A DEEP DIVE INTO THE DIGITAL DIVIDE

DAVID DUTWIN, SSRS | TRENT D. BUSKIRK, UMASS - BOSTON

NISS/WSS WORKSHOP ON INFERENCE FROM NONPROBABILITY SAMPLES

September 25, 2017

DDutwin@ssrs.com | 484-840-4406 | @ddutwin
Trent.Buskirk@umb.edu | 781-964-4997 | @trentbuskirk

"THEY COULDN'T BE ANYMORE DIFFERENT"

- Just what is it exactly that "prevents" people from having the Internet?
- Most Internet panels do not offer solutions to cover non-Internet panelists…can we model them in?
- A range of articles have reported single-survey documentation of the digital divide…what about across "all" surveys?

THE PROCESS

STEP 1: FIND DATASETS

Survey	Year	N	Total Number of Variables
American Identity & Representation Survey	2012	1,702	~130
American National Election Survey	2012	5,914	~1,100
BRFSS	2015	434,382	~300
General Social Survey	2014	1,238	~840
National Health Interview Survey	2015	33,672	~1,300
Outlook on Life Survey	2012	2,294	~400
Survey of Consumer Attitudes	2013	2,013	~270
Survey on Public Participation in the Arts	2012	4,708	~770
Pew Science Survey	2014	2,002	~80
Pew Libraries and Technology Survey	2015	1,003	~130
Pew Public Survey	2014	1,501	~110
Pew Civic Engagement Survey	2012	2,251	~100
Pew Gaming Survey	2015	2,001	~120
Pew Gender Survey	2014	1,835	~80
Pew Connectivity Survey	2013	1,801	~130

STEP 2: FLAG SIGNIFICANT VARIABLES

Survey	Year	N	Total Number of Variables	Step 1 Sig. Vars.
American Identity & Representation Survey	2012	1,702	~130	15
American National Election Survey	2012	5,914	~1,100	105
BRFSS	2015	434,382	~300	12
General Social Survey	2014	1,238	~840	100
National Health Interview Survey	2015	33,672	~1,300	70
Outlook on Life Survey	2012	2,294	~400	42
Survey of Consumer Attitudes	2013	2,013	~270	12
Survey on Public Participation in the Arts	2012	4,708	~770	38
Pew Science Survey	2014	2,002	~80	28
Pew Libraries and Technology Survey	2015	1,003	~130	10
Pew Public Survey	2014	1,501	~110	11
Pew Civic Engagement Survey	2012	2,251	~100	14
Pew Gaming Survey	2015	2,001	~120	9
Pew Gender Survey	2014	1,835	~80	6
Pew Connectivity Survey	2013	1,801	~130	21

Rules:

1. At least 10% Difference

2. Lower Bound > 10%

3. Non-Demographic

542 Variables Found

DIMENSIONS OF THE DIGITAL DIVIDE

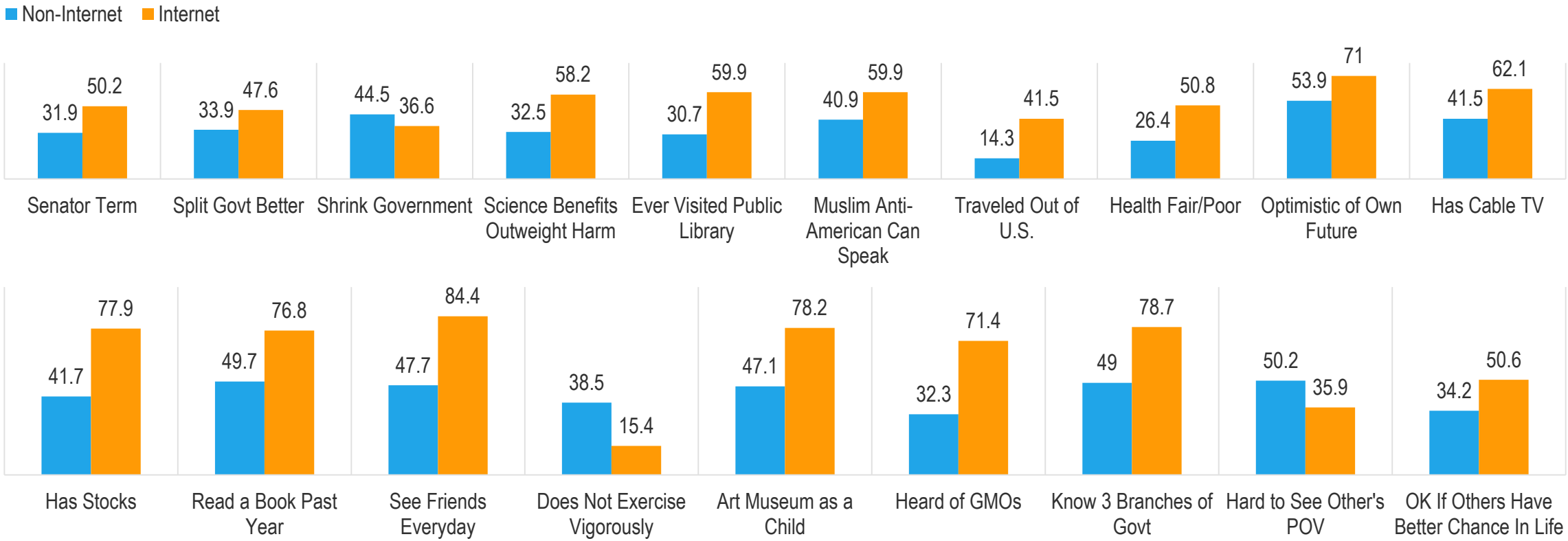
Technology Ownership	6
Technology: Attitudes Toward	15
Activities: Community/Membership/Cultural	52
Americanism	12
Abortion	7
Government Role/Setup	19
Government Financial	21
Candidate Qualities	10
Self/Family/HH Financial	38
Political Knowledge	19
Knowledge (non-Political)	23

Health Care	21
Self-Health Status Physical	60
Self-Health Status Mental	14
Religion	7
Science	34
Privacy/Openness	20
Tolerance of other groups	62
Trust/Efficacy	9
Political Participation	42
Environment/Energy	11

- Variables coded for topic
- Krippendorf's alpha = .76
- Of course, number of variables in each content category confounded with topics of the source surveys
- Still a number of dimensions seem apparent:
  - Fear of technology
  - Isolationism/Fear of other groups
  - Conservativism
  - Age
  - SES
  - American-centric
  - Low political knowledge
  - Religiosity/Distrust of science

A FEW DIFFERENCES OF NOTE

INTERNET/NON-INTERNET HOUSEHOLDS



DATA REDUCTION STEP 2

- As a second step of data reduction we applied a recursive feature elimination step using fuzzy random forests to select the top 20% of the Final Subset of Variables remaining after step 1.
- The forests used 2500 trees and the default number of variables randomly selected at each node for tree branching.
- The importance of each variable was based on the mean decrease in the accuracy of predicting non-internet households.

STEP 3: DATA REDUCTION ROUND 2

Survey	Year	N	Total Number of Variables	Step 1 Sig. Vars.	Max Final Vars. Step 2	Final Non-Demo
American Identity & Representation Survey	2012	1,702	~130	15	5	1
American National Election Survey	2012	5,914	~1,100	105	11	4
BRFSS	2015	434,382	~300	12	5	2
General Social Survey	2014	1,238	~840	100	21	4
National Health Interview Survey	2015	33,672	~1,300	70	9	3
Outlook on Life Survey	2012	2,294	~400	42	7	2
Survey of Consumer Attitudes	2013	2,013	~270	12	5	1
Survey on Public Participation in the Arts	2012	4,708	~770	38	18	2
Pew Science Survey	2014	2,002	~80	28	6	1
Pew Libraries and Technology Survey	2015	1,003	~130	10	5	1
Pew Public Survey	2014	1,501	~110	11	5	1
Pew Civic Engagement Survey	2012	2,251	~100	14	5	1
Pew Gaming Survey	2015	2,001	~120	9	4	1
Pew Gender Survey	2014	1,835	~80	6	4	1
Pew Connectivity Survey	2013	1,801	~130	21	6	1

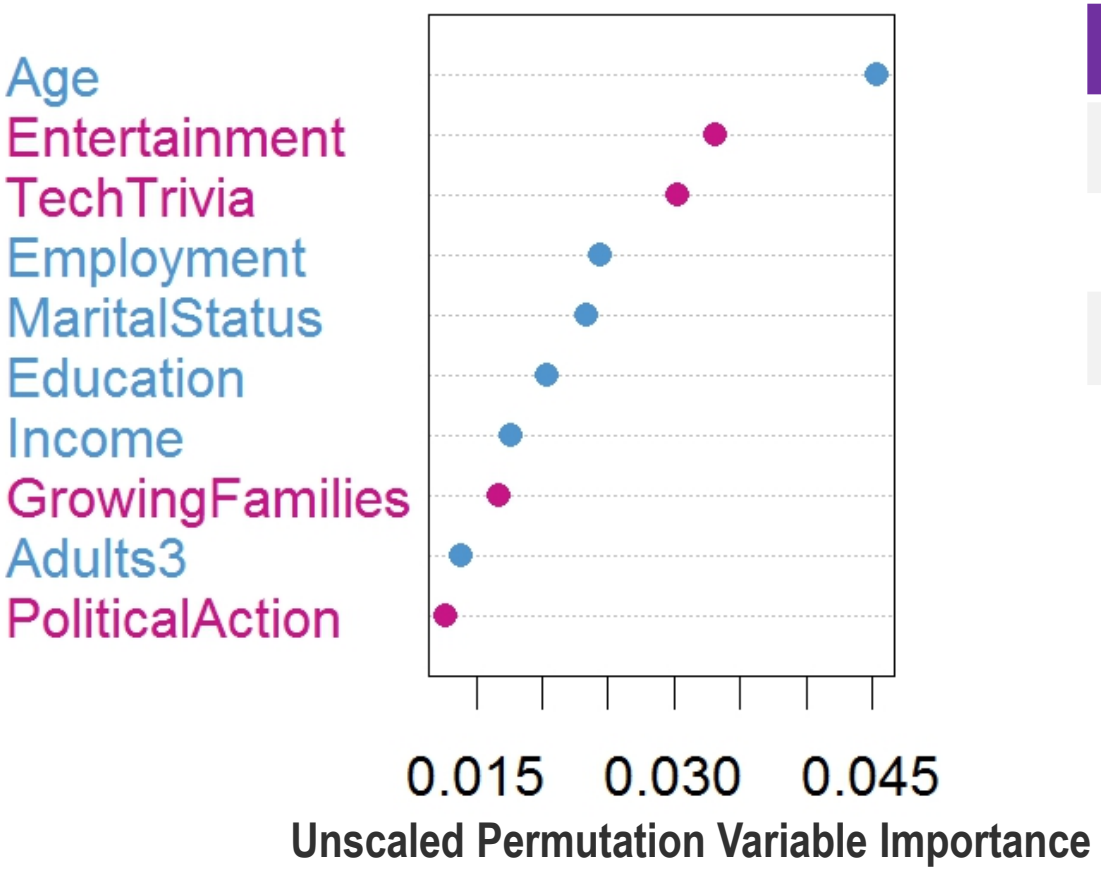
COMBINING VARIABLES FROM SEPARATE DATASETS

- 26 Candidate variables concurrently administered in the SSRS Omnibus in April, 2017
- Oversample waves interviewed only those without Internet utilization
- Baseline Internet measure plus smartphone/tablet follow-up
- N = 1,373 with Internet, 1,058 without
- Dataset thus “balanced” with respect to outcome of interest to facilitate predictive models and variable selection.
- Wanted to identify half dozen “non-demographic” variables that were important for predicting non-internet households while including standard demographic variables.

USING ALL 38 VARIABLES TAKEN TOGETHER AS ONE MODULE, WE APPLIED:

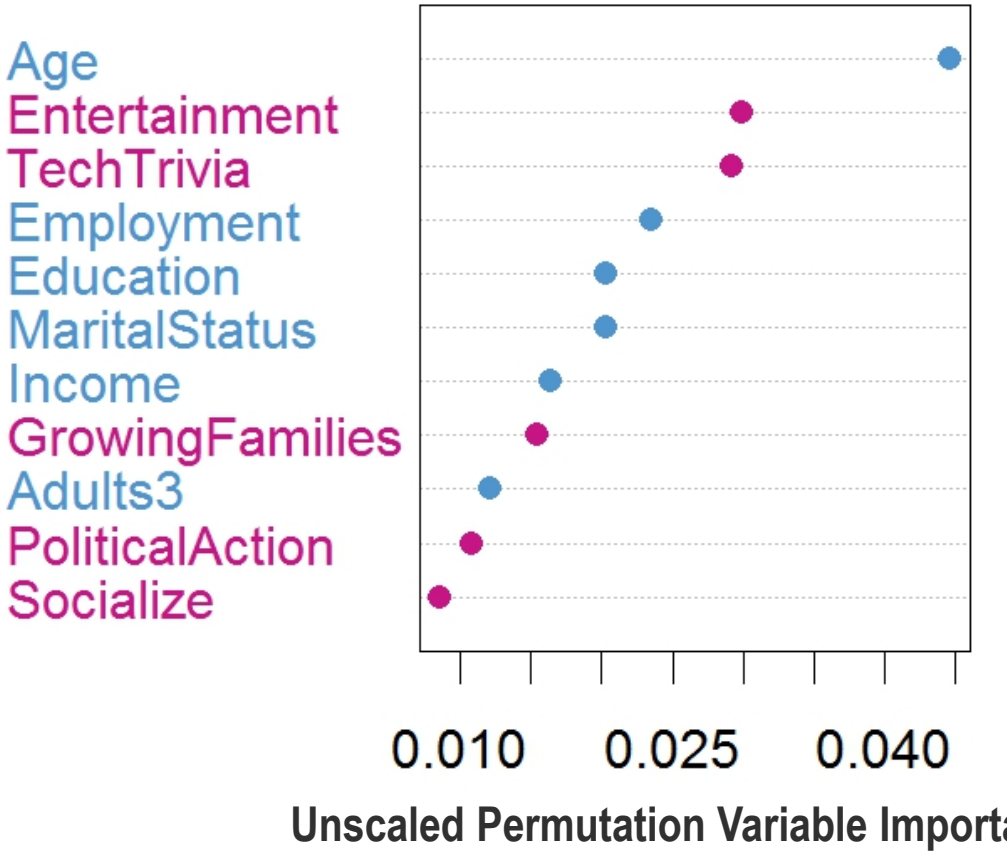
- A recursive feature elimination procedure based on random forest models to identify the top 10, 11 and 12 predictors.
- The number of predictors varied as we wanted to identify the top 5 or 6 “non-demographic predictors” to use in our propensity models for predicting NON-INTERNET Households
- The recursive feature elimination was performed using Fuzzy Forests which iterate a series of classification-based random forests to identify the top-most important features.
- The estimates of variable importance for this method are less biased in the presence of correlated predictors when compared to using a single random forest model (see Conn et al., 2015)
- The tuning parameters for the fuzzy forest method were determined using a grid search that was based on testing overall model misclassification error across a series of 4 levels of forest size and three levels of the “mtry” parameter that controls the number of variables considered for splitting each node in the trees. For tractability the nodesize was set to 5 for all forests.
- A total of 10 independent iterations were conducted per combination of forest size, mtry and number of variables to be selected.
- The combination of forest size and mtry that produced the smallest error, per number selected, were used to create the final models to determine the best variables for predicting non-internet status.

SELECTION OF THE 10 MOST IMPORTANT PREDICTORS



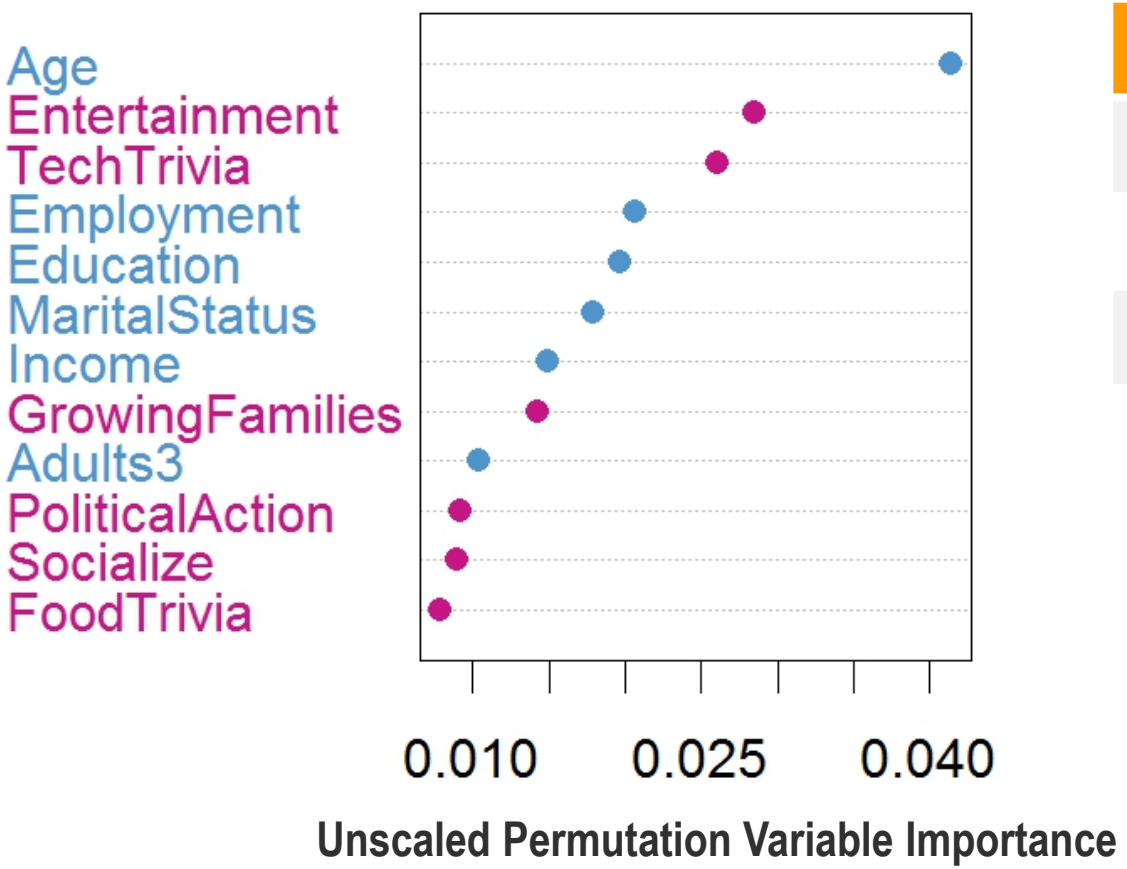
Statistic	Estimate
Overall Accuracy	79.6%
True Positive Rate	79.8%
True Negative Rate	79.5%

SELECTION OF THE 11 MOST IMPORTANT PREDICTORS



Statistic	Estimate
Overall Accuracy	80.1%
True Positive Rate	80.5%
True Negative Rate	79.8%

SELECTION OF THE 12 MOST IMPORTANT PREDICTORS



Statistic	Estimate
Overall Accuracy	80.2%
True Positive Rate	80.9%
True Negative Rate	79.6%

TOP 6 NON-DEMS

If we just used the top 6 Non-Dems to Predict Non-Internet Households…

based on a random forest model with 1000 tree and mtry=2

- Entertainment
- Tech Trivia
- Growing Families
- Political Action
- Socialize
- Food Trivia

Statistic	Estimate
Overall Accuracy	75.4%
True Positive Rate	70.5%
True Negative Rate	79.2%

- Age
- Employment
- Education
- Marital Status
- Income
- Adults

Statistic	Estimate
Overall Accuracy	75.6%
True Positive Rate	74.0%
True Negative Rate	76.8%

LOOKING FORWARD TO A REGRESSION MODEL

Variable Ranking	Forest	Regression 6	Regression 7
Tech Ownership		3	1
Tech Trivia	2	1	5
Growing Families	3	2	4
Political Action	4		
Entertainment	1	4	2
Food Trivia	6		7
Socialize	5	5	3
Political Trivia		6	6

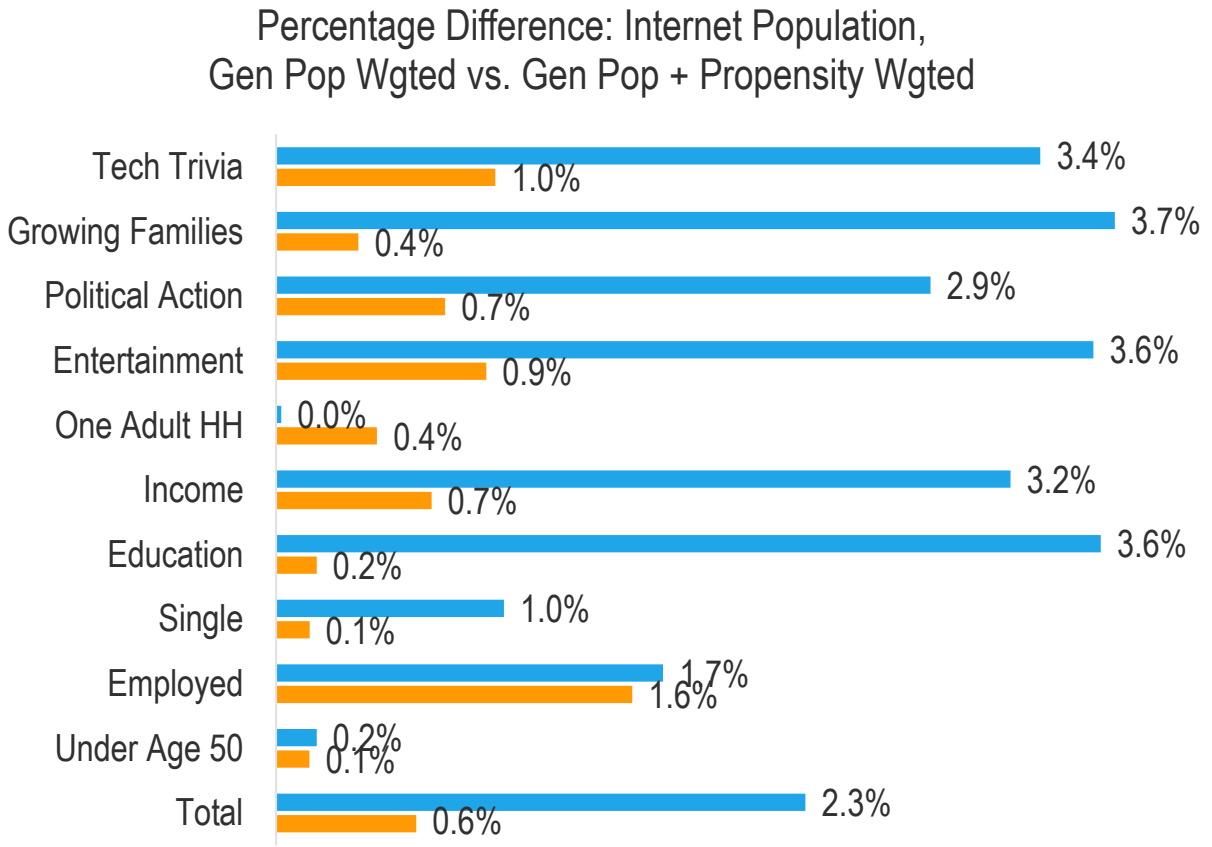
Cases Correctly Classified	80%
R^2, Demographics	.41
R^2, Full Model	.55
R^2, Final Model	.54

METHOD TO ASSESS THE EFFECTIVENESS OF THE MODEL

- Random Forest model allows using a master dataset for modelling and testing.
- Development of testing weights:
  1. Single full omnibus wave weighted using Buskirk and Best (2012) and raking by agexgender, ethnicityxforeignborn, education, phone status, genderxregion, marital status, and population density. Deff = 1.75. This establishes the non-internet estimate: 15.8%
  2. Full data to the same procedures as #1, with the addition of internet/non-internet. This provides "gold standard" estimates of test measures.
  3. Internet respondents only, weighted to the same procedure as #1. This provides impact of raking the internet-only population to full population parameters.
  4. Internet respondents only, weighted as #1 with non-Internet propensity weight as part of the baseweight. This compares to #3 above in a test of bias reduction.

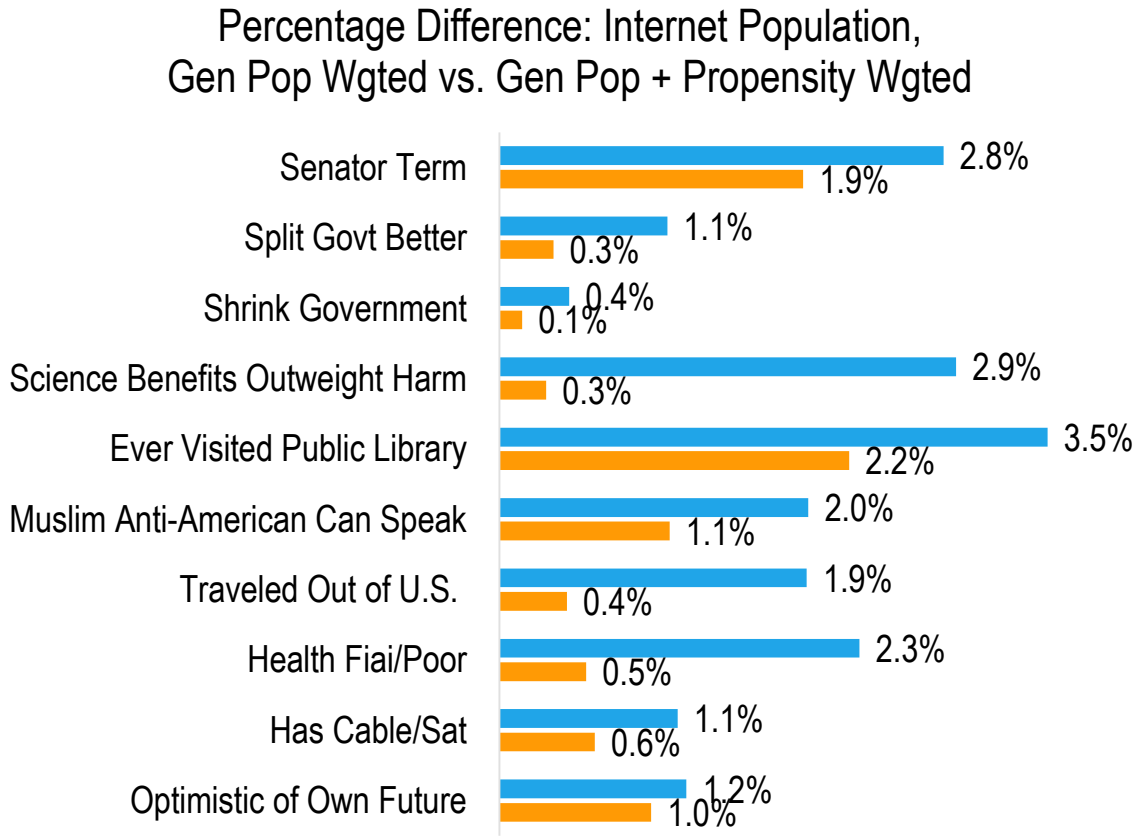
ASSESSING THE EFFECTIVENESS OF THE MODEL

Propensity Weight Variables	Full Pop	Int Pop	Int Pop w/ Propensity	Int Pop Diff	Int Pop Prop Diff	Int Pop % Diff	Int Pop % Diff
Tech Trivia	64%	68%	63%	5%	1%	3%	1%
Growing Families	79%	82%	79%	5%	0%	4%	0%
Political Action	32%	34%	32%	9%	2%	3%	1%
Entertainment	64%	67%	63%	6%	1%	4%	1%
One Adult HH	20%	20%	21%	0%	2%	0%	0%
Income	37%	34%	38%	9%	2%	3%	1%
Education	35%	31%	34%	10%	1%	4%	0%
Single	29%	30%	29%	3%	1%	1%	0%
Employed	59%	61%	57%	3%	3%	2%	2%
Under Age 50	54%	54%	54%	0%	0%	0%	0%
Total				5.0%	1.4%	2.3%	0.6%



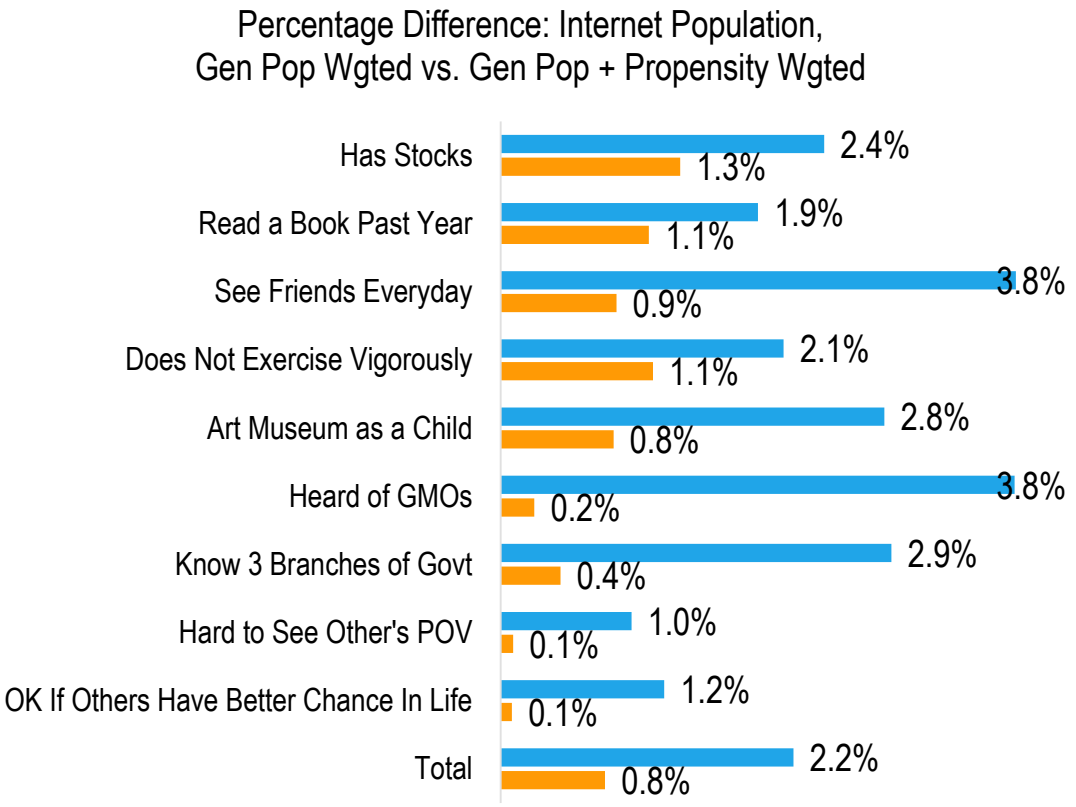
ASSESSING THE EFFECTIVENESS OF THE MODEL

Propensity Weight Variables	Full Pop	Int Pop	Int Pop w/ Propensity	Int Pop Diff	Int Pop Prop Diff	Int Pop % Diff	Int Pop % Diff
Senator Term	47%	50%	49%	6%	4%	3%	2%
Split Govt Better	45%	46%	46%	2%	1%	1%	0%
Shrink Government	43%	44%	43%	1%	0%	0%	0%
Science Benefits > Harm	55%	58%	55%	5%	1%	3%	0%
Ever Visited Public Library	35%	32%	33%	10%	6%	3%	2%
Muslim Anti-American Can Speak	57%	59%	58%	3%	2%	2%	1%
Traveled Out of U.S.	63%	61%	62%	3%	1%	2%	0%
Health Fair/Poor	22%	20%	22%	10%	2%	2%	1%
Has Cable/Sat	67%	68%	67%	2%	1%	1%	1%
Optimistic of Own Future	68%	69%	67%	2%	1%	1%	1%
Senator Term	47%	50%	49%	6%	4%	3%	2%



ASSESSING THE EFFECTIVENESS OF THE MODEL

Propensity Weight Variables	Full Pop	Int Pop	Int Pop w/ Propensity	Int Pop Diff	Int Pop Prop Diff	Int Pop % Diff	Int Pop % Diff
Has Stocks	53%	55%	51%	4.5%	2.5%	2.4%	1.3%
Read a Book Past Year	73%	74%	71%	2.6%	1.5%	1.9%	1.1%
See Friends Everyday	79%	82%	79%	4.8%	1.1%	3.8%	0.9%
Does Not Exercise Vigorously	19%	17%	18%	10.9%	5.9%	2.1%	1.1%
Art Museum as a Child	73%	76%	74%	3.8%	1.1%	2.8%	0.8%
Heard of GMOs	65%	69%	65%	5.8%	0.4%	3.8%	0.2%
Know 3 Branches of Govt	74%	77%	74%	3.9%	0.6%	2.9%	0.4%
Hard to See Other's POV	38%	37%	38%	2.5%	0.2%	1.0%	0.1%
OK If Others Have Better Chance In Life	31%	30%	31%	3.9%	0.3%	1.2%	0.1%
Total				4.6%	1.7%	2.2%	0.8%
Has Stocks	53%	55%	51%	4.5%	2.5%	2.4%	1.3%



CONCLUSIONS

CONCLUSIONS

- Age and economic status go a long way in explaining who does not have the Internet. But the story is much richer than that…
- Non-Internet households are at some level, ISOLATED households, be it socially, media-connectedness, etc. They are isolated behaviorally, attitudinally, and in worldly knowledge.
- Many, many variables have strong relationships to non-Internet use. Making a model both fully specified and parsimonious means choosing among a range of variables that all serve well in discriminating by Internet use.
- "Signal strength" for predicting non-internet households is not limited to one content category specifically; meaningful variables from across a wide array of content areas were identified.

CONCLUSIONS

- Variables tested on average had point estimates where the non-internet population was 74% off (benchmark / test estimate) from the estimates of the internet population.
- Just taking the internet population, weighted to the full population, bias is reduced to 21%
- Standard raking the internet population to full population targets reduces this bias to 5%.
- Adding the propensity weight reduces this to 1.4%
- The propensity weight does not add significant bias at a 1.17 propensity design effect.

CONTACT US

DAVID DUTWIN

DDUTWIN@SSRS.COM

484-840-4406

@DDUTWIN

TRENT BUSKIRK

TRENT.BUSKIRK@UMB.EDU

781-964-4997

@TRENTBUSKIRK