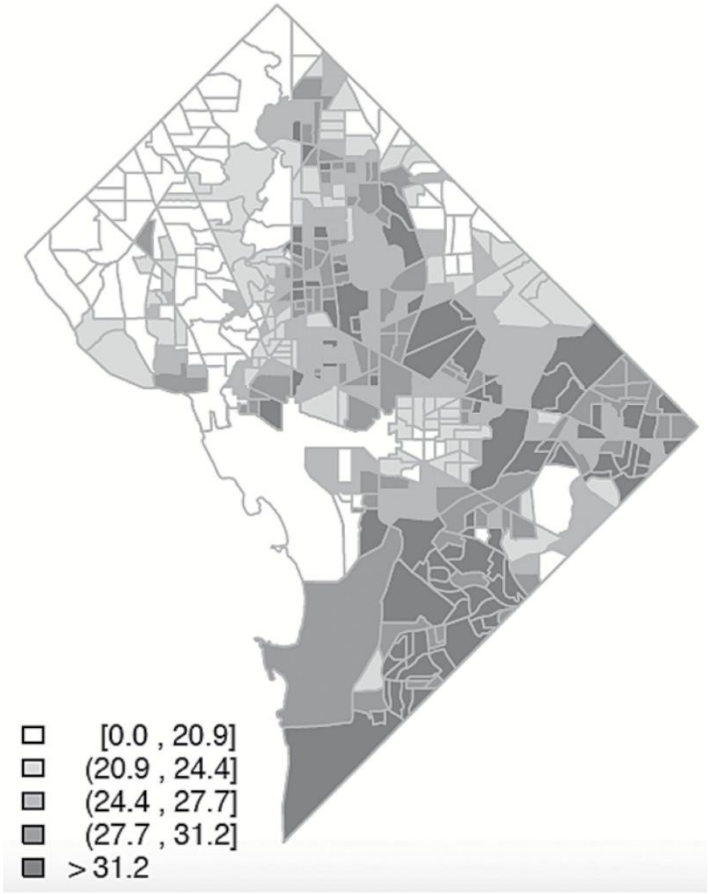
Constructing an Address-level Low Response Score for Address Based Sampling frames

American Association for Public Opinion Research Annual Conference
New Orleans, LA, May 19th, 2017

Joe McMichael & Joe Murphy

Inspired by the Census Low Response Score (LRS)

- Predict mail return rates for the Decennial Census
  - Understand self-respondent behavior
- Kaggle competition
  - Census Return Rate Challenge
  - Sought to improve earlier methods
  - Top models employed machine learning ensemble methods
- Low Response Score
  - Model using top rank ordered predictors from competition



Low Response Score (LRS)

Census Planning Database

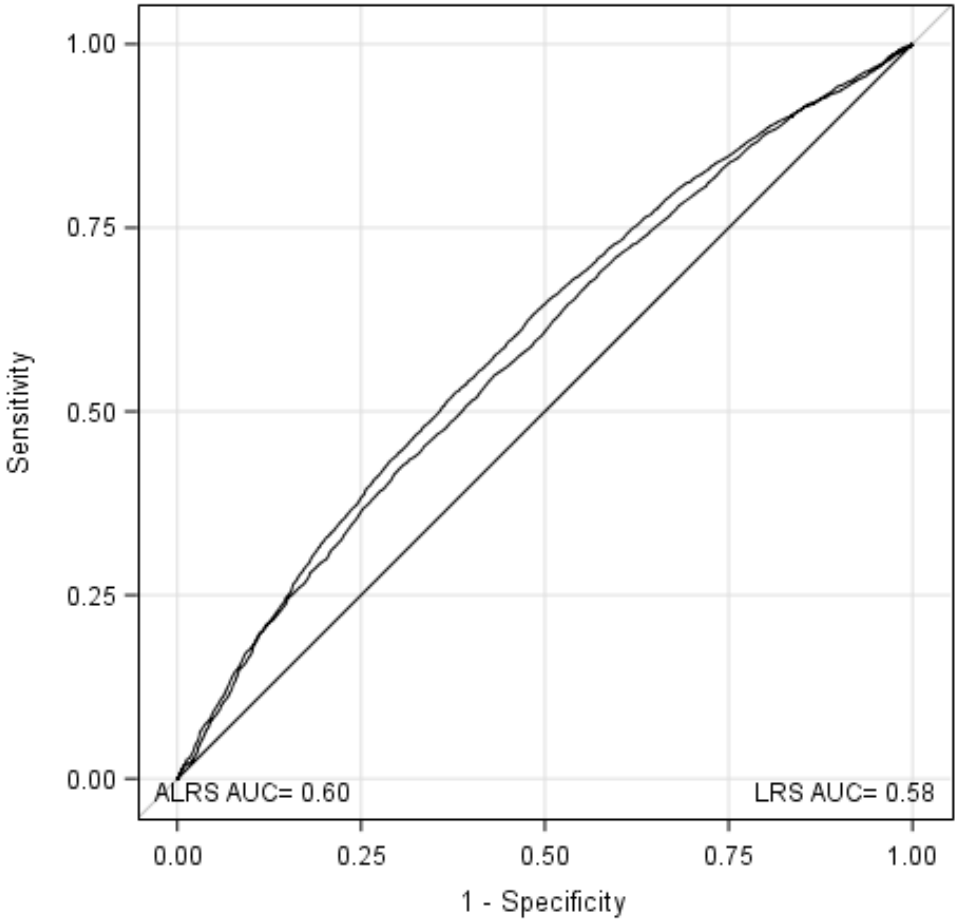
- Census block group level
- Census tract level

Image from: The Low Response Score (LRS)A Metric to Locate, Predict, and Manage Hard-to-Survey Populations
Public Opin Q. 2016;81(1):144-156. doi:10.1093/poq/nfw040

Address-level Low Response Score

Response Propensity for Mail Surveys

ROC Curve - Can we build a better model? – Not yet.



Area Under the Curve (AUC)

- Census LRS = 0.58
- ALRS = 0.60

Methods

Enhanced ABS Frame – Primary Data Sources

- CDS (ABS foundation)
- Geocode addresses
- Area-level demographics
  - Census PDB
  - Decennial Census
  - ACS
- Address & person-level
  - Acxiom InfoBase
  - Many sources (black-box)
  - Completeness varies
  - Accuracy varies

- City-style, PO Box, etc
- Vacancy status
- Single vs. Multi-family
- Census block group demo
- Child age group
- Adult name
- Adult age
- Adult race
- Income
- Education
- Subscribe to Cat Fancy (joking)
- many others

Survey Data

Survey A (used for building model)

- Mail survey
- Single state (a large one)
- Sample - 73,000
- Respondents - 26,500
- 36% return rate

Survey B

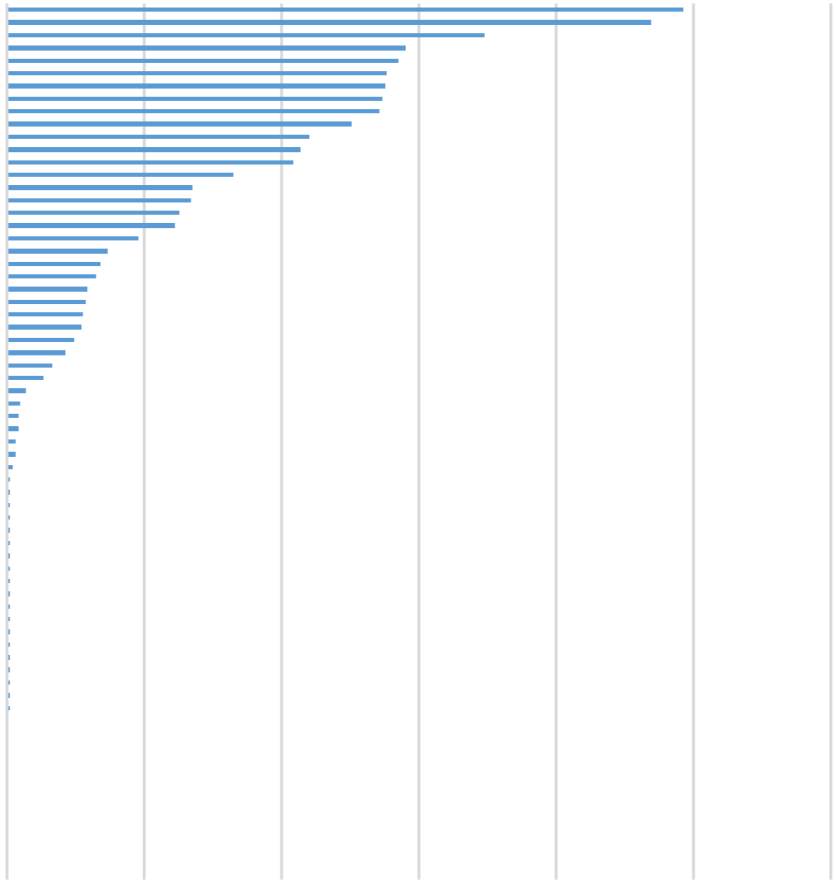
- Multi-mode national survey
- Consider mail screener only
- Sample – 18,000
- Respondents - 8,000
- 44% return rate

Model building

- Using Survey A

- Random forest for variable selection
  - Python 3, SciKit, RandomForestClassifier
  - 205 predictors

Variable Importance



Highest Ranked Variables

- LRS (PDB)
- DOB age 65+ (Acxiom)

- High rise (CDS)
- Vacant (CDS)

- % white alone (Decennial Census)
- PersonicX generation categories (Acxiom)
- % College Grad (ACS)
- has a surname (Acxiom)
- has DOB (Acxiom)
- Hispanic Surname (Acxiom)

- has child (Acxiom)
- DOB 40-49 (Acxiom)
- DOB 50-59 (Acxiom)
- Black Surname (Acxiom)
- % Other Language (ACS)

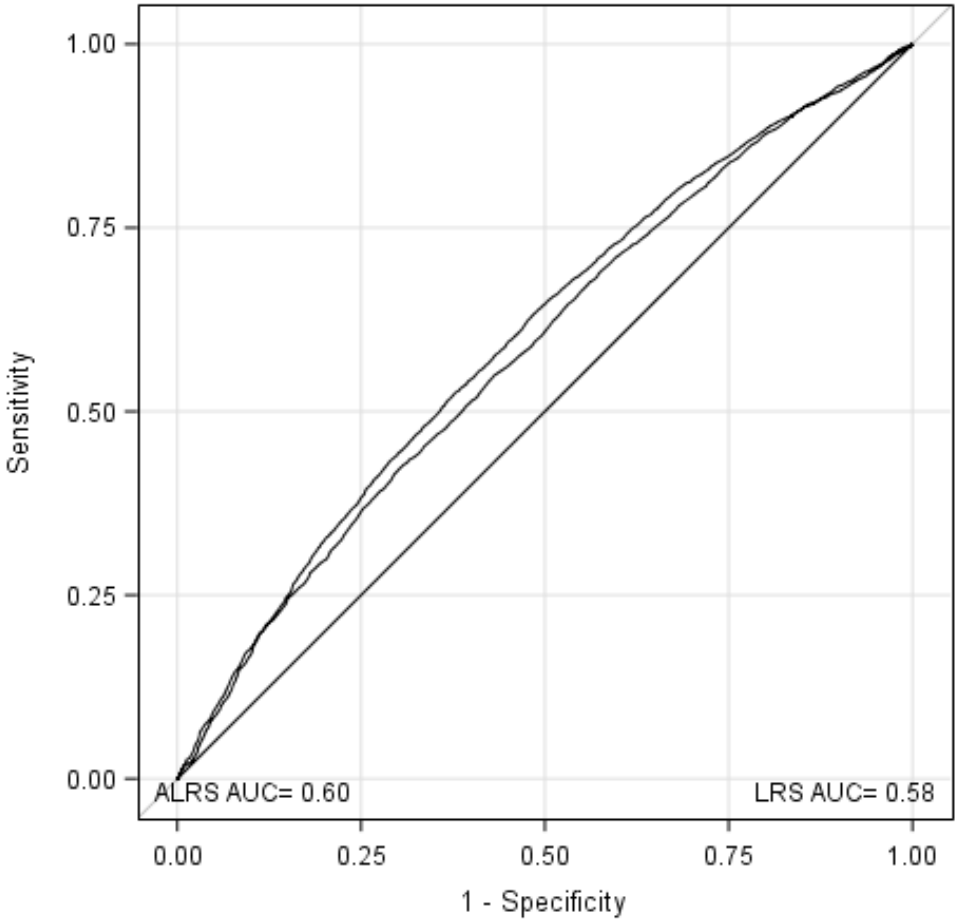
Model building

- Using Survey A
- Random forest for variable selection
  - Python 3, SciKit, RandomForestClassifier
  - 205 predictors
- Logistic Regression
  - Top 90 from random forest
  - Split into training and validation data
  - Stepwise variable selection
  - ASE of model with validation data
  - SAS HPLOGISTIC

Evaluation

- Applied the final model to Survey B
- Compared ROC and AUC

ROC Curve - Can we build a better model? – Not yet.



Area Under the Curve (AUC)

- Census LRS = 0.58
- ALRS = 0.60

Summary

Census LRS performs almost as well ALRS

- Result surprising given address-level demographic models are doing well (references below)
- Findings should be considered preliminary

Next Steps:

- Feature Engineering
- Explore other ensemble methods
- Add additional data sources

McMichael, J. P., & Ridenhour, J.L. (2017). **Improving Demographic Information for Address Based Sampling (ABS) frames.** Presented at the American Association for Public Opinion Research (AAPOR) annual conference, New Orleans, LA.

Ridenhour, J. L., & McMichael, J. P. (2017). **Propensity Stratification with Auxiliary Data for Address-Based Sampling Frames.** Presented at the American Association for Public Opinion Research (AAPOR) annual conference, New Orleans, LA.

More Information

Joe McMichael
Research Statistician
919.485.5519
mcmichael@rti.org