gesis Leibniz Institute for the Social Sciences

Using Predictive Modeling to Identify Nonresponse

J.-P. Kolb, B. Weiß and C. Kern

Bigsurv Conference, Barcelona
October 24 to October 27, 2018

Mitglied der Leibniz Gemeinschaft

Motivation

- Nonresponse is a substantial problem for data quality.
- Understanding panel-nonresponse is crucial to correct for systematic dropout patterns.
- Social science panels such as the GESIS Panel collect a constantly growing amount of variables (e.g. survey data like socio-demographics, but also paradata, etc.).
- Recently, researchers (e.g. Lugtig and Blom in 2018) have highlighted the importance of paradata to explain nonresponse.
- When we use all this information, we would have many more variables than observations. ($\Rightarrow p > n$)

$\Rightarrow$ We use statistical learning techniques to predict nonresponse.

Research questions

We focus on identifying panelists-at-risk of dropping out. We use statistical learning to compare variable importances and prediction performance.

1. Can we find evidence for the importance of paradata?
2. Does adding panel management information increase the performance of the prediction models?
3. How well perform various statistical learning techniques, especially ensemble methods (Lasso, conditional trees, random forest, gradient boosting)?

Data: The GESIS Panel

The GESIS Panel[1] is a German *probability-based mixed-mode access panel* ($n \approx 4,800$):

- Probability-based: (multi-stage) random sample based on German population registers.
- Mixed-mode: Web- and mail-based survey - unified mode design
- Access Panel: Free data collection services for social, economic and behavioral sciences.

Bi-monthly data collection - 32 waves - recent wave fe.

[1] GESIS (2018): GESIS Panel - Standard Edition. GESIS Data Archive, Cologne. ZA5665 Data file Version 24.0.0, doi:10.4232/1.13001

gesis Leibniz Institute for the Social Sciences

Background information

Groups of predictors

- Survey data (SD)
- Paradata (PARA)
- Panel management data (PM)

Example: Panel management data

- We carry out a detailed panel management and maintain intensive contact with the panellists.
- All contacts are recorded in a panel protocol.
- 117 complaints, 347 technical problems, 836 changes (in name, e-mail etc.), 2385 general feedback

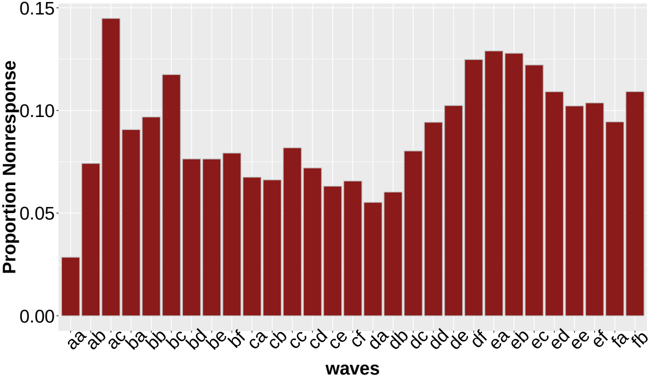
Outcome - unit nonresponse

Does the panelist fall into one of the following AAPOR categories?

211	Refusal
212	Break-off: Q too incomplete to process
319	Nothing ever returned
2112	Explicit refusal
211221	Logged on to survey, no item complete

gesis Leibniz Institute for the Social Sciences

Nonresponse in the GESIS Panel waves



- Example for wave ed: 470 nonrespondents of 4307 Panelists → 11 % nonresponse

Outcome and predictor variables

Outcome		Explanation
Nonresponse		For wave ed (August 2017)

Predictors group[2]		Examples
Survey data	SD	Age, gender, educational attainment, distance to large city, hh size, Health (HLTH)
Survey participation	SP	Mode of invitation, cohort, membership in other panels
Survey attitudes	SA	Interesting Q, obligated to participate
Panel management	PM	Nr. contacts, complaints, tech. problems
Paradata	PARA	Panel experience, response latency

[2] Mostly from wave ec (6/2017).

Log-likelihood function (logistic regression)[3]

Maximize:

$$l(\beta) = \sum_{i=1}^{n} \left[y_i x_i \beta - \log\left(1 + e^{x_i \beta}\right) \right]$$

- x_i is the i-th row of an matrix of n observations with p predictors
- β is the column vector of the regression coefficients
- y_i is the binary outcome

[3] Pareira et al. 2015

Logit model - parameter estimates

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-47.53519	12.63762	-3.76141	0.00017
PARA:Panel exper.	-14.86417	0.92414	-16.08426	0.00000
SD:Female TRUE	0.61924	0.20707	2.99040	0.00279
SD:Degree of urbanisation	0.02926	0.00643	4.55184	0.00001
PARA:Latency	0.01775	0.00702	2.52758	0.01149
WELL:Feeling depressed	-0.00852	0.00492	-1.72924	0.08377
SD:Age^2	0.00001	0.00000	4.50808	0.00001

- Panel experience (respondent participated in preceding waves) is by far the most important parameter
- Logit model - DF 3019 / AIC 817.32

Lasso model[4]

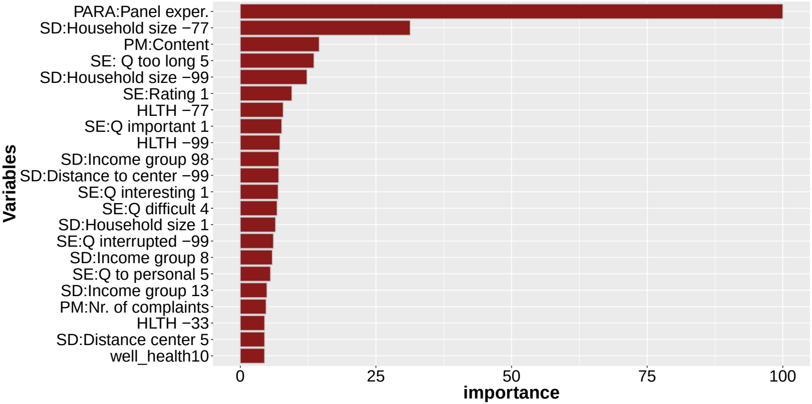
Maximize:

$$l_\lambda(\beta) = \sum_{i=1}^{n} \left[y_i x_i \beta - log\left(1 + e^{x_i \beta}\right)\right] - \lambda \sum_{j=1}^{p} |\beta_j|$$

- x_i is the i-th row of an matrix of n observations with p predictors
- β is the column vector of the regression coefficients
- y_i - binary outcome
- λ - shrinkage parameter

[4] Pareira et al. 2015

Importance Lasso regression



Statistical learning techniques

Parametric methods (logit model, lasso)

- **X** dependent on specification
- Linearity, additivity
- $\rightarrow$ causation

Tree-based methods (Conditional Inference Trees[5], random forests[6], gradient boosting[7])

- "Built-in" feature selection
- No predefined flexible functional form
- $\rightarrow$ prediction

---

[5]Hothorn et al. 2006; [6] Liaw & Wiener 2002; [7] Chen & Guestrin 2016

Decision trees: Recursively partition the predictor space into disjoint regions R_j

$$T(x;\Theta) = \sum_{j=1}^{J} \gamma_j I(x \in R_j)$$

with tree parameters $\Theta = \{R_j, \gamma_j\}$.

Conditional inference trees

- Partitioning based on permutation tests

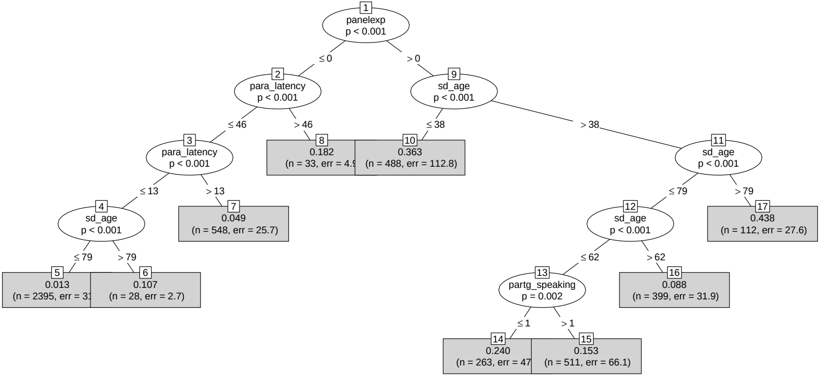
Random forests (RF)

- Grow many (decorrelated) trees using bootstrapping

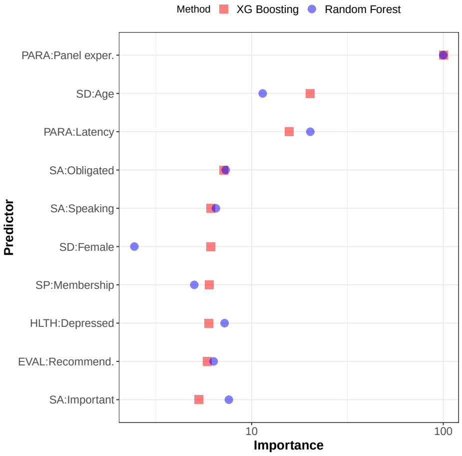
Gradient boosting (GBM)

- Sequence of (e.g.) trees using updated residuals

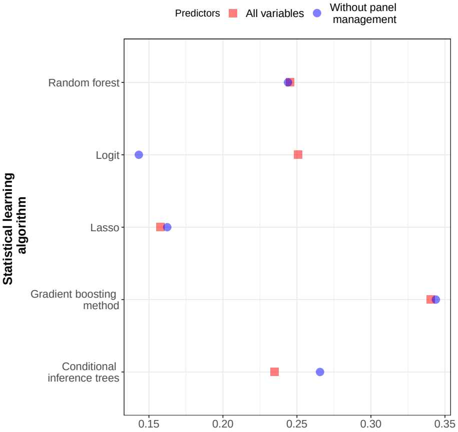
Conditional inference tree



RQ1: The importance of paradata

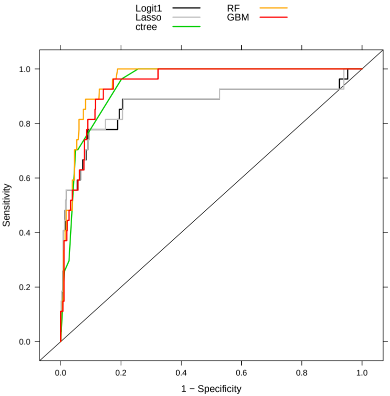


RQ2: Impact of panel management information



gesis Leibniz Institute for the Social Sciences

RQ3: ROC Curve & classification metrics



w/o panel management data

	Acc.	Kappa	Sens.	Spec.
Logit1	0.07	-0.02	0.82	0.02
lasso	0.93	0.23	0.18	0.98
ctree	0.93	0.22	0.16	0.99
RF	0.99	0.95	0.93	1.00
GBM	0.99	0.88	0.82	1.00

	Acc.	Kappa	Sens.	Spec.
Logit1	0.05	-0.03	0.67	0.02
lasso	0.97	0.54	0.43	0.99
ctree	0.95	0.12	0.07	1.00
RF	0.99	0.89	0.83	1.00
GBM	0.99	0.87	0.80	1.00

Conclusion

- Paradata, survey evaluation & attitudes are important for predicting nonresponse.
- Including panel management information does not increase the Kappa measure (or accurancy) substantially.
- (Extreme) Gradient boosting and Random forest show best performance.

gesis Leibniz Institute for the Social Sciences

Planned further steps

We plan to...

- apply models to other waves of the GESIS Panel.
- consider more meta information, e.g. question types or content of the previous questionnaire.
- use recent wave for prediction on coming wave.
- consider dropout patterns (Panelists leave the panel if they have not responded three times in a row)

gesis Leibniz Institute for the Social Sciences

Thank you very much for your attention!

Questions? Comments?

Jan-Philipp Kolb <Jan-Philipp.Kolb@gesis.org>

gesis Leibniz Institute for the Social Sciences

References

- Chen T. and Guestrin C. (2016) XGBoost: A scalable tree boosting system. Technical report, `https://arxiv.org/abs/1603.02754`.
- Hothorn T., Hornik K., and Zeileis, A. (2006) Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics, 15(3):651–674.
- Liaw A. and Wiener M. (2002) Classification and regression by randomForest. R News, 2(3):18–22.
- Lugtig P. and Blom A. (2018) It's the process stupid! Using Machine Learning to understand the relation between paradata and panel dropout
- Pereira J., Bastoa M. and da Silva A. (2015) The logistic lasso and ridge regression in predicting corporate failure