# Using Machine Learning to Classify ORS Task Data

**Rebecca Hu**

Civic Digital Fellow

Office of Compensation and Working Conditions

Demo Day

August 8, 2019

# Project Overview

- Goal: Classify work task data from the Occupational Requirements Survey (ORS) into General Work Activities (GWA)
- Phase 1: Use Occupational Information Network (O*NET) data to train a machine learning model to classify work tasks into GWA
- Phase 2: Use the trained model to classify ORS work task data into GWA

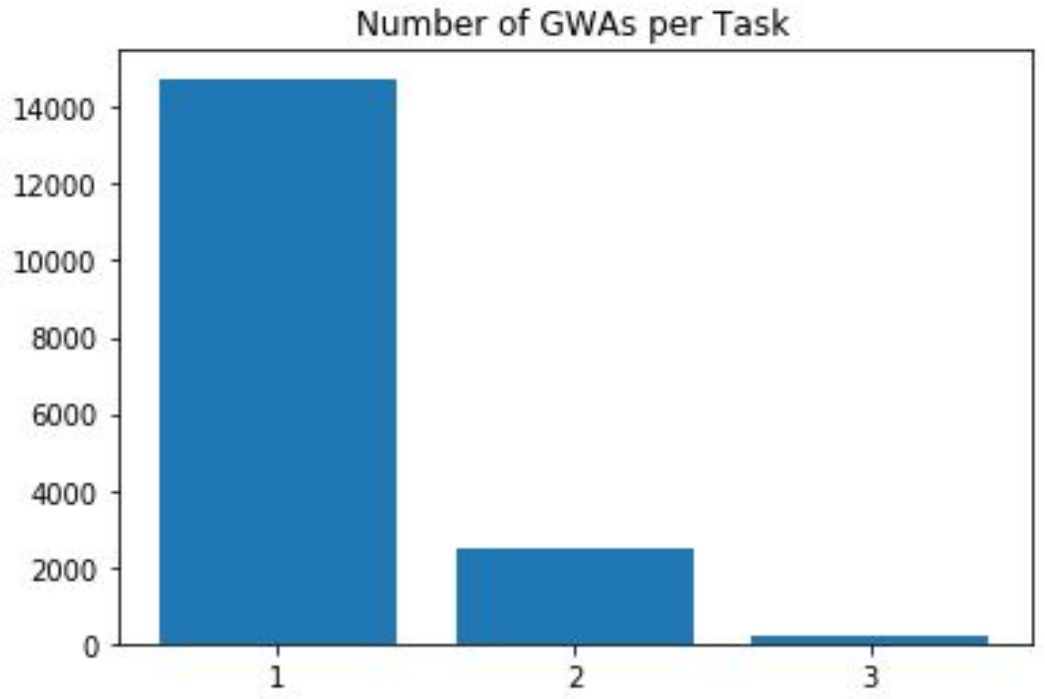
# The Data

| | Task | GWA |
|---|---|---|
| 0 | Review and analyze legislation, laws, or publi... | Analyzing Data or Information |
| 1 | Review and analyze legislation, laws, or publi... | Provide Consultation and Advice to Others |
| 2 | Direct or coordinate an organization's financi... | Guiding, Directing, and Motivating Subordinates |
| 3 | Confer with board members, organization offici... | Communicating with Supervisors, Peers, or Subo... |
| 4 | Analyze operations to evaluate performance of ... | Analyzing Data or Information |
| 5 | Direct, plan, or implement policies, objective... | Making Decisions and Solving Problems |
| 6 | Direct, plan, or implement policies, objective... | Developing Objectives and Strategies |
| 7 | Direct, plan, or implement policies, objective... | Guiding, Directing, and Motivating Subordinates |
| 8 | Prepare budgets for approval, including those ... | Guiding, Directing, and Motivating Subordinates |
| 9 | Direct or coordinate activities of businesses ... | Guiding, Directing, and Motivating Subordinates |
| 10 | Negotiate or approve contracts or agreements w... | Resolving Conflicts and Negotiating with Others |
| 11 | Review reports submitted by staff members to r... | Analyzing Data or Information |
| 12 | Appoint department heads or managers and assig... | Guiding, Directing, and Motivating Subordinates |
| 13 | Direct human resources activities, including t... | Guiding, Directing, and Motivating Subordinates |
| 14 | Prepare or present reports concerning activiti... | Documenting/Recording Information |
| 15 | Implement corrective action plans to solve org... | Resolving Conflicts and Negotiating with Others |
| 16 | Coordinate the development or implementation o... | Guiding, Directing, and Motivating Subordinates |
| 17 | Direct non-merchandising departments, such as ... | Guiding, Directing, and Motivating Subordinates |
| 18 | Deliver speeches, write articles, or present i... | Communicating with Persons Outside Organization |
| 19 | Serve as liaisons between organizations, share... | Communicating with Supervisors, Peers, or Subo... |

- A subset of the O*NET data
- Pure text
- One column of task data
- One column of GWA labels

# Phase 1: O*NET

- Merging the Data
- EDA
- Pre-processing
- Classification Model
- Evaluation



# Pre-processing

- Remove numbers, punctuation, stopwords
- Stemming
- TF-IDF
- N-grams

```
['Maintain accurate, complete, and correct student records as required by laws, district policies, and administrative regulati
ons.']
**********************************************************************
Stemmed words and n-grams:
 ['accur', 'accurate complet', 'accurate complete correct', 'accurate complete correct stud', 'accurate complete correct stude
nt record', 'administr', 'administrative regul', 'complet', 'complete correct', 'complete correct stud', 'complete correct stu
dent record', 'complete correct student records requir', 'correct', 'correct stud', 'correct student record', 'correct student
records requir', 'correct student records required law', 'district', 'district polici', 'district policies administr', 'distri
ct policies administrative regul', 'law', 'laws district', 'laws district polici', 'laws district policies administr', 'laws d
istrict policies administrative regul', 'maintain', 'maintain accur', 'maintain accurate complet', 'maintain accurate complete
correct', 'maintain accurate complete correct stud', 'polici', 'policies administr', 'policies administrative regul', 'recor
d', 'records requir', 'records required law', 'records required laws district', 'records required laws district polici', 'regu
l', 'requir', 'required law', 'required laws district', 'required laws district polici', 'required laws district policies admi
nistr', 'student', 'student record', 'student records requir', 'student records required law', 'student records required laws
district']
```

# Classification Model

- After benchmarking some models, we decided to use the Logistic Regression implementation from sk-learn as our baseline
- Use One-vs-Rest technique for multi-labels
  - Oversample the target class and under-sample the rest
  - Train 41 unique classifiers
  - Combine results for final predictions

# Model Results

- Because the labels of the O*NET data overlap, it is difficult to gage how well the model is truly performing.
- By looking through the results, in general, it seems the model does a decent job about predicting related labels
- It over-labels GWA, thus favoring recall over precision, which is preferred for our purposes.

| Tasks | GWA 1 | GWA 2 | GWA 3 | Predicted 1 | Predicted 2 | Predicted 3 |
|---|---|---|---|---|---|---|
| Inspect and maintain vehicle supplies and equipment, such as gas, oil, water, tires, lights, or brakes, to ensure that vehicles are in proper working condition. | Inspecting Equipment, Structures, or Material | Repairing and Maintaining Mechanical Equipment | None | Monitor Processes, Materials, or Surroundings | Inspecting Equipment, Structures, or Material | Repairing and Maintaining Mechanical Equipment |
| Provide applicants with assistance in completing application forms, such as those for job referrals or unemployment compensation claims. | Assisting and Caring for Others | None | None | Staffing Organizational Units | Assisting and Caring for Others | None |

# Phase 2: ORS

- (ORS data is private so it is not displayed here)
- Clean ORS data
- Use pre-trained models to predict labels
- ORS has no “ground-truth” GWA labels

# Challenges

- O*NET labels are similar, and their descriptions overlap
- The majority of tasks are only labelled with one label, but the model will predict many similar labels
- Thus, quantitative metrics are unreliable
- O*NET and ORS data tasks are written differently
- ORS data has no GWA labels

# Contact Information

**Rebecca Hu**

Civic Digital Fellow

Office of Compensation and Working Conditions

www.bls.gov/ORS

Hu.Rebecca@bls.gov