Unemployment Forecasting Challenge

Predicting Unemployment with Demographic and Labor Force Characteristics

Emma Pérez, Edgar Chicurel and Elena Yustres

14/03/2023

The Plan

01

Introduction:

Descriptive Statistics
and Data Pre-Processing

02

Our Model(s)

03

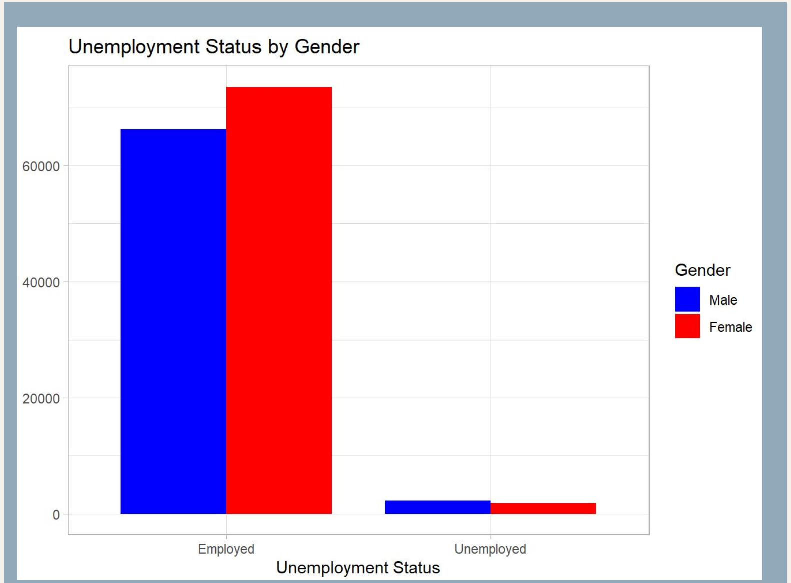
Conclusions & Limitations

01

Introduction:

Descriptive Analysis and Data Pre-Processing

The Current Population Survey (CPS): Descriptive Statistics



The dataset is severely unbalanced toward employed individuals !

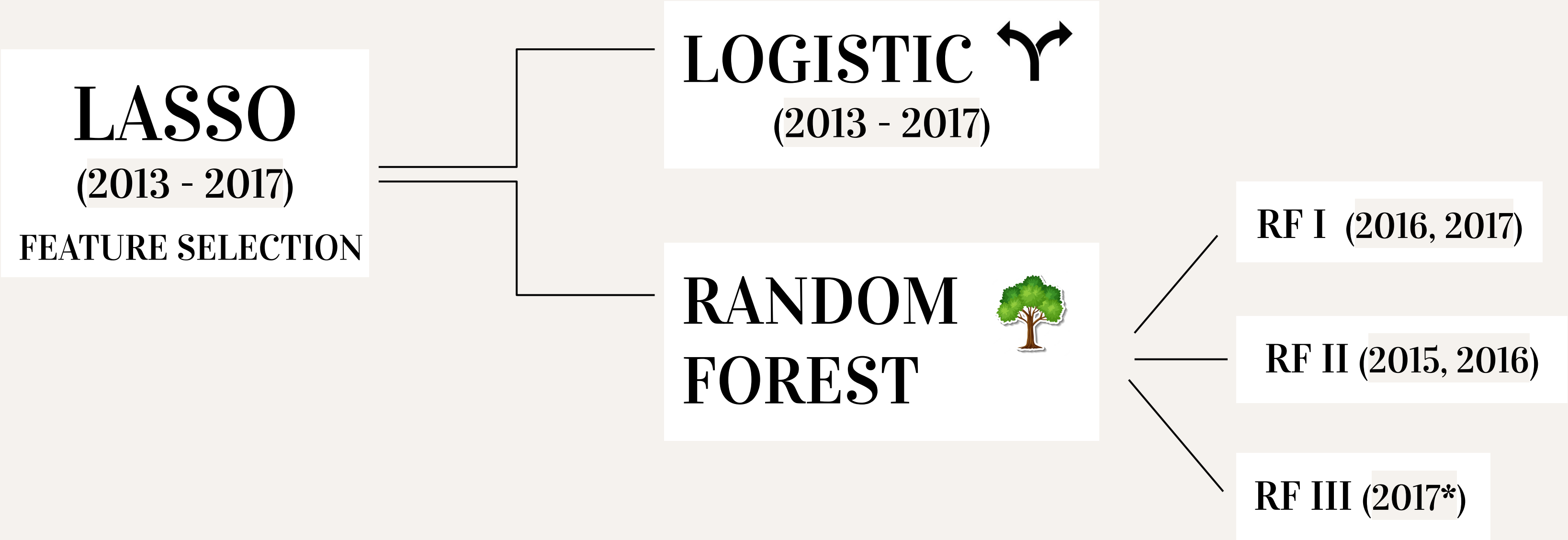
Pre-Processing

- Removed variables with more than 50% of NAs or more than 50% of 0s
- Removed all nominal income variables
- For income ranking variables, removed text “th%”
- Transformed values like “99”, “66”, “Not in Universe”, “Didn’t Respond” to NAs
- Transformed categorical variables to factors and removed variables with many categories, e.g. “state”
- Remove the three variables which **perfectly predict ′unem′**

65 variables

02 Our Model(s)

Modeling: Selecting Relevant Variables and Predicting Individual Unemployment



LASSO
(2013 - 2017)
FEATURE SELECTION

LOGISTIC
(2013 - 2017)

RANDOM FOREST

RF I (2016, 2017)

RF II (2015, 2016)

RF III (2017*)

LASSO

- *Elastic Net* models: regularization incorporates penalty term that encourages model sparsity and prevents overfitting

- Preliminary feature selection: choose variables most strongly associated with *unemployment*

- Obtain 7 relevant variables

1 Presence in **labor force**

2 Someone receiving of **food stamps** in the household

3 Having **private health** insurance

4 **Full/part-time** work

5 **Weeks** worked **previous year**

6 Marginal **tax** rate

7 Person in **poverty**

Logistic

- Data from 2013 to 2017

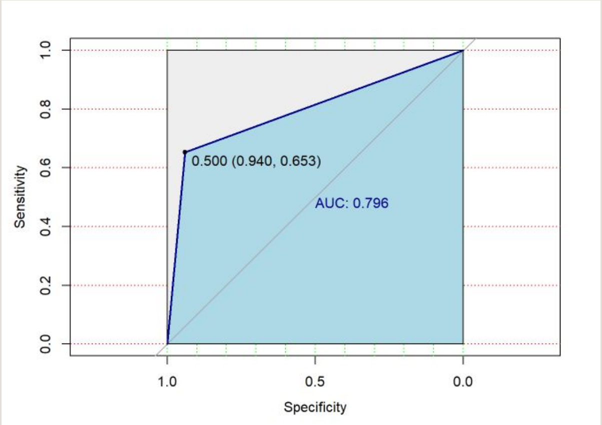
- Modelling probability of unemployment based on a set of predictor variables

LASSO
(7 relevant variables)

→ **LOGISTIC** →

+ selected
socio-economic variables:

8 Age	9 Education
10 Gender	11 Total family income
12 Family in poverty	13 Income from wage & salary



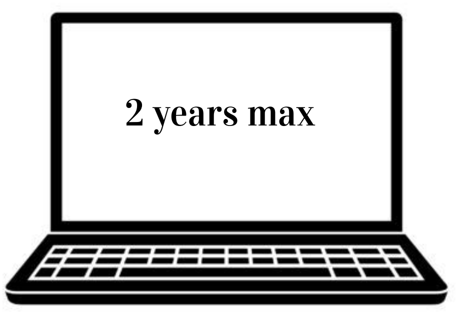
ROC curve

Random Forest I & II

- Improve Logistic AUC □ Machine Learning: **Random Forest**
- Same predictors as Logistic and type = “Classification”
- First try: 2016 & 2017
- Second try: 2015 & 2016

Almost identical AUC

Computational restrictions:

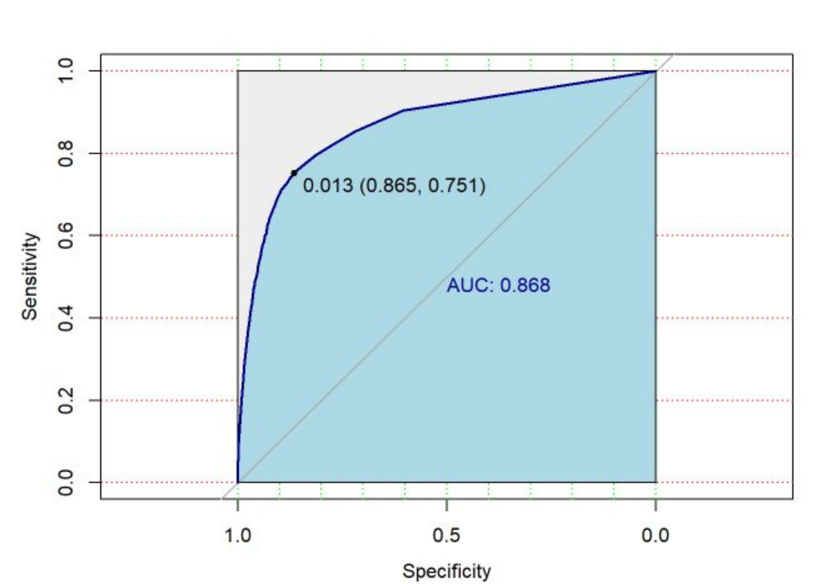


2 years max

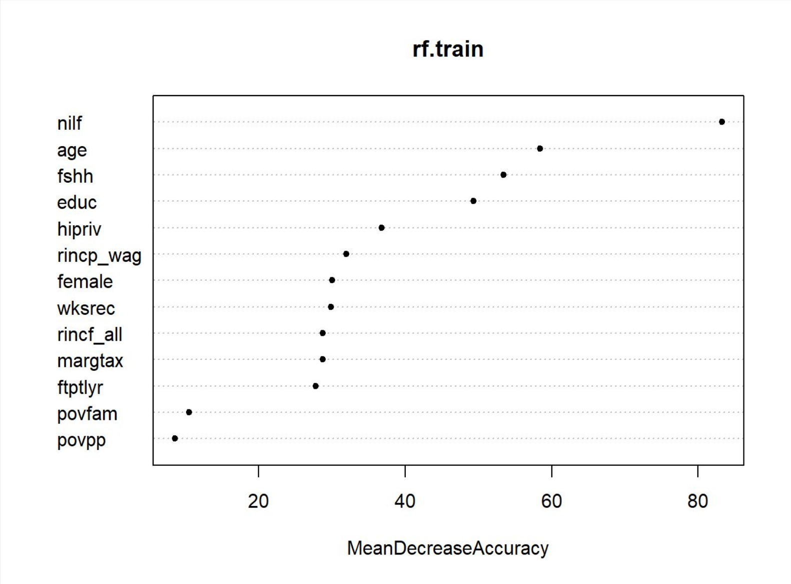
Confusion Matrix, 10% threshold

Prediction	0	1
0	132043	2071
1	4309	1508

Random Forest I & II



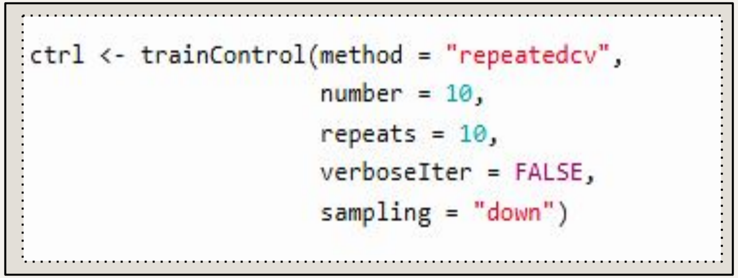
ROC curve



Variable importance

RF III: Random Forest With Undersampling

- Try to improve even more with resampling methods
- Undersampling: reducing number of employed people in the training set



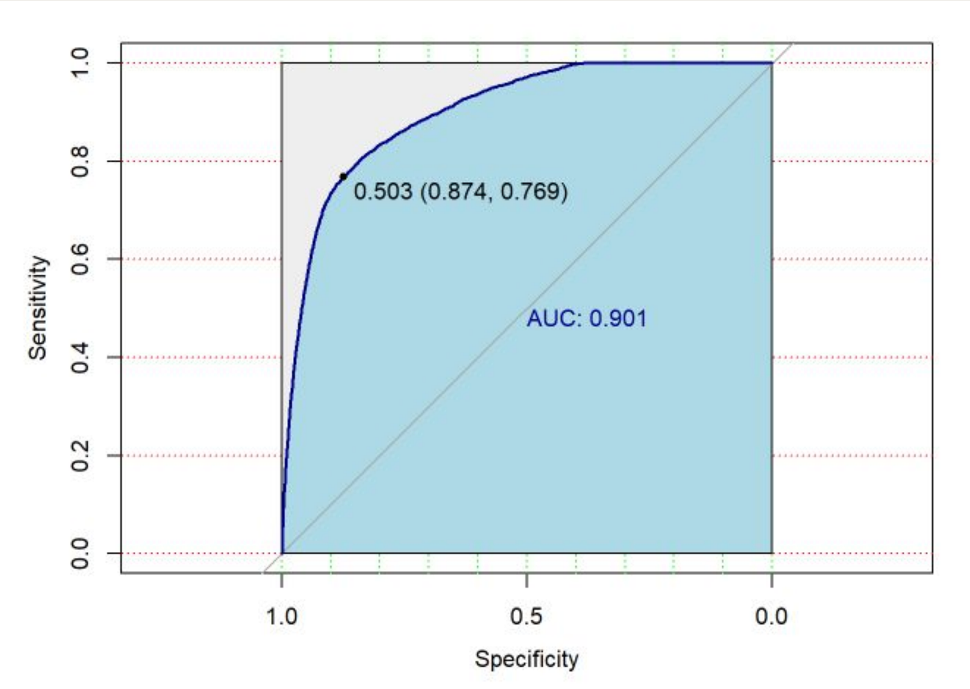
```
ctrl <- trainControl(method = "repeatedcv",
                     number = 10,
                     repeats = 10,
                     verboseIter = FALSE,
                     sampling = "down")
```

- Prediction for 2018 improved a lot:

Prediction	0	1
0	119087	826
1	17265	2753

Threshold = 0.5
Sensitivity = 0.87
Specificity = 0.77

Random Forest III



ROC curve

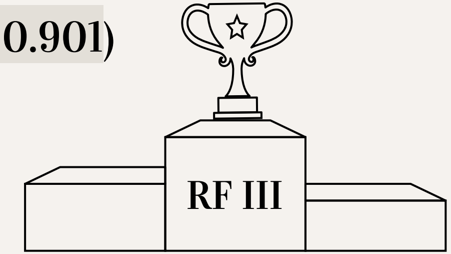
03

Conclusions & Limitations

Conclusions

- Clean □ Select □ LASSO □ Logistic □ RF □ RF with undersampling

- Our winner: RF model with 13 variables and down-sampling (AUC = 0.901)



- Challenges:
  - Cleaning the data - many variables in the CPS dataset with different classifications
  - Restrictions on computational capabilities preventing inclusion of data from more years
    - ⬆ Sample size, models performed (slightly) better

Q & A

Thank you for your attention!