Predicting Future Unemployment

Mountain Plains Business Conference

Robert Winslow

Ocotober 5, 2024

Context

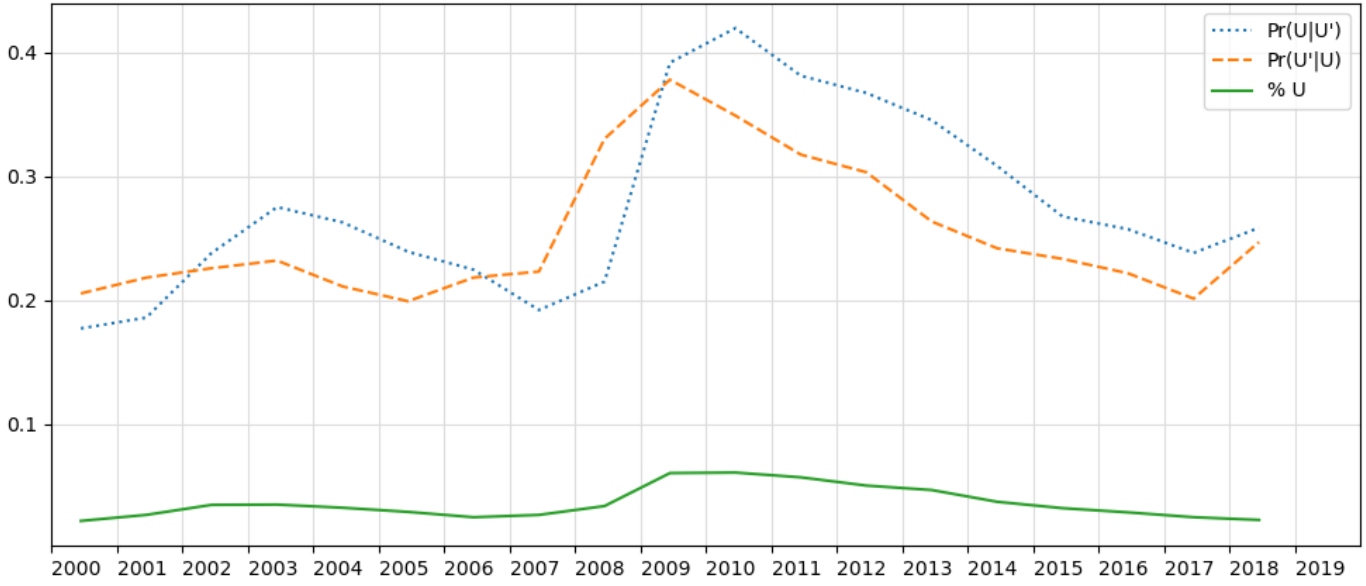
- 2022: Machine Learning Competition
- Binary prediction about whether each person will be unemployed in one year's time.
- The competition's scoring metric placed equal weight on accurate predictions of unemployment and accurate predictions of non-unemployment:
  - $GF \equiv \frac{\#\text{ Correctly Predicted Unemployed}}{\#\text{ Unemployed}} \cdot \frac{1}{2} + \frac{\#\text{ Correctly Predicted Not Unemployed}}{\#\text{ Not Unemployed}} \cdot \frac{1}{2}$
- Data is drawn from the CPS outgoing rotation groups
  - people aged 20-64
  - years 1999-2018 (Mebdi Competition covered years 2008-2014)
    - 1.4 million observations, roughly 3% of whom will be unemployed in one year's time.

Model Accuracy Overview

	LASSO	Ridge	Gradient Boosted Decision Trees	Simple Ensemble	ENU Ensemble
Balanced Accuracy (GF)	**73.4**	**73.4**	**73.0**	**73.7**	**73.8**
Will Be Unemployed	**73.2**	**73.8**	**65.9**	**71.0**	**70.8**
Employed→Unemployed	56.3	57.2	44.7	52.6	52.3
NILF→Unemployed	78.5	79.3	72.3	77.1	77.0
Unemployed→Unemployed	100	100	99.8	100	100
Won't Be Unemployed	**73.6**	**73.0**	**80.1**	**76.3**	**76.7**
Employed→Employed	78.2	77.8	85.4	81.1	81.8
NILF→Employed	40.7	39.4	47.1	42.2	39.1
Unemployed→Employed	0	0	0.7	0	0.5
Employed→NILF	64.1	63.6	71.4	66.8	64.6
NILF→NILF	74.9	73.9	79.5	77.4	78.2
Unemployed→NILF	0	0	1.3	0	0.6

Current Unemployment Predicts Future Unemployment

- Most of the signal comes from Current Employment Status
- A single-variable model using only current employment status can achieve a score of $GF = 64\%$ by itself.
  - This heuristic faired even better in the competition sample.



Which Variables Are Most Important to the Model?

- Permutation Importance:
  1. Fit a model and evaluate predictions.
  2. Permute a feature or set of features.
  3. Make predictions with permuted X, and re-evaluate.
  4. Take the difference in scores.
- In the simple ensemble, the most important groups of features are:

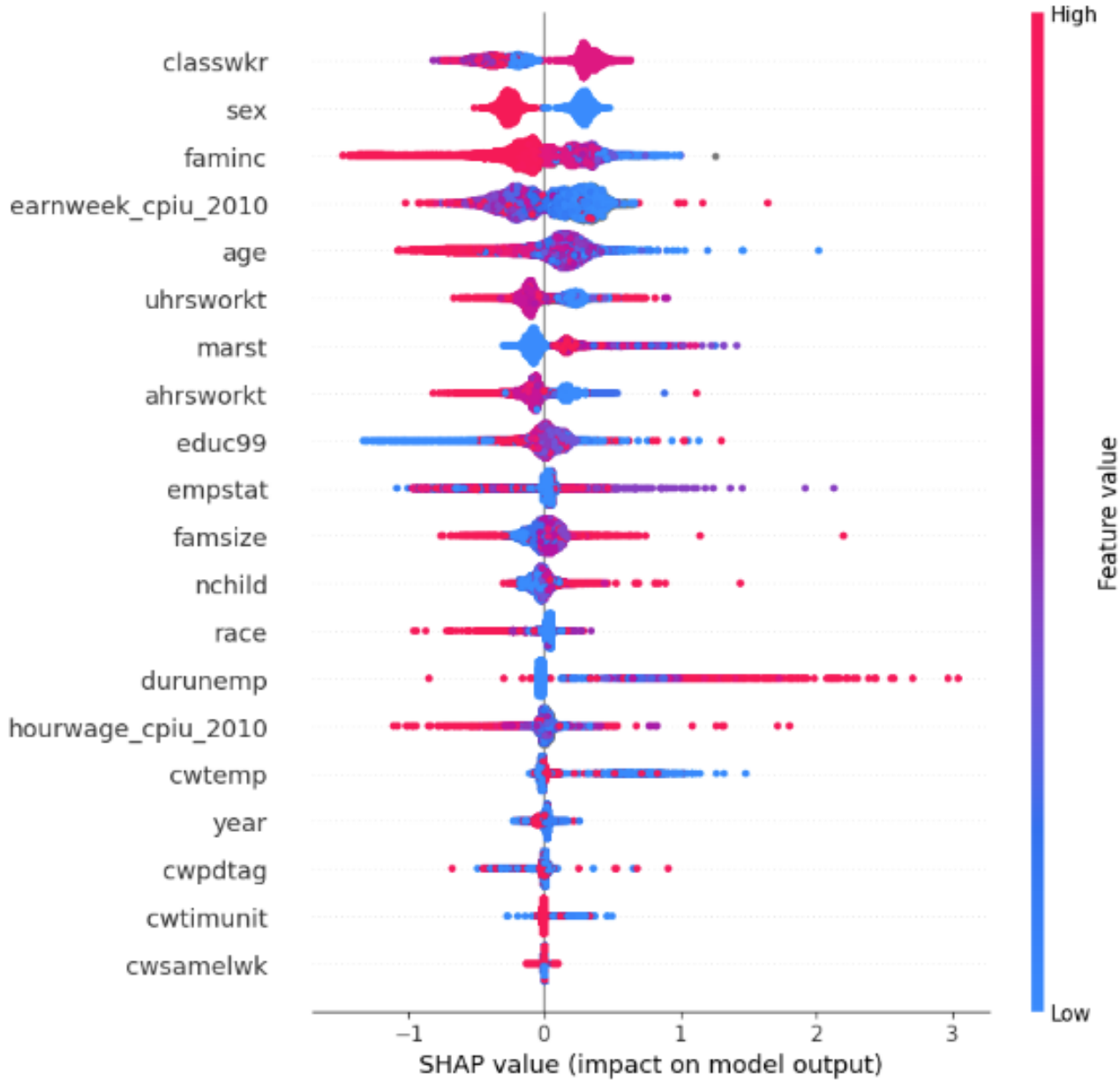
	LASSO	Ridge	Gradient Boosted Decision Trees	Simple Ensemble
Histories	0.050	0.043	0.044	0.042
Employment Status	0.034	0.038	0.056	0.035
Class of Worker	0.019	0.024	0.007	0.015
Work Status	0.018	0.025	0.002	0.013
Time Period	0.012	0.012	0.013	0.013
Earnings/hourly wages	0.008	0.013	0.005	0.007

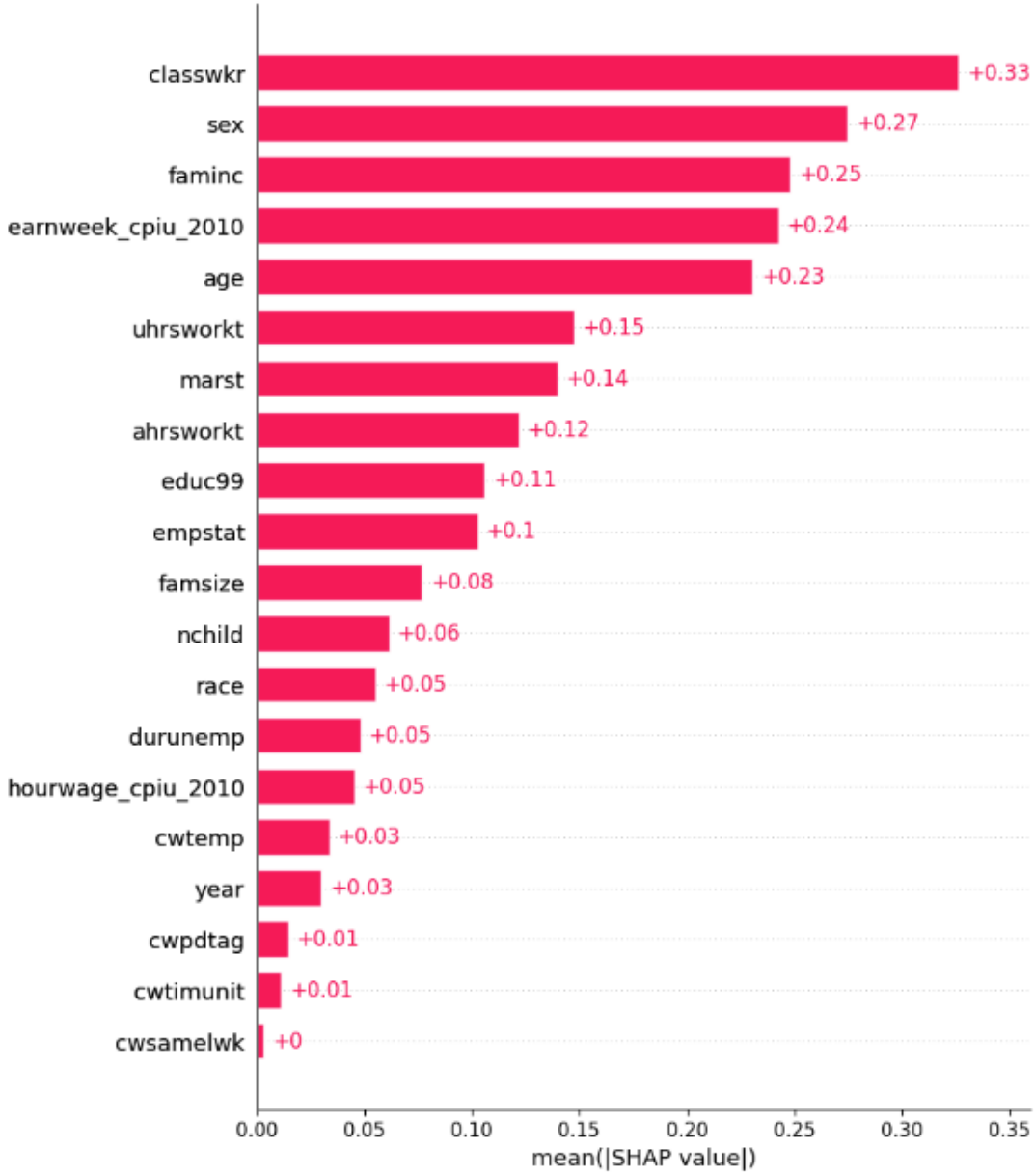
So What Can I do With This?

- Consistent with evidence that workers persistently differ in their attachment to employment.
  - Ex: (Victoria, Menzio, Wiczer. 2021): Cluster analysis seperates workers into types corresponding to stability of employment, and type is not forecast by demographics or industry.

So What Can I do With This?

- Or maybe I can use the structure of CPS data to weigh in on the use of administrative data
  - adminstrative data is high quality but lacks some of the detail of survey questions
  - CPS has both a short panel and very specific questions.
    - Does introducing a short panel diminish the importance of those "extra" questions?


classwkr
sex
faminc
earnweek_cpiu_2010
age
uhrsworkt
marst
ahrsworkt
educ99
empstat
famsize
nchild
race
durunemp
hourwage_cpiu_2010
cwtemp
year
cwpdtag
cwtimunit
cwsamelwk
−1
0
1
2
3
SHAP value (impact on model output)
High
Feature value
Low


classwkr
sex
faminc
earnweek_cpiu_2010
age
uhrsworkt
marst
ahrsworkt
educ99
empstat
famsize
nchild
race
durunemp
hourwage_cpiu_2010
cwtemp
year
cwpdtag
cwtimunit
cwsamelwk
+0.33
+0.27
+0.25
+0.24
+0.23
+0.15
+0.14
+0.12
+0.11
+0.1
+0.08
+0.06
+0.05
+0.05
+0.05
+0.03
+0.03
+0.01
+0.01
+0
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
mean(|SHAP value|)