Title: Machine Learning Methods for Modeling in Psychology

Instructors: David Goretzko & Philipp Sterner

Abstract:

In this workshop, we will introduce the basics of machine learning (ML) modeling with a specific focus on applications in psychological research and assessment. After a general introduction clarifying key concepts, we will discuss a) different resampling strategies to ensure a valid performance evaluation of complex modeling pipelines, b) tree-based supervised learning algorithms as state-of-the-art approaches for tabular data, c) regularization to avoid overfitting and boost generalizability, d) hyperparameter-tuning to optimize the performance of ML algorithms, e) different performance metrics for classification and regression tasks, and f) how to set up benchmarks for model selection.

Furthermore, we will present two extensions to ML modeling that are particularly relevant for psychologists – interpretable machine learning (IML) and cost-sensitive learning (CSL). While complex ML algorithms often achieve high predictive accuracy, they frequently produce "black- box" models that lack human interpretability. IML addresses this by offering tools such as local explanations, feature importance metrics, and visualizations of marginal effects, making these models more understandable. In high-stakes psychological contexts, such as suicide prevention, classification errors carry unequal consequences (e.g., regarding money spent, time lost, suffering for patients). Traditional ML models – optimized purely with regard to accuracy – fail to consider these differing misclassification costs. CSL provides methods to account for varying costs, making it a critical approach for mitigating the impact of misclassifications.

Throughout the course, participants will apply the discussed concepts and modeling strategies through hands-on exercises in R. These exercises will guide them through the entire ML modeling cycle, including setting up key components of a modeling pipeline, evaluating various algorithms in a benchmarking experiment, selecting the most suitable model, and interpreting its predictions. The exercises will also contain practical applications of IML and CSL methods, showcasing how and where these methods can be implemented in the ML modeling process.

Prerequisites:

(1) Familiarity with R software for data analysis and

(2) A good understanding of (parametric) regression modeling. Participants are asked to bring their own laptops.

Assignment: Active participation

Credits: 4 workshop days

Literature:

We do not expect that you prepare readings in advance. Much of the workshop will be based on the following articles and books.

Bischl, B., Sonabend, R., Kotthoff, L., & Lang, M. (Eds.). (2024). *Applied machine learning using mlr3 in R*. CRC Press.

Hastie, T. (2009). The elements of statistical learning: data mining, inference, and prediction.

Henninger, M., Debelak, R., Rothacher, Y., & Strobl, C. (2023). Interpretable machine learning for psychological research: Opportunities and pitfalls. *Psychological Methods*. https://doi.org/10.1037/met0000560

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103). Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Pargent, F., Schoedel, R., & Stachl, C. (2023). Best practices in supervised machine learning: A tutorial for psychologists. *Advances in Methods and Practices in Psychological Science*, *6*(3), 25152459231162559.

Probst, P., Boulesteix, A. L., & Bischl, B. (2019). Tunability: Importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research*, *20*(53), 1-32.

Sterner, P., Goretzko, D., & Pargent, F. (2023). Everything has its price: Foundations of cost-sensitive machine learning and its application in psychology. *Psychological Methods*. https://doi.org/10.1037/met0000586

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, 14(4), 323–348. https://doi.org/10.1037/a0016973