CE Diary Autocoder
Demo

Michell Li
Data Science Intern
Civic Digital Fellow
August 6, 2019

What is Machine Learning?

- Definition: Machine learning is programming computers to **optimize a performance criterion** using example data or **past experience**

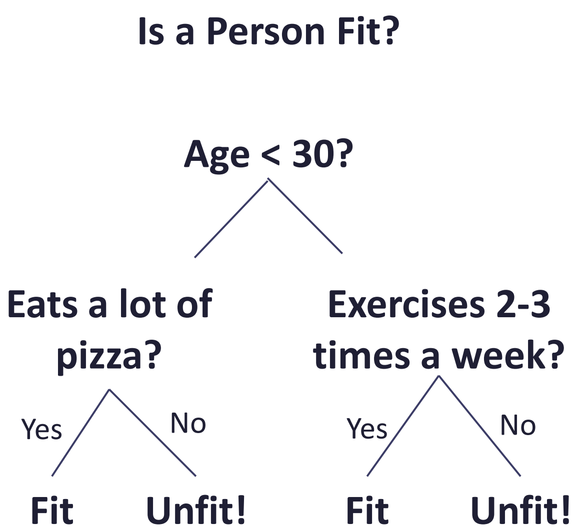
- Goal: Build a model that is a **good and useful approximation** to the data

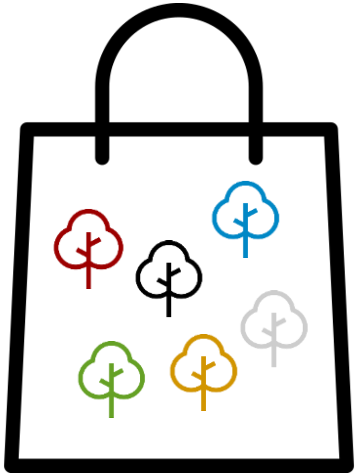
Machine Learning is Like Gardening

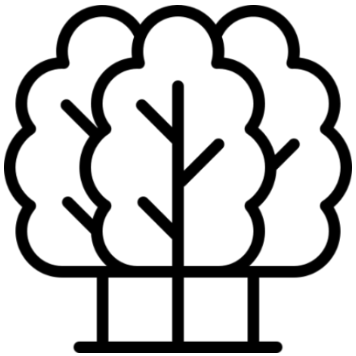
What is a Random Forest Model?

Random forest models are **bagged decision tree** models that split on a **subset of features** at each split.

Decision Tree:







Motivation

The Bureau of Labor Statistics wants to automatically assign item codes in the Diary survey.

The process is currently labor intensive and expensive.

The existing autocoder is a rule based system that needs to account for special cases, which leads to inaccuracies.

Creating a new autocoder using machine learning can: **reduce costs, improve accuracy.**

Contributions

Models

Four Models: ECLO, EFDB, EOTH, EMLS

Data and Analysis

Spell Checker created using Levenshtein, Jaro Winkler and QWERTY distance

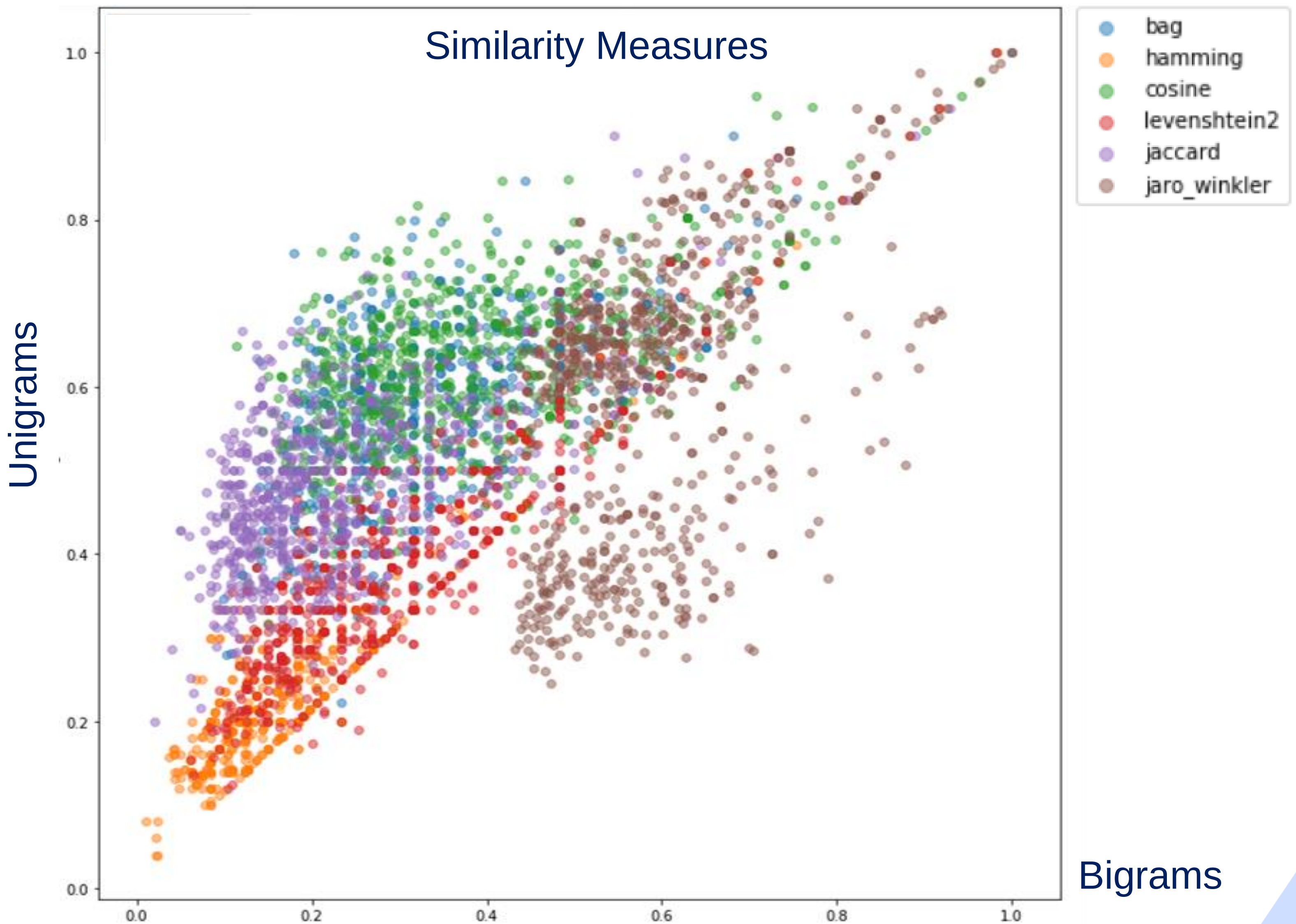
Module for analyzing Unclassifiable diary items

Analytics Dashboard UI using Dash

Data Exploration

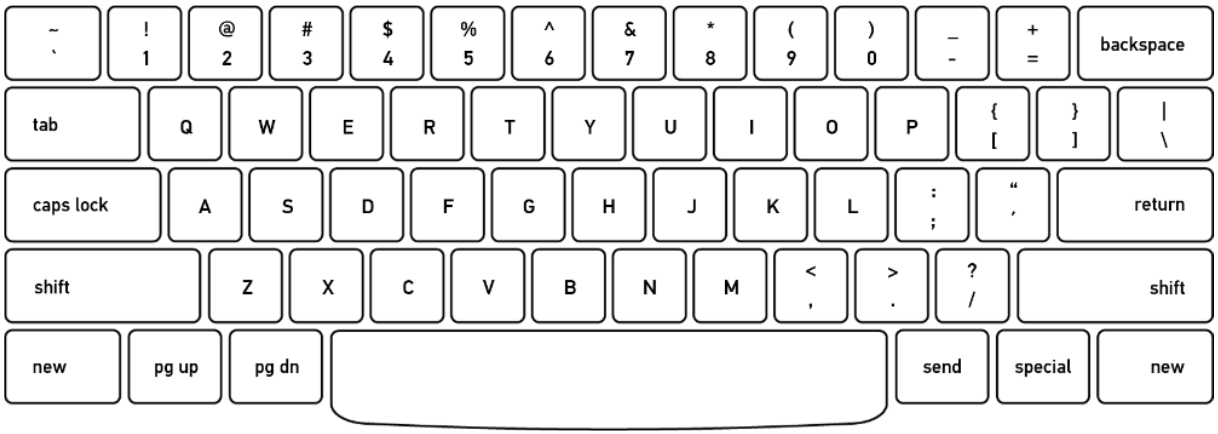
Total: 753
(not including >900000)

Dealing with Dirty Data



Spell Checker

LABOT vs LAOBR to LABOR



Transposition/Replace QWERTY Penalty:
T -> R = log(1)
B <-> O = log(4)

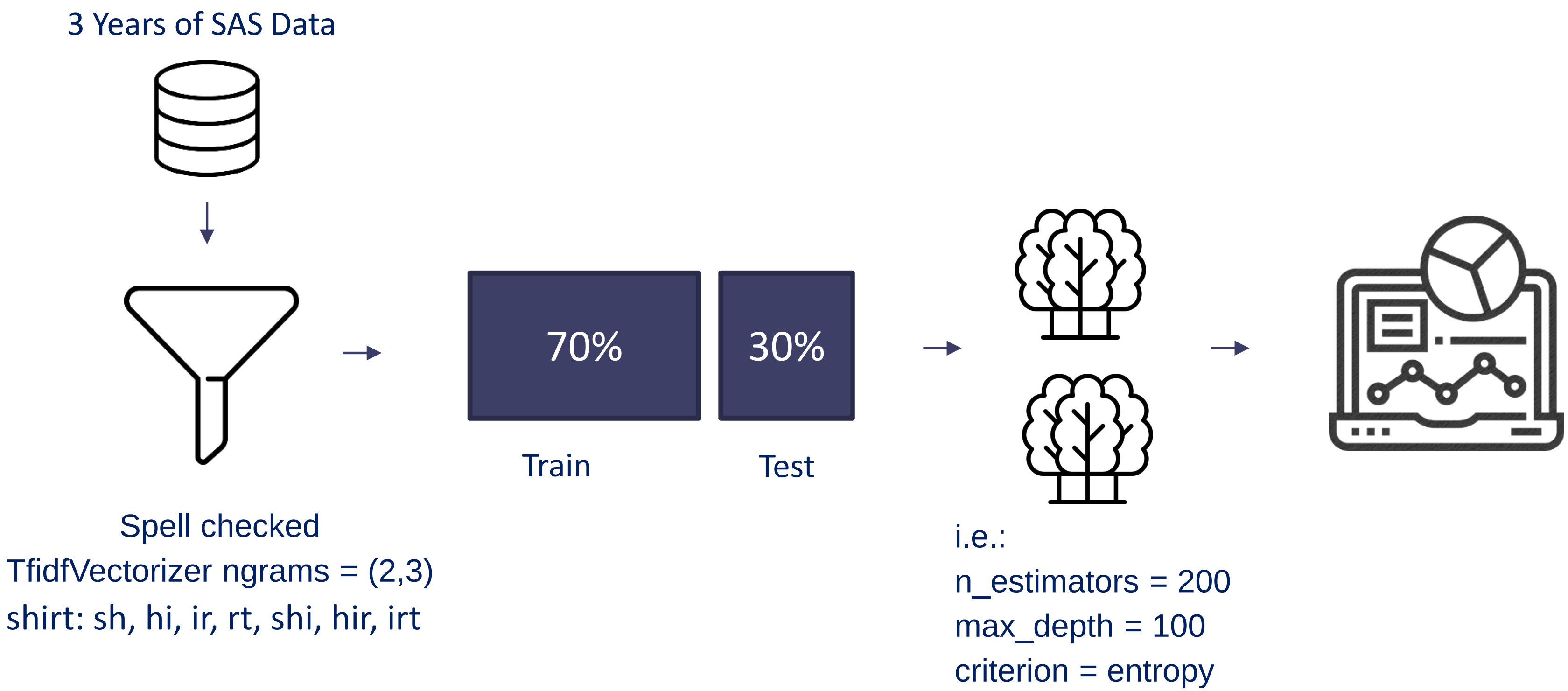
```json
{
  "spagh": "spaghetti",
  "spagheti": "spaghetti",
  "spaghettie": "spaghetti",
  "spaghtti": "spaghetti",
}
```

Insertion/Deletion Penalty: 1

Model Architecture

Item Descriptions are **spell checked and vectorized** before being split into training and testing data sets.

The data sets are then fed into respective Random Forest Models and **Item predictions are reflected on the dashboard.**

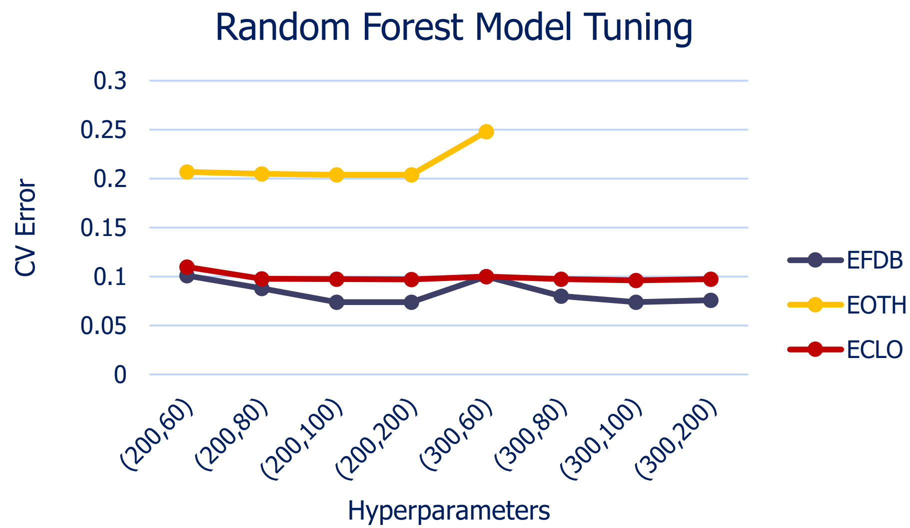


Train Model

Each RF model is trained using KFold cross validation.

Hyperparameter settings:

n_estimators: [200, 300]
max_depth: [60, 80, 100, 200]
criterion: [‘gini’, ‘entropy’]

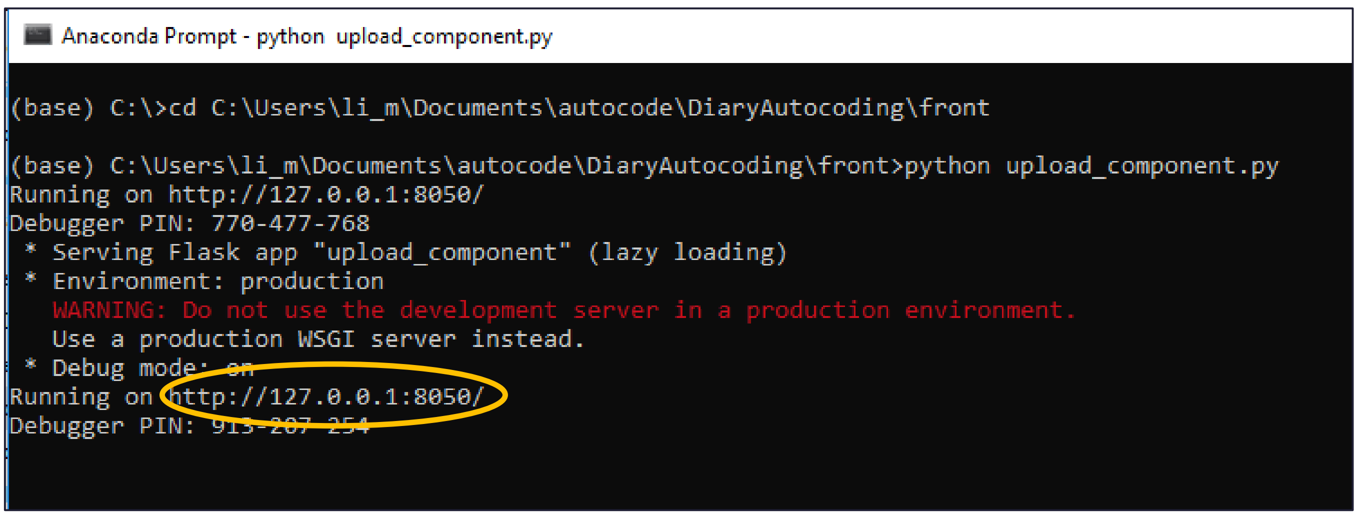


Total # of features

Other models tried:
Logistic Regression, SVM, Decision Tree

Demo

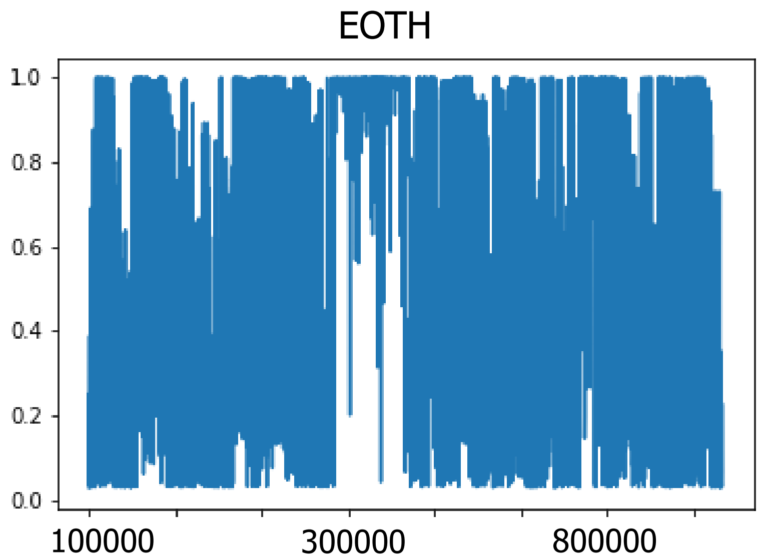
- Installation:
  - Anaconda/Python required
  - Pip install autocoder package
- Dashboard
- Download to Excel

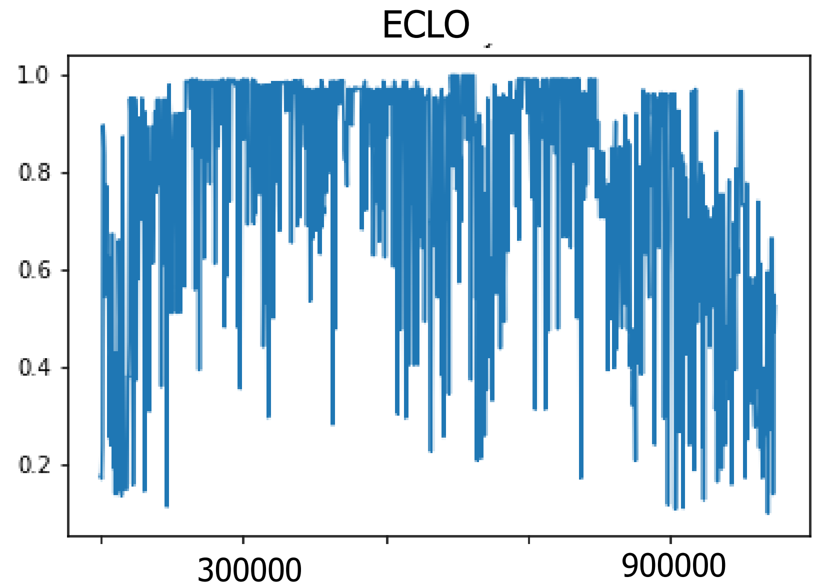


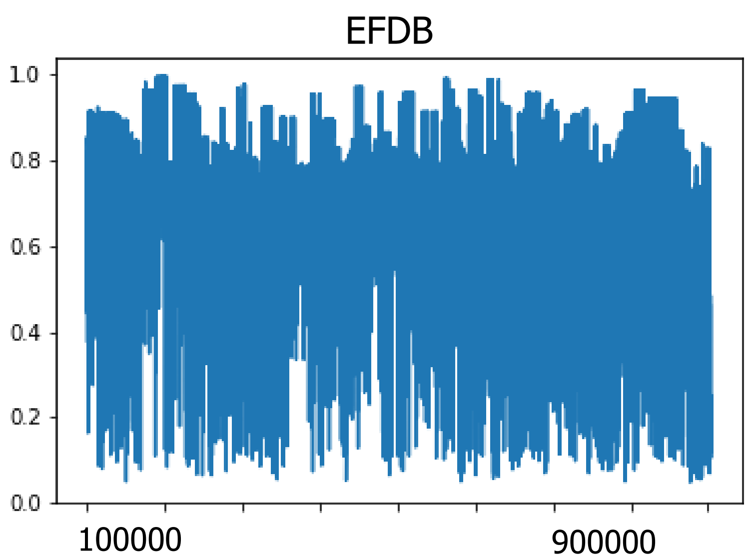
Evaluation

Model	Precision	Recall	F1 Score	Accuracy
Clothing	0.91	0.9	0.9	0.9
Food and Beverages	0.95	0.94	0.94	0.94
Meals	0.98	0.98	0.98	0.98
Other	0.85	0.83	0.83	0.83

Item Code Probability Distributions

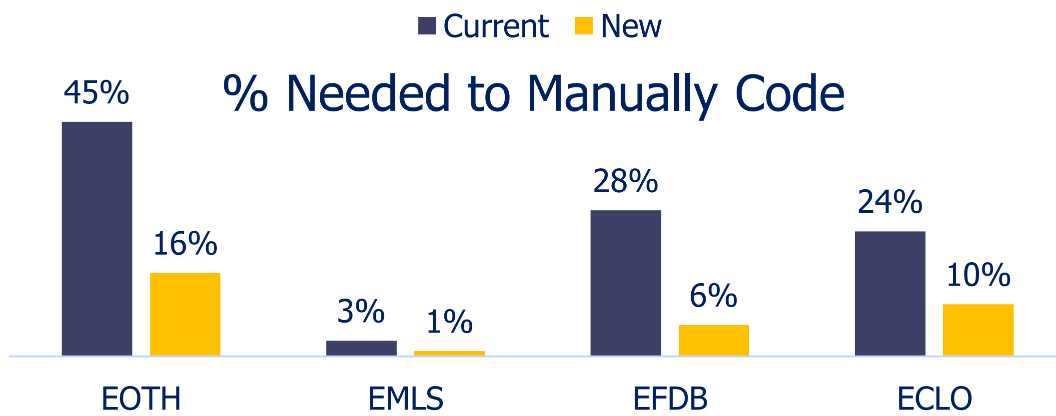




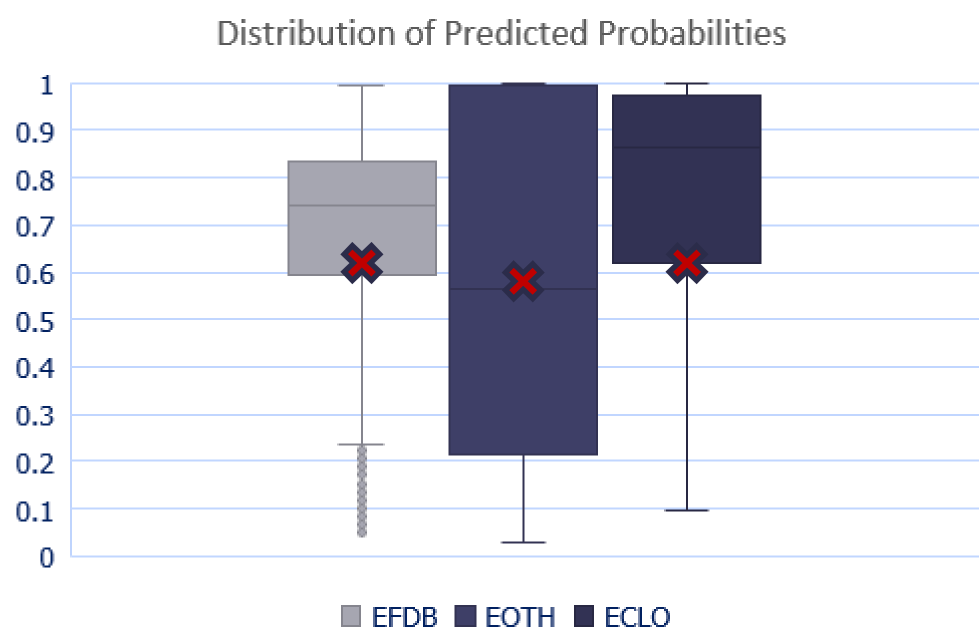


EOTH and EFDB show greater variability in predicted probabilities whereas ECLO shows greater confidence in its predictions.

Evaluation



Current
New
% Needed to Manually Code
45%
16%
3%
1%
28%
6%
24%
10%
EOTH
EMLS
EFDB
ECLO



Distribution of Predicted Probabilities
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
EFDB
EOTH
ECLO

Threshold Evaluation

Above Threshold

Model	Precision	Recall	F1 Score	Accuracy
Clothing	0.97	0.96	0.97	0.97
Food and Beverages	0.99	0.99	0.99	0.99
Other	0.97	0.97	0.97	0.97

Below Threshold

Model	Precision	Recall	F1 Score	Accuracy
Clothing	0.802	0.7	0.71	0.7
Food and Beverages	0.87	0.85	0.85	0.85
Other	0.79	0.73	0.72	0.73

Takeaways

1. Even though **EFDB** has the most number of unique item codes (185), **the codes are the most separable** (least overlap); **ECLO** item codes have **the most overlap**
2. Thresholds can be lowered (less manual coding) to a certain extent without accuracy compromise
3. Misspellings does not impact accuracy that much
   - (ECLO accuracy +0.4% after misspellings fixed)

Future Work

This autocoder is a **promising proof of concept** for use of machine learning at the BLS.

- **Moving away from Census NPC coding:** Semi-supervised learning methods can predict item codes using past data's target variable without the need for its own

- **Reducing model error:** boosting methods, fix misspellings, dimensionality reduction, more training data (more CPUs), include store name

- **Thresholding for manual classification:** One vs. Rest Classifier ROC curve to determine specific threshold

- **Incorporating BLS expertise:** Diary specific vector embeddings using Genism's Word2Vec

Thank You!

Questions?

Appendix

Ethics of Machine Learning

Machine Learning uses **past data sets to predict outcomes**, meaning a model is only as good as its data.

Machine Learning in the Government

Bias and discrimination: [Government Crime Classification Tool Racially Biased]
Erosion of Privacy: [Chinese Government Launches Social Credit System], [NYC Patrolling Officers Wear Body Cameras]

Machine Learning at the BLS

Models have **preferences**.

```
"pepperoi": "pepper",
```

Automation contributes **to workforce displacement**.

Automation automates **human biases**.

ItemType	ItemCode	ItemDescription
EOTH	821132	ACRYLIC NAILS
EOTH	870048	NAILS
EOTH	314120	COMMON NAILS

Under what circumstances is autocoding worthwhile?