logistic	0.99
XGBoost	0.97
NeuralNet	0.99
SVM	0.99

Confusion Matrix

- Visualize correlation between predicted and “True” labels
- Correct labels on the diagonal
- Off-diagonal indicate mispredictions
- Most classes had no error, but visualizing those that did…


Confusion Matrix Heatmap for Classes with Errors in DEPT Logistic Model
True
Predicted
AA012
AA022
AA023
AA033
AB014
AB015
AC011
AC031
AC041
AC043
AD011
AD013
AD015
AF011
AG011
AG021
EE011
HH021
HH031
HH032
HK022
HL031
HL042
HM012
HM014
RE011
x

Confusion Matrix Heatmap for Classes with Errors in DEPT Logistic Model

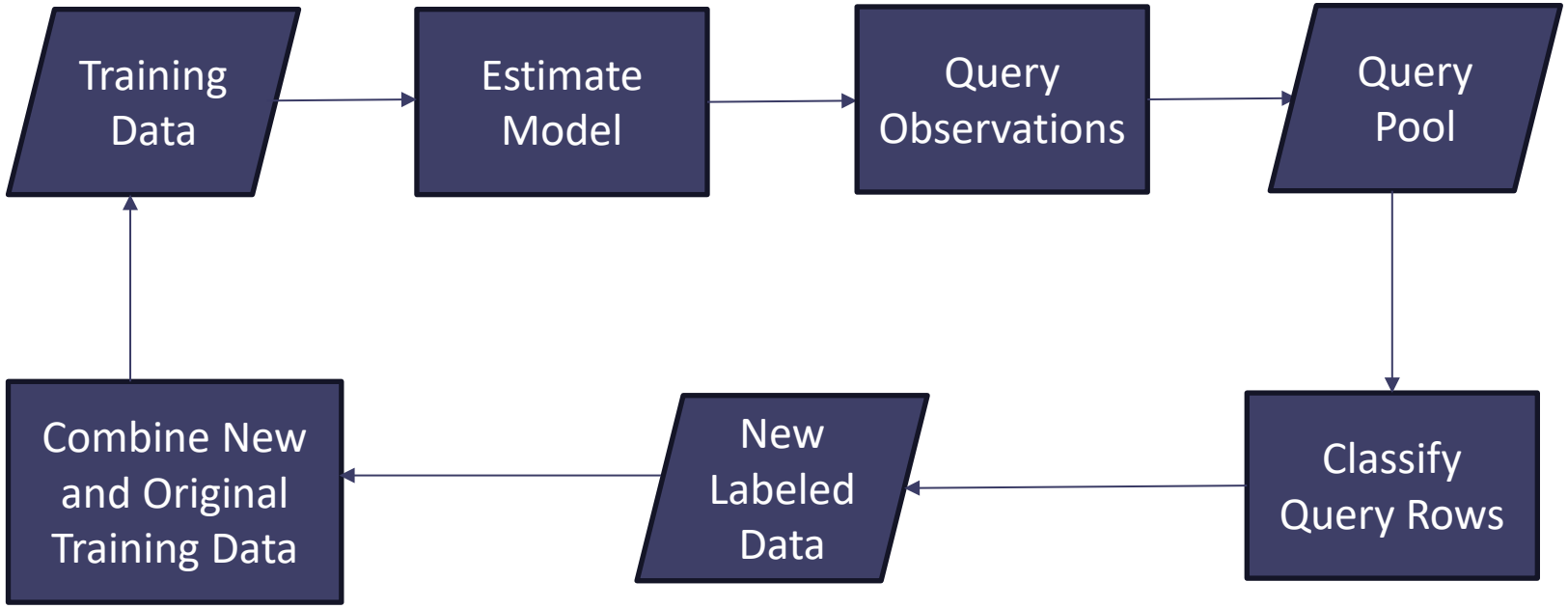


Active Learning

Active Learning Overview

- "Active learning" is a set of machine learning techniques that identifies the data that might most improve a model
- Maximizes the value of labeling data
  - Reduces requirements for human coding
- Potential to maintain accuracy of a model in production

Active Learning Flow



Query Methods

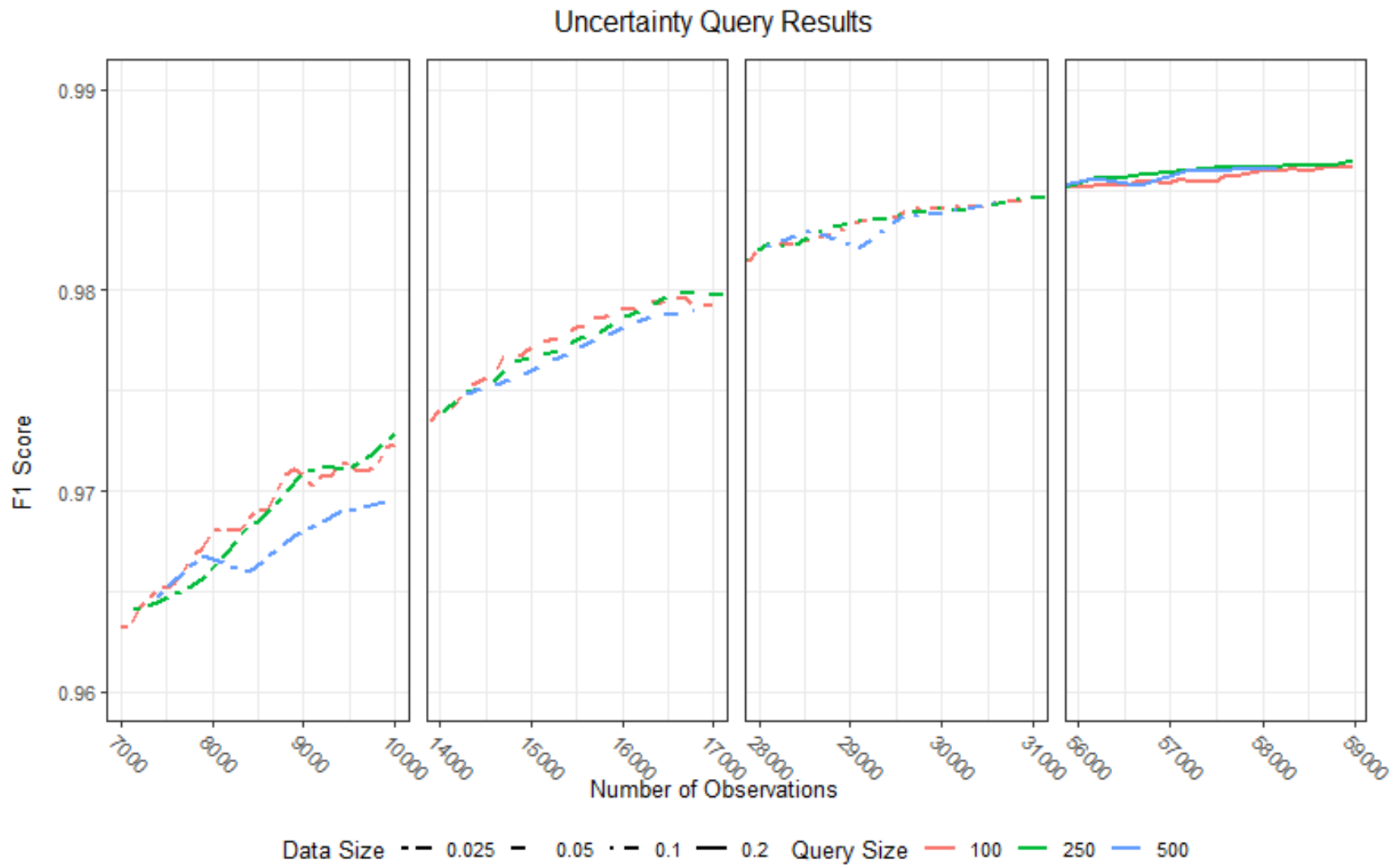
- Least Confidence Sampling
  - Lowest p-values for the predicted class overall
  - $U(x) = 1 - P(\hat{x}|x)$
- Margin Sampling
  - Lowest difference in p-value between predicted and second-best class
  - $M(x) = P(\hat{x}_1|x) - P(\hat{x}_2|x)$
- Entropy Sampling
  - Highest amount of spread among all classes
  - $H(x) = -\sum_k p_k \log(p_k)$

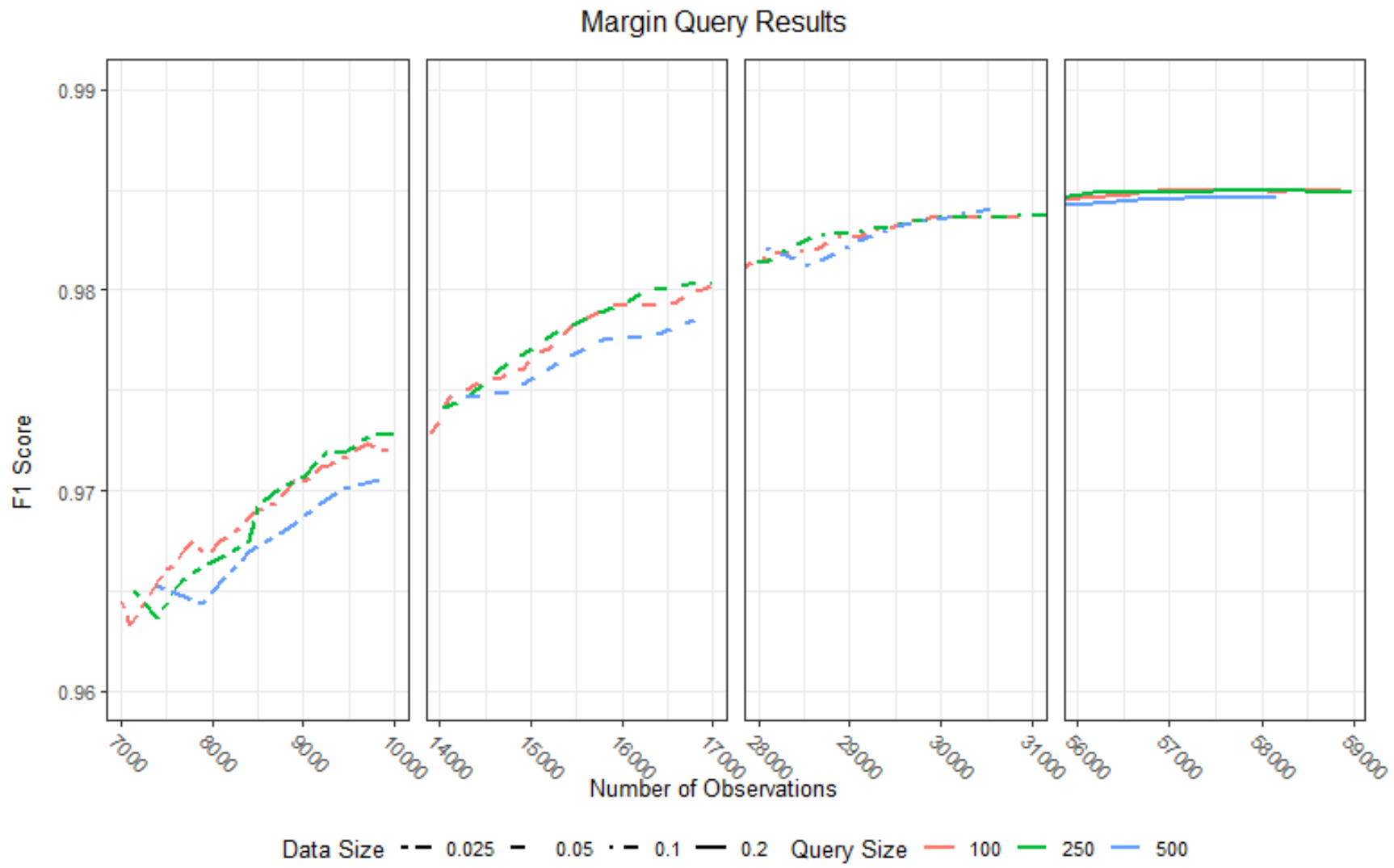
Pool-based Learning

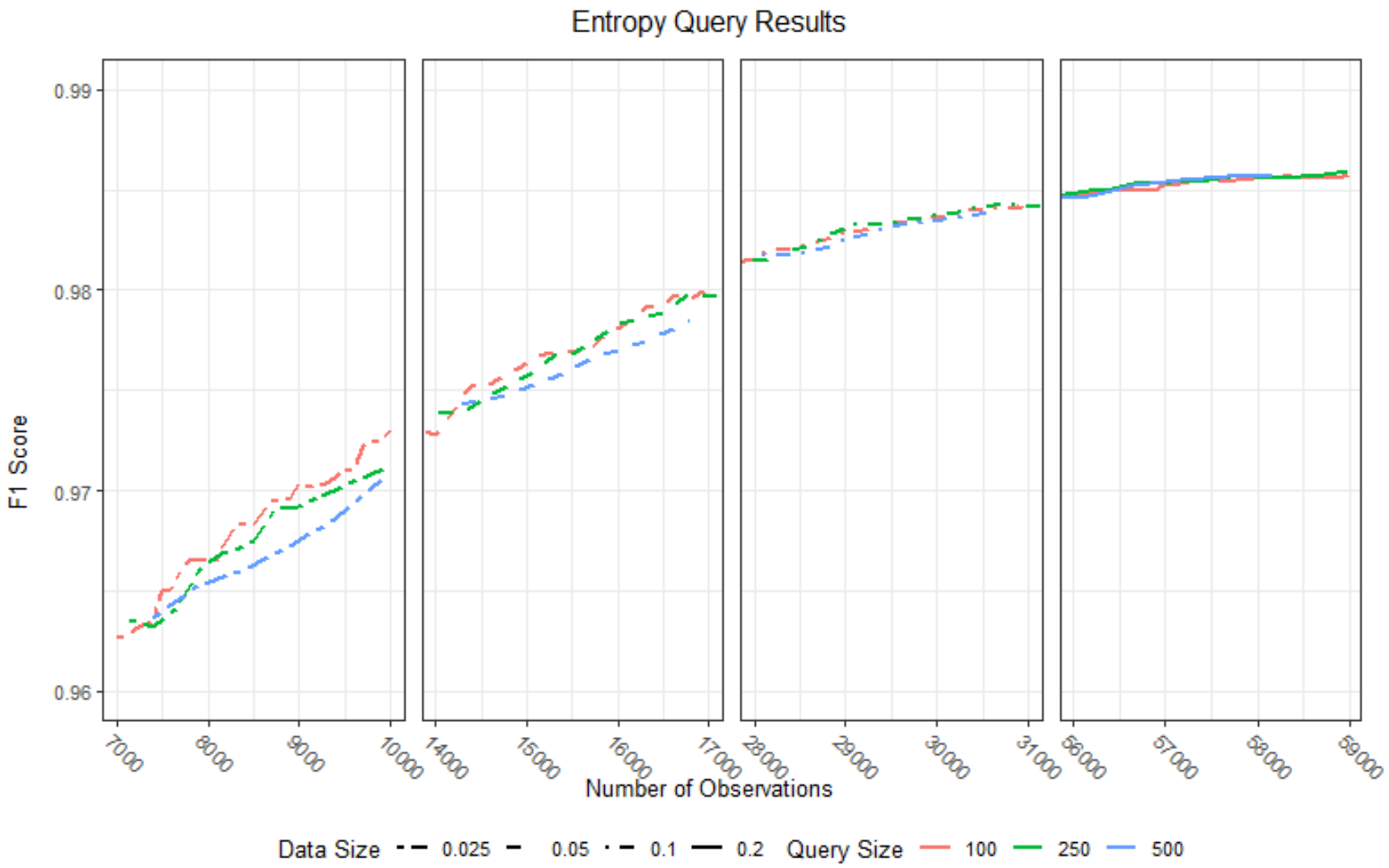
- Active learning ideally done one observation at a time
  - Updates “value” measures after each training observation is added
  - Minimizes redundant labeling
- Pool-based learning
  - Query many observations between model iterations
  - Larger batches mean fewer runs of the model and less back-and-forth with labelers but sub-optimal labeling

Active Learning Simulation

- Run simulations to evaluate pool-size and query method
- Split labeled data into 80% training and 20% test data
- Randomly sample original training data for initial simulation training data
- Query remaining training data and treat observations as newly labeled
- Evaluate improvements based on test data and iterate queries


Uncertainty Query Results
F1 Score
0.99
0.98
0.97
0.96
7000
8000
9000
10000
14000
15000
16000
17000
28000
29000
30000
31000
56000
57000
58000
59000
Number of Observations
Data Size
0.025
0.05
0.1
0.2
Query Size
100
250
500


Margin Query Results
F1 Score
0.99
0.98
0.97
0.96
7000
8000
9000
10000
14000
15000
16000
17000
28000
29000
30000
31000
56000
57000
58000
59000
Number of Observations
Data Size
0.025
0.05
0.1
0.2
Query Size
100
250
500


Entropy Query Results
F1 Score
0.99
0.98
0.97
0.96
7000
8000
9000
10000
14000
15000
16000
17000
28000
29000
30000
31000
56000
57000
58000
59000
Number of Observations
Data Size
0.025
0.05
0.1
0.2
Query Size
100
250
500

Grocery Data

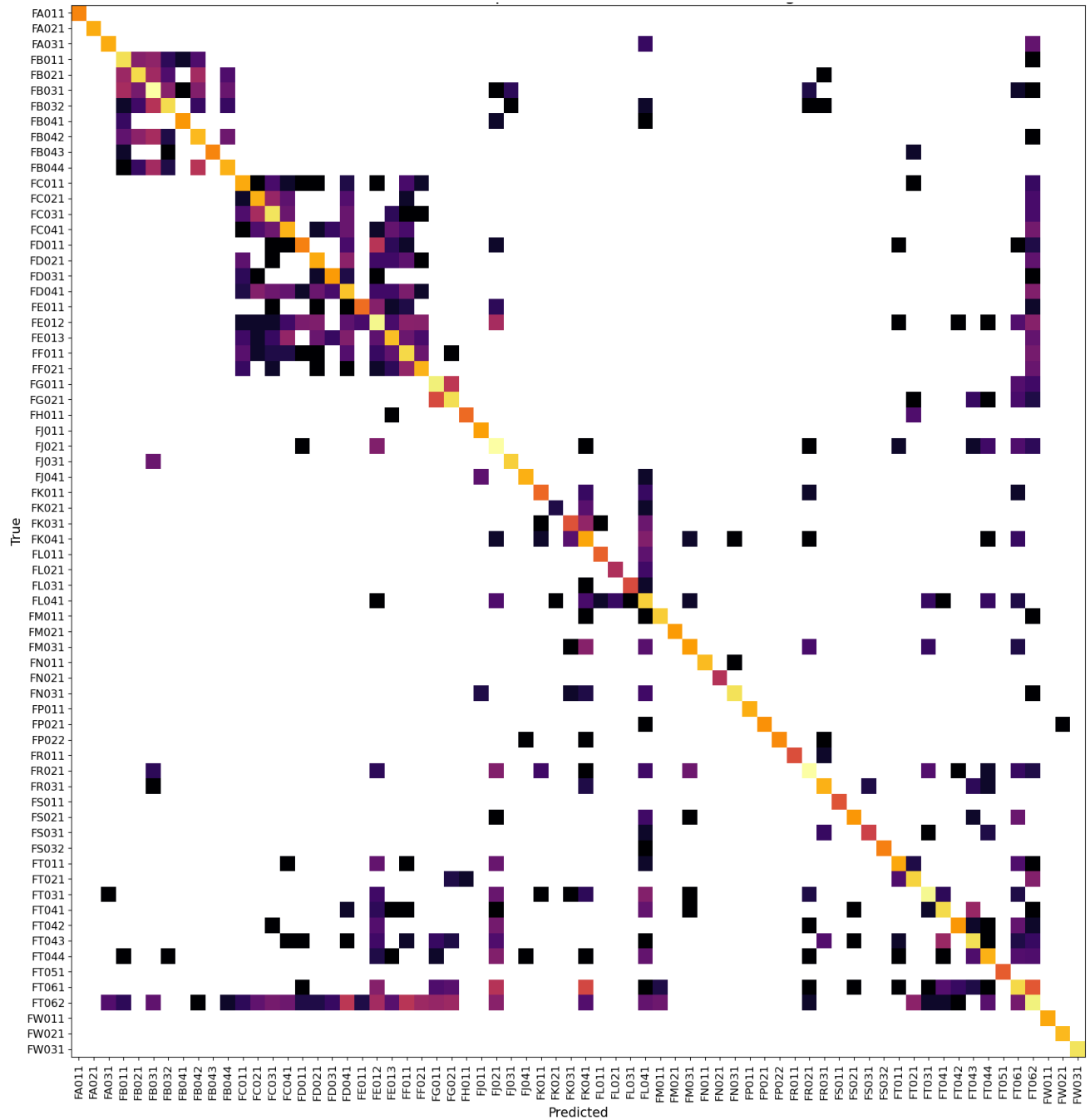
Extension to Groceries

- Extending the same method to research grocery data
- Evaluated ML Classifiers
- Explored term frequency-inverse document frequency (TF-IDF) as alternative to a BoW

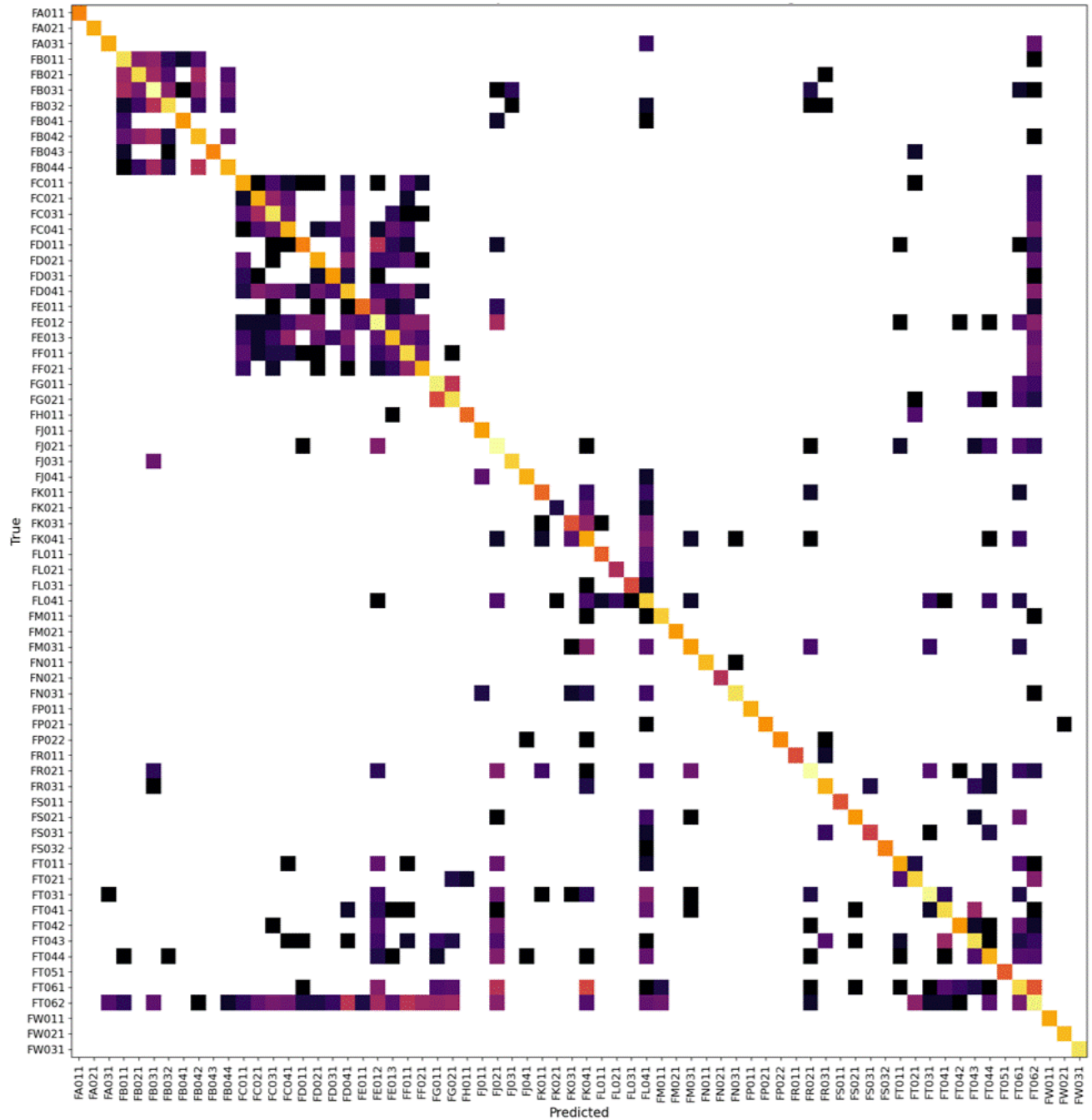
F1-Scores Grocer

	GROCER (TF-IDF)	GROCER (BoW)
Logistic	0.89	0.89
XGBoost	0.88	0.96
NeuralNet	0.97	0.97
SVM	0.88	0.96

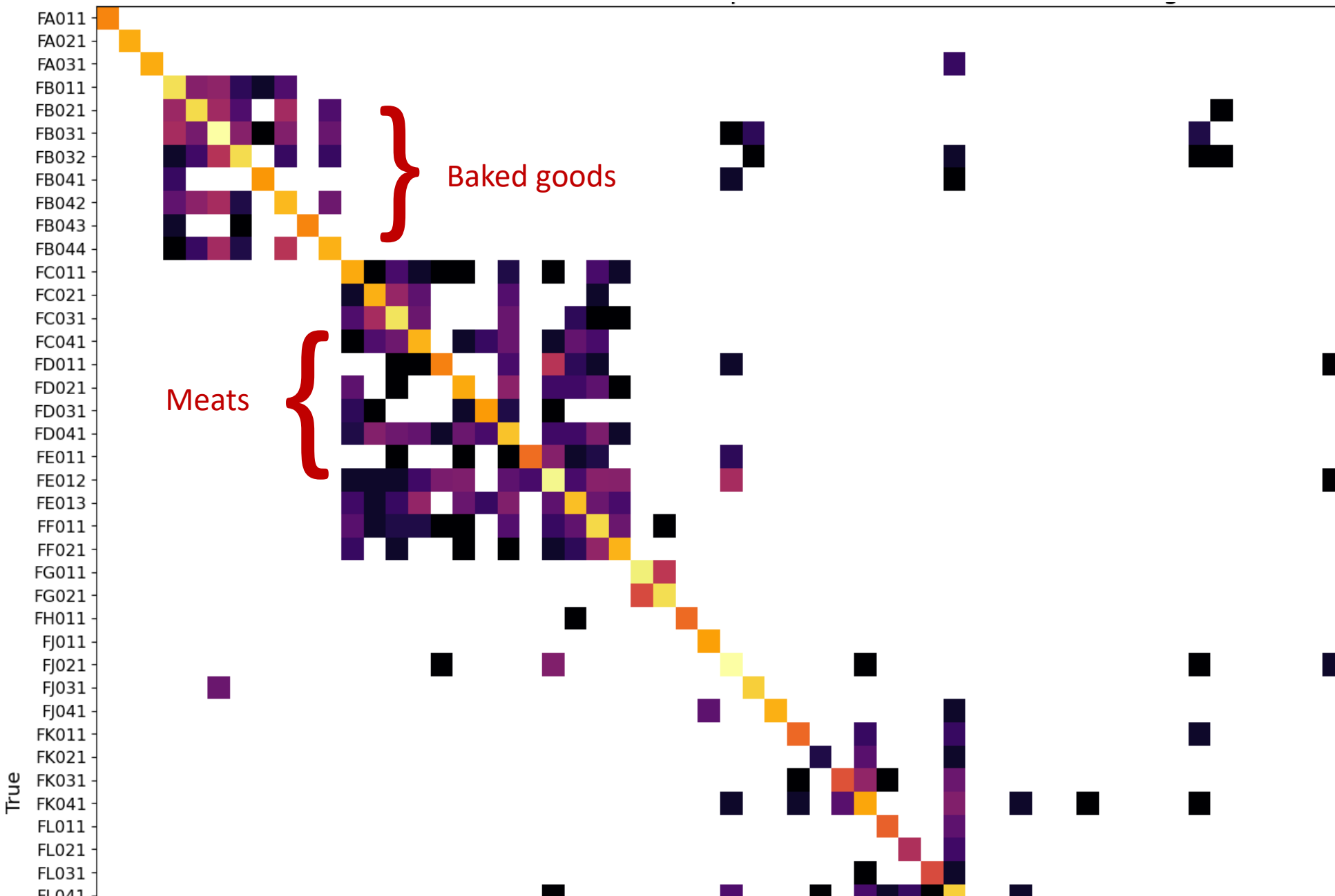
Confusion Matrix Heatmap for Food Classes in GROCER Logistic Model



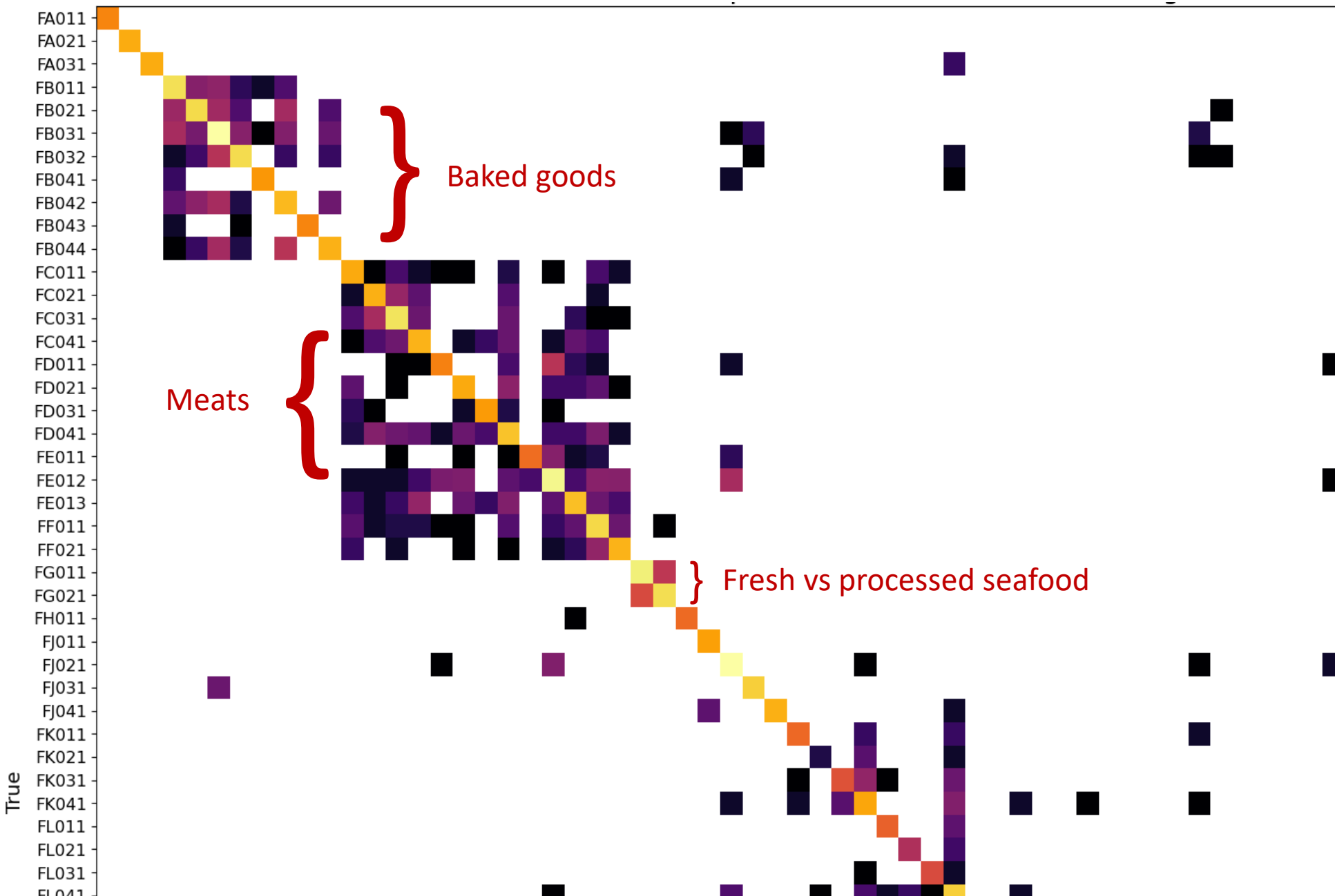
Confusion Matrix Heatmap for Food Classes in GROCER Logistic Model



Confusion Matrix Heatmap for Food Classes in GROCER Logistic Model



Confusion Matrix Heatmap for Food Classes in GROCER Logistic Model



Conclusions

- Simpler models perform well—suggesting direct relationships between individual words and classifications
- XGBoost performed well, but required tuning
- Small training datasets perform well
- For active learning: Less of a disadvantage for larger pool size with larger existing training data, and no clear advantage to a query method

Contact Information

Brendan Williams
Senior Economist
Consumer Price Index Program
www.bls.gov/cpi
202-691-5414
Williams.Brendan@bls.gov