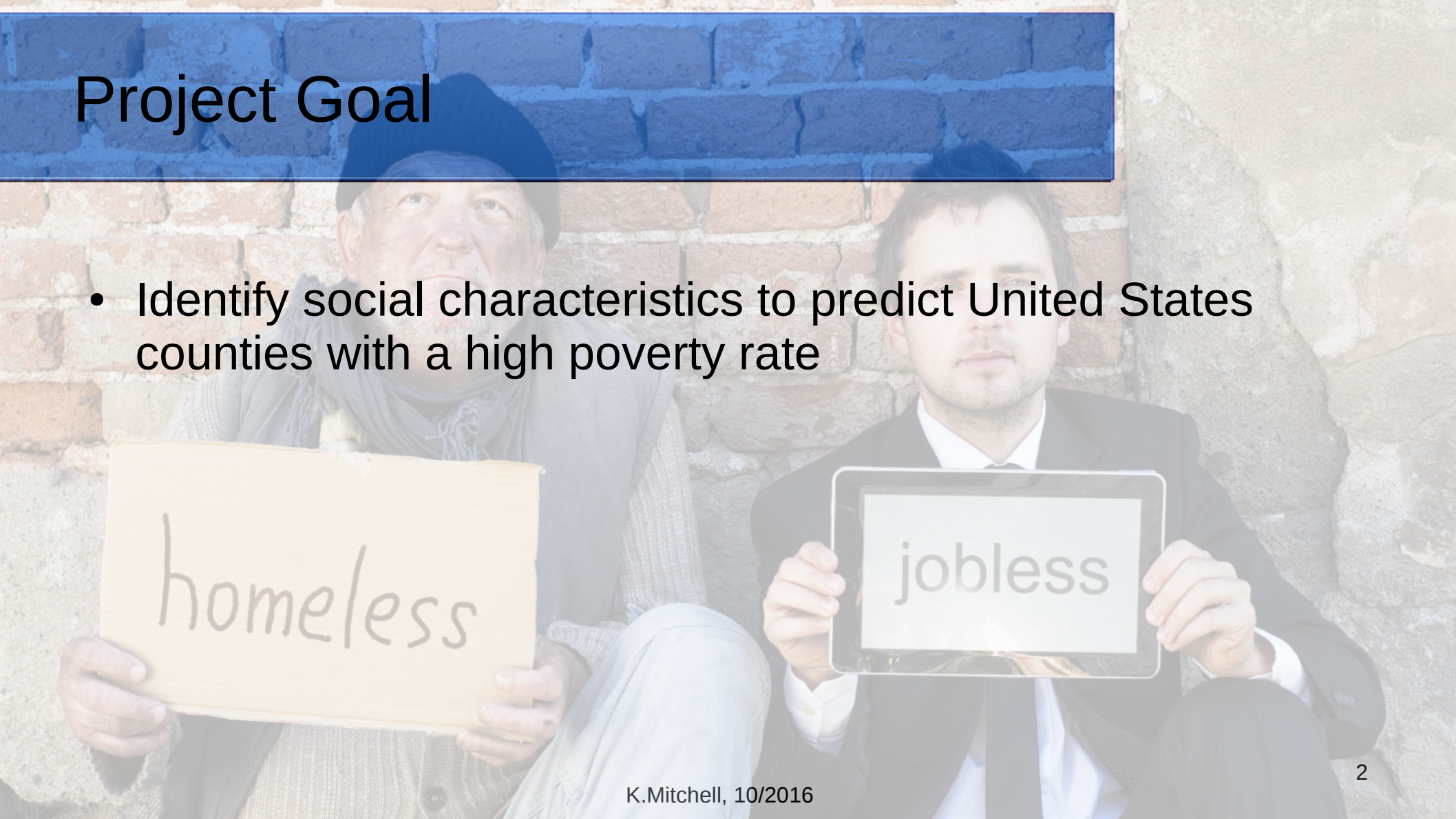
Poverty Extent via Community Factors

Classification Project

Kimberley Mitchell
kim@who-knows.com

Project Goal

- Identify social characteristics to predict United States counties with a high poverty rate



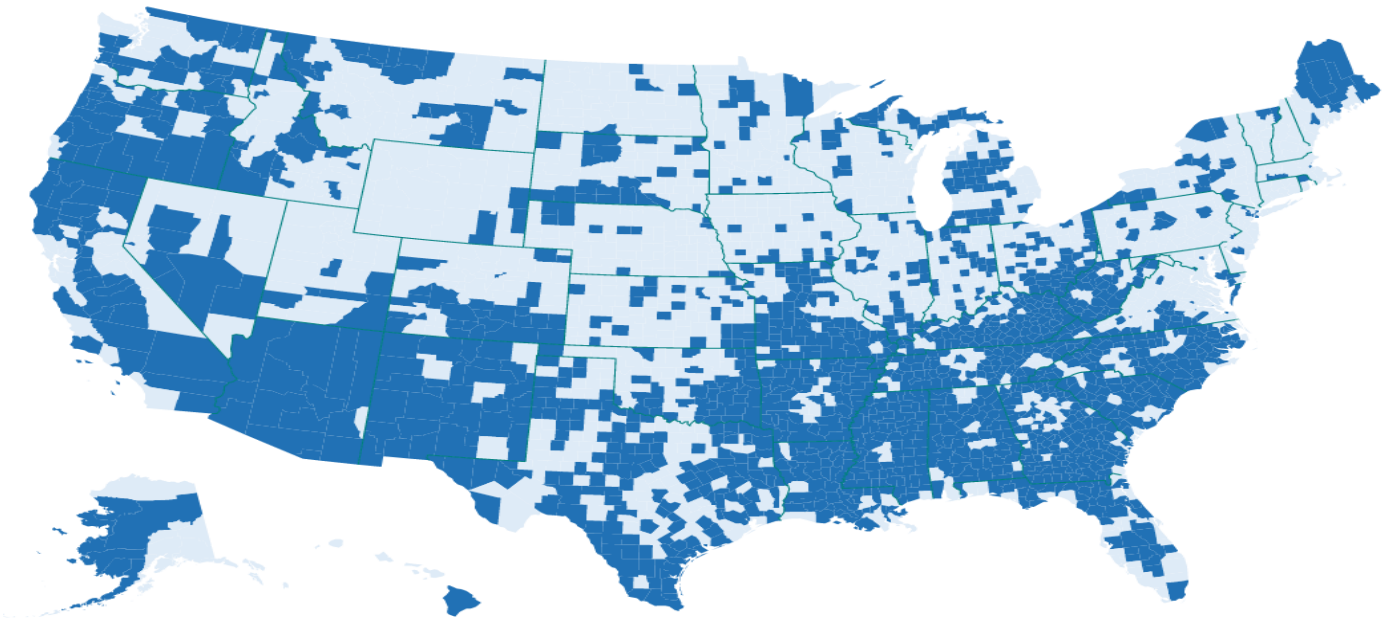
Data: Dependent Variable

- Poverty rate:
  - Per United States county
  - Binned: above (1) or below (0) median
  - Median county poverty rate = 16.0%

- Interactive poverty rate map:
  - http://mitchki.com/D3/poverty.html

- **Source:** 2010-2014 American Community Survey 5-Year Estimates, aggregated by county, http://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml

Label: High / Low Poverty



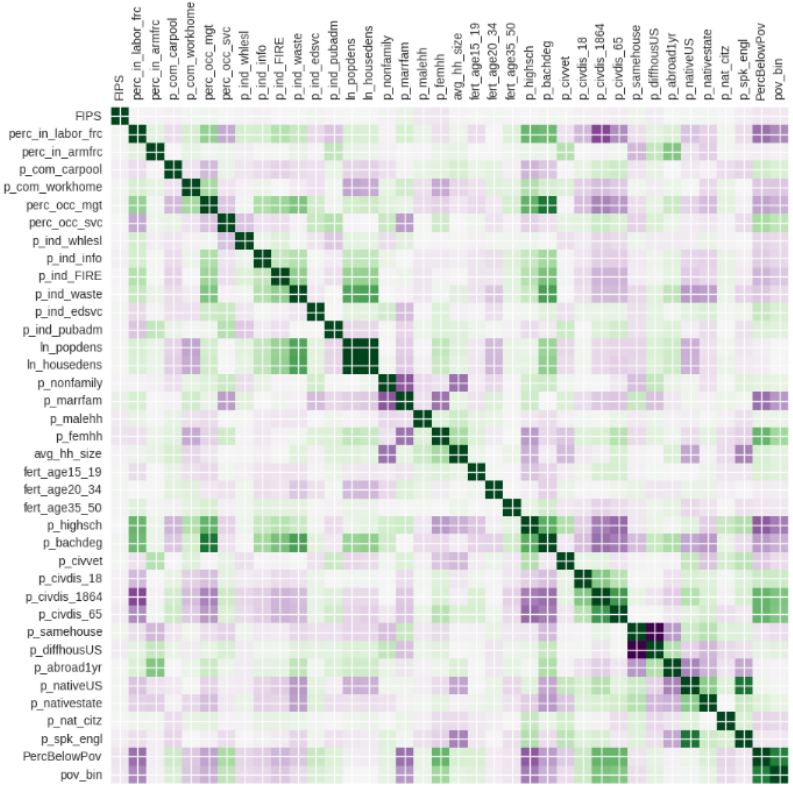
Feature Categories

- Household configuration
- Commute type
- Occupations, Industries
- Workforce characteristics
- Rural /Urban (by population density)
- Population transcience
- Citizenship, language, fertility
- Educational attainment

Methods

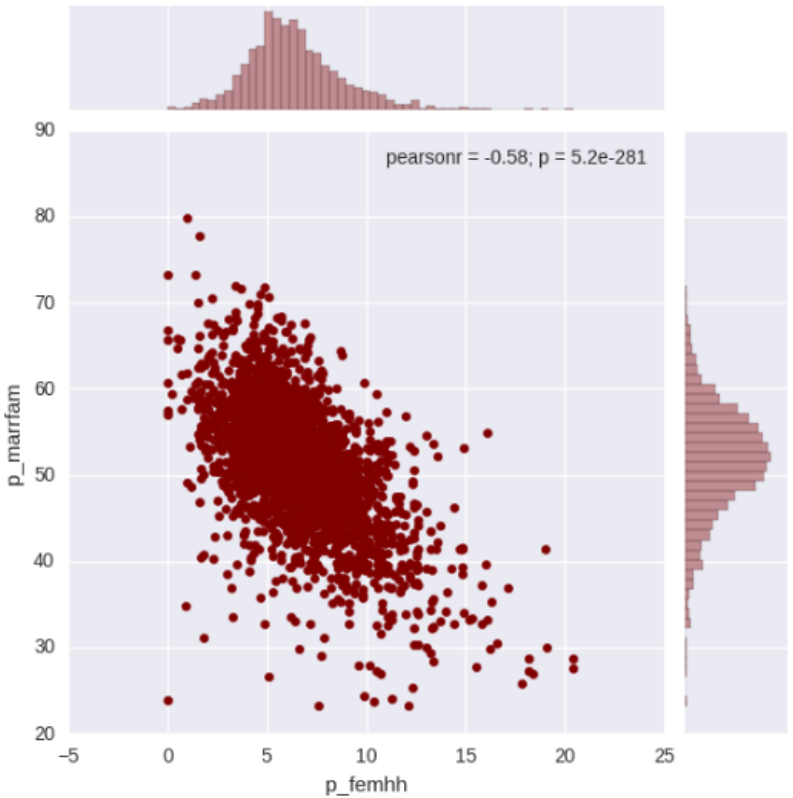
- Exploratory Feature Analysis
- Models
  - Logistic Regression (also w/ regularization)
  - Naive Bayes Classification
  - Random Forest Classification

Exploratory Feature Analysis



Heat map –
Feature / Label correlations

Exploratory Data Analysis



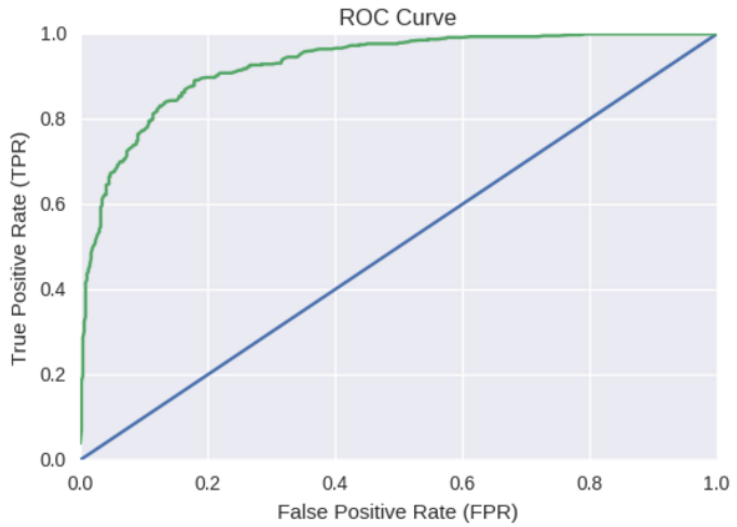
Joint plot –

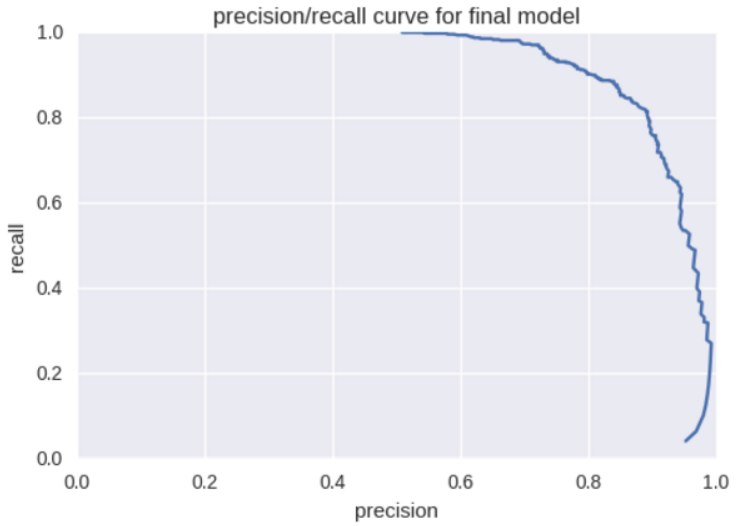
Tool to explore joint distributions

Model Development Strategy

- Initial model - .80 to .85 accuracy score
- Reduced feature set via:
  - Logistic classification p-values & regularization
  - Random trees feature importance
- Estimated logistic coefficients

Random Forest Results



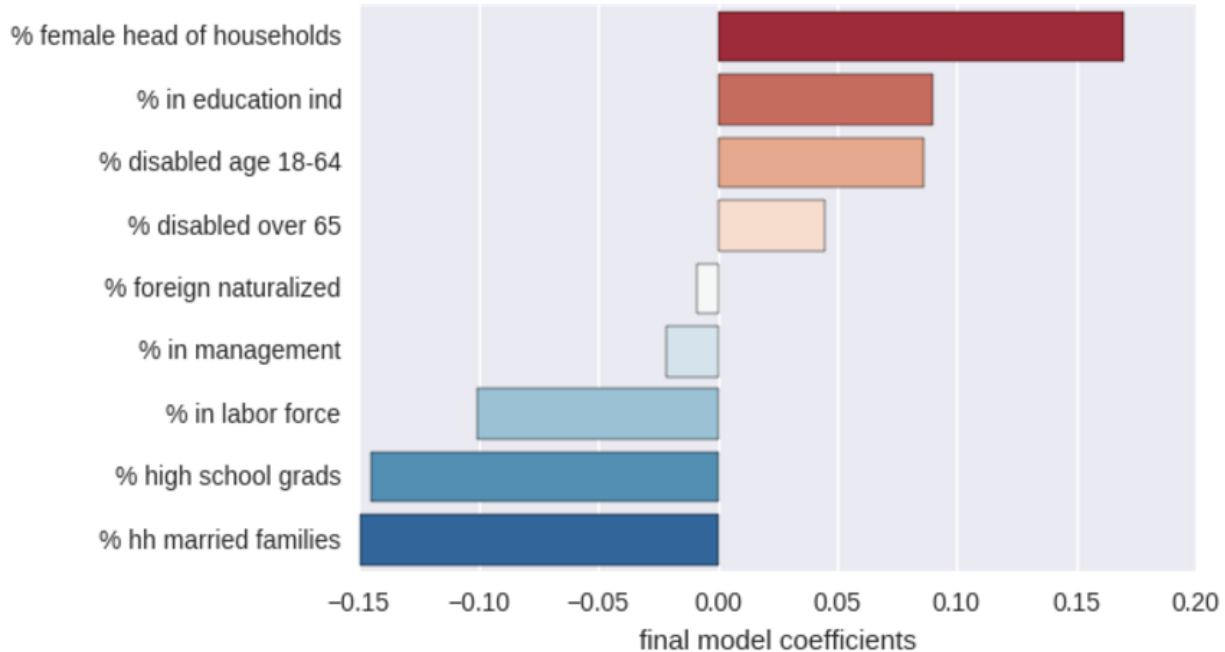


Random Forest Classifier (sklearn)

Accuracy: 0.855

AUC: 0.927

Logit Coefficient Results



Logit (statsmodels) from Random Forest feature importances
- Coefficients show change in odds of above-median poverty

Conclusions

- From a few key factors, we can predict high / low poverty rates.

- Future work:
  - Check trends for consistency over time
  - Check if same factor universe can predict other measures of well-being

Questions???

Kimberley Mitchell
kim@who-knows.com