Using machine learning models to predict follow-up survey participation in a panel study

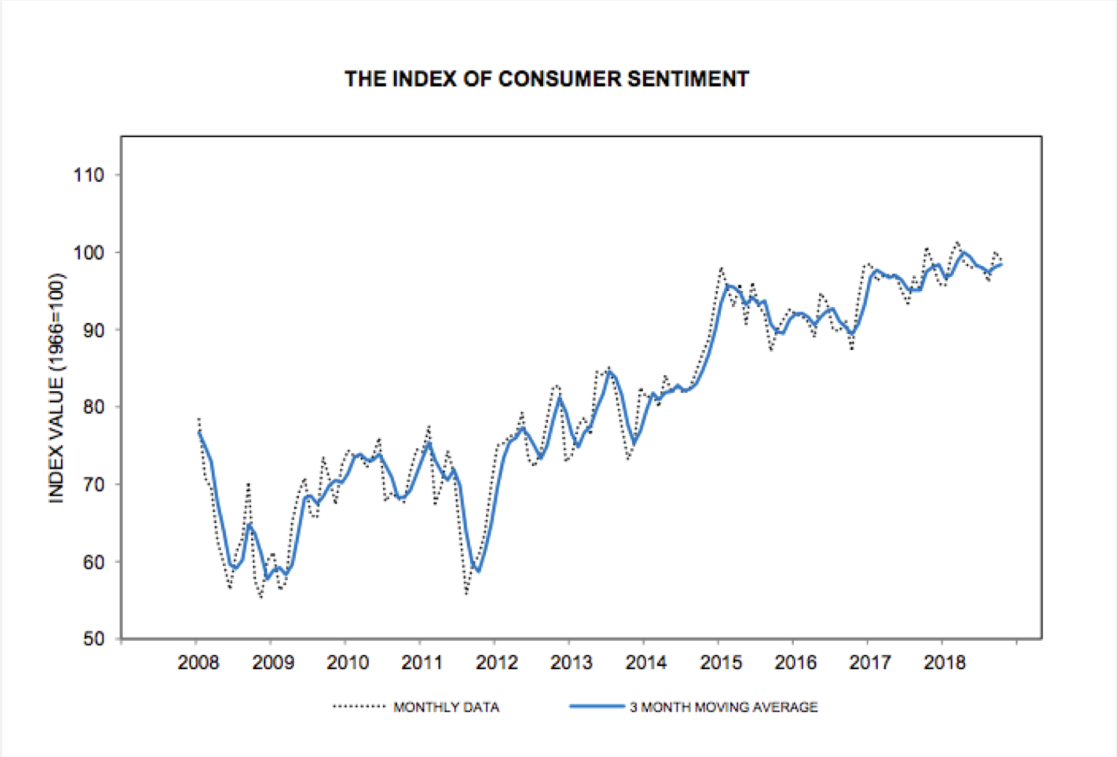
Mingnan Liu

Yichen Wang

Can we predict who will respond to the 2nd interview in a panel phone survey?

Surveys of Consumers

- ***Survey setup***
  - A monthly telephone survey with 500 interviews
  - First interview →(6 months)→ Second interview
- ***Data used***
  - Core questions cover 3 broad areas of consumer sentiment:
    - Personal finances
    - Business conditions
    - Buying conditions
  - Demographics
  - Interviewer evaluation
    - Respondent's attitude
    - Understanding
- ***Predictors***
  - 39 predictors (26 categorical, 13 continuous)



surveys of consumers
UNIVERSITY OF MICHIGAN

Why should we care?

- Not everyone will response
- Resources constrains

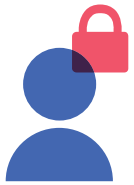
- Reduced sample size
- Poorer estimates

- Nonresponse bias

What are the causes?

- Non-location

- Non-contact



- Refusal

Lynn, Peter. (2018). Tackling Panel Attrition. In D. L. Vannette & J. A. Krosnick (Eds.), *The Palgrave Handbook of Survey Research* (pp. 143–153). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-54
Lepkowski, J. M., & Couper, M. P. (2002). Nonresponse in the second wave of longitudinal household surveys. *Survey Nonresponse*, 259–272.

What do we know?

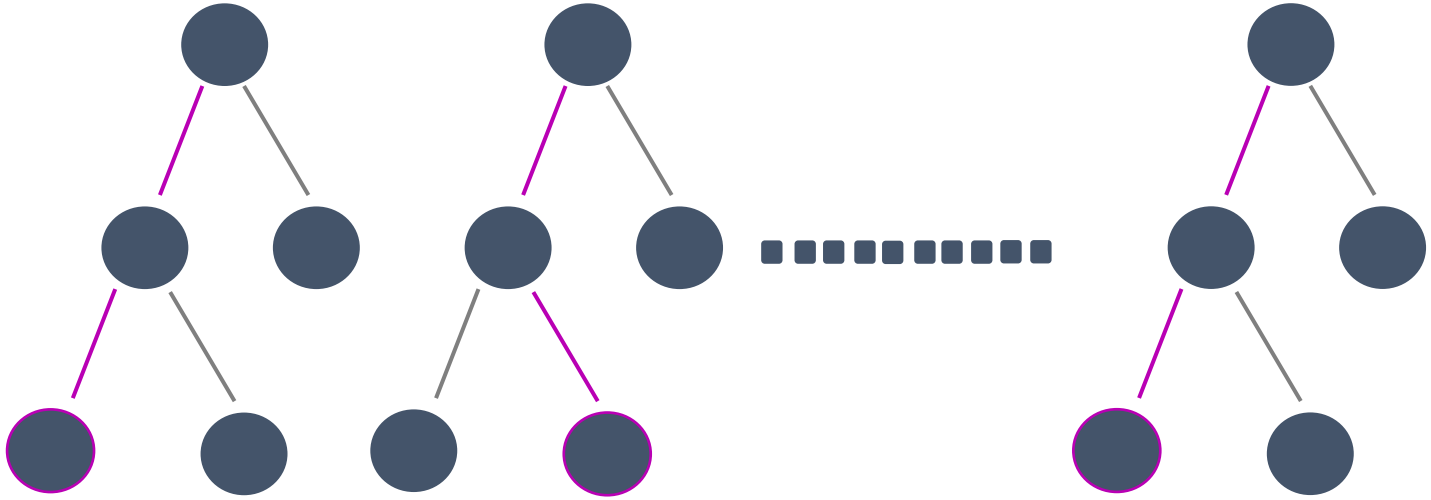
- Sample characteristics
  - Demographics
  - characteristics
- Data collection mode
  - Face-to-face
  - Mixed-mode
- Paradata
  - Call records
- Design features
  - Incentives
  - Interview length
  - Address update
  - Interviewer consistency

Limitations

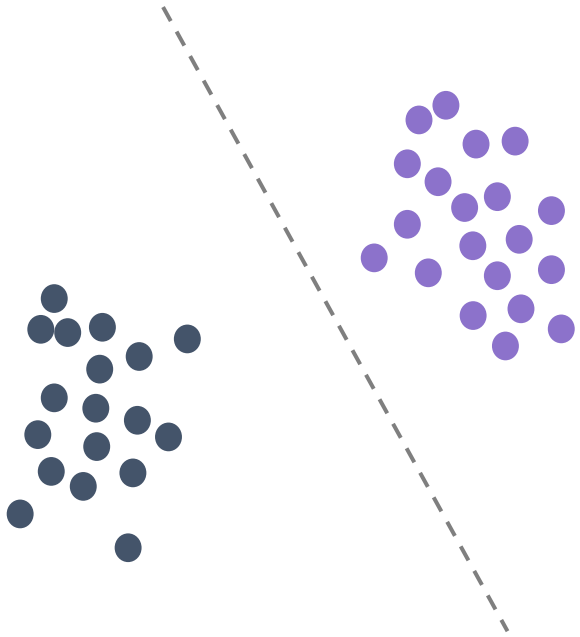
- Mean effects of treatments
  - Success metric for different techniques on attrition is the overall sample mean impact
- Targeted approach more effective

Decision tree

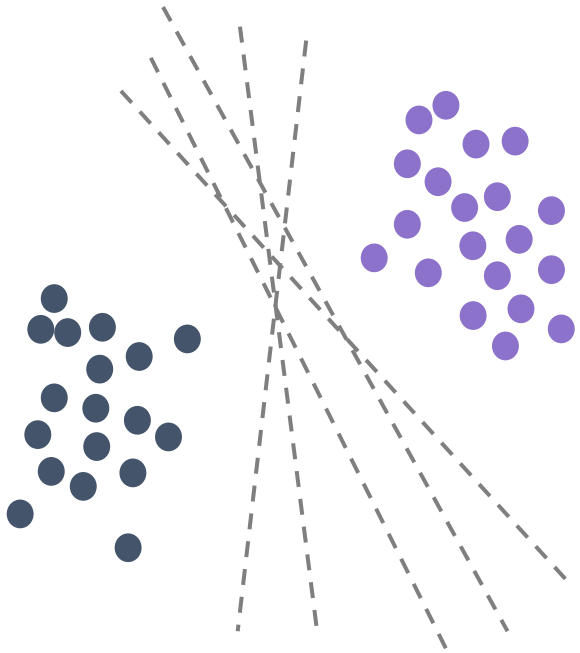
Random forests



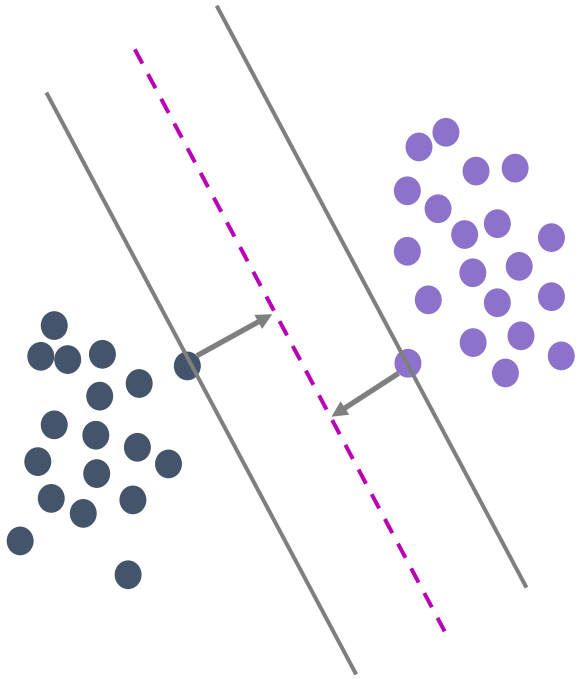
Support Vector Machine (SVM)

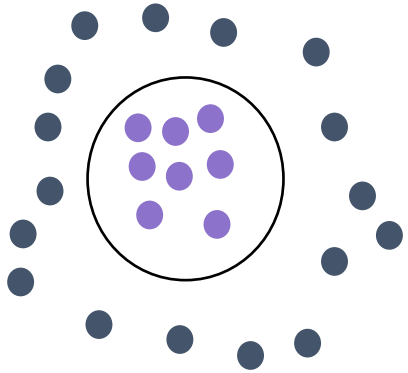


Support Vector Machine (SVM)



Support Vector Machine (SVM)





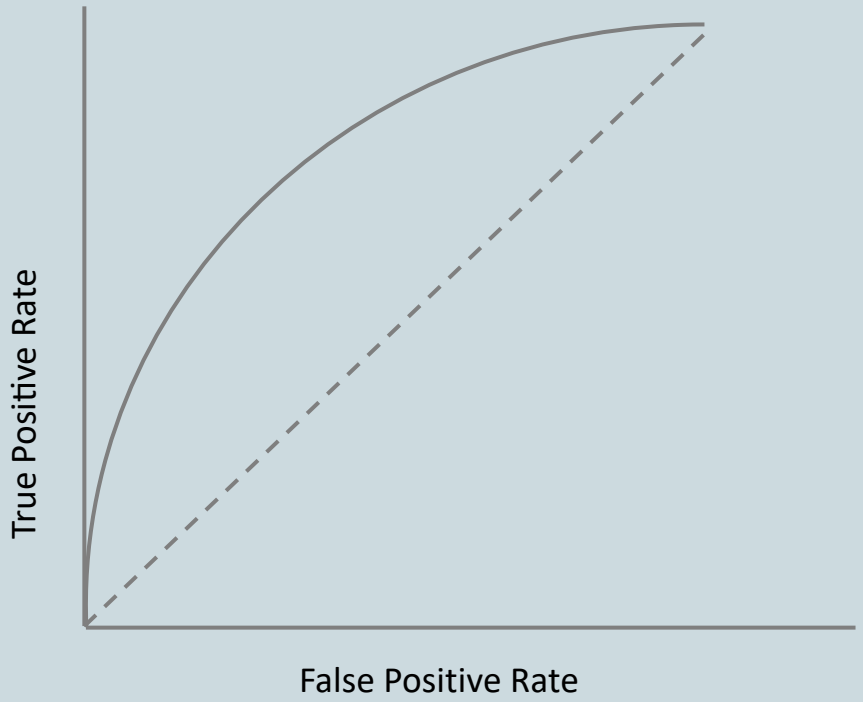
LASSO

- Shrinkage method: shrinks the coefficient estimates of some predictors towards zero or forces them to be exactly zero
  - Esp. those don’t really predict the outcome

Evaluation criteria

- Accuracy = $\frac{a+b}{a+b+c+d}$
- Sensitivity = $\frac{a}{a+c} = \frac{\#\ of\ true\ wave\ 2\ Rs}{\#\ of\ Rs\ predicted\ to\ be\ wave\ 2\ Rs}$
- Specificity = $\frac{d}{b+d} = \frac{\#\ of\ true\ wave\ 2\ non-Rs}{\#\ of\ Rs\ predicted\ to\ be\ wave\ 2\ non-Rs}$
- Balanced accuracy = average (sensitivity+specificity)
- Area under the ROC curve

	Model prediction	
Actual	Wave 2 *R*s	Wave 2 non-*R*s
Wave 2 *R*s	a (true +)	c (false -)
Wave 2 non-*R*s	b (false +)	d (true -)



Results

	Random forest	SVM	LASSO	Logistic regression
Accuracy			☑	
Sensitivity			☑	
Specificity	☑			
Balanced accuracy			☑	
Area under ROC	☑		☑	

The most predictive variables are …

- Education level
- Interviewer evaluation
  - Respondent's attitude
  - Understanding

Future studies

- Expand predictors
- More waves
- Survey modes

Thank you

mingnanliu@gmail.com