Machine Learning in Survey Research: Modeling Nonresponse and Completion Conditions from a Prediction Perspective

Christoph Kern

University of Mannheim

ITACOSM Conference 6/6/2019

Introduction

- Machine learning (ML) methods provide a vast set of tools for exploring and analyzing diverse data
- Comprise flexible/ non-parametric methods that adapt to complex data structures
- Focus on out-of-sample prediction performance

- ML increasingly used by survey researchers in various contexts (Buskirk et al., 2018; Kern et al., 2019)
- A promising *supplement* in the survey methods toolkit?

$\rightarrow$ This talk highlights **two applications of prediction methods in survey research**

Study I: Predicting Panel Nonresponse

Joint work with Bernd Weiß (GESIS - Leibniz Institute for the Social Sciences), Jan-Philipp Kolb (GESIS - Leibniz Institute for the Social Sciences)

Study I – Data

GESIS Panel[1]

- Probability-based mixed-mode panel of the general population in Germany
- Recruitment in 2013, bi-monthly surveys since 2014 ($\sim$4900 panelists)
- $\sim$20min each wave, includes external studies and longitudinal core study
- Online (web surveys) and offline (mail) mode
  - About 62% online and 38% offline respondents

→ **Outcome: Non-participation in (each) next wave**

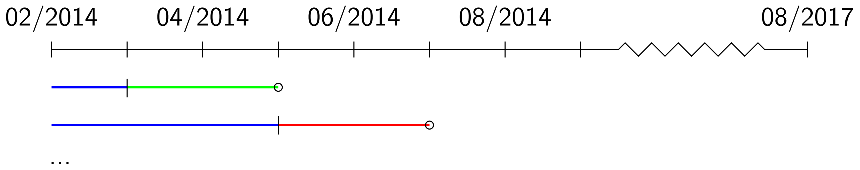
- Complete or partial interview with sufficient information (0) vs. else (1)
- Sample: Excluding "ineligible" panelists per wave

[1] https://www.gesis.org/en/gesis-panel/

Study I – Temporal CV

Longitudinal configuration

- Compare methods/ performance by repeatedly mimicking usage of model in real world
- Temporal Cross-Validation via `triage` (`Python`)[2]
  - Start with first complete GESIS panel wave (Feb 2014)
  - End with most recent wave up to date (August 2017)
  - Time between waves (update frequency, label timespan): 2 months



$\rightarrow$ 20 train and 20 test matrices

[2] `https://github.com/dssg/triage`

Study I – Features

- Block I: Time-invariant
  - Demographics from welcome survey
  - Survey cooperation in welcome survey
- Block II: Time-variant
  - Response status and survey evaluation last wave
- Block III: Time-variant (aggregated)
  - Response status and survey evaluation over last two and three waves
- Block IV: Time-variant (aggregated)
  - Response status and survey evaluation over all previous waves

→ Feature group strategies: `all`, `leave-one-out`

Study I – Methods

- Penalized Logistic Regression
  - Logit regression plus lasso/ ridge penalty on model complexity (Tibshirani 1996)
- Decision Trees
  - Split predictor space into subregions τ_m with associated constants γ_m (Breiman et al. 1984)

$$\mathcal{T}(x;\Theta) = \sum_{m=1}^{M} \gamma_m I(x \in \tau_m)$$

- Random Forest, ExtraTrees
  - Grow an ensemble of decorrelated trees (Breiman 2001, Geurts et al. 2006)

$$\hat{f}_B(x) = \frac{1}{B}\sum_{b=1}^{B} \mathcal{T}_b(x;\Theta_b)$$

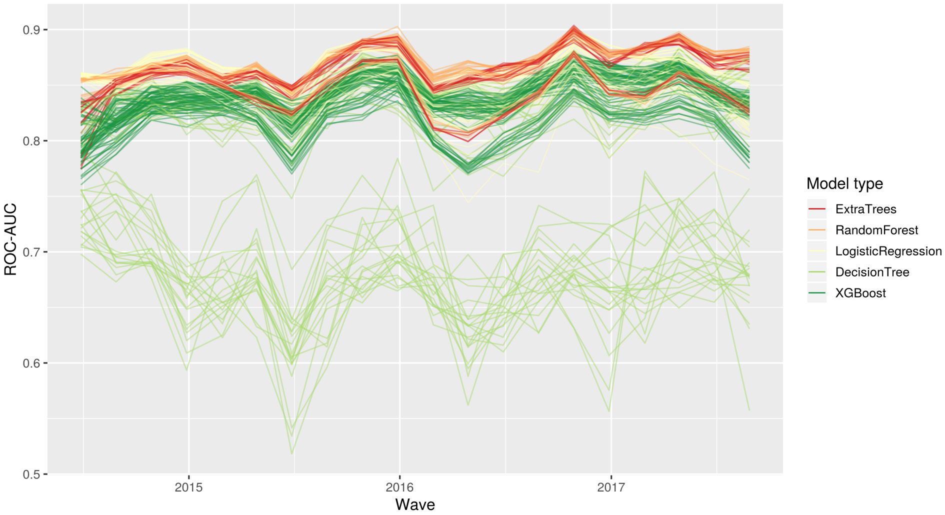
- Extreme Gradient Boosting (XGBoost)
  - Build a sequence of trees using updated pseudo-residuals (Chen and Guestrin 2016)

$$\hat{f}_T(x) = \sum_{t=1}^{T} \mathcal{T}(x;\Theta_t)$$

$\rightarrow$ 3600 models to train ($20 \times 5 \times 36$)

Study I – Results

Figure 1: ROC-AUCs for all waves and models

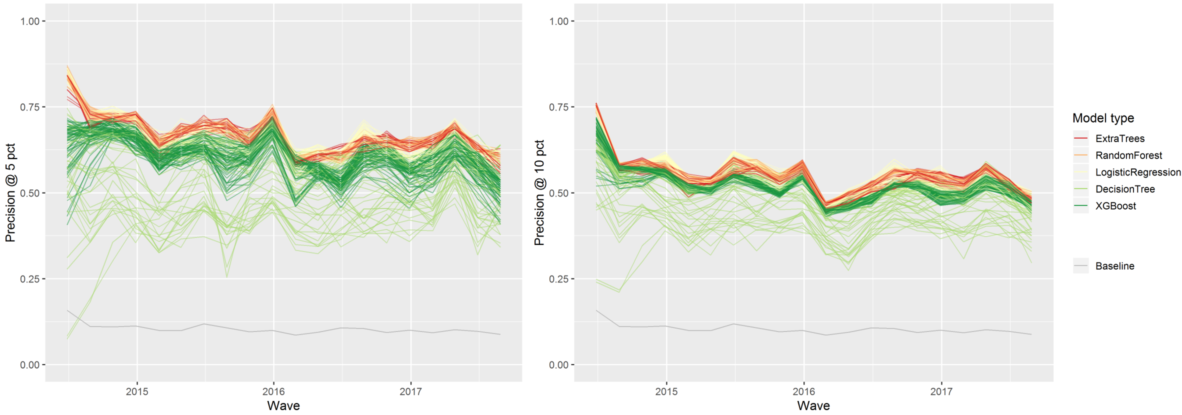


Study I – Results

Figure 2: Precision at top K for all waves and models

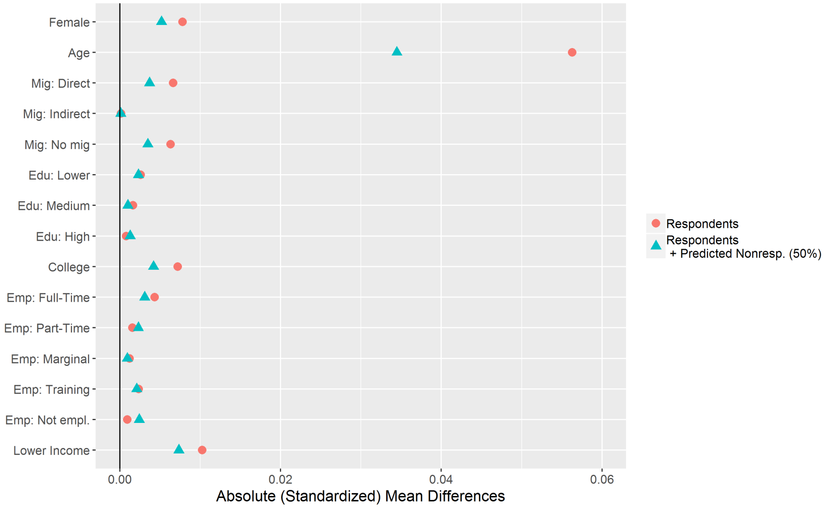
(a) Precision @ 5 pct

(b) Precision @ 10 pct



Study I – Results

Figure 3: Differences between active panel population, respondents and potential respondents (RF)



Study II: Predicting completion conditions in mobile web surveys

Joint work with Jan Karem Hoehne (University of Mannheim), Stephan Schlosser (University of Goettingen), Melanie Revilla (RECSM-Universitat Pompeu Fabra)

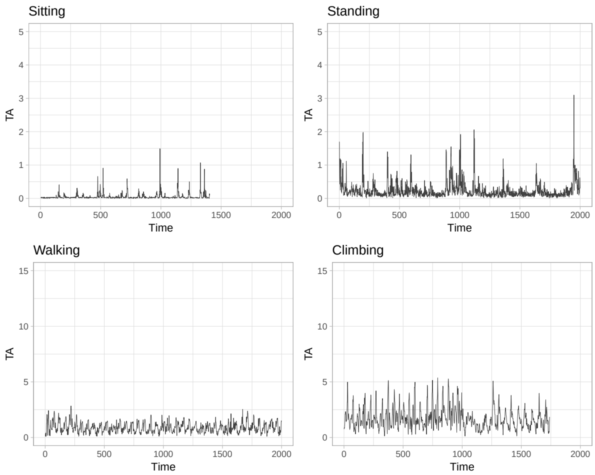
Study II – Introduction

Utilizing acceleration data from smartphone sensors and ML to infer completion conditions

1. Can we accurately predict respondents completion conditions by using acceleration data?
2. Do respondents with different completion conditions differ in terms of response behavior?

→ SurveyMotion (Höhne and Schlosser, 2019)

Figure 4: Examples of total acceleration profiles



Study II – Data

Training data: Lab experiment

- Data collected in August 2017 at the University of Goettingen
- 89 university students
- Completed mobile web survey in one of four experimental groups
  1. First group was seated in front of a desk
  2. Second group stood at a fixed point
  3. Third group walked along an aisle
  4. Fourth group climbed stairs

Prediction: Cross-sectional web survey

- Data collected in December 2017 at the University of Goettingen
- 2,357 respondents
- 61.6% smartphone respondents
  - Acceleration data available for 97,2% of smartphone respondents

Study II – Methods

Variables

- Outcome
  - 4 class outcome: sitting, standing, walking, climbing stairs
  - 2 class outcome: **moving** (walking, climbing stairs), **not moving** (sitting, standing)
- Predictors
  - Aggregated TA measurements

Training and evaluation

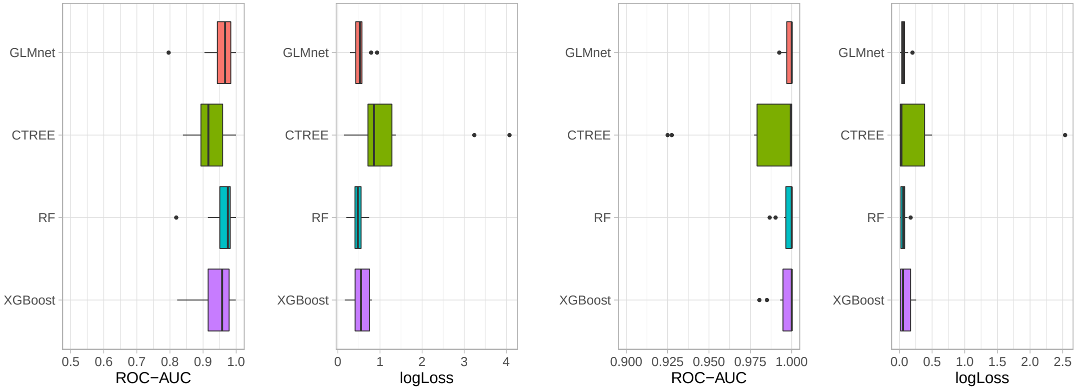
- ML methods
  - Elastic net (GLMnet; Friedman et al. 2010)
  - Conditional Inference Trees (CTREE; Hothorn and Zeileis 2015)
  - Random Forests and Extremely Randomized Trees (RF; Wright and Ziegler 2017)
  - Extreme Gradient Boosting (XGBoost; Chen and Guestrin 2016)
- 10-Fold Cross-Validation (grouped by respondent IDs)

Study II – Results

Figure 5: Cross-Validation results (training set)

(a) 4 class outcome

(b) 2 class outcome



Study II – Results

Figure 6: Class predictions in web survey

Study II – Results

Table 1: Mixed effects regressions modeling completion time[3]

	Dependent variable			
	Completion time			
	(1)	(2)	(3)	(4)
Moving	0.906	0.801	0.799	0.649
se	(0.330)	(0.334)	(0.334)	(0.372)
p	0.007	0.017	0.017	0.081
Matrix			28.742	28.720
se			(0.983)	(0.983)
p			0.000	0.000
Moving × Matrix				0.567
se				(0.623)
p				0.363
Constant	13.033	13.134	7.386	7.391
se	(3.857)	(3.853)	(0.466)	(0.466)
Demographic controls		*X*	*X*	*X*
Observations	11,029	10,688	10,688	10,688
Bayesian Inf. Crit.	68,040.810	65,779.330	65,744.750	65,752.300

[3] Completion time outliers excluded based on .05 and .95 quantile.

Discussion

- ML can be used to study topics in survey research from a prediction perspective...
- ...and to derive insights from new data types and measures

- Study I
  - Promising prediction performance over panel waves
  - Targeting predicted nonrespondents may reduce systematic nonresponse
- Study II
  - Low rate of respondents with predicted high motion levels
  - Modest differences in response behavior between motion groups

Contact: c.kern@uni-mannheim.de

References I

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Monterey, CA: Brooks/Cole Publishing.

Buskirk, T. D., Kirchner, A., Eck, A., and Signorino, C. S. (2018). An introduction to machine learning methods for survey researchers. *Survey Practice*, 11(1).

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Technical report, `https://arxiv.org/abs/1603.02754`.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1):3–42.

Höhne, J. K. and Schlosser, S. (2019). SurveyMotion: What can we learn from sensor data about respondents' completion and response behavior in mobile web surveys? International Journal of Social Research Methodology.

Hothorn, T. and Zeileis, A. (2015). partykit: A modular toolkit for recursive partytioning in R. *Journal of Machine Learning Research*, 16:3905–3909.

Kern, C., Klausch, T., and Kreuter, F. (2019). Tree-based machine learning methods for survey research. *Survey Research Methods*, 13(1):73–93.

References II

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Wright, M. N. and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17.