Original Article

Predictive Modeling With Psychological Panel Data

Florian Pargent[1] and Johannes Albert-von der Gönna[2]

[1]Department of Psychology, LMU, Munich, Germany
[2]Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities (LRZ), Garching, Germany

Abstract: Longitudinal panels include several thousand participants and variables. Traditionally, psychologists analyze only a few variables – partly because common unregularized linear models perform poorly when the number of variables (p) approaches the number of observations (N). Predictive modeling methods can be used when $N \approx p$ situations arise in psychological research. We illustrate these techniques on exemplary variables from the German GESIS Panel, while describing the choice of preprocessing, model classes, resampling techniques, hyperparameter tuning, and performance measures. In analyses with about 2,000 subjects and variables each, we predict panelists' gender, sick days, an evaluation of US President Trump, income, life satisfaction, and sleep satisfaction. Elastic net and random forest models were compared to dummy predictions in benchmark experiments. While good performance was achieved, the linear elastic net performed similar to the nonlinear random forest. Elastic nets were refitted to extract the ten most important predictors. Their interpretation validates our approach, and further modeling options are discussed. Code can be found at https://osf.io/zpse3/.

Keywords: predictive modeling, machine learning, elastic net, random forest, panel data

In various scientific fields and for many real-world applications, it is increasingly not only the availability of growing sets of data, but also the use of data-driven, computationally demanding analysis strategies that allow for novel, valuable insights. Researchers in psychology have to date predominantly applied null hypothesis testing in the context of unregularized linear models to rather small datasets. This can be attributed to a strong focus on a framework termed "explanatory modeling" (Shmueli, 2010). Despite common claims, it is questionable whether this approach allows for accurate predictions of future experiences and behavior (Yarkoni & Westfall, 2017), a task which is also considered an integral part of psychological science. As new data sources get unlocked and increasing amounts of data become available, psychologists should embrace statistical techniques that allow for improved analyses when answering predictive research questions. Although most psychological research will probably not deal with millions of observations in the near future, a certain and imminent "big data" scenario is the number of variables (p) approaching the number of observations (N). Some of the richest data currently available to psychologists originates from open probability-based panels like the GESIS Panel in Germany (Bosnjak et al., 2018). Such longitudinal panels offer high-quality datasets from a big representative sample of the general population, including several thousand participants as well as variables ($N \approx p$). Along with longitudinal core studies, open research panels include additional studies submitted by researchers from various disciplines from the social, political, and economic sciences. Traditionally, only small subsets of the variables included in these panels are considered in the majority of scientific studies (e.g., White, Alcock, Wheeler, & Depledge, 2013). However, researchers might sometimes want to consider a larger number of predictors (e.g., hundreds of different questionnaire items), especially when addressing predictive research questions. Methods from machine learning and predictive modeling provide efficient techniques to deal with this type of high-dimensional information when theory-driven variable reduction is not intended. Adopting these tools and discovering new research strategies beyond the scope of the current methodology would enable psychologists to perform panel analyses that include a large number of available variables. That might lead to unforeseen discoveries and novel insights, which could guide future explanatory modeling attempts.

In this paper, we will demonstrate how predictive modeling methods can be used to analyze psychological panel data. Referring to Kuhn and Johnson (2013), we define predictive modeling as the process of developing and evaluating a statistical model that generates an accurate prediction of some target variable, based on a series of predictor variables. While unregularized linear regression models are predictive models, their predominant usage in psychological research differs from the predictive modeling approach in two important ways:

(1) Bias–variance tradeoff: During their methods training, most psychologists get familiarized with the idea that statistical models should always approximate the underlying data generating process. Assuming the true relationship between variables is linear, unregularized linear regression allows for an unbiased estimation of population parameters which is useful for the development and testing of psychological theories. What many psychological researchers do not know is that an unbiased property can be a disadvantage when the modeling goal is to achieve a maximum of predictive performance. It can be shown that the expected prediction error of a predictive model is a function of both the bias (how well can a model, on average, approximate the true relationship?) and the variance (how stable are the predictions when using different random samples for model estimation?). In the machine learning literature, this phenomenon is described as the "bias–variance tradeoff" (Geman, Bienenstock, & Doursat, 1992; James, 2003), as higher predictive accuracy can often be achieved by slightly increasing the bias in favor of a larger reduction in variance (or the other way around). For this reason, the field of machine learning has developed predictive algorithms that minimize a combination of bias and variance, two of which are introduced in the Methods section.
(2) Performance evaluation: In psychology, statistical models are typically assessed based on some measure of model fit (e.g., in-sample R^2, AIC/BIC, χ^2 model tests) evaluated on the same data used for model estimation (Yarkoni & Westfall, 2017). While this might be an appropriate strategy to find a model whose internal structure is a good representation of the unknown data generating process in the population, it is insufficient to evaluate predictive performance. For a realistic estimate of how accurate the model can predict new observations, resampling methods like cross-validation have to be used, which will be introduced later.

In the context of panel data, a predictive framework can be used to pursue different scientific goals. The most direct use of predictive modeling is in applied work, where predictions of new observations are the ultimate goal (e.g., suicide prevention, selection of personnel). It has been noted that longitudinal panel data in contrast to cross-sectional or time series data has the potential for more accurate prospective predictions of individual outcomes, by pooling information across participants (Hsiao, 2007). Unfortunately, predictive models based on panel data are generally of no direct use for practitioners, as responses to the large variety of panel questions which would be needed to compute real-world predictions are never available in concrete applications in educational, organizational, or clinical psychology. However, it has been suggested that insights of great theoretical interest can be gained with predictive methods by reframing research questions focused on understanding human experience and behavior (Yarkoni & Westfall, 2017). From a statistical perspective, Shmueli (2010) notes that predictive modeling can serve many functions for the purpose of developing and testing theories. The following might be especially relevant for psychological panels:

Because of the large number of variables combined with big representative samples and a longitudinal structure, panel data have a great capacity for constructing and testing sophisticated psychological hypotheses (Hsiao, 2007). Panel researchers have developed interpretable explanatory models that respect longitudinal structures as well as important confounding variables. However, deciding which variables to include in such models is not an easy task when facing thousands of possible predictors. Using predictive models in combination with variable importance measures (which quantify the impact of each predictor) can reveal important predictive variables. This might generate new hypotheses or suggest ways of improving existing models. Also, open probability panels include questions from different research projects. Thus, predictive modeling can be used to develop new questionnaire measures to predict interesting criteria in practice, by uncovering a small number of important variables from a full panel dataset (Chapman, Weiss, & Duberstein, 2016). Finally, predictive modeling can identify promising topics to develop new explanatory models. As Shmueli (2010) puts it, "knowledge of (un)predictability is a fundamental component of scientific knowledge." Depending on whether some phenomenon can be reliably predicted based on the large variety of variables in a panel dataset, constructing a reasonable explanatory model might be achievable or not.

No matter which goal is pursued, the workflow of any predictive analysis can usually be divided into three main parts: In a first step, the available dataset has to be processed in preparation of the actual analysis (i.e., preprocessing). To deal with the large number of variables included in panel data, different strategies are possible. Feature selection methods might be used to reduce the entire set of variables to a number that can be handled by the predictive model (Guyon & Elisseeff, 2003). Alternatively, some machine learning algorithms have internal feature selection mechanisms and can be used with the complete set of predictors. Longitudinal panel data are prone to include missing values as the same subjects are encouraged to take part in many surveys over a course of several years. Imputing missing values can thus be crucial to maximize predictive performance.

After preprocessing, the typical modeling process consists of comparing a series of algorithms in order to find

the best performing predictive model (i.e., benchmarking). This usually includes both linear and nonlinear methods from machine learning, as different model classes can vary with respect to their performance based on specific characteristics of the data. During this process, resampling methods are used to tune hyperparameters and to estimate predictive performance of the different models.

When the best performing predictive algorithm has been found, one is often also interested in the interpretation of the final model fitted to the complete dataset. In this context, interpretation refers to an understanding of how certain predictors influence the predictions resulting from the predictive model (Ribeiro, Singh, & Guestrin, 2016). Many nonlinear machine learning model classes can be hard to interpret. However, there exist an increasing number of tools to extract information about the importance of different predictor variables either globally for the whole population (Friedman, 2001) or locally for single observations (Goldstein, Kapelner, Bleich, & Pitkin, 2015; Ribeiro et al., 2016). All of these analysis steps can require significant computational resources with regards to processing time and memory consumption. It is therefore often necessary to rely on large-scale computing clusters or high-performance GPU infrastructure to complete all parts of the described analysis workflow in a reasonable amount of time.

In an attempt to convince psychologists of the great potential of predictive modeling for a large variety of research questions, we predict different target variables in a large-scale longitudinal panel. For these analyses, we will follow, document, and present the approach described above: We choose target variables from the categories demographic variables, health, political attitude, and life satisfaction to exemplify the procedure. As predictors, the available dataset provides a large number of survey items from different psychological scales as well as demographic variables. The primary goal of our analyses is to maximize the performance of generalizable predictions. All decisions occurring during preprocessing, model optimization, and model comparison, as well as the interpretation of the best performing models are subject to this requirement. When suitable, different competing approaches are discussed.

Methods

Dataset and Preprocessing

We used data of the GESIS Panel, an open probability-based mixed-mode (62% online, 38% by mail) longitudinal panel representing the adult German-speaking population permanently residing in Germany (Bosnjak et al., 2018). The analyzed standard edition of the dataset, which can be requested for scientific use from the GESIS Leibniz Institute for the Social Sciences (https://www.gesis.org/en/en/gesis-panel/gesis-panel-home/), included 7,592 variables of 7,599 panelists (GESIS, 2017). For all analyses, we conducted a series of preprocessing steps: The panel collects a large amount of metadata, administrative variables, and questions for quality assessment. We removed this data for our analyses, although some variables (e.g., time spent on each page of the online questionnaire) might be promising candidates for future predictive research. We also removed all variables from the first three survey waves, as they only include a subsample of the full panel. This leaves all content questions from the recruitment interview and the welcome survey of the panel in 2013, as well as content variables from 20 bimonthly survey waves collected between February 2014 and June 2017. We further removed questions with open response format and items that were not presented to all panelists (e.g., due to some sort of experimental manipulation or nested response structures). In a next preprocessing step, we removed all subjects that did not participate in all regular survey waves. Then, we removed all variables for which the percentage of missing values for the remaining participants was higher than one standard deviation above the mean. In addition to the official missing value scheme provided by the panel, special coding labels (mostly used for response categories such as "I do not know" or "does not apply") were also treated as missing values. Most survey responses in the panel are given on ordinal response scales. However, nominally scaled items (e.g., educational levels) are also numerically coded in the dataset, and there is no automatic way of distinguishing both types of data. As manually checking the structure of several thousand variables is impractical, we treated all predictors as numerical variables in our analyses. Deleting selected categories assured that survey items with a Likert-type scale format (e.g., coded from 1 to 5) were not distorted by some special category (e.g., coded with 96 or higher). Certain demographic variables in the panel were collected multiple times by different studies (e.g., being regarded important moderators in many psychological applications). While keeping the data from the recruitment interview, we manually removed all duplicate versions of gender and year of birth (assuming invariance over time).

Target Variables

In addition to the preprocessing steps described above, additional item responses were removed depending on the different target variables. For each analysis, panelists with missing values on the target variable had to be excluded. With an exception for the analysis of *Gender*,

we always excluded all items from the same and all later waves as the target variable. Prior to modeling, all predictor variables in the reduced datasets were standardized and constant variables were removed. The amount of missing responses in the predictor variables ranged from 2.67% to 3.39% across all targets. Missing values were imputed by random draws from a histogram based on the nonmissing values of the respective predictor variable. Descriptive statistics for all target variables are presented in Table 1. We also report official variable IDs. The first two letters of the ID represent the number of the respective panel wave in increasing order. For detailed variable descriptions, refer to the codebook which can be downloaded from the panel homepage (GESIS, 2017).

Table 1. Descriptive statistics of target variables

Target	Statistics
Gender	Female: 1,222, male: 1,182
Sick Days	None: 667, at least one: 902
Trump	Very negative: 1,164
	Negative: 698
	Neither nor: 390
	Positive/very positive: 138
Income	M = 8.36, SD = 3.62, N = 2,145
Life Satisfaction	M = 7.04, SD = 1.94, N = 2,389
Sleep Satisfaction	M = 6.45, SD = 2.38, N = 2,380

Note. For binary targets, the positive class in the classification task is reported first. Data was coded from 1 to 15 for Income and from 0 to 10 for Life Satisfaction and Sleep Satisfaction.

Gender

ID: a11d054a

For all panelists, gender was recorded by the interviewer in the recruiting telephone interview of the panel. The positive class in this retrospective binary classification task refers to female panelists. The final dataset included 2,404 panelists and 2,341 predictor variables.

Sick Days

ID: ebbm175a

In the last available wave collected from April to June 2017, panelists reported the number of certified sick days in the preceding 12 months. Although the exact number of days is not available in the standard version of the dataset, it is recorded whether the person reported at least one sick day or no sick days for the last year ("no sick days" is the positive class here). We removed all participants who reported that they did not have a job. The final dataset included 1,569 panelists and 2,201 predictor variables.

Trump

ID: eazz118a

Between February and April of 2017, panelists were asked how they evaluate the election of Donald Trump as President of the United States on a fully labeled 5-point Likert scale ("*very negative*," "*negative*," "*neither nor*," "*positive*," "*very positive*"). For the analyses, the categories "*positive*" and "*very positive*" were collapsed due to the low frequency of the highest category. The final dataset included 2,390 persons and 2,091 predictors.

Income

ID: dfzh055b

During December 2016 and February 2017, panelists reported their personal income (after taxation) on 16 increasing categories ranging from "under 300 €" to "5,000 € and more." We removed all reports on individual income from earlier waves. We also removed all panelists reporting in the target variable that they did not have their own income. The final dataset included 2,145 persons and 1,969 predictors.

Life Satisfaction

ID: ebaw248a

In the last panel wave, panelists rated the overall satisfaction with their life on an 11-point Likert scale with labeled endpoints ("*fully unsatisfied*," "*fully satisfied*"). We excluded earlier items with the words "satisfaction" or "happiness" in their variable name. The final dataset included 2,389 persons and 1,975 predictors.

Sleep Satisfaction

ID: ebaw222a

For 13 consecutive panel waves between April 2015 and June 2017, panelists reported how satisfied they were with their sleep on an 11-point Likert scale with labeled endpoints ("*fully unsatisfied*," "*fully satisfied*"). We used the responses from the last wave as target variable and kept all earlier versions of the question among the predictors. The final dataset included 2,380 persons and 2,201 predictors.

Machine Learning Algorithms

Elastic Net

Linear regression is probably the most widespread statistical method in the social sciences. Yet the precision of traditional unregularized linear models depends heavily on the ratio of available information (i.e., the sample size) to the number of predictors whose parameters need to be estimated (i.e., regression coefficients). As this ratio decreases, the estimates of unregularized linear models become highly unstable. By penalizing model coefficients, the variance of the resulting predictions can be greatly reduced at the cost

of a comparatively small increase in bias. Such regularized linear models like ridge regression and the LASSO (Tibshirani, 1996) were introduced to allow for accurate predictions based on thousands of predictors (even with $p > N$), by shrinking the size of regression coefficients toward zero. While ridge regression retains all predictors, the LASSO shrinks some coefficients all the way to zero, thereby potentially resulting in a sparser solution with a smaller number of "active" predictor variables. The elastic net combines both penalties, promising even better results, especially in cases with more predictor variables than observations and with groups of highly correlated predictors (Zou & Hastie, 2005). Importantly, coefficients for the remaining predictors can be interpreted similarly to unregularized ordinary least squares linear regression models, as the elastic net is still a linear model. This makes it a powerful, yet familiar tool for the psychological researcher.

Random Forest

While the elastic net only addresses linear relationships between variables and interaction terms are typically not included, the random forest is a highly nonlinear model class which has proven to be a rather versatile and accurate machine learning algorithm. A random forest (Breiman, 2001) combines the predictions of several hundreds of decision trees grown on bootstrap samples of the original dataset. This averaging process greatly increases precision by reducing the variance of the – quite unstable – predictions of single trees. With a high number of trees, the random forest generally achieves smooth representations of nonlinear relationships despite the underlying, simple tree structure. Moreover, the tree structure enables the random forest to capture complex nonlinear interactions between a high number of predictors. In contrast to other powerful algorithms (like support vector machines, gradient boosting, or neural networks), the random forest requires little tuning of additional hyperparameters while achieving comparable predictive performance in many scenarios. Although the effect of single predictor variables is not directly interpretable because of the nonlinearity of the algorithm and the lack of a simple prediction equation, it is still possible to compute variable importance measures quantifying the influence of each predictor. While early measures have been shown to overestimate the importance of variables with many unique values or high correlations with truly important predictors, new developments have been suggested to combat some of these issues (Nembrini, König, & Wright, 2018; Strobl, Boulesteix, Kneib, Augustin, & Zeileis, 2008; Strobl, Boulesteix, Zeileis, & Hothorn, 2007). Random forests can – in principle – work with missing values in the data, but existing software implementations do currently not offer the efficiency necessary to apply this approach to large-scale datasets.

Featureless Learner

In predictive modeling, model accuracy can often only be judged properly by comparing different algorithms. A simple method to evaluate whether the performance of an algorithm can be considered "above chance" is by comparison with a baseline "dummy" algorithm. Such an algorithm should not use any information included in the predictor variables and therefore only make uninformed guesses. This featureless learner will constantly predict the mean of the target variable in the training set for regression settings, while the label of the majority class in the training set is predicted for classification tasks.

Nested Resampling

The standard approach in psychology to assess the fit of regression models is by use of the coefficient of determination R^2. This estimate for the fraction of variance explained by a linear regression model is most commonly computed on the observed data. In cases, where model predictions are supposed to generalize for an assumed population or when the model should allow for future predictions with new sample data, in-sample R^2 tends to overestimate the predictive out-of-sample accuracy of the model (Larson, 1931; Yin & Fan, 2001).

With an increase of computational resources, it is feasible to better estimate the performance of statistical models on new data by means of resampling (i.e., the repeated use of random subsamples of the observed data). This has become the standard approach in machine learning. The most common resampling strategy is tenfold cross-validation (10-CV; Kohavi, 1995). For a guide on when to use different resampling strategies, see Bischl, Mersmann, Trautmann, and Weihs (2012). When applying 10-CV, the complete dataset is randomly divided into 10 equally sized parts. Each subsample serves as a test set which is predicted based on a model estimated on a training set, consisting of the combined sample of the remaining nine parts. The average of the 10 resulting performance estimates is used as an estimate for the full model based on the whole dataset. To further reduce the variance of the performance estimation, cross-validation can be repeated multiple times to average the performance across cross-validation estimates for different partitions of the data.

The presence of hyperparameters (i.e., parameters whose values have to be determined prior to model estimation) in most machine learning algorithms further complicates the correct estimation of the out-of-sample predictive performance for the final model which is based on the whole sample. Suitable hyperparameter values have to be identified in a data-driven way before the final predictive model can be estimated with a given algorithm. If hyperparameter

values were to be determined just once for the whole dataset and then applied in all CV iterations, predictive performance would typically be overestimated, as all observations in the test sets already have been used to find the optimal values. When the final model would be used to predict new and unobserved samples, it has to be expected to perform worse than what has been estimated by the resampling scheme in which the hyperparameters were overly adapted to the respective test observations. This bias can be avoided by repeating the tuning of hyperparameters in each iteration of the resampling process, which is sometimes termed "nested resampling" (Bischl et al., 2012). The outer resampling loop is used to evaluate predictive performance. Then, an additional inner resampling loop is created within each of the outer model estimations to optimize hyperparameter values for the respective fold. The outer resampling scheme simulates the performance on unseen data, which is comparable to the algorithm later being trained on the whole dataset with hyperparameter values tuned by the same strategy as being used in the inner resampling. In the present study, 10-CV with 3 repetitions was used to gain an estimate of the predictive performance of the models and simple 10-CV to guide the tuning of hyperparameters. Thus, before making predictions for one test set in the outer resampling loop, the corresponding training set is again divided into ten parts. The algorithm is fitted with certain hyperparameter configurations on each of the ten inner training sets and predictive performance is measured on each of the ten inner test sets. To make predictions for the outer test set, the model is then trained on the outer training set with the specific combination of hyperparameter values which achieved the best average performance on the inner test sets.

The same partitions of the data were used for all algorithms in the outer resampling during the benchmark process to ensure comparable performance estimates. For binary and ordinal classification tasks, resampling in the inner and outer loop was stratified, which guarantees the class distributions of the whole sample to be reproduced in each fold.

A grid with 40 different values was used when tuning the α parameter of the elastic net and the *mtry.perc* parameter of the random forest. The α parameter determines the mixing of the elastic net penalty. The LASSO penalty is returned for $\alpha = 0$ while $\alpha = 1$ is equal to ridge regression. For efficiency reasons, tuning of the second penalization parameter λ of the elastic net – which controls the amount of regularization – was handled internally by the cv.glmnet() function of the glmnet package (10-CV with 300 different lambdas were used to find the largest value within one standard error of the best average performance). In this case, a third compute loop is employed on the lowest level to obtain the best value for λ within the inner loop, which is used to obtain the optimal value for α, and the outer loop is used to estimate predictive performance, as described above. In a random forest, the ratio parameter *mtry.perc* determines the number of randomly drawn predictor variables at each split of a tree. For all target variables, *mtry.perc* was tuned on the full range between 0 (rounded up to a single predictor) and 1 (all possible predictors). Grid values were selected on a quadratic scale to reduce the number of models with higher numbers of predictors which require a longer computation time as more variables are considered repeatedly when searching for the optimal split-points in the tree-growing algorithm. The benchmark for random forests involves only the inner loop (for obtaining the best value for *mtry.perc*) and the outer loop (for estimating predictive performance).

When using the elastic net or the random forest, missing values in the predictor variables had to be imputed before model estimation. While we used the simple histogram method described earlier, see Jerez et al. (2010) for a demonstration of more complex imputation methods that might improve predictive performance but also increase computation time. To avoid overly optimistic performance estimates, it is good practice to include important preprocessing steps into the resampling process as well. Instead of imputing all missing values for the complete dataset once, imputation was performed each time a model was trained.

Performance Measures

Predictive modeling heavily relies on performance measures (or loss functions) to quantify the similarity between predictions of an algorithm and the true labels. First, all predictive algorithms use a loss function in adjusting model coefficients during the estimation (i.e., training) process. While some algorithms allow the use of a custom loss function for model estimation, this intrinsic loss can usually not be altered by the practitioner. However, practitioners have to choose the second performance measures which are used in the inner loop of the nested resampling strategy to guide the tuning of optimal hyperparameters. Third and most important, a performance measure is always used in the outer loop of nested resampling to quantify the performance of the predictive algorithm on new, unseen data. A common strategy is to choose a performance measure for the outer resampling based on the problem at hand or widely used standards. The same measure can also be used in inner resampling to encourage configurations of hyperparameters that are optimized in the same manner.

This paper covers three different predictive scenarios: binary classification, ordinal classification, and regression. For the binary targets *Gender* and *Sick Days*, we use the mean misclassification error $\text{MMCE} = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$ as

loss function in the inner and outer resampling loops (I represents the indicator function which is equal to one if the expression in the bracket is true and zero otherwise). This is the most common performance measure for classification as it often quantifies the error of interest when the prediction of class labels is required in practical applications. As secondary measures in the outer resampling, we also report sensitivity and specificity to better judge the performance of the different algorithms, as MMCE alone can give misleading results under some circumstances. Brier score and AUC are popular alternative measures to consider (Ferri, Hernández-Orallo, & Modroiu, 2009). For ordinal targets with a small number of different values, it is less clear what can be regarded as the standard approach (Cardoso & Sousa, 2011). We treat the target *Trump* as a classification problem and use the mean absolute error $\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$ based on the resulting 4×4 contingency table of predicted and true labels (i.e., confusion matrix), in which predictions and true labels are coded from one to the maximum number of categories. In this way, the ordinal structure is reflected by penalizing predictions which lie further away from the true label more strongly. We further report multiclass *MMCE* in the outer loop. We consider the targets *Income*, *Life Satisfaction*, and *Sleep Satisfaction* to be regression problems and use the standard measure, in this case the mean squared error $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. As secondary measure, we also report $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$, which social scientists are generally more familiar with. Note that when evaluated on test sets, R^2 can take on negative values (which frequently happens when particularly flexible models are "overfitted" to small samples). Thus, resampled R^2 should not be interpreted as a ratio of explained variance. Similar to the idea of comparing a model's predictive performance to the featureless learner, the interpretation of negative R^2 is that the predictive model provides worse predictions for new observations than a simple model which constantly predicts the mean target value in the test set.

Computational Resources

The reported analyses were implemented relying on the computational resources of the Leibniz Supercomputing Centre (LRZ). For data pre- and postprocessing, the institution's RStudio Server web interface, a cluster of multiple RStudio Server nodes, was used. Model optimization and performance estimation were scheduled via a workload manager to run on the CoolMUC-2 Linux Cluster, a massively parallel processing cluster with more than 10,000 CPU cores. Each of the 30 hyperparameter optimization iterations (resulting from 3 repetitions of 10-CV in the outer resampling loop) was parallelized for all the different machine learning algorithms applied to the different target variables. While a featureless learner can be trained within seconds, this process of inner resampling and model tuning can take several hours (or even days) for more sophisticated models. In total, the equivalent of 263 days of computing time was necessary to perform the reported benchmarks. The statistical programming language R (R Core Team, 2017) was used for all calculations. The package mlr (Bischl et al., 2016) was used to perform model optimization and benchmarking. The packages glmnet (Friedman, Hastie, & Tibshirani, 2010) and ranger (Wright & Ziegler, 2017) provided the discussed machine learning techniques. Submission to the parallel cluster infrastructure was supported by the package batchtools (Lang, Bischl, & Surmann, 2017). Finally, the manuscript was prepared with the help of the packages knitr (Xie, 2015) and papaja (Aust & Barth, 2018). The code used to run all analyses can be downloaded from https://osf.io/zpse3/.

Results

Benchmark Results

Table 2 shows the aggregated performance estimates for the classification targets *Gender*, *Sick Days*, and *Trump* while the results of the regression targets *Income*, *Life Satisfaction*, and *Sleep Satisfaction* are presented in Table 3.

Table 2. Benchmark results for binary and ordinal classification tasks

	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	0.32
SPEC	1.00	0.86	0.83
Trump			
MAE	0.79	0.64	0.67
SD (MAE)	0.00	0.03	0.03
MMCE	0.51	0.49	0.48

Note. Mean MMCE and MAE across test sets are computed in each repetition of repeated cross-validation. The mean of the aggregated measures from all repetitions is presented as final performance estimate. *SD* (MMCE) and *SD* (MAE) reflect the standard deviations of the primary performance measure, computed across all test sets (10×3). MAE = mean absolute error; MMCE = mean misclassification error; SENS = sensitivity; SPEC = specificity.

Table 3. Benchmark results for regression tasks

	Featureless	Elastic net	Random forest
Income			
MSE	13.14	5.60	5.02
SD (*MSE*)	0.91	0.58	0.45
R^2	−0.01	0.57	0.62
Life Satisfaction			
MSE	3.76	1.98	1.99
SD (*MSE*)	0.37	0.26	0.24
R^2	−0.01	0.47	0.47
Sleep Satisfaction			
MSE	5.69	2.11	2.07
SD (*MSE*)	0.45	0.20	0.21
R^2	−0.01	0.62	0.63

Note. Mean *MSE* and R^2 across test sets are computed in each repetition of repeated cross-validation. The mean of the aggregated measures from all repetitions is presented as final performance estimate. *SD* (*MSE*) reflects the standard deviation of the primary performance measure, computed across all test sets (10 × 3). *MSE* = Mean squared error; R^2 = coefficient of determination.

To evaluate whether predictive performance differed between algorithms, the tables include the standard deviation of the primary performance measure across all 10 × 3 test sets in the outer resampling loops. For all target variables, predictive performance of the elastic net and the random forest seemed to be superior to the featureless learner, although for *Sick Days* and *Trump* the improvement in accuracy was quite small. Good performance was achieved for the prediction of *Gender* with an impressive misclassification error rate below 5%. The predictive performances of the elastic net and the random forest were highly similar for all targets considering the standard deviation between test sets (the elastic net showed slightly better performance in the MAE measure for *Trump* while the random forest achieved slightly better performance for *Income*). As should be the case, R^2 was always close to zero for the featureless learner.

Variable Importance

Our benchmark results did not show a convincing advantage of the random forest for any target variable. When nonlinear models like the random forest do not achieve clearly superior predictive performance, the linear elastic net is usually preferred in practice due to the higher interpretability of the model. Therefore, a final elastic net model was trained on the complete dataset for each target. These models correspond to regularized versions of logistic and multinomial regressions for binary and ordinal targets, and a regularized version of ordinary linear regression for continuous targets (Friedman et al., 2010). Note that multinomial regression is not a true ordinal model. The predictive model treats *Trump* as a nominal variable, and the ordinal structure is only considered in the performance evaluation with the MAE. Tuning was performed with the same inner resampling strategy already used in the prior benchmark experiment. Estimates of standardized regression coefficients were extracted from all final models to determine the most important predictors. The importance of a predictor was defined as the absolute coefficient estimate for binary classification and regression or as the sum of absolute estimates per variable for ordinal classification. The ten most important variables for each target in decreasing order are reported in Tables 4, 5, 6, 7, 8 and 9, labeled with the official variable IDs. Table notes include the tuning result for the hyperparameter α as well as the total number of predictor variables with nonzero coefficients. The ungrouped elastic net penalty was used for the multinomial regression model of the ordinal target *Trump* as it provides faster model estimation. This leads to a separate coefficient estimate for each combination of predictor variable and item category, which partly explains the high number of variables with nonzero values. As mentioned earlier, some variables were repeatedly included in different waves of the panel. Thus, there is a certain

Table 4. Top ten variable importance for target gender

ID	IMP	Name
caav061c	−0.50	Height in cm
bdao061a	0.50	Shaving: Legs
bdao098c	−0.49	Height in cm
a11c042a	−0.43	Affinity for technology
a11d096b	−0.34	Personal income
caav062a	−0.28	Weight
ebbm167a	−0.27	Weight
bdao032b	0.25	Care products: Make-up, incl. o.e.
bdao032a	0.25	Care products: Make-up
bdao053a	−0.22	Shaving: Face

Notes. Tuned α = 0.21. Number of predictors with nonzero coefficient = 264.

Table 5. Top ten variable importance for target sick days

ID	IMP	Name
dfaw099a	0.08	Satisfaction: Work
caav059a	0.07	Health insurance
a11d056b	−0.07	Year of birth
dbzc032a	−0.06	Important in life: Family
eaaw146a	−0.06	Social contacts constrained
dfaw116a	−0.04	Social contacts constrained
eaaw145a	−0.04	Physical pain
dazb009a	−0.04	Importance: Leisure time
cazb033a	0.04	Comparator finances
dbbg115a	0.04	Paying rent/mortgage on time

Notes. Tuned α = 0.21. Number of predictors with nonzero coefficient = 86.

Table 6. Top ten variable importance for target Trump

ID	IMP1	IMP2	IMP3	IMP4	Name
dcax047a	0.04	0.02	−0.01	−0.06	Candidate orientation: Cem Özdemir
bcaj085a	−0.08	0.00	0.00	0.04	Vote for: AfD
a11d072d	−0.01	−0.03	0.00	0.05	Country of birth (GER, EU, other)
dfbo058a	−0.03	0.00	0.00	0.05	Satisfaction with democracy (–)
dcax046a	0.00	0.03	0.00	−0.05	Candidate orientation: Sigmar Gabriel
dbzy063a	−0.07	0.00	0.00	0.00	Attitude toward Islam: constrained practice
dcax045a	0.03	0.00	0.00	−0.04	Candidate orientation: Angela Merkel
dfbo086a	0.04	0.00	0.00	−0.03	Trust in newspapers
dfbo077a	−0.04	0.00	0.00	0.03	Foreigners should marry own nationality
a12d021b	−0.05	0.00	0.01	0.00	Federal state (GER), west/east

Notes. Tuned α = 0.08. Number of predictors with nonzero coefficients = 369. Negatively coded variables are flagged with (–).

Table 7. Top ten variable importance for target income

ID	IMP	Name
a11d054a	−0.65	Gender
cfzh090c	0.52	Household income
bfzh089c	0.38	Household income
a11d086b	0.36	Vocational or professional training
dezh060a	−0.35	Employment situation
caav053a	0.35	Number of registered cars
ddaw157a	0.24	Satisfaction: Income
ddbk075a	0.19	Extra money per month for sustainable energy
bfzh084a	−0.18	Household size (one person, more than one)
cfba067a	0.18	Self-comparison (GER): financial wealth

Notes. Tuned α = 0.59. Number of predictors with nonzero coefficient = 70.

Table 8. Top ten variable importance for target life satisfaction

ID	IMP	Name
eaat037a	0.09	Positive life changes
debl223a	−0.08	General standard of living: Feel good (–)
dazb026a	0.08	Feeling: Enjoyed life
eazb026a	0.07	Feeling: Enjoyed life
dbzc032a	0.06	Important in life: Family
dcaw184a	−0.06	Social contacts constrained
eazb021a	−0.06	Feeling: Depressed
deaw258a	0.05	Feeling: Relaxed
ddbg141a	−0.05	Overall living standard (–)
dbbh132a	−0.05	Self-description: Far away from everything

Notes. Tuned α = 0.10. Number of predictors with nonzero coefficient = 82. Negatively coded variables are flagged with (–).

Table 9. Top ten variable importance for target sleep satisfaction

ID	IMP	Name
eaaw123a	0.48	Satisfaction: Sleep
dfaw093a	0.28	Satisfaction: Sleep
deaw246a	0.24	Satisfaction: Sleep
dbaw226a	0.21	Satisfaction: Sleep
ddaw156a	0.15	Satisfaction: Sleep
dcaw161a	0.14	Satisfaction: Sleep
cfaw121a	0.12	Satisfaction: Sleep
daaw114a	0.09	Satisfaction: Sleep
cbaw103a	0.08	Satisfaction: Sleep
ceaw139a	0.05	Satisfaction: Sleep

Note. Tuned α = 0.54. Number of predictors with nonzero coefficient = 14.

likelihood for these items with comparable contents to appear together in a single prediction model.

When predicting *Gender*, the repeated variables height and weight can be observed with comparable impacts on the prediction model. Also, notice the importance of personal income – a seemingly stable relationship that is also evident when predicting *Income*.

In the model for *Sick Days*, the interpretation of variable *caav059a* (asking what kind of health insurance the respondents have) has to be carried out with caution. While the highest category of this nominal item is "no health insurance," the possible impact of the additionally included distinctions between mandatory, voluntary, statutory, and private health insurances (idiosyncrasies of the German healthcare system) is not clear. A similar problem arises for the nominal variable *cazb033a* (which person or group do respondents compare themselves to the most).

In the model for *Trump*, note that the Alternative for Germany (German: Alternative für Deutschland, AfD) is a right-wing political party, currently the third-largest parliamentary group in the German parliament (Bundestag). Cem Özdemir, Sigmar Gabriel, and Angela Merkel are well-known, moderate German politicians of the Green Party, Social Democratic Party, and Christian Democratic Union, respectively.

While a gender–income relationship has already become apparent in the model for *Gender*, it seems necessary to highlight that gender is the most prominent variable when predicting *Income* – even more important than the reported household income. Speaking of which, one could argue for

the advanced removal of these variables, as they are closely related to the target. Yet, as we have described certain criteria for removal above, this reflects a general challenge when using automatic variable selection procedures. Lastly, both variables describing the vocational or professional training level and the employment situation should be interpreted carefully as they can be considered to reflect an ordinal tendency at best.

For the prediction of *Life Satisfaction*, it is noteworthy that both variables *debl223a* and *ddbg141a* suggest a positive relationship between economic status and general satisfaction with life.

In the predictive model for *Sleep Satisfaction*, it is remarkable that of only 14 nonzero predictors selected by the elastic net, 11 were lagged versions of the target variable included in earlier panel waves. This suggests a great stability in the temporal pattern of sleep satisfaction. Naturally, more recent reports (e.g., *eaaw123*) seemed to better predict current sleep satisfaction than earlier reports (e.g., *cbaw103*).

Discussion

Large-scale longitudinal panels are among the biggest and most high-quality psychological data sources available today. They monitor a large representative sample of the general population over a long period of time. Considering the associated costs for the scientific community, research ethics demand to use these carefully collected data in the most effective way. With regard to research questions of predictive nature and approaches in search of important variables in panel studies, we demonstrated that predictive modeling methods are well suited to assist with these goals. While using a challenging dataset with several thousand variables that were treated as predictors in an automated fashion, we achieved promising predictive performances for a variety of different target variables. With the adopted analysis strategy, regularized linear elastic net models resulted in highly similar performance compared to the random forest as a nonlinear model, including interactions. The benchmarking results and models presented above suggest that the analytic approach chosen in this work is well suited to explore real-world relationships of self-reported variables common in psychological research. Judging by the most important predictors for our exemplary targets, the final elastic net models seem to identify highly convincing variables with regard to their content. This is, for instance, impressively demonstrated by the various popular themes emerging in the *Trump* task, for example, a general satisfaction with democracy, attitudes toward Islam, or the trust in newspapers. Thus, these analytic strategies might also be useful to uncover formerly unknown or unforeseen relationships in increasingly large datasets. In order to adopt machine learning techniques for their own studies, many psychological researchers will have to familiarize themselves with new modeling strategies and acquire additional computer skills. However, we believe that with an increasing availability of data – including but not limited to psychological panels – these methods will be a necessary addition to psychology's traditional methodological toolkit with a focus on unregularized linear modeling and null hypothesis significance testing.

Further Modeling Options

Even though our benchmark results showed good predictive accuracy for most analyzed variables, several strategies could be employed to further increase the performance:

The elastic net as well as the random forest can handle a very large number of predictor variables and can be used with more predictors than observations (e.g., *Sick Days*), which would not be possible with unregularized linear models. With sample sizes of only about 2,000 observations, a preliminary reduction of predictor variables in a data-driven fashion might further improve predictive accuracy. The two most common generally applicable strategies for selecting important variables in predictive modeling (i.e., feature selection) are filter and wrapper methods (Guyon & Elisseeff, 2003). When applying filter methods, predictors are ranked with respect to some univariate measure of feature importance (e.g., the linear correlation with the target variable), while the optimal number of variables chosen from this ranking is typically considered a hyperparameter that can, again, be tuned together with other hyperparameters of the chosen algorithm. In contrast, common wrapper methods build increasingly complex models by adding the variables that show the highest gain in test set performance – an approach similar to stepwise regression strategies.

As described earlier, we relied on a simple imputation scheme to deal with missing values in predictor variables. In theory, it would be possible to choose more complex strategies like multiple imputation or nonlinear, model-based imputations, for example, replacing missing values with the predictions of a single decision tree (Jerez et al., 2010). However, with several hundreds of predictor variables, such approaches would tremendously increase the required computation time.

In some of our classification problems (e.g., *Trump*), the relative frequency of one (or more) classes was quite low compared to the others (i.e., imbalanced classification). Achieving good predictive performance is challenging in these cases, as algorithms tend to predict the label of the larger classes too frequently. One strategy to address this problem is threshold tuning (Sheng & Ling, 2006). Usually,

an observation is classified as belonging to the positive class in a binary classification task if the estimated probability for this observation is greater than 0.5. When the relative frequency of observations in this positive class is high, increasing the cutoff (and finding its optimal value by resampling) can result in better predictive performance. Alternatively, oversampling can be used to fight imbalanced classification by synthetically generating additional cases for the underrepresented class (Chawla, Bowyer, Hall, & Kegelmeyer, 2002).

Finally, more complex nonlinear methods than the random forest could be used for the prediction tasks we presented or multiple algorithms could be stacked into an ensemble (Rokach, 2010). Possible candidates might be support vector machines, gradient boosting, or neural networks (see Hastie, Tibshirani, & Friedman (2009) for an introduction). In contrast to the random forest, some of these algorithms have many more hyperparameters and require excessive tuning to be effective. As tuning on a grid of different hyperparameter values becomes rapidly inefficient in higher dimensions, more complex tuning strategies like model-based optimization (Bischl et al., 2017) have to be used in order to obtain optimal tuning results with a limited number of model evaluations. However, when accepting this increase in methodological complexity and computational resources, impressive predictive performance can be achieved by some of these algorithms (Chen & Guestrin, 2016), even in data situations too scarce to effectively use methods like deep neural networks.

If nonlinear algorithms should turn out to achieve superior predictive performance in psychological research, the interpretation of resulting models will become more difficult. Most of these algorithms do not rely on simple regression coefficients that could be used to evaluate the impact of important predictors. Although random forests provide nonlinear measures of variable importance, those can give different results in practice and several peculiarities have to be carefully considered which we can only mention briefly: Standard implementations (as in the ranger package) can only provide measures which mainly reflect the marginal effects of predictors and suffer from variable selection bias (Strobl et al., 2007). For unbiased (un)conditional importance measures (which can be interpreted more like the coefficients in regression models), an alternative random forest algorithm has to be used (Strobl et al., 2008) which has much longer runtimes and is not guaranteed to achieve comparable predictive performance. Thus, which measure to use in an application has to be determined by balancing computational resources with the intended interpretation imposed by the concrete research question. Apart from algorithms for which variable importance measures can be computed, finding model-agnostic strategies to understand how certain predictors influence specific predictions is an important research field in machine learning (Friedman, 2001; Goldstein et al., 2015; Ribeiro et al., 2016). By employing such methods, researchers and practitioners can place greater confidence in increasingly complex models and do not have to rely on predictions based on a "black box," as they can at least partly determine which variables drive predictive performance and how these functional relationships might look like.

Longitudinal Data

Although based on a panel dataset, our analyses rely on predictive models that do not explicitly take the longitudinal structure into account. In general, this does not pose any problems considering that the successful prediction of future events is one of the main reasons for the current success of machine learning methods and the employed performance evaluation scheme is not biased by the temporal structure of the data. However, when using repeated measurements of variables with identical content, this strategy implies that the final model can only be used to predict observations from the exact same time points used in the training data. Consider the target *Sleep Satisfaction* for which we achieved a solid performance predicting self-ratings in the last panel wave collected from April to June 2017 while including comparable variables from earlier time points as predictors. While our final model highly suggests that past ratings of sleep satisfaction are strong indicators of future sleep satisfaction, it can not – in contrast to proper longitudinal methods – be directly used to predict sleep satisfaction for a later wave (e.g., collected from July to August 2017). This could only be achieved by employing a model class which includes an explicit time component that could be used to generalize over time periods not explicitly trained upon. Recently, specific machine learning methods have been proposed to deal with a longitudinal panel structure in particular. These methods are based on the linear (mixed) models heavily used in panel data analysis, with the distinction that some fixed effects are modeled in a nonlinear fashion by either decision trees (Sela & Simonoff, 2012) or neural networks (Crane-Droesch, 2017). Within the model-based recursive partitioning framework, another method builds decision trees with a generalized linear mixed model in each node (Fokkema, Smits, Zeileis, Hothorn, & Kelderman, 2018). As panel data become increasingly "big," such methods that allow for an unbiased estimation of linear time or group effects might be very promising tools for panel data researchers, depending on the goal of the analysis. In any case, the effective adoption of these methods requires some knowledge about

nonlinear modeling as well as machine learning concepts like resampling and performance evaluation which were described in this paper.

Conclusion

State-of-the-art predictive modeling analyses pose unique challenges which are usually not encountered by psychological researchers when analyzing smaller datasets with traditional methods. By demonstrating all necessary steps based on a series of exemplary analyses, the aim of this study was to encourage more researchers to include predictive modeling into their methodological toolkit. Additional information about predictive modeling can be found in the modern classic (Hastie et al., 2009) and its more accessible companion (James, Witten, Hastie, & Tibshirani, 2013). As a unified interface to perform predictive modeling in R, we recommend the mlr package (Bischl et al., 2016), which also provides a detailed tutorial on many important topics.

References

Aust, F., & Barth, M. (2018). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., … Jones, Z. M. (2016). mlr: Machine learning in r. *Journal of Machine Learning Research, 17*, 1–5. Retrieved from http://jmlr.org/papers/v17/15-066.html

Bischl, B., Mersmann, O., Trautmann, H., & Weihs, C. (2012). Resampling methods for meta-model validation with recommendations for evolutionary computation. *Evolutionary Computation, 20*, 249–275. https://doi.org/10.1162/EVCO_a_00069

Bischl, B., Richter, J., Bossek, J., Horn, D., Thomas, J., & Lang, M. (2017). *mlrMBO: A modular framework for model-based optimization of expensive black-box functions*. Retrieved from http://arxiv.org/abs/1703.03373

Bosnjak, M., Dannwolf, T., Enderle, T., Schaurer, I., Struminskaya, B., Tanner, A., & Weyandt, K. W. (2018). Establishing an open probability-based mixed-mode panel of the general population in Germany: The GESIS panel. *Social Science Computer Review, 36*, 103–115. https://doi.org/10.1177/0894439317697949

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

Cardoso, J. S., & Sousa, R. (2011). Measuring the performance of ordinal classification. *International Journal of Pattern Recognition and Artificial Intelligence, 25*, 1173–1195. https://doi.org/10.1142/S0218001411009093

Chapman, B. P., Weiss, A., & Duberstein, P. R. (2016). Statistical learning theory for high dimensional prediction: Application to criterion-keyed scale development. *Psychological Methods, 21*, 603. https://doi.org/10.1037/met0000088

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. Retrieved from http://dl.acm.org/citation.cfm?id=1622407.1622416

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). New York, NY: ACM. https://doi.org/10.1145/2939672.2939785

Crane-Droesch, A. (2017). *Semiparametric panel data models using neural networks*. Retrieved from https://arxiv.org/abs/1702.06512

Ferri, C., Hernández-Orallo, J., & Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters, 30*, 27–38. https://doi.org/10.1016/j.patrec.2008.08.010

Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., & Kelderman, H. (2018). Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behavior Research Methods, 50*, 2016–2034. https://doi.org/10.3758/s13428-017-0971-x

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*, 1189–1232. Retrieved from http://www.jstor.org/stable/2699986

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*, 1–22. Retrieved from http://www.jstatsoft.org/v33/i01/

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation, 4*, 1–58. https://doi.org/10.1162/neco.1992.4.1.1

GESIS. (2017). *GESIS panel – standard edition (version 21.0.0, data file ZA5665)*. Cologne, Germany: GESIS Data Archive. Retrieved from https://doi.org/10.4232/1.12829

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics, 24*, 44–65. https://doi.org/10.1080/10618600.2014.907095

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. New York, NY: Springer.

Hsiao, C. (2007). Panel data analysis – advantages and challenges. *TEST, 16*, 1–22. https://doi.org/10.1007/s11749-007-0046-x

James, G. M. (2003). Variance and bias for general loss functions. *Machine Learning, 51*, 115–135. https://doi.org/10.1023/A:1022899518027

James, G. M., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York, NY: Springer.

Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine, 50*, 105–115. https://doi.org/10.1016/j.artmed.2010.05.002

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence – volume 2* (pp. 1137–1143). San Francisco, CA: Morgan Kaufmann. Retrieved from http://dl.acm.org/citation.cfm?id=1643031.1643047

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26) New York, NY: Springer.

Lang, M., Bischl, B., & Surmann, D. (2017). Batchtools: Tools for r to work on batch systems. *The Journal of Open Source Software, 2*, 135. https://doi.org/10.21105/joss.00135

Larson, S. C. (1931). The shrinkage of the coefficient of multiple correlation. *Journal of Educational Psychology, 22*, 45–55. https://doi.org/10.1037/h0072400

Nembrini, S., König, I. R., & Wright, M. N. (2018). The revival of the Gini importance? *Bioinformatics, 34*, 3711–3718. https://doi.org/10.1093/bioinformatics/bty373

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). New York, NY: ACM. https://doi.org/10.1145/2939672.2939778

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review, 33*, 1–39. https://doi.org/10.1007/s10462-009-9124-7

Sela, R. J., & Simonoff, J. S. (2012). RE-EM trees: A data mining approach for longitudinal and clustered data. *Machine Learning, 86*, 169–207. https://doi.org/10.1007/s10994-011-5258-3

Sheng, V. S., & Ling, C. X. (2006). Thresholding for making classifiers cost-sensitive. In A. Cohn (Ed.), *Proceedings of the 21st National Conference on Artificial Intelligence* (Vol. 1, pp. 476–481). Boston, MA: AAAI Press. Retrieved from http://dl.acm.org/citation.cfm?id=1597538.1597615

Shmueli, G. (2010). To explain or to predict? *Statistical Science, 25*, 289–310. https://doi.org/10.1214/10-STS330

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics, 9*, 307. https://doi.org/10.1186/1471-2105-9-307

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics, 8*, 25. https://doi.org/10.1186/1471-2105-8-25

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological), 58*, 267–288. Retrieved from http://www.jstor.org/stable/2346178

White, M. P., Alcock, I., Wheeler, B. W., & Depledge, M. H. (2013). Would you be happier living in a greener urban area? A fixed-effects analysis of panel data. *Psychological Science, 24*, 920–928. https://doi.org/10.1177/0956797612464659

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software, 77*, 1–17. https://doi.org/10.18637/jss.v077.i01

Xie, Y. (2015). *Dynamic documents with R and knitr*, (2nd ed.). Boca Raton, FL: Chapman; Hall/CRC. Retrieved from https://yihui.name/knitr/

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12*, 1100–1122. https://doi.org/10.1177/1745691617693393

Yin, P., & Fan, X. (2001). Estimating r 2 shrinkage in multiple regression: A comparison of different analytical methods. *Journal of Experimental Education, 69*, 203–224. https://doi.org/10.1080/00220970109600656

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

History
Received March 15, 2018
Revision received August 10, 2018
Accepted August 31, 2018
Published online February 22, 2019

Florian Pargent
Department of Psychology
LMU
Leopoldstr. 13
80802 München
Germany
florian.pargent@psy.lmu.de