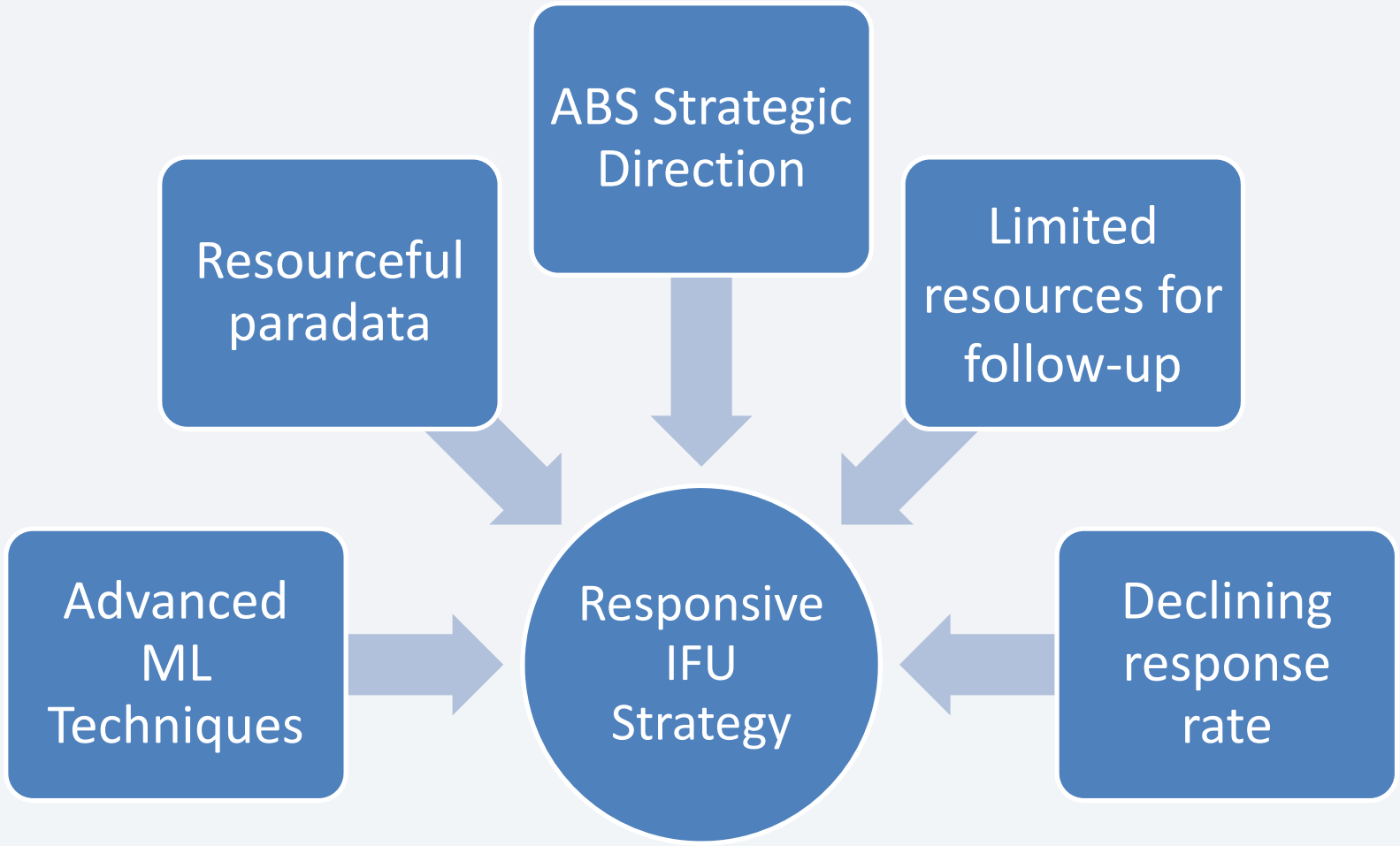
Using Predicted Response Propensities with the Random Forests Method to Direct a Responsive Intensive Follow-up Strategy

Summer Wang

Australian Bureau of Statistics

Australian Bureau of Statistics
Informing Australia's important decisions

Setting the Scene



A Responsive IFU Strategy Using Response Propensity

- First Phase

Identifying self-responders, AKA., gold provider (GP) units and delaying the IFU efforts towards them

- Second Phase

Subgrouping non-respondents and allocating different call-caps to each of the subgroups

- Final Phase

Identifying a list of potential non-respondents with no further IFU attempts to be made towards them

Predicting Response Propensity

- Determine the ***influencing factors*** to predict the RP
  - Framework of the Factors Influencing RP
- Determine the ***modelling method*** to predict the RP
  - Logistic regression model
  - Classification and Regression Trees (CART)
  - Random Forests Method

Framework of Factors Influencing RP

Influencing Factors	Available info for Business Surveys within the ABS
Area Characteristics	State, Industry, legal structure, sector classification
Business Characteristics	Size of the business Significance level Weighting contribution towards estimates Benchmark Employee Numbers Whether the business has changed address/ownership/structure
Respondent Characteristics	Whether there is a regular contact Whether they are in sample of other ABS surveys, etc. Business' response history, etc.
Interviewer Characteristics	NA
Interviewer Observations	Whether has an answering machine in use, etc.
Collection Design Features	Number of calls made by interviewer Timing of calls made by interviewer Whether telephone interview or not Interview length Outcome of previous call attempt Reminder letters (RL1, RL2, and RL3), etc.

The Random Forests Method

Advantages

- It relaxes the assumptions regarding the form of the propensity models and adapts to the size and complexity of the underlying data at hand
- It tends to generate more stable estimates compared to those generated from a single tree

Key challenges

- it is harder to interpret the robustness of the results because it doesn't have diagnostics based on statistical theory
- It also fails to provide a framework in which distributional results can be easily determined (Mentch and Hooker, 2015).

Simulation Study – Dataset & Modelling Process

- The Dataset
  - Pools together 4 cycles of the biannual Average Weekly Earnings (AWE) survey data from 2015 to 2016 in a cross-sectional approach
  - Includes 22,748 observations from 8,422 units
  - Contains 18 predictors, including 7 of business characteristics, 1 of the survey results, and another 10 of paradata information
- The Modelling Method - A Forest of Regression Trees
  - Parameter tuning, i.e., 500 trees with mtry=6, node size=10
  - Data training – the 10-Fold Cross Validation
  - Run separately at three main phases of the IFU period, namely the Reminder Letter (RL) 2, RL3 and IFU end
  - Modelling results – response propensity scores (0-1) produced at each phase for each unit

Simulation Study – Model Fitness & Prediction Accuracy

		Predicted Response Outcome			
		Non-response	Response	Total	Classification Error
Model 1 Response status by RL2	Non-response	5349	3063	8412	36.4%
	Response	3063	11273	14336	21.4%
	Total	8412	14336	22748	26.9%
Model 2 Response status by RL3	Non-response	2450	2000	4450	44.9%
	Response	2000	16298	18298	10.9%
	Total	4450	18298	22748	17.6%
Model 3 Final response status	Non-response	1334	580	1914	30.3%
	Response	580	20254	20834	2.8%
	Total	1914	20834	22748	5.1%

Simulation Study – Retrospective Analysis

- The Analysis Process
  - Using predicted RP to conduct the proposed Responsive IFU Strategy to the May 2018 cycle
  - New IFU effort required and final response rate achieved were simulated and compared to the results in reality with current IFU process
- The Analysis Results

	Current Follow-up Strategy	Proposed Follow-up Strategy
Nonresponse Units	163	358
New Response Rate	95.39%	89.87%
Response Rate Drop	0.00%	5.63%
Contacts Made	3362	2615
Contacts Saved	0	747
% Contacts Saved	0%	22.22%

Live Trial of the Gold Provider (GP) Strategy

- The Annual Agriculture Survey
  - Sample size – around 27,500
  - Response rate – around 80%
- The Live Trial of the GP strategy
  - GP units being identified for the entire survey sample
  - Entire survey sample being split into two homogenous sub-groups: control group and treatment group
  - The control group will receive the standard IFU practice during the data collection process
  - The treatment group will receive a live trial of the GP strategy
    - The Non-GP units will have "normal" IFU practice
    - the GP units will not have any IFU action taken until the RL3 stage
    - the IFU resources saved from the GP units will be re-allocated to the Non-GP units
  - Live results so far
    - response rates of GP units are much higher than that of the Non-GP units
    - The overall response rate of the treatment group is higher than the control group

Questions?

Feel free to contact me via

summer.wang@abs.gov.au

Australian Bureau of Statistics
Informing Australia's important decisions