Using the Household Pulse Survey to Identify Potential External Benchmarks of Economic, Social, and Health Well-Being

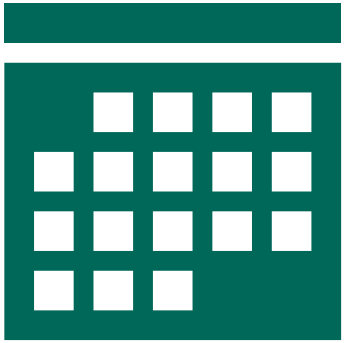
Session A-1: Leveraging Probability Online Survey Panels for Federally Sponsored Statistical Data Collection
FCSM 2023 Research and Policy Conference
10/24/2023

Priyam Patel, MSPH, Lauren M. Rossen, Ph.D., M.S., Katherine Irimata, Ph.D., Morgan Earp, Ph.D.
Division of Research and Methodology,
National Center for Health Statistics,
Centers for Disease Control and Prevention

Background

- The U.S. Census Bureau created the Household Pulse Survey (HPS) to produce data on the social and economic effects of coronavirus on American households
  - Broad range of topics
  - Questions are developed in collaboration with multiple federal agencies
- Does HPS collect social and economic well-being indicators that trend with health and are also predictive of the prevalence of future health outcomes?

Scope



- Data collection began in April 2020
  - Two weeks on, two weeks off
  - Analysis focused on weeks 46-61 (June 2022 – September 2023)
- Twelve collection cycle phases, thirteenth is ongoing
  - Three major phases, ten subphases
  - Changes in content and administration between phases
  - Core demographic household characteristics and additional topics fielded in partnership with other federal agencies

Data Analysis Steps

Evaluate predictors for missingness and relevance

Assess predictive importance through variable selection methods

Train models and compare metrics

Methods

- Explored variable selection methods to identify important predictors of selected health and health-related outcomes
- Trained conditional random forest models from the most analytically informative variables to predict selected outcomes
- Compared model performance to evaluate the impact of predictors and assess the generalizability of models

Outcomes of Interest

- **ANXIOUS**
  - Over the last 2 weeks, how often have you been bothered by feeling nervous, anxious, or on edge?
- **DOWN**
  - Over the last 2 weeks, how often have you been bothered by feeling down, depressed, or hopeless?
- **INTEREST**
  - Over the last 2 weeks, how often have you been bothered by having little interest or pleasure in doing things?
- **WORRY**
  - Over the last 2 weeks, how often have you been bothered by the not being able to stop or control worrying?
- **INSURANCE**
  - Are you currently covered by any of the following types of health insurance or health coverage plans?
- **LONGCOVID**
  - Did you have any symptoms lasting 3 months or longer that you did not have prior to having coronavirus or COVID-19?

Variable Selection Methods

- Ridge regression (L2 regularization)
  - Coefficients are shrunk toward zero according to a penalty term, with less important features being nearly eliminated
  - Features cannot be completely eliminated
- XGBoost feature importance
  - Gain metric represents the contribution of a feature to the model
  - Calculated as the improvement in the Gini impurity
- Random forest feature importance (impurity-corrected)
  - Similar to measuring improvement in Gini impurity, but adjusts for potential bias from high cardinality features

Variable Selection: Predictors of Selected Outcomes

Variables	Anxiety	Feeling Down	Interest	Insurance	Long COVID	Worry
Access to infant formula	x	x	x	x	x	x
Childcare arrangements and cost	x	x	x		x	x
COVID-19 vaccinations and long COVID symptoms and impact						
Demographics		x	x	x	x	
Education, specifically K-12 enrollment						
Employment		x	x	x	x	
Food sufficiency			x			
Housing security	x	x			x	x
Household spending, including energy expenditures and consumption	x		x	x		x
Inflation concerns and changes in behavior due to increasing prices	x	x	x			x
Physical and mental health		x	x		x	x
Rental assistance from state and local governments				x		
Transportation, including behavioral changes related to the cost of gas						

Model Results

Anxiety

Performance Metric	Ridge	XGBoost	Random Forest
Accuracy	0.70	0.73	0.72
F1	0.75	0.77	0.76
Balanced Accuracy	0.69	0.72	0.71
Precision	0.73	0.76	0.75
Sensitivity/Recall	0.77	0.79	0.78
Specificity	0.62	0.65	0.65

Down

Performance Metric	Ridge	XGBoost	Random Forest
Accuracy	0.72	0.73	0.73
F1	0.64	0.67	0.66
Balanced Accuracy	0.70	0.72	0.71
Precision	0.71	0.71	0.71
Sensitivity/Recall	0.58	0.63	0.62
Specificity	0.82	0.80	0.80

Worry

Performance Metric	Ridge	XGBoost	Random Forest
Accuracy	0.71	0.73	0.73
F1	0.67	0.72	0.71
Balanced Accuracy	0.71	0.73	0.72
Precision	0.73	0.73	0.73
Sensitivity/Recall	0.62	0.70	0.69
Specificity	0.79	0.77	0.76

Interest

Performance Metric	Ridge	XGBoost	Random Forest
Accuracy	0.72	0.73	0.73
F1	0.67	0.68	0.68
Balanced Accuracy	0.71	0.72	0.72
Precision	0.71	0.72	0.71
Sensitivity/Recall	0.63	0.65	0.64
Specificity	0.79	0.79	0.79

Model Results

Long COVID

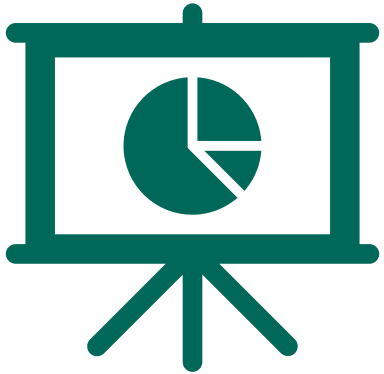
Performance Metric	Ridge	XGBoost	Random Forest
Accuracy	0.85	0.85	0.85
F1	0.64	0.66	0.65
Balanced Accuracy	0.74	0.75	0.74
Precision	0.93	0.89	0.86
Sensitivity/Recall	0.49	0.52	0.52
Specificity	0.99	0.98	0.97

Insurance

Performance Metric	Ridge	XGBoost	Random Forest
Accuracy	0.96	0.96	0.96
F1	0.98	0.98	0.98
Balanced Accuracy	0.50	0.52	0.51
Precision	0.96	0.96	0.96
Sensitivity/Recall	0.99	0.99	0.99
Specificity	0.00	0.03	0.03

Conclusions

- Identified topics of interest that can be used to inform the selection of data sources and potential predictors for use in small area and temporal models
- Evaluated the validity of predictive indicators by developing models and assessing model performance
  - High performing social and economic well-being indicators can be used to support forecasting, explanatory modeling, small area estimation, and other types of analyses aiming to improve the accuracy and timeliness of the estimates NCHS produces



Acknowledgements

- Bailey Sutton, Health Resources and Services Administration
- Jonaki Bose, NCHS Division of Health Interview Statistics
- Jim Dahlhamer, NCHS Division of Health Interview Statistics
- Trent Buskirk, Bowling Green State University
- Paul Scanlon, NCHS Division of Research and Methodology

Thank you!

Priyam Patel qkn6@cdc.gov

For more information, contact CDC
1-800-CDC-INFO (232-4636)
TTY: 1-888-232-6348 www.cdc.gov