What Predicts Whether a Person Will Be Unemployed Next Year?

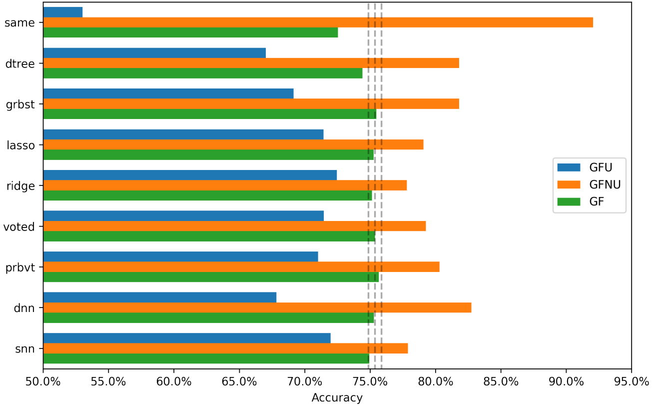
Robert Winslow

2024

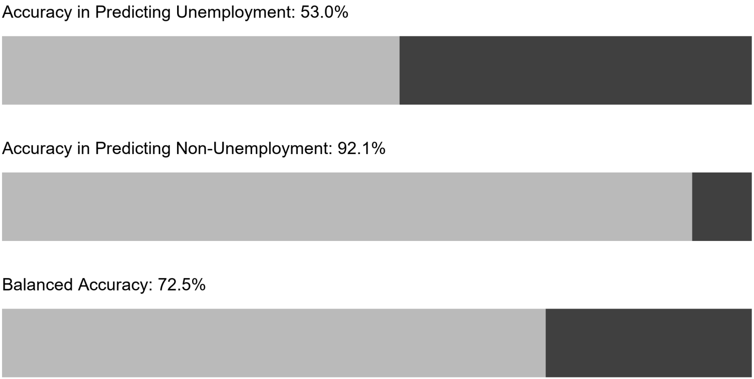
Details About the Task

- Binary prediction about whether each person will be unemployed in one year's time.
- Unbalanced data: Only 5 percent of individuals 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 2008-2014

Score Comparison for Different Models



Scores for Simple Heuristic: Assume Empstat Doesn't Change



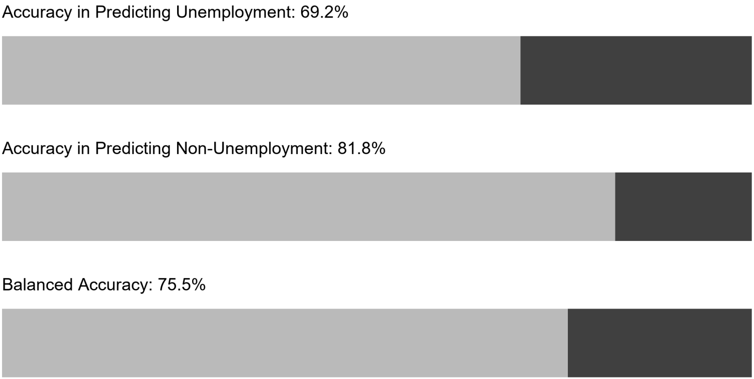
Scores for Decision Tree

Accuracy in Predicting Unemployment: 67.0%

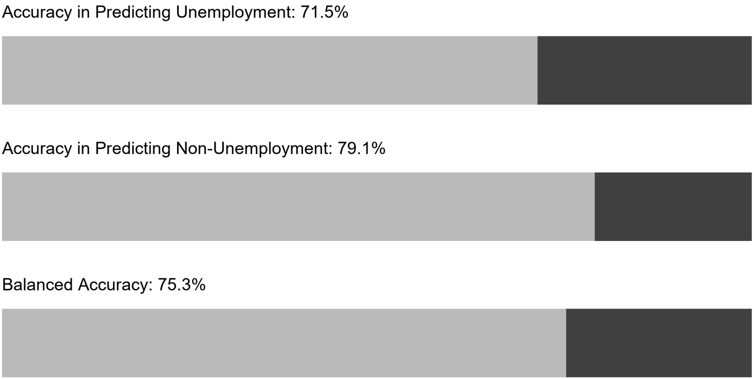
Accuracy in Predicting Non-Unemployment: 81.8%

Balanced Accuracy: 74.4%

Scores for Gradient Boosted Decision Tree



Scores for Regularized Regression (Lasso)



Scores for Regularized Regression (Ridge)

Accuracy in Predicting Unemployment: 72.5%

Accuracy in Predicting Non-Unemployment: 77.8%

Balanced Accuracy: 75.1%

Scores for 2-out-of-3 Vote

Accuracy in Predicting Unemployment: 71.5%

Accuracy in Predicting Non-Unemployment: 79.3%

Balanced Accuracy: 75.4%

Scores for 2-out-of-3 Vote (With Gradient Boosting)

Accuracy in Predicting Unemployment: 71.0%

Accuracy in Predicting Non-Unemployment: 80.3%

Balanced Accuracy: 75.7%

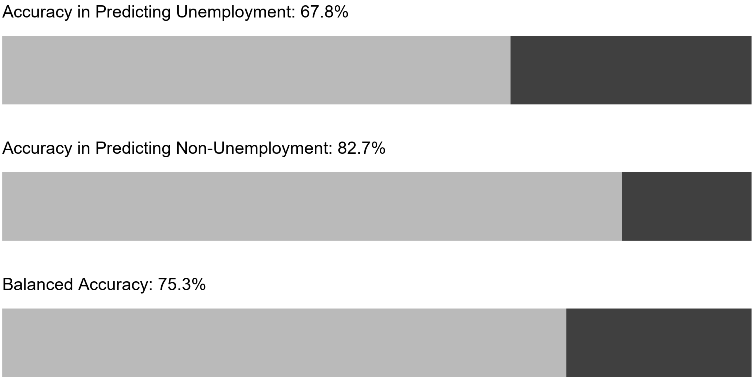
Scores for Shallow Neural Net

Accuracy in Predicting Unemployment: 72.0%

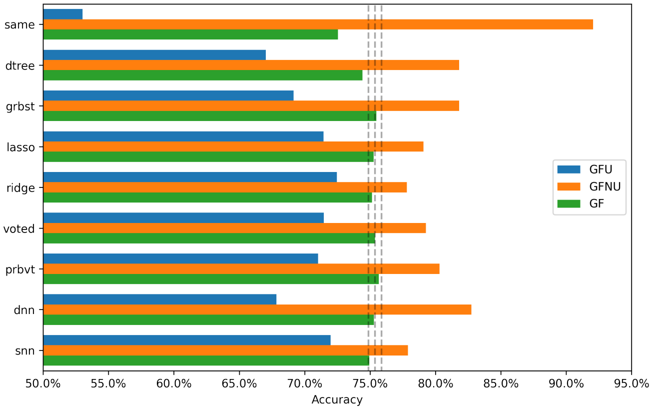
Accuracy in Predicting Non-Unemployment: 77.9%

Balanced Accuracy: 74.9%

Scores for Deep Neural Net



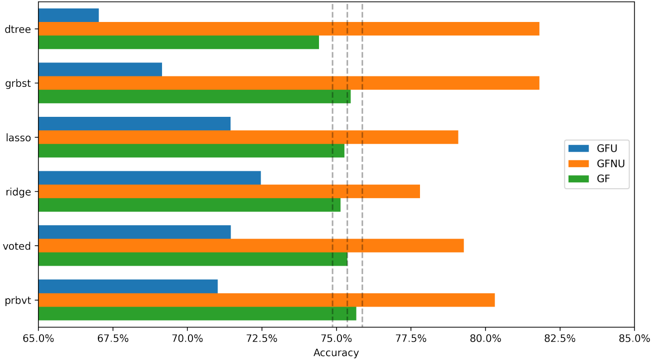
Score Comparison for Different Models



Does Adding Extra Features from CPS Help?

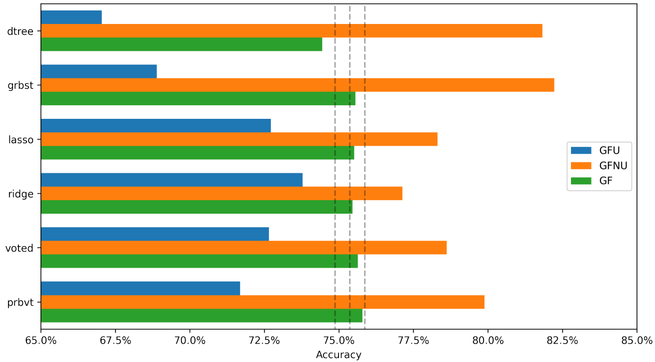
Does Adding Extra Features from CPS Help?

Using only variables from the MEBDI sample:



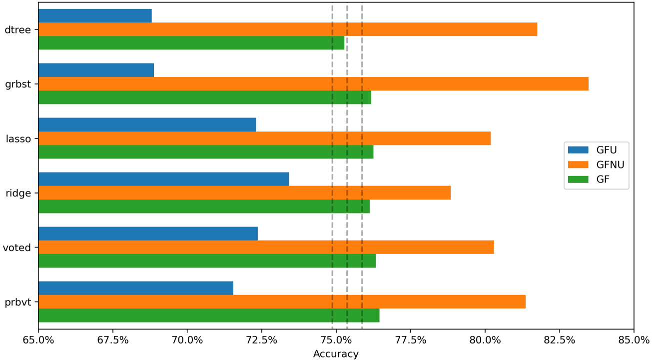
Does Adding Extra Features from CPS Help?

Using additional variables from IPUMS CPS:



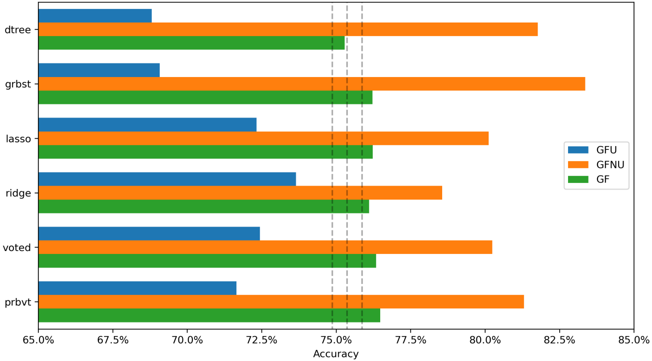
Does Adding Extra Features from CPS Help?

Using worker's data from prior three months as well:

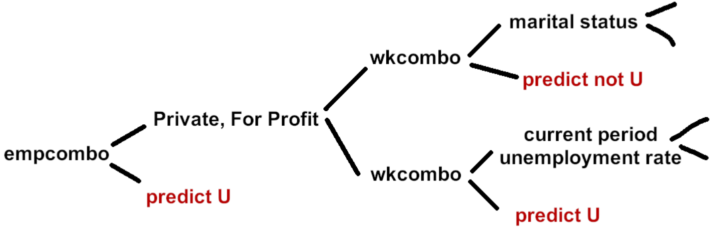


Does Adding Extra Features from CPS Help?

Using spousal characteristics as well:



An Example Small Decision Tree



LASSO Regularizes Linear Regressions

With LASSO, the objective is:

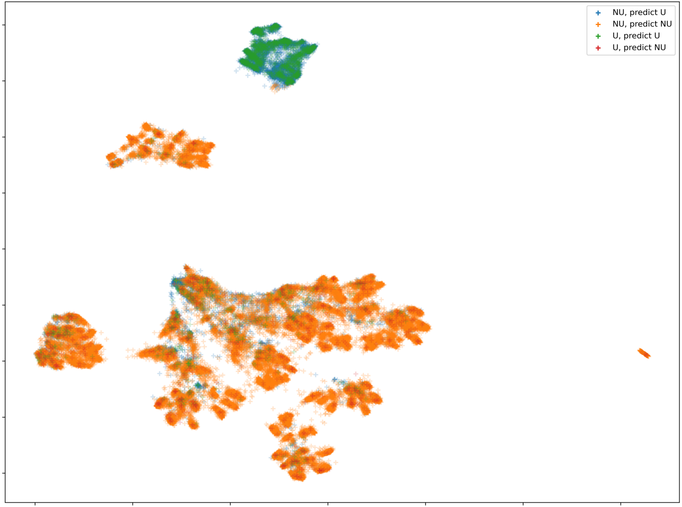
$$\min_{\beta} \sum (X\beta - y)^2 + \alpha \sum |\beta|$$

Some Important Coefficients in Lasso

Variable Description	Contribution to GF
Works at Private, for Profit Firm	0.00081015
Number of Jobs	0.000588447
Family Income $10,000 - 12,499	0.000505645
Family Income $5,000	0.000409072
Not a Citizen	0.000382732
Actual Hours Worked Last Week	0.000363647
Family Income $7,500 - 9,999	0.000335916
fFamily Income $20,000 - 24,999	0.000259002
Family Income $12,500 - 14,999	0.000239303
Industry: Construction	0.000228201
Family Income $15,000 - 19,999	0.000187173
"wnftlook" >12 months	0.000185251
Usual Hours Worked	0.000157677

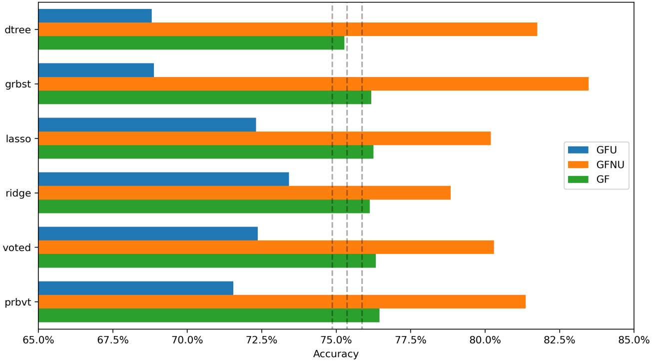
UMAP (Uniform manifold approximation and projection)

Dimension Reduction of Data:



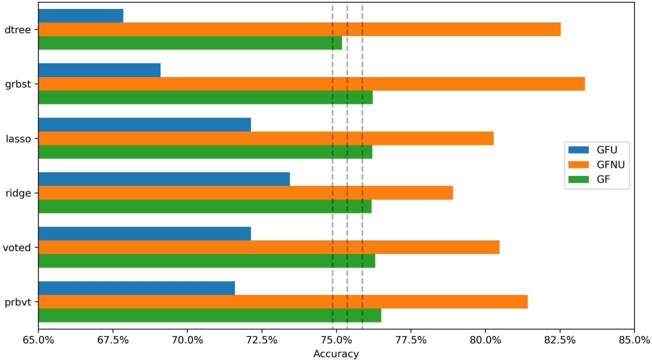
UMAP Doesn't Improve Scores Much

Scores with Extra IPUMS variables, including worker history:



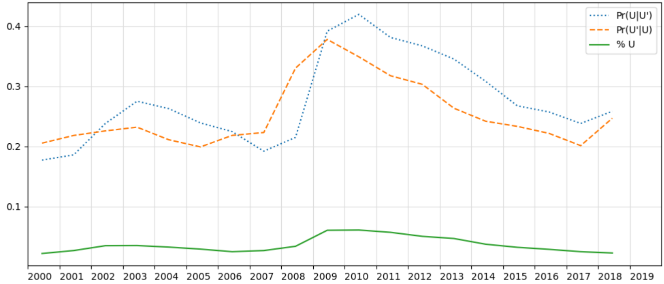
UMAP Doesn't Improve Scores Much

Same, with UMAP embeddings included as features:



Analysis by Year.

Unemployment Persistence by Year



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.

LASSO Importance for Years 2001-2003

On data from 2001-2003, model scores GF=.706, GFU=.711

	GF Importance		GFU Importance
empstat	0.031	class_selfemp	0.025
durunemp	0.009	classwkr	0.023
wkcombo	0.008	sex	0.013
ahrsworkt	0.008	multjob	0.013
wkstat	0.008	occg	0.011
classwkr	0.007	uhrsworkt	0.009
class_selfemp	0.007	indg	0.008
sex	0.006	wnftlook	0.006
relate	0.005	racec	0.005
absent	0.004	relate	0.004

LASSO Importance for Years 2004-2006

On data from 2004-2006, model scores GF=.714, GFU=.701

	GF Importance		GFU Importance
empstat	0.035	empstat	0.016
wkstat	0.021	classwkr	0.015
absent	0.020	region	0.013
durunemp	0.014	durunemp	0.011
ahrsworkt	0.011	statefip	0.011
classwkr	0.011	occg	0.011
sex	0.008	class_selfemp	0.010
class_selfemp	0.008	wksworkorg	0.010
empsame	0.008	sex	0.009
racec	0.007	racec	0.009

LASSO Importance for Years 2007-2009

On data from 2007-2009, model scores GF=.741, GFU=.752

	GF Importance		GFU Importance
empstat	0.030	classwkr	0.048
classwkr	0.026	class_selfemp	0.043
wkstat	0.025	year	0.022
class_selfemp	0.024	marst	0.016
year	0.011	region	0.010
durunemp	0.010	multjob	0.010
marst	0.006	statefip	0.009
wkcombo	0.006	empstat	0.008
ahrsworkt	0.005	uhrsworkt	0.008
region	0.005	racec	0.007

LASSO Importance for Years 2010-2012

On data from 2010-2012, model scores GF=.766, GFU=.733

	GF Importance		GFU Importance
empstat	0.033	marst	0.015
wkstat	0.031	faminc	0.014
durunemp	0.017	racec	0.012
classwkr	0.009	occg	0.009
racec	0.008	classwkr	0.007
class_selfemp	0.006	indg	0.007
wkcombo	0.006	multjob	0.006
ahrsworkt	0.006	year	0.005
marst	0.005	yngch	0.004
occg	0.003	nmothers	0.004

LASSO Importance for Years 2013-2015

On data from 2013-2015, model scores GF=.725, GFU=.686

	GF Importance		GFU Importance
empstat	0.027	durunemp	0.010
wkstat	0.023	occg	0.008
durunemp	0.017	class_selfemp	0.007
ahrsworkt	0.008	year	0.007
class_selfemp	0.006	marst	0.007
classwkr	0.004	union	0.005
year	0.003	wksworkorg	0.005
union	0.002	classwkr	0.003
racec	0.002	racec	0.003
uhrsworkt	0.002	indg	0.003

LASSO Importance for Years 2016-2018

On data from 2001-2003, model scores GF=.691, GFU=.634

	GF Importance		GFU Importance
empstat	0.045	faminc	0.023
durunemp	0.016	empstat	0.021
faminc	0.014	durunemp	0.021
absent	0.011	marst	0.015
ahrsworkt	0.009	class_selfemp	0.011
class	0.008	region	0.010
wkcombo	0.007	wnlook	0.008
wkstat	0.006	nchildren	0.008
region	0.006	union	0.008
wnlook	0.006	nmothers	0.008