Predictive Modeling with Psychological Panel Data

Florian Pargent (Florian.Pargent@psy.lmu.de)
Ludwig-Maximilians-Universität München

Johannes Albert-von der Gönna (j.avdg@lrz.de)
Leibniz Supercomputing Center of the Bavarian Academy of Sciences and Humanities

Introduction

Why Predictive Modeling?

> *"Predictions in psychology are statements about the likelihood that a certain behavior will occur or that a given relationship will be found. [. . .] When different explanations are put forward to account for some behavior or relationship, they are usually judged by how well they can make accurate and comprehensive predictions."* *(Gerrig 2013)*

Yarkoni and Westfall (2017):

- predictive claims rarely evaluated by suitable statistics
- pose more psychological questions as predictive analyses

Aim of Study

- show how to use predictive modeling in psychological research:
  1. prediction is the applied goal
  2. exploratory research to identify patterns in data
- demonstrate with psychological panel data:
  - high quality samples
  - wide range of variables
  - accessible for scientific use

6 case studies: predicting demographics, political attitudes, and health-related variables with (most available) panel items

Workflow of Predictive Analyses

Preprocessing

- preselection of predictors
- coding scheme

Benchmarking

- dummy vs. linear vs. nonlinear models
- nested resampling
  - performance evaluation
  - hyperparameter tuning
  - missing value imputation

Model Interpretation

- final model fit on complete data
- variable importance

Methods

Dataset and Preprocessing

GESIS Panel Dataset (Bosnjak et al. 2018; GESIS 2017)

- representative sample of Germany
- 20 bimonthly surveys (February 2014 - June 2017)

Preprocessing

- remove administrative variables, metadata, items for quality assessment, open questions, task specific variables
- code special response categories as missing values
- remove panelists not participating in all waves
- remove variables with more than 1 SD of missing values

Final prediction tasks include 1569 – 2404 panelists and 1969 – 2341 predictor variables

Target Variables

Target	Statistics
Gender	female: 1222, male: 1182
Sick Days	none: 667, at least one: 902
Trump	very negative: 1164
	negative: 698
	neither nor: 390
	positive/very positive: 138
Income	M: 8.36, SD: 3.62, N: 2145
Life Satisfaction	M: 7.04, SD: 1.94, N: 2389
Sleep Satisfaction	M: 6.45, SD: 2.38, N: 2380

Note. Data coded from 1 to 15 for Income and from 0 to 10 for Life Satisfaction and Sleep Satisfaction.

Predictive Models

Featureless Learner

- classification: majority vote
- regression: mean prediction

Elastic Net (Zou and Hastie 2005)

- regularized linear model
- two hyperparamers (λ, α)

Random Forest (Breiman 2001)

- nonlinear model with complex interactions
- one hyperparameter (*mtry*)

Performance Measures

Binary Classification

- $MMCE = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$
- $Sensitivity$, $Specificity$

Ordinal Classification

- $MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$
- based on 4x4 confusion matrix

Regression

- $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$

Nested Resampling Strategy

Outer Resampling (Performance Evaluation)

- repeated crossvalidation: 10 folds, 3 repetitions

Inner Resampling (Hyperparameter Tuning)

- 10-fold crossvalidation
- grid tuning
- histogram imputation

Computational Resources

- more than 65 days of serial computing time
- CoolMUC-2 linux cluster at the LRZ
- *R* packages *mlr* (Bischl et al. 2016) and *batchtools* (Lang, Bischl, and Surmann 2017)

Results

Benchmark Results: Classification

	Featureless	Elastic Net	Random Forest
Gender			
MMCE	0.49	0.04	0.05
SD(MMCE)	0.00	0.01	0.01
SENS	1.00	0.95	0.95
SPEC	0.00	0.96	0.94
Sick Days			
MMCE	0.43	0.39	0.39
SD(MMCE)	0.00	0.03	0.03
SENS	0.00	0.27	0.27
SPEC	1.00	0.86	0.87
Trump			
MAE	0.79	0.65	0.70
SD(MAE)	0.00	0.03	0.02
MMCE	0.51	0.49	0.49

Benchmark Results: Regression

	Featureless	Elastic Net	Random Forest
Income			
MSE	13.14	5.68	5.65
SD(MSE)	0.93	0.73	0.64
R-squared	-0.01	0.56	0.57
Life Satisfaction			
MSE	3.76	1.98	2.03
SD(MSE)	0.53	0.31	0.30
R-squared	0.00	0.47	0.46
Sleep Satisfaction			
MSE	5.69	2.11	2.27
SD(MSE)	0.46	0.34	0.34
R-squared	-0.01	0.63	0.60

Variable Importance: Gender

IMP	Name
-0.50	Height in cm
0.50	Shaving: Legs
-0.49	Height in cm
-0.43	Affinity for technology
-0.34	Personal income
-0.28	Weight
-0.27	Weight
0.25	Care products: Makeup, incl. o.e.
0.25	Care products: Makeup
-0.22	Shaving: Face

Note. Tuned alpha = 0.21. Nonzero coefficients = 264.

Variable Importance: Trump

IMP1	IMP2	IMP3	IMP4	Name
0.04	0.02	-0.01	-0.06	Candidate orientation: Cem Oezdemir
-0.08	0.00	0.00	0.04	Vote for: AfD
-0.01	-0.03	0.00	0.05	Country of birth (GER, EU, other)
-0.03	0.00	0.00	0.05	Satisfaction with democracy (-)
0.00	0.03	0.00	-0.05	Candidate orientation: Sigmar Gabriel
-0.07	0.00	0.00	0.00	Attitude towards Islam: constrained practice
0.03	0.00	0.00	-0.04	Candidate orientation: Angela Merkel
0.04	0.00	0.00	-0.03	Trust in newspapers
-0.04	0.00	0.00	0.03	Foreigners should marry own nationality
-0.05	0.00	0.01	0.00	Federal state (GER), west/east

Note. Tuned alpha = 0.08. Nonzero coefficients = 369.

Discussion

Summary of Results

- high predictive performance for some targets (e.g. *Gender*)
- low or near chance performance for others (e.g. *Sick Days*)

-> **promising for applied prediction**

- important predictors highly plausible, even for targets with low estimated performance
  - e.g. *Trump*: top 10 include familiar topics, automatically selected by the *elastic net* out of 2000+ predictors

-> **promising for exploratory research**

- with the current setup, no performance gain with nonlinear models (similar performance for *elastic net* and *random forest*)

Keep Psychologists Competitive

- high demand for predictive solutions by practicioners:
  - personnel selection
  - detection/treatment of mental disorders
  - marketing
  - ...
- empower psychologists to cooperate with computer/data scientists on psychological research questions:
  - personality prediction from facebook (Segalin et al. 2017)
  - depression markers from instagram (Reece and Danforth 2017)
- teach (basic) predictive methods:
  - machine learning techniques
  - programming skills

Thanks! Questions?

|

Appendix

Variable Importance: Sick Days

IMP	Name
0.08	Satisfaction: Work
0.07	Health insurance
-0.07	Year of birth
-0.06	Important in life: Family
-0.06	Social contacts constrained
-0.04	Social contacts constrained
-0.04	Physical pain
-0.04	Importance: Leisure time
0.04	Comparator finances
0.04	Paying rent/mortgage on time

Note. Tuned alpha = 0.21. Nonzero coefficients = 86.

Variable Importance: Income

IMP	Name
-0.65	Gender
0.52	Household income
0.38	Household income
0.36	Vocational or professional training
-0.35	Employment situation
0.35	Number of registered cars
0.24	Satisfaction: Income
0.19	Extra money per month for sustainable energy
-0.18	Household size (one person, more than one)
0.18	Self-comparison (GER): financial wealth

Note. Tuned alpha = 0.59. Nonzero coefficients = 70.

Variable Importance: Life Satisfaction

IMP	Name
0.09	Positive life changes
-0.08	General standard of living: feel good (-)
0.08	Feeling: Enjoyed life
0.07	Feeling: Enjoyed life
0.06	Important in life: Family
-0.06	Social contacts constrained
-0.06	Feeling: Depressed
0.05	Feeling: Relaxed
-0.05	Overall living standard (-)
-0.05	Self-description: far away from everything

Note. Tuned alpha = 0.10. Nonzero coefficients = 82.

Variable Importance: Sleep Satisfaction

IMP	Name
0.48	Satisfaction: Sleep
0.28	Satisfaction: Sleep
0.24	Satisfaction: Sleep
0.21	Satisfaction: Sleep
0.15	Satisfaction: Sleep
0.14	Satisfaction: Sleep
0.12	Satisfaction: Sleep
0.09	Satisfaction: Sleep
0.08	Satisfaction: Sleep
0.05	Satisfaction: Sleep

Note. Tuned alpha = 0.54. Nonzero coefficients = 14.

Quellen I

Bischl, Bernd, Michel Lang, Lars Kotthoff, Julia Schiffner, Jakob Richter, Erich Studerus, Giuseppe Casalicchio, and Zachary M. Jones. 2016. "mlr: Machine Learning in R." *Journal of Machine Learning Research* 17 (170): 1–5. http://jmlr.org/papers/v17/15-066.html.

Bosnjak, Michael, Tanja Dannwolf, Tobias Enderle, Ines Schaurer, Bella Struminskaya, Angela Tanner, and Kai W. Weyandt. 2018. "Establishing an Open Probability-Based Mixed-Mode Panel of the General Population in Germany: The Gesis Panel." *Social Science Computer Review* 36 (1): 103–15. doi:10.1177/0894439317697949.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32. doi:10.1023/A:1010933404324.

Gerrig, R. J. 2013. *Psychology and Life*. 20th ed. Boston: Pearson.

GESIS. 2017. *GESIS Panel - Standard Edition* (version 21.0.0, Data file ZA5665). GESIS Data Archive: Cologne. doi:10.4232/1.12829.

Quellen II

Lang, Michel, Bernd Bischl, and Dirk Surmann. 2017. "Batchtools: Tools for R to Work on Batch Systems." *The Journal of Open Source Software* 2 (10). doi:10.21105/joss.00135.

Reece, Andrew G., and Christopher M. Danforth. 2017. "Instagram Photos Reveal Predictive Markers of Depression." *EPJ Data Science* 6 (1): 15. doi:10.1140/epjds/s13688-017-0110-z.

Segalin, Cristina, Fabio Celli, Luca Polonio, Michal Kosinski, David Stillwell, Nicu Sebe, Marco Cristani, and Bruno Lepri. 2017. "What Your Facebook Profile Picture Reveals About Your Personality." In *Proceedings of the 2017 Acm on Multimedia Conference*, 460–68. MM '17. New York, NY, USA: ACM. doi:10.1145/3123266.3123331.

Quellen III

Yarkoni, Tal, and Jacob Westfall. 2017. "Choosing Prediction over Explanation in Psychology: Lessons from Machine Learning." *Perspectives on Psychological Science* 12 (6): 1100–1122. doi:10.1177/1745691617693393.

Zou, Hui, and Trevor Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2). Blackwell Publishing Ltd: 301–20. doi:10.1111/j.1467-9868.2005.00503.x.