INSTITUTE FOR EMPLOYMENT RESEARCH
The Research Institute of the Federal Employment Agency

EVALUATING MACHINE LEARNING ALGORITHMS TO DETECT INTERVIEWER FALSIFICATION

Silvia Schwanhäuser
Joseph W. Sakshaug
Yuliya Kosyakova
Natalja Menold
Peter Winker

ESRA 2021 Conference

July 16, 2021

THE PROBLEM WITH

INTERVIEWER FALSIFICATION

“ ‘Interviewer falsification’ means the ***intentional*** *departure from the designed interviewer guidelines or instructions,* ***unreported*** *by the interviewer, which could result in the* ***contamination*** *of data.”*

American Association for Public Opinion Research (AAPOR) 2003: 1

(AAPOR 2003; DeMatteis et al. 2020)

THE PROBLEM WITH

INTERVIEWER FALSIFICATION

“ ‘Interviewer falsification’ means the ***intentional*** *departure from the designed interviewer guidelines or instructions,* ***unreported*** *by the interviewer, which could result in the* ***contamination*** *of data.”*

American Association for Public Opinion Research (AAPOR) 2003: 1

- Fabrication of complete interviews
- Fabrication of single items
- Fabrication of few interviews
- Miscoding of respondents’ answers
- Deviations from selection rules
- …

Differences in

… detection probability

… influence on data quality

… verification costs

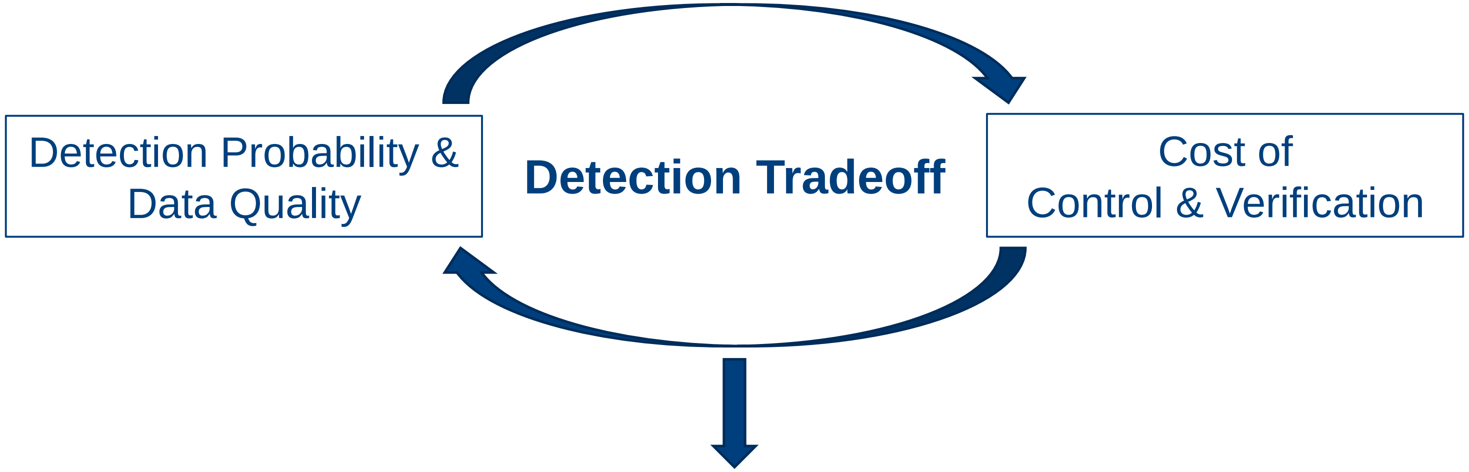
(AAPOR 2003; DeMatteis et al. 2020)

THE PROBLEM WITH

INTERVIEWER FALSIFICATION

*“ ‘Interviewer falsification’ means the **intentional** departure from the designed interviewer guidelines or instructions, **unreported** by the interviewer, which could result in the **contamination** of data.”*

American Association for Public Opinion Research (AAPOR) 2003: 1



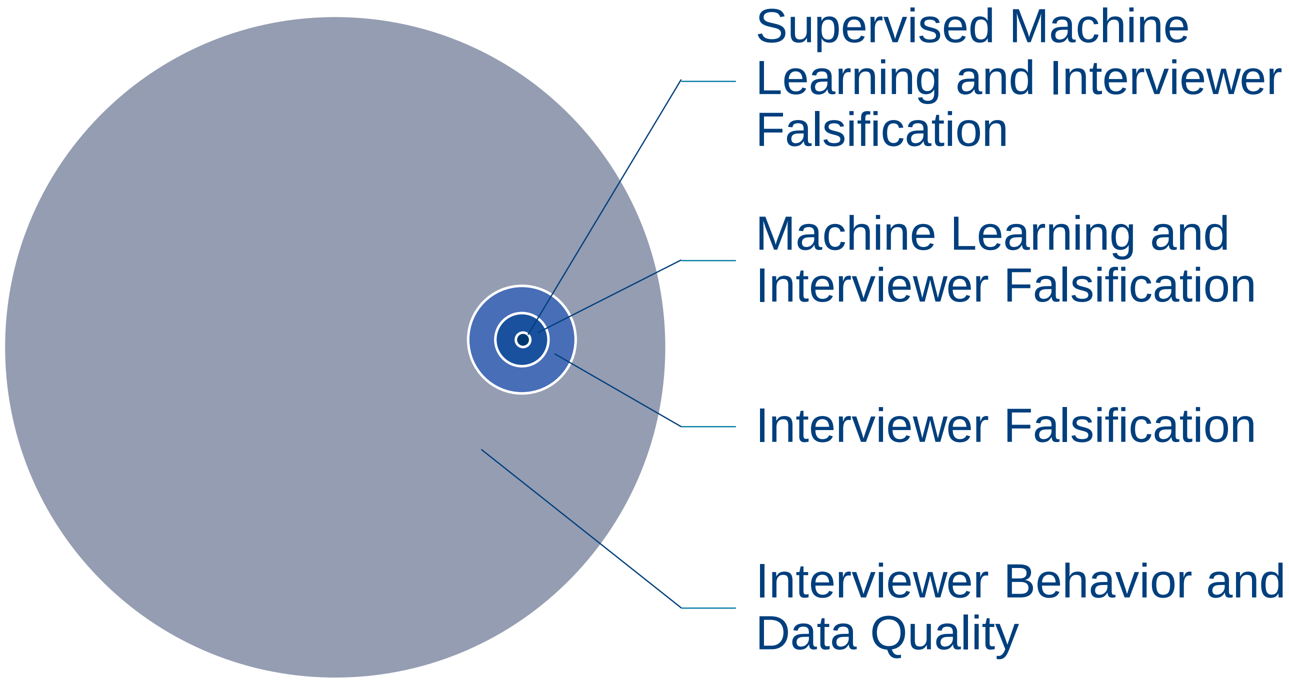
How can we improve this tradeoff?

(AAPOR 2003; DeMatteis et al. 2020)

CAN WE USE MACHINE LEARNING TO DETECT INTERVIEWER FALSIFICATION?

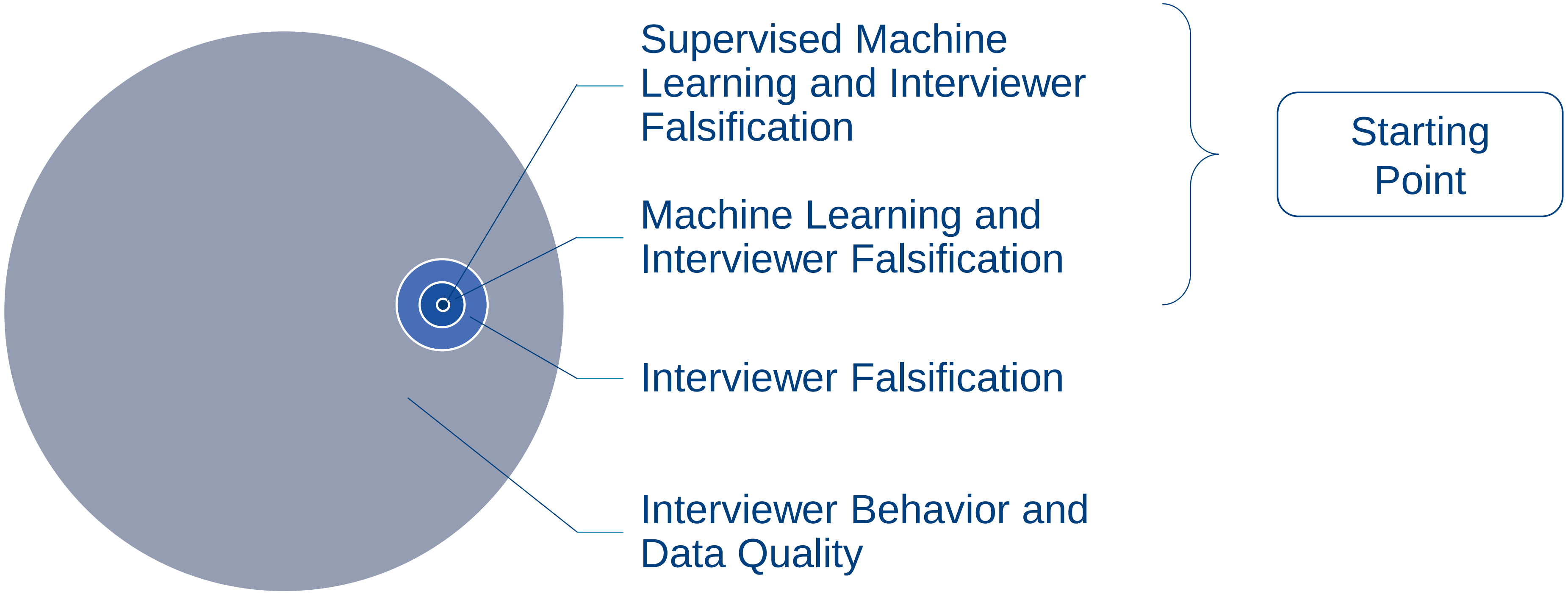
PREVIOUS LITERATURE ON

INTERVIEWER FALSIFICATION



PREVIOUS LITERATURE ON

INTERVIEWER FALSIFICATION



PREVIOUS LITERATURE ON

INTERVIEWER FALSIFICATION

Machine Learning and Interviewer Falsification

- Unsupervised Machine Learning
  - Data Mining and Outlier Detection (e.g., Weinauer 2019; Murphy et al. 2005)
  - Cluster Algorithms (e.g., Bergmann, Schuller, and Malter 2019; Haas and Winker 2014; Menold et al. 2013; Bredl, Winker, and Kötschau 2012)
  - Principal Component Analysis (e.g., Blasius and Thiessen 2013, 2012)

INTERVIEWER FALSIFICATION

Machine Learning and Interviewer Falsification

- Unsupervised Machine Learning
  - Data Mining and Outlier Detection (e.g., Weinauer 2019; Murphy et al. 2005)
  - Cluster Algorithms (e.g., Bergmann, Schuller, and Malter 2019; Haas and Winker 2014; Menold et al. 2013; Bredl, Winker, and Kötschau 2012)
  - Principal Component Analysis (e.g., Blasius and Thiessen 2013, 2012)
- Supervised Machine Learning
  - Regression Techniques (Li et al. 2009)
  - Tree-based Methods (Birnbaum et al. 2013)

INTERVIEWER FALSIFICATION

Machine Learning and Interviewer Falsification

- Unsupervised Machine Learning
  - Data Mining and Outlier Detection (e.g., Weinauer 2019; Murphy et al. 2005)
  - Cluster Algorithms (e.g., Bergmann, Schuller, and Malter 2019; Haas and Winker 2014; Menold et al. 2013; Bredl, Winker, and Kötschau 2012)
  - Principal Component Analysis (e.g., Blasius and Thiessen 2013, 2012)
- Supervised Machine Learning
  - Regression Techniques (Li et al. 2009)
  - Tree-based Methods (Birnbaum et al. 2013)

Lack of suitable data → Lack of literature using supervised machine learning → Lack of evaluation of different supervised algorithms

STUDIES DEALING WITH

INTERVIEWER FALSIFICATION

Real-World Data	Experimental Data
⊖ Seldom available	⊕ High intern validity
⊖ Few falsifications	⊕ Balanced falsification ratio
⊖ Uncertain falsification status	⊕ Certain falsification status
⊕ High external validity	⊖ Low external validity
⊕ Real conditions and motivations	⊖ Selective Groups (mostly Students)

STUDIES DEALING WITH

INTERVIEWER FALSIFICATION

Real-World Data	Experimental Data
⊖ Seldom available ⊖ Few falsifications ⊖ Uncertain falsification status ⊕ High external validity ⊕ Real conditions and motivations	⊕ High intern validity ⊕ Balanced falsification ratio ⊕ Certain falsification status ⊖ Low external validity ⊖ Selective Groups (mostly Students)

INTERVIEWER FALSIFICATION

Real-World Data	Experimental Data
⊖ Seldom available ⊖ Few falsifications ⊖ Uncertain falsification status ⊕ High external validity ⊕ Real conditions and motivations	⊕ High intern validity ⊕ Balanced falsification ratio ⊕ Certain falsification status ⊖ Low external validity ⊖ Selective Groups (mostly Students)

Combine real-world data with experimental data

DATA

IAB-BAMF-SOEP Survey of Refugees in Germany

- Annual longitudinal household panel (starting 2016)
- **Target population**: asylum-seekers and adult household members
- **Mode**: computer-assisted personal interviewing (CAPI)
- **Interviewer**: 98 trained interviewers
- **Falsifications**: 351 (7.3%) complete falsifications out of 4,816 interviews

(Brücker et al. 2016; Brücker et al. 2017; IAB 2017; Kosyakova et al. 2019; Haas and Winker 2016, 2014; Storfinger and Winker 2013; Menold et al. 2013)

DATA

IAB-BAMF-SOEP Survey of Refugees in Germany

- Annual longitudinal household panel (starting 2016)
- **Target population**: asylum-seekers and adult household members
- **Mode**: computer-assisted personal interviewing (CAPI)
- **Interviewer**: 98 trained interviewers
- **Falsifications**: 351 (7.3%) complete falsifications out of 4,816 interviews

Experimental Data

- 2011 conducted cross-sectional experiment at the University of Giessen, Germany
- **Respondents**: students from the University of Giessen
- **Mode**: paper-and-pencil interviews (PAPI) which were tape recorded
- **Interviewers**: 78 trained students
- **Falsifications**: 710 (50 %) complete falsifications out of 1,420 interviews

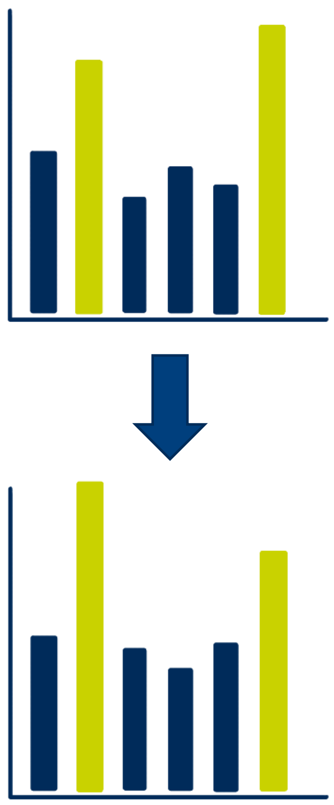
(Brücker et al. 2016; Brücker et al. 2017; IAB 2017; Kosyakova et al. 2019; Haas and Winker 2016, 2014; Storfinger and Winker 2013; Menold et al. 2013)

APPROACH

1. **Feature selection**:
   Identification of appropriate features available for both datasets ⇨ Falsification Indicators
2. **Dataset shifting**:
   Address possible problem of dataset shifting due to the different data sources
3. **Algorithm selection**:
   Identification of appropriate algorithms applicable in the context of binary classification problems
4. **Performance evaluation**:
   Training, testing and comparison of different model results
5. **Tuning models**:
   Improving model performance
6. **Final evaluation**

FEATURE SELECTION

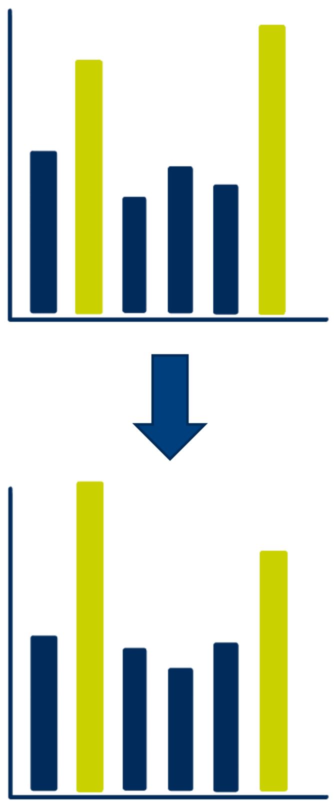
Aim: Identify comparable features between the different datasets, which allow a discrimination between falsified and real interviews



(Hood and Bushery 1997; AAPOR 2003; Bredl et al. 2012; Menold et al. 2013)

FEATURE SELECTION

Aim: Identify comparable features between the different datasets,
which allow a discrimination between falsified and real interviews



Falsification indicators …

… are derived from rational (answering) behaviors of falsifiers

… allow measurement of systematic differences between real and falsified data

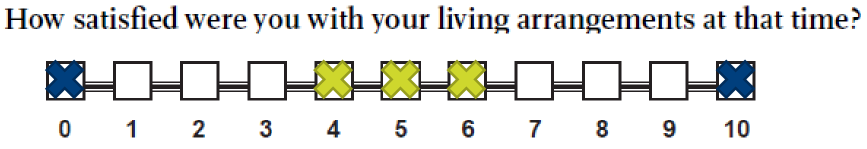
… are not easily manipulated by falsifiers

… are comparable between different datasets

(Hood and Bushery 1997; AAPOR 2003; Bredl et al. 2012; Menold et al. 2013)

FALSIFICATION INDICATORS

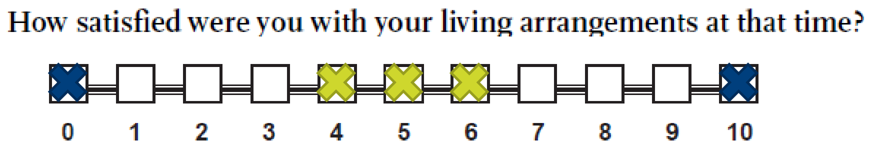
- **Extreme responses**: Lower share of extreme responses on rating scales for falsifiers
- **Middle responses**: Higher share of middle responses on rating scales for falsifiers



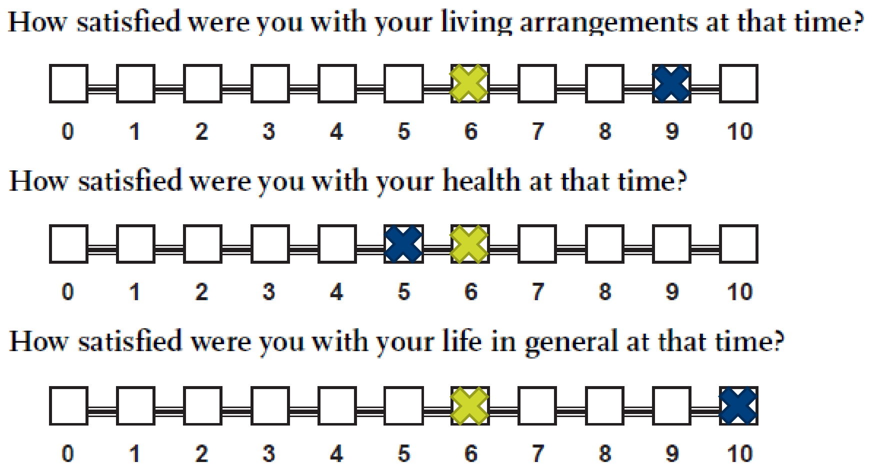
(Schäfer et al. 2005; Bredl et al. 2012)

FALSIFICATION INDICATORS

- **Extreme responses**: Lower share of extreme responses on rating scales for falsifiers
- **Middle responses**: Higher share of middle responses on rating scales for falsifiers



- **Non-Differentiation**: Lower standard deviation across item scales for falsifiers



(Schäfer et al. 2005; Bredl et al. 2012)

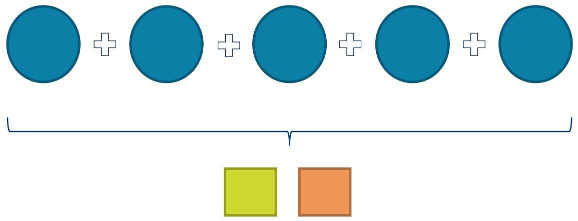
FALSIFICATION INDICATORS

Indicator	Abbreviation	Description
Acquiescent responding	ACQ	Share of positive connotation (“Agree/Strongly Agree”) independent of content
Benford’s Law	BFL	Decreasing distribution of leading digit for numeric quantities
Interview duration	DUR	Duration of completed interviews
Extreme responses	ERS	Share of extreme responses to rating scales
Item nonresponse	INR	Item nonresponse rate within an interviewer’s workload of closed-ended questions
Non-Differentiation	ND	Standard deviation within an item scale
Middle category responses	MRS	Share of middle responses to rating scales
Primacy effects	PRIM	Share of choosing the first two categories in non-ordered answer option lists
Recency effects	RECE	Share of choosing the last two categories in non-ordered answer option lists
Rounding	ROUND	Share of rounding numbers in numerical open-ended questions
Semi-Open responses	SOR	Share of responses to “other” in semi-open-ended question

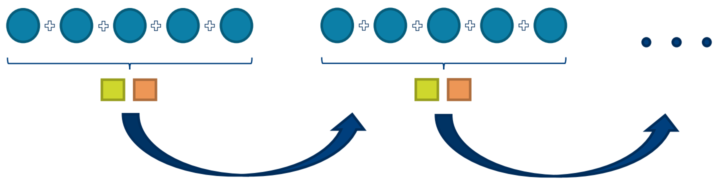
(Reuband 1990; Hood and Bushery 1997; Schäfer et al. 2005; Bredl et al. 2012; Menold et al. 2013)

ALGORITHMS – REGRESSION MODELS

- **Logistic Regression**
  Models the probability of the binary output (falsification status) by fitting a linear combination of input variables (features) into a logistic function



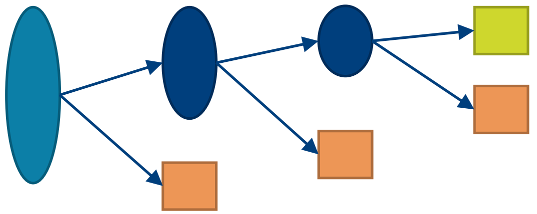
- **Boosted Logistic Regression**
  Ensemble of logistic regression models, sequentially applied to reweight the training data and prediction through weighted majority vote



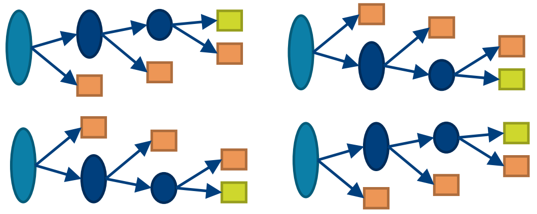
(Hastie et al. 2009; Friedman et al. 2000)

ALGORITHMS – TREE-BASE METHODS

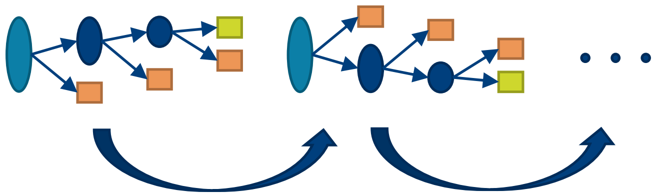
- **Simple Decision Tree**
  Processes input (features) by making a series of logical decisions comprised in different branches leading to the output (falsification status) according to the combination of decisions/splits



- **Random Forest**
  Ensemble of multiple Decision Trees with random feature selection for each Decision Tree



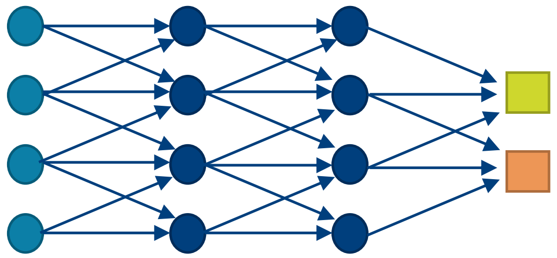
- **XGBoost (Tree Boosting)**
  Ensemble of multiple Decision Trees, sequentially applied to perform iterative optimization



(Lantz 2013; Hastie et al. 2009; Friedman et al. 2000)

ALGORITHMS – DEEP LEARNING

- **Neural Networks**
  Models connection between input (features) and output (falsification status) by weighting the input according to an activation function and processing the weighted information through (multiple) nodes and layers

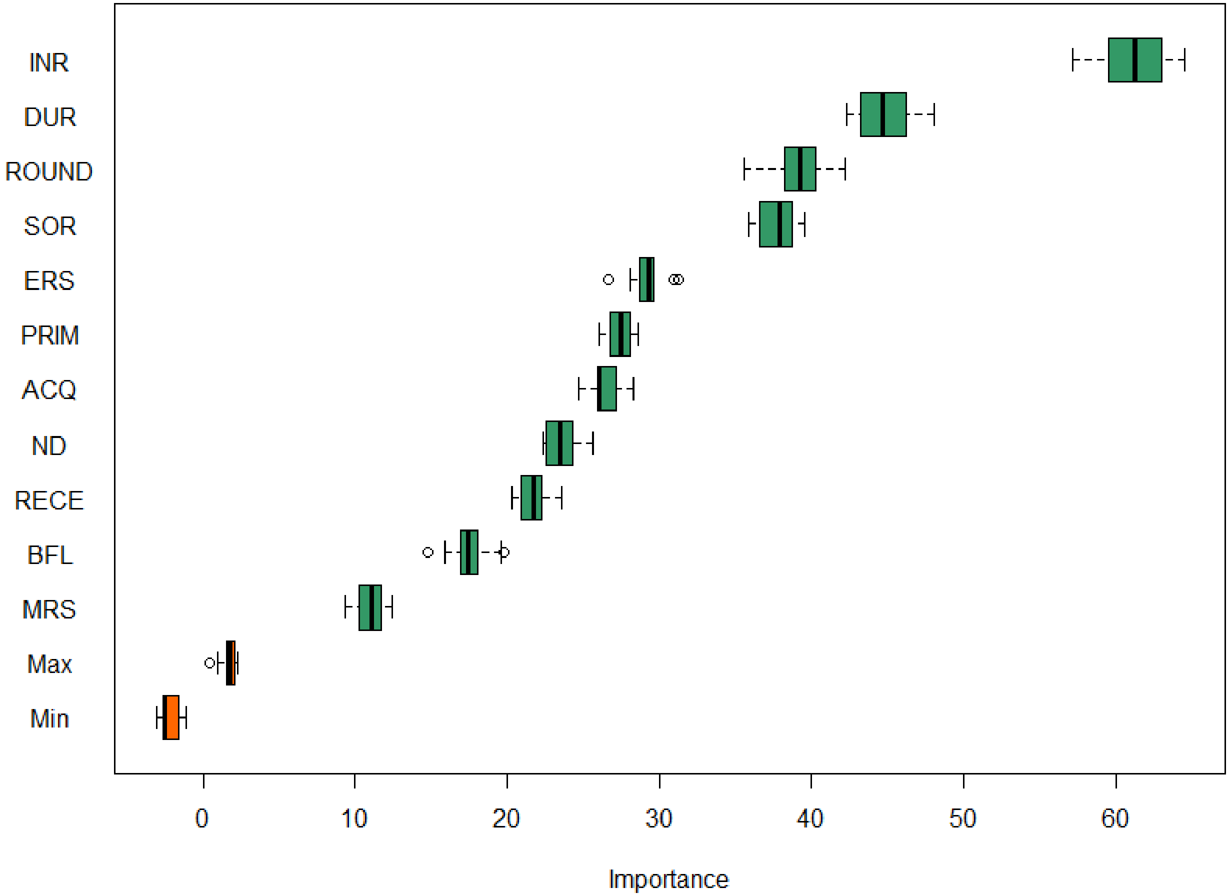


(Lantz 2013; Hastie et al. 2009)

PRELIMINARY RESULTS

FEATURE SELECTION

Feature importance according to Boruta-Algorithm



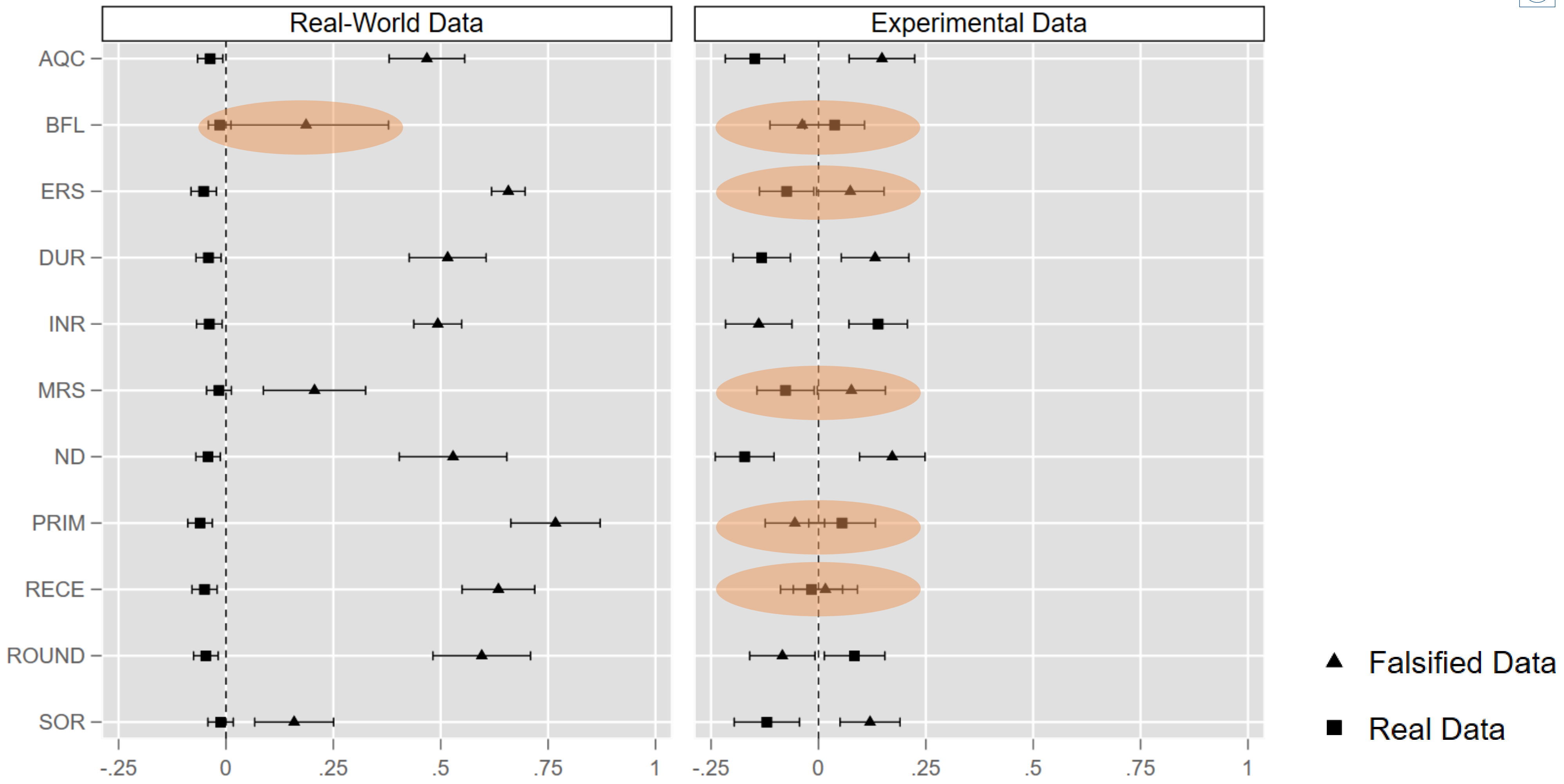
Boruta-Algorithm =
wrapper algorithm built on Random Forest

- Iterations of different feature combinations
- Defines feature importance for the model accuracy of Random Forest
- Adds randomness, by creating mixed copies of features

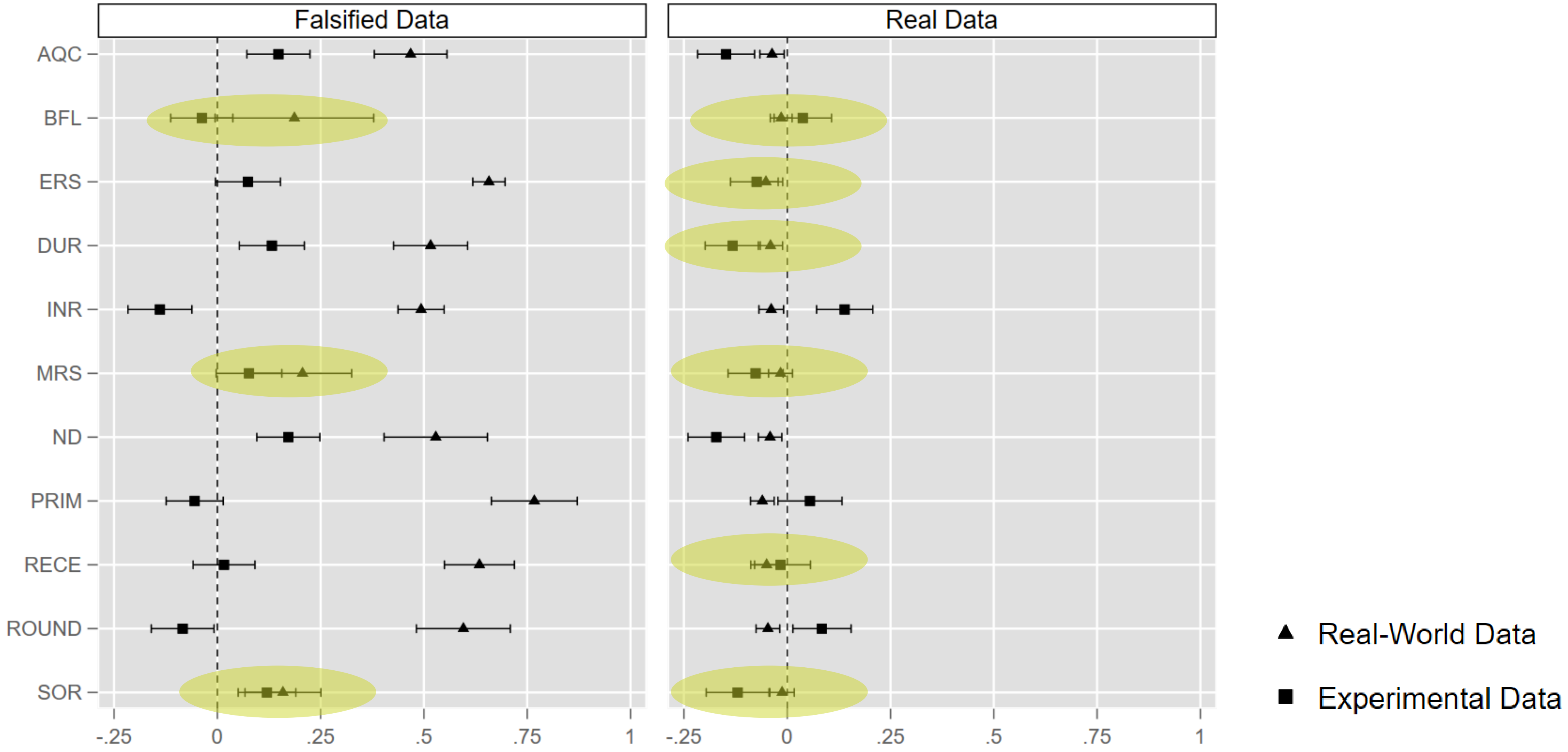
Advantage:

Captures all circumstances in which a feature is important

DATASET SHIFTING

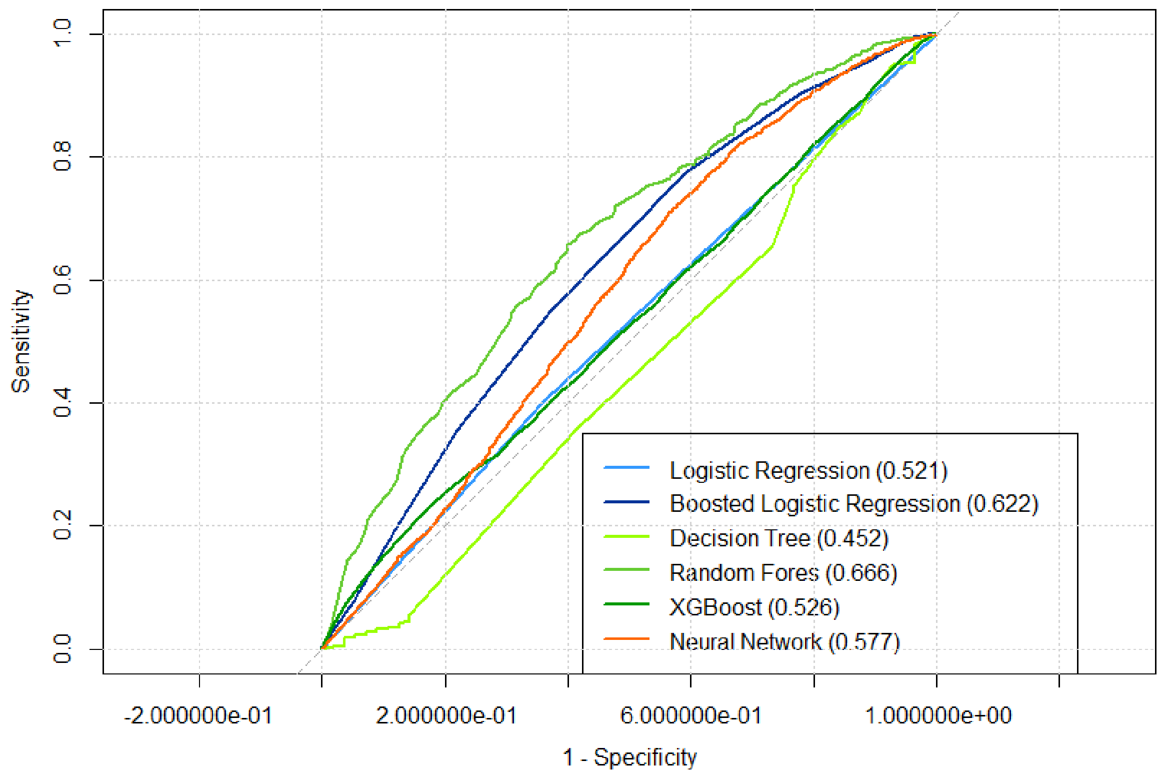


DATASET SHIFTING



COMPARISON OF ALGORITHMS

Training Data: ROC (Receiver Operating Characteristic) curve comparing different algorithms



⇨ **ROC curve** visualizes tradeoff between sensitivity and specificity

Sensitivity:

Proportion of real interviews, correctly classified as real interviews

Specificity:

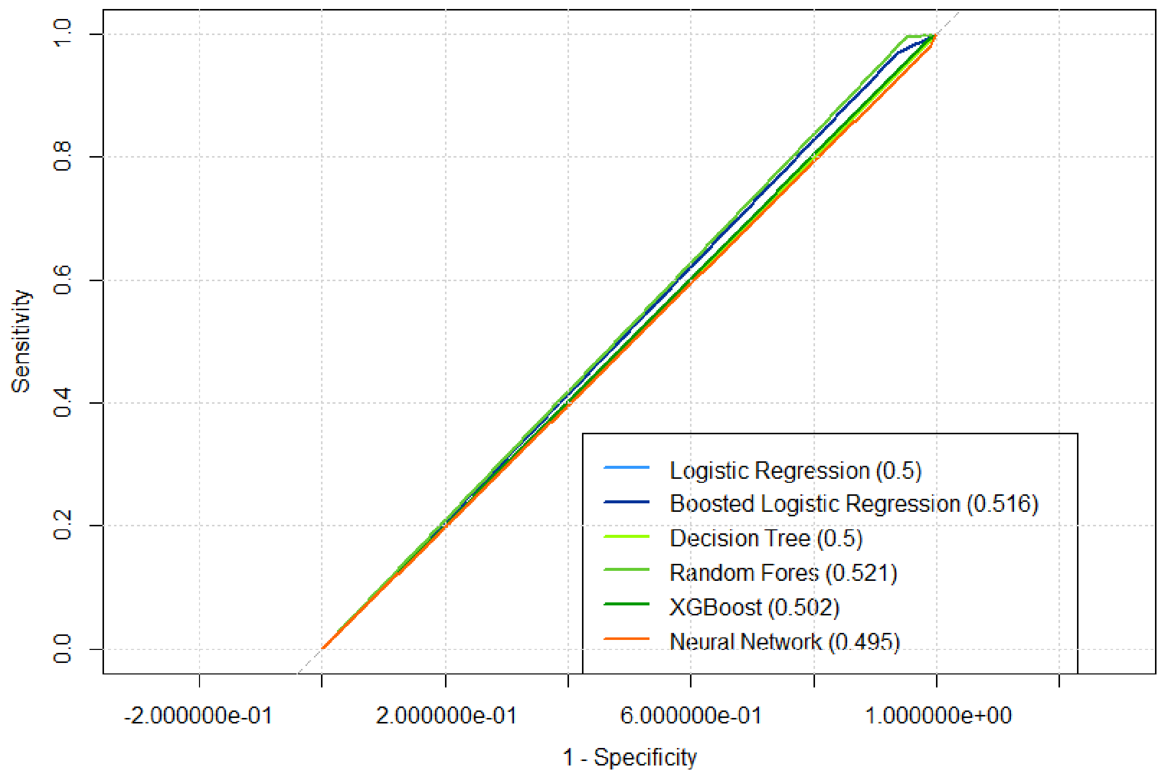
Proportion of falsifications, correctly classified as falsifications

AUC:

Area under the ROC curve

COMPARISON OF ALGORITHMS

Test Data: ROC (Receiver Operating Characteristic) curve comparing different algorithms



⇨ **Best Performance:**
Random Forest

⇨ **Possible problems:**
1. Unbalanced data
2. Overfitting
3. Dataset shifting

OUTLOOK AND DISCUSSION

- Still many problems to address:
  - Features → can we increase the number of features?
  - Data shifting → which form data shifting is important for us?
  - Algorithms → which algorithms should we add?
  - Tuning → how can we tune the models without running into overfitting?

- Further starting points for research:
  - Falsification forms → can we simulate further falsifications forms (e.g. partial falsifications)?
  - Falsification share → what happens if we change the share of falsifications?

CONTACT

Silvia Schwanhäuser

✉ Silvia.Schwanhaeuser2@iab.de

☎ +49 911 179 2770

LITERATURE

- AAPOR – American Association for Public Opinion Research (2003): Interviewer Falsification in Survey Research: Current Best Methods for Prevention, Detection and Repair of Its Effects. American Association for Public Opinion Research (AAPOR). Online available https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/falsification.pdf.
- Bergmann, M.; Schuller, K.; Malter, F. (2019): Preventing Interview Falsifications During Fieldwork in the Survey of Health, Ageing and Retirement in Europe (SHARE). Longitudinal and Life Course Studies, 10(4): 513–30.
- Birnbaum, B.; Borriello, G.; Flaxman, A.D.; DeRenzi, B.; Karlin A.R. (2013): Using Behavioral Data to Identify Interviewer Fabrication in Surveys.
- Blasius, J.; Thiessen, V. (2012): Assessing the Quality of Survey Data. London: Sage.
- Blasius, J.; Thiessen, V. (2013): Detecting Poorly Conducted Interviews. In: Peter Winker, Natalja Menold und Rolf Porst (Ed.): Interviewers' Deviations in Surveys: Peter Lang, 67–88.
- Bredl, S.; Winker, P.; Kötschau, K. (2012): A statistical approach to detect interviewer falsification of survey data. Survey Methodology - Statistics Canada 38(1): 1–10.
- Brücker, H.; Rother, N.; Schupp, J.; Babka von Gostomski, C.; Böhm, A.; Fendel, T.; Friedrich, M.; Giesselmann, M.; Holst, E.; Kosyakova, Y.; Kroh, M.; Liebau, E.; Richter, D.; Romiti, A.; Schacht, D.; Scheible, J.A.; Schmelzer, P.; Siegert, M.; Sirries, S.; Trübswetter, P.; Vallizadeh, E. (2016): IAB-BAMF-SOEP-Befragung von Geflüchteten: Flucht, Ankunft in Deutschland und erste Schritte der Integration. IAB-Kurzbericht, 2016(24).

LITERATURE

- Brücker, H.; Rother, N.; Schupp, J. (2017): IAB-BAMF-SOEP-Befragung von Geflüchteten 2016: Studiendesign, Feldergebnisse sowie Analysen zu schulischer wie beruflicher Qualifikation, Sprachkenntnissen sowie kognitiven Potenzialen. Berlin: DIW Berlin, German Institute for Economic Research.
- DeMatteis, J.M.; Young, L.J.; Dahlhamer, J.; Langley, R.E.; Murphy, J.; Olson, K.; Sharma, S. (2020): Falsification in Surveys. American Association for Public Opinion Research, September 28. Online available https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/AAPOR_Data_Falsification_Task_Force_Report.pdf.
- Friedman, J.; Hastie, T.; Tibshirani, R. (2000): Additive Logistic Regression: A Statistical View of Boosting (with discussion). Annals of Statistics 28: 337–07
- Haas, S. de; Winker, P. (2014): Identification of Partial Falsifications in Survey Data. IOS Press 30: 271–81.
- Haas, S. de; Winker, P. (2016): Detecting Fraudulent Interviewers by Improved Clustering Methods - the Case of Falsifications of Answers to Parts of a Questionnaire. Journal of Official Statistics 32: 1.
- Hastie, T.; Tibshirani, R.; Friedman, J. (2009): The elements of statistical learning: Data Mining, Inference, and Prediction. Springer, New York.
- Hood, C.; Bushery, J.M. (1997): Getting More Bang from the Reinterview Buck. Identifying "At Risk" Interviewers. Proceedings of the Survey Research Method Section, American Statistical Association, 820–24.
- IAB – Institute for Employment Research. (2017): Revidierter Datensatz Der IAB-BAMF-SOEP- Befragung von Geflüchteten. Nuremberg.

LITERATURE

- Kosyakova, Y.; Olbrich, L.; Sakshaug, J.; Schwanhäuser, S. (2019): Identification of interviewer falsification in the IAB-BAMF-SOEP Survey of Refugees in Germany. FDZ-Methodenreport. Institute for Employment Research (IAB).
- Lantz, B. (2013): Machine Learning with R: Expert techniques for predictive modeling. Pack Publishing Ltd.
- Li, J.; Brick, M.; Tran, B.; Singer, P. (2009): Using Statistical Models for Sample Design of a Reinterview Program. Proceedings of the Survey Research Method Section, American Statistical Association, 4681–95.
- Menold, N.; Winker, P.; Storfinger, N.; Kemper, C.J. (2013): A Method for Ex-Post Identification of Falsification in Survey Data. In: : Peter Winker, Natalja Menold und Rolf Porst (Ed.): Interviewers' Deviations in Surveys: Peter Lang, 25–47.
- Murphy, J.; Eyerman, J.; McCue, C.; Hottinger, C.; Kennet, J. (2005). Interviewer Falsification Detection Using Data Mining.
- Reuband, K.-H. (1990): Interviews, Die Keine Sind: "Erfolge" Und "Mißerfolge" Beim Fälschen Von Interviews. Kölner Zeitschrift für Soziologie und Sozialpsychologie 42: 706–33.
- Schäfer, C.; Schräpler, J.P.; Müller, K.R.; Wagner, G.G. (2005): Automatic Identification of Faked and Fraudulent Interviews in Surveys by Two Different Methodes. Discussion Paper. Deutsches Institut für Wirtschaftsforschung (DIW).
- Storfinger, N.; Winker, P. (2013): Assessing the Performance of Clustering Methods in Falsification using Bootstrap. In: Peter Winker, Natalja Menold und Rolf Porst (Ed.): Interviewers' Deviations in Surveys: Peter Lang, 46–65.
- Weinauer, M. (2019). Be a Detective for a Day: How to Detect Falsified Interviews with Statistics. Statistical Journal of the IAOS 35(4): 569–75.

APPENDIX

APPENDIX | FALSIFICATION INDICATORS

Differences between real data and falsified data, separate for experimental data and real-world data

Mean	Experimental Data			Real-World Data		
	Falsified	Real	Diff. Group Mean	Falsified	Real	Diff. Group Mean
ACQ	0.15	-0.15	**0.30 (0.000)**	0.47	-0.04	**0.51 (0.000)**
BFL	-0.04	0.04	-0.08 (0.150)	0.19	-0.02	**0.20 (0.000)**
DUR	0.07	-0.07	**0.15 (0.004)**	0.66	-0.05	**0.71 (0.000)**
ERS	0.13	-0.13	**0.26 (0.000)**	0.52	-0.04	**0.56 (0.000)**
INR	-0.14	0.14	**-0.28 (0.000)**	0.49	-0.04	**0.53 (0.000)**
MRS	0.08	-0.08	**0.15 (0.004)**	0.21	-0.02	**0.22 (0.000)**
ND	0.17	-0.17	**0.34 (0.000)**	0.53	-0.04	**0.57 (0.000)**
PRIM	-0.06	0.06	-0.11 (0.039)	0.77	-0.06	**0.83 (0.000)**
RECE	0.02	-0.02	0.03 (0.544)	0.63	-0.05	**0.68 (0.000)**
ROUND	0.08	0.08	**-0.17 (0.002)**	0.60	-0.05	**0.64 (0.000)**
SOR	0.12	-0.12	**0.24 (0.000)**	0.16	-0.01	**0.17 (0.002)**