Increasing survey response rates and decreasing costs by combining numeric and text mining strategies on survey paradata

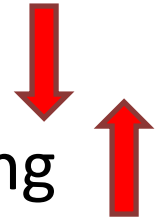
Sudip Bhattacharjee (Presenter)
Senior Research Fellow, US Census Bureau;
Professor, University of Connecticut

Nevada Basdeo, US Census Bureau
Ugochukwu Etudo, University of Connecticut
Joseph Kang, US Census Bureau
John Charles Langer, US Census Bureau
Saman A. Senadeera, US Census Bureau

6th WORKSHOP: ADVANCES IN ADAPTIVE AND RESPONSIVE SURVEY DESIGN: FROM THEORY TO PRACTICE

All views expressed are those of the authors and not necessarily those of the U.S. Census Bureau. All results have been reviewed to ensure no confidential data have been disclosed.

Research Problem from American Community Survey (ACS) Operations

- Multi-dimensional problem
  - Response rates are declining
  - Data collection costs are rising 
  - Respondent burden cannot increase beyond a threshold when field representative contacts respondents who have not completed the survey
- Leads to a multi-objective optimization problem with conflicting objectives
  - We present first steps to solve this problem

Research question and solution approach

- **Predict response propensity, decrease cost, and not increase respondent burden -- using both numeric and textual information (structured and unstructured)**
  - CHI: Structured
  - Case Notes: free form text, unstructured
  - Combine and use CHI (Contact History Information) and Case Notes information
- CHI based response propensity modeling for ACS (new)
- Case Notes based response propensity model (new)

Data and merging CHI with Case Notes

- 2017 ACS CHI and Case Notes (focus of current analysis)

- Each CHI record was merged with zero-many Case Notes associated with that contact attempt
- Challenge: CHI and Case Notes captured on different systems
  - Merged on control number, date and timestamp
  - Timestamp does not match
  - Manually verified large (>400) samples to identify pattern of linkage
  - Custom linkage algorithm based on control number, date and proximity of timestamp between CHI and Case Notes

Distribution of CHI codes in 2017 ACS

Group	Outcome Code	Percent	% HH with Code
1 - new/salvaged	200	1.00%	3.00%
1 - new/salvaged	583	0.02%	0.02%
2 - completed	201	11.51%	45.20%
3 - incomplete	203	0.49%	1.92%
3 - incomplete	206	5.07%	19.46%
3 - incomplete	202	54.34%	70.89%
3 - incomplete	204	0.90%	2.49%
3 - incomplete	208	1.84%	3.77%
4 - Type A non-interview	213	0.20%	0.73%
4 - Type A non-interview	214	0.08%	0.30%
4 - Type A non-interview	216	1.10%	3.62%
4 - Type A non-interview	217	0.08%	0.28%
4 - Type A non-interview	218	2.29%	7.28%
4 - Type A non-interview	219	2.19%	7.28%
5 - Type B/C non-interview	233	0.08%	0.30%
5 - Type B/C non-interview	229	0.06%	0.23%
5 - Type B/C non-interview	240	0.48%	1.76%

Group	Outcome Code	Percent	% HH with Code
5 - Type B/C non-interview	241	0.42%	1.54%
5 - Type B/C non-interview	243	0.25%	0.92%
5 - Type B/C non-interview	244	0.04%	0.15%
5 - Type B/C non-interview	245	0.04%	0.16%
5 - Type B/C non-interview	248	0.42%	1.51%
5 - Type B/C non-interview	253	0.28%	0.99%
5 - Type B/C non-interview	254	0.48%	1.73%
5 - Type B/C non-interview	255	0.07%	0.26%
5 - Type B/C non-interview	258	0.06%	0.22%
6 - Vacant/Temp. Occupied	301	6.23%	24.11%
6 - Vacant/Temp. Occupied	501	0.16%	0.61%
6 - Vacant/Temp. Occupied	305	0.81%	1.82%
7 - mailed/onine/given to fld rep	308	2.69%	4.57%
7 - mailed/onine/given to fld rep	309	2.93%	9.94%
7 - mailed/onine/given to fld rep	310	0.61%	1.25%
Other	313	0.82%	2.17%
Other	580	1.96%	6.51%

n=774,000

Predictive Model: Logistic Regression

- **Predict:** Final outcome – 201 (completed) vs 218 (refused)
- Prediction based on information from 1st contact only (personal or telephone)
- **Model 1:** Predictors: Case Notes
  - Field Rep. Case Notes (textual data)
- **Model 2:** Predictors: Admin and CHI variables
  - Census region, interview period (month), roster indicator (Administrative control vars.)
  - Partial/unable to conduct (H##), Language
  - Concern/behavior/reluctance (B##)
  - Noncontact telephone (P##), noncontact personal visit (V##)
  - Strategies attempted (Q##)
  - Includes interaction effects
- **Model 3:** Predictors: Admin and Case Notes
- **Model 4:** Predictors: Admin, CHI variables and Case Notes

Primer: Natural Language Processing (NLP)

- NLP is a method designed to give mathematical meaning to human language
- Extremely powerful
- NLP can be used to identify
  - Parts of speech (Noun, Verb)
  - Entities (Person, Place, Organization)
  - Tokens of related words (San and Francisco probably combined to San Francisco)
  - Probability of words appearing in a corpus of text
  - Probability of word associations
    - e.g. ‘England’ and ‘Queen’ should be more common than pairing ‘England’ and ‘Sunshine’

Cleaning

[common words, punctuation, etc.]

Create terms

- parts of speech
- Stemming
- others

Calculate term frequency (TF)

term

document

Predicting with textual data

Predictive models

- Logistic regression
- Random forest
- etc.

Weighted matrix (TF-IDF)

weighted term

document

normalized matrix

Weighted value of term (IDF)

[Varies inversely with number of documents the term appears in]

Term Frequency–Inverse Document Frequency (TF-IDF) primer

Term Document Matrix					
	term				
doc	cat	dog	grooming	best	total terms/doc
A	4	12	0	7	3
B	4	9	10	2	4
C	6	0	8	1	3

$$\text{IDF(t)} = \log\_e\left(\frac{Total\ number\ of\ documents}{Number\ of\ documents\ with\ term\ t\ in\ it}\right)$$

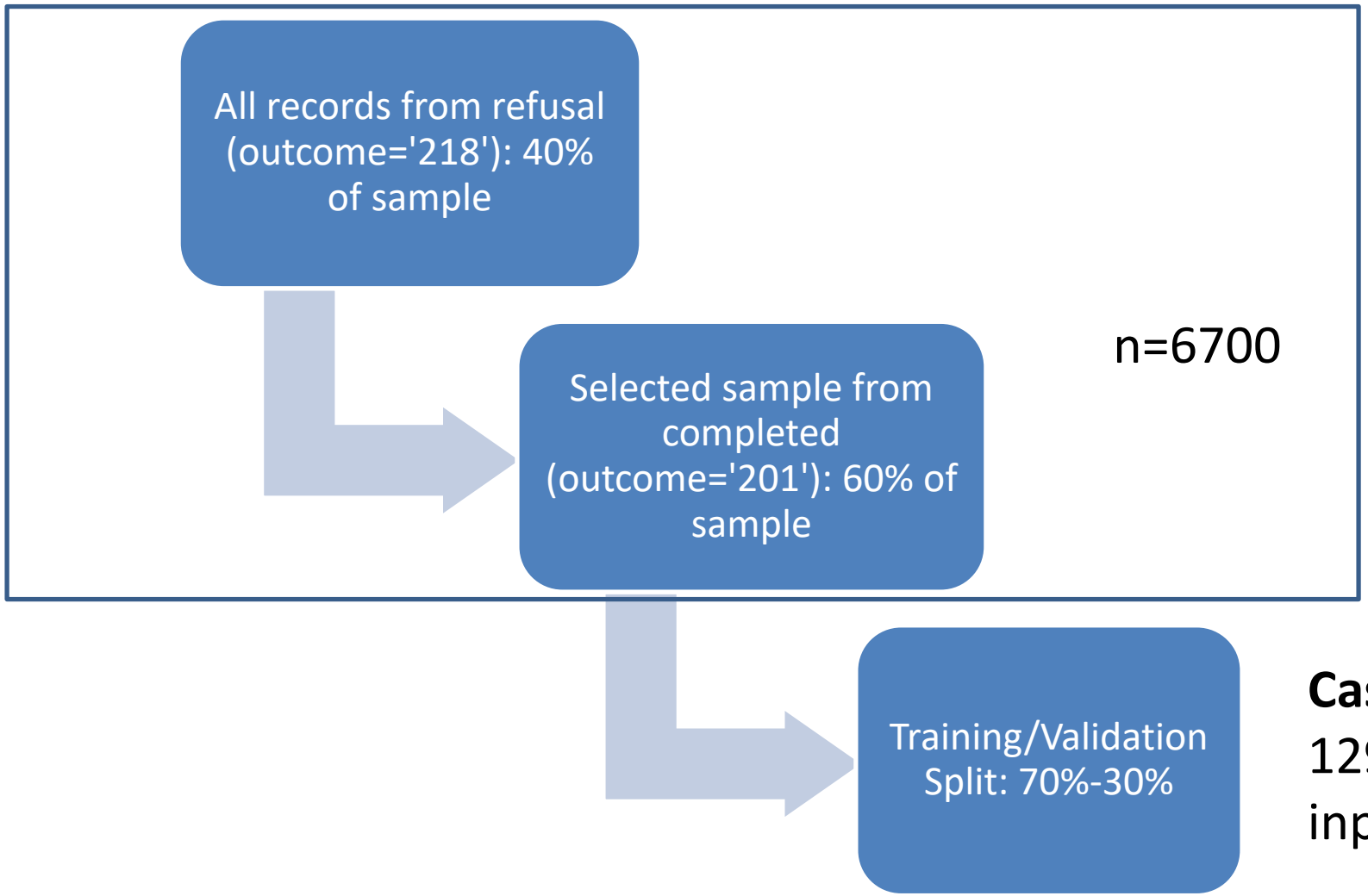
IDF			
cat	dog	grooming	best
1	1.405465	1.405465	1

$$\text{TF} = \frac{Number\ of\ times\ term\ t\ appears\ in\ a\ document}{Total\ number\ of\ terms\ in\ the\ document}$$

Term Frequency Matrix				
doc	cat	dog	grooming	best
A	1.333	4	0	2.333
B	1	2.25	2.5	0.5
C	2	0	2.666667	0.333

TF-IDF [TF]*[IDF]				
doc	cat	dog	grooming	best
A	1.333333	5.62186	0	2.333333
B	1	3.162296	3.513663	0.5
C	2	0	3.747907	0.333333

Prediction model dataset setup



Case Notes:
129,000 TF-IDF input vectors

Comparing Text-only vs. CHI-only model results

Model 1: Case Notes

	Predicted Outcome		Accuracy
Actual Outcome	**1**	**0**	0.6337
1	0.5213	0.0777	
0	0.2886	0.1124	

Model 2: Admin and CHI variables

	Predicted Outcome		Accuracy
Actual Outcome	**1**	**0**	0.628
1	0.5498	0.0405	
0	0.3315	0.0782	

Top 30 terms associated with completion			
nice	spanish	letter_brochures	mother
left_business	return	opened	internet_spending
nis	partial	scheduled	ltr_note
geo	Monday	left_materials	personal_letter
left_intro	around	vehicles	mis
transmit	son	accessor	trailer
env	agreed	note_online	
left_online	today	kids	

Top 30 terms associated with refusal			
refused	shut	gate	card_brochure
interested	government	tell	participate
send	she	reluctant	explaining
send_internet	it	intercom	shut_door
doorman	male	big	place
answered_door	online	find_nobody	lft_pkt
refusal	line	explain	
closed_door	attempted	census_letter	

Model metrics comparison

Model	Description	Accuracy	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value
1	Case Notes	**0.634**	0.870	0.280	0.644	0.591
2	Admin + CHI Variables	0.628	**0.931**	0.191	0.624	**0.659**
3	Admin + Case Notes	0.602	0.752	0.387	0.639	0.520
4	Admin + CHI Variables + Case Notes	0.611	0.756	**0.404**	**0.646**	0.534

Conclusion and next steps

- Promising results from Case Notes predictive model
  - Used only 1st contact information to predict eventual outcome
  - Used only 2017 ACS paradata
  - Available 2015 and 2016 datasets – to augment for next set of analysi
- Fine-tuning number of terms to use in model
  - Also different types of vectorization methods
- Working on models where information taken from 1st and 2nd contact
  - Challenge: Decreases training dataset size
- Investigating other models including Markov decision models

BACKUP

Model metrics comparison

Model 1: Case Notes			
	Predicted Outcome		Accuracy
Actual Outcome	1	0	0.6337
1	0.5213	0.0777	
0	0.2886	0.1124	

Model 2: Admin and CHI variables			
	Predicted Outcome		Accuracy
Actual Outcome	1	0	0.628
1	0.5498	0.0405	
0	0.3315	0.0782	

Model 3: Admin and Case Notes			
	Predicted Outcome		Accuracy
Actual Outcome	1	0	0.6023
1	0.4437	0.1466	
0	0.2511	0.1586	

Model 4: Admin, CHI and Case Notes			
	Predicted Outcome		Accuracy
Actual Outcome	1	0	0.6114
1	0.4461	0.1442	
0	0.2444	0.1653	