Machine Learning Assisted Complex Survey Weights

Stas Kolenikov

Abt Associates Inc.

FCSM Research & Policy Conference
November 2021

Application: a national evaluation project

- Evaluation of a labor market program
- Funding to ~50 local programs
- Evaluation of the initiative to better understand program implementation, participant outcomes, and return on investment (ROI)

Sample

Sample and description	Sample Size
Baseline: All participants	~16,000
SSN Sample (subset of Baseline): Have non-missing SSN	~11,000
Survey Sample Frame (subset of SSN Sample): Have non-missing SSN and complete contact information (email, phone, and address) after lookup	~10,000
Selected Survey Sample (subset of Survey Sample Frame): Sample selected for survey	~8,000
Survey Respondent Sample (subset of Selected Survey Sample): Respondents to the participant survey	~2,500

Role of sampling weights

- Analysis weights are necessary to obtain approximately unbiased estimates of statistical quantities obtained in a complex survey design.
- Weights typically incorporate:
  - base selection probability
  - nonresponse adjustments
  - calibration adjustments
- Weights provide protection against informative sampling designs, i.e., designs where the survey outcomes are correlated with the design variables.

Weight construction

1. Model response propensity for each stage
   - Stage 3, random sampling: selection probabilities are known
   - All other stages: quasi-selection, need to model propensities
2. Form a product of inverse propensities / inverse sampling probabilities, and
3. Calibrate to the population totals.

Weight construction: tradition

1. Model response propensity for each stage
   – Stage 3, random sampling: selection probabilities are known
   – All other stages: quasi-selection, need to model propensities

   Stepwise logistic

2. Form a product of inverse propensities / inverse sampling probabilities, and

   Demographics

3. Calibrate to the population totals.

Weight construction: MLAW

1. Model response propensity for each stage
   - Stage 3, random sampling: selection probabilities are known
   - All other stages: quasi-selection, need to model propensities

GBM

2. Form a product of inverse propensities / inverse sampling probabilities, and
3. Calibrate to the population totals.

Demographics
+ prognostic scores

Match/response propensity

- Machine learning prediction: library(h2o)
  - Random forests (10-fold validation, depth 5 to 10, 20 cases per leaf, 1000 trees, 5 variables in each tree)
  - GBM (1000 trees, learning rate 0.05 annealed by 0.995)
- GBM score ⇒ predictor in mixed logistic model with location random effects

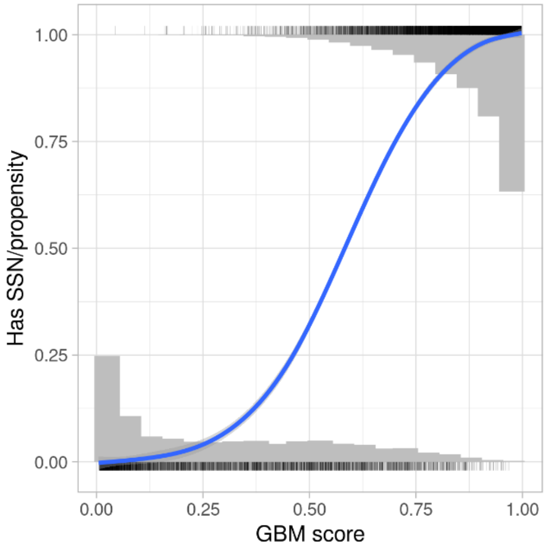
GBM models: match and response

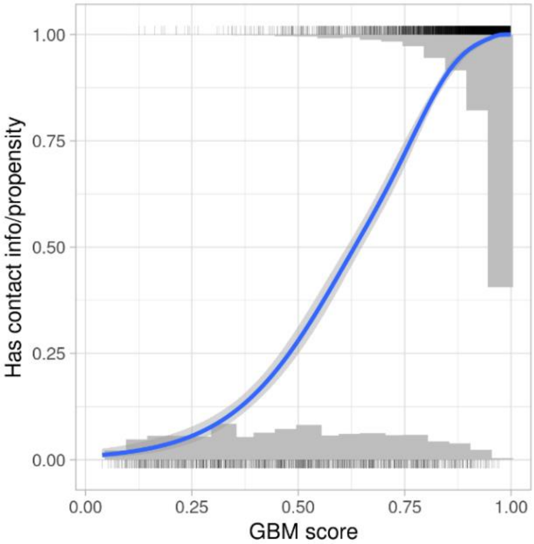
Outcome	Cross-validated AUC	Gini	Variable importance					
			Highest	Relative importance	Second highest	Relative importance	Third highest	Relative importance
1. Has SSN	0.9693	0.9386	Log (starting wage)	0.3915	SOC	0.2195	age	0.1389
2. Has contact info	0.9648	0.9295	age	0.2964	Log (starting wage)	0.2228	SOC	0.1906
4. Survey response	0.9348	0.8696	age	0.2648	SOC	0.2285	Log (starting wage)	0.1632

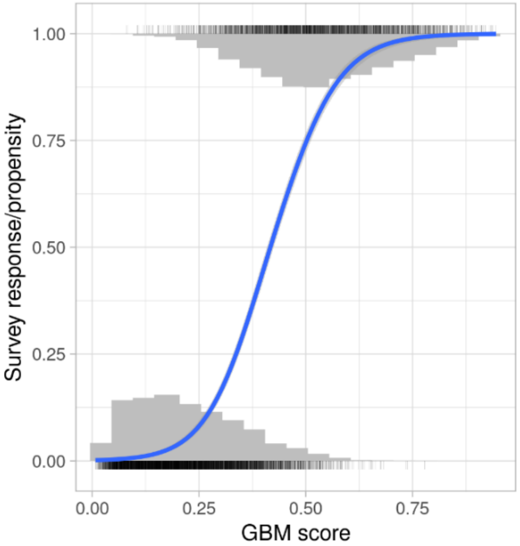
GBM models: outcomes

Outcome	Cross-validated R2	Error rate	Variable importance					
			Highest	Relative importance	Second highest	Relative importance	Third highest	Relative importance
Same occupation	0.566	10.1%	age	0.268	SOC	0.256	Log (starting wage)	0.171
Program status	0.708	3.4%	Last contact code	0.230	age	0.227	SOC	0.225
Earnings	0.405		Log (starting wage)	0.406	SOC	0.235	age	0.192

Resulting ML propensities







Did ML help?

Outcome	Score range	Baseline	Demo weight
Program completion	[0, 0.2)	38.25%	47.91%
	[0.2, 0.5)	18.91%	12.85%
	[0.5, 1]	42.84%	39.23%
Earnings	Bottom	11.39%	11.85%
	2	24.71%	25.59%
	3	28.03%	29.53%
	4	17.77%	18.33%
	Top	18.11%	14.70%

Parting thoughts

- Complementary propensity models and implementation
  - Arguably fewer lines of code than model selection with logistic regression
  - … except when you need to find the right tuning parameters
- Improvements in population representation due to model calibration

BOLD THINKERS DRIVING REAL-WORLD IMPACT

For comments or questions about this presentation, please contact:

Stas Kolenikov
Principal Scientist, DSET
Stas_Kolenikov@abtassoc.com
@StatStas

abtassociates.com