Machine learning and assessment

“When we raise money it’s AI, when we hire it’s machine learning, and when we do the work it’s logistic regression.”

(Tweet by bio-statistician Daniella Witten; original author unknown)

# Abstract

Modern regression techniques (a.k.a. ‘artificial intelligence’, AI, or ‘machine learning’, ML; refered to in this editorial as ‘AI/ML’) provide unprecedented flexibility to capture non-linear associations, and/or the effects of many potential predictors at once, while at the same time providing built-in overfitting control.

The focus of AI/ML on optimizing the predictive accuracy of future observations, instead of explanation of the training data, is beneficial for the field of psychological assessment, which has prediction as one of its core tasks.

Sophisticated AI/ML techniques (like random forests, support vector machines, artificial neural networks) have shown superior performance in terms of predictive accuracy in many applications and benchmark experiments.

At the same time, compared to simpler methods like a single decision tree or a penalized GLM, the gains in predictive accuracy provided by sophisticated methods may often be marginal, because simpler methods capture most of the explainable variance (signal) already. In addition, apparent gains in predictive accuracy estimated through cross-validation may be swamped by practical aspects of a data problem, like population drift, measurement error, and a need for explainability.

In this editorial, we aim to show what may be gained and lost from psychological assessment by the use of simpler versus more complex statistical techniques in a prediction problem.

# Introduction

Machine learning and artificial intelligence are by now familiar buzzwords in many fields of empirical research. In the field of psychological assessment, we also see increasing interest and application of methods from these fields - this is also true for our journal (Ammannato & Chiesi, 2020; Elleman et al., 2020; Finkelman et al., 2020; Fokkema & Greiff, 2018; e.g., Opitz et al., 2021; Park et al., 2020; Sauer et al., 2016). We believe ML and AI methods have the potential to contribute to our field. The buzz around these methods, however, can appear reminiscent of the tale of the emperor’s new clothes. We believe when it comes to ML and AI, the emperor is in fact wearing clothes, but their new- and fanciness may be somewhat of a mirage. Many of the powerful techniques used in machine learning (e.g., cross validation, regularization, ensembling) have long been known and fruitfully applied in more traditional statistical analyses, in psychometrics and psychological assessment.

This editorial therefore has two aims:

1. We give a very short introduction to several modern statistical learning methods that provide optimal predictive accuracy and/or interpretability, and to discuss how they relate to more traditional regression techniques.
2. We discuss how novel methods and insights from the field of AI/ML may contribute to the continuing improvement of our field, and to indicate possible pitfalls.

We will use the term “statistical learning” to refer to both more traditional and more recent tools for data analysis, because there is no clear distinction between statistics, artificial intelligence and machine learning. Of note, we believe newer data-analytic methods can only contribute to our field as long as their use is firmly rooted in existing statistical knowledge about sampling, estimation and uncertainty. For the sake of brevity, in this editorial we focus on methods for statistical prediction, a.k.a. supervised learning methods. Thus, unsupervised learning methods (e.g., factor analysis, clustering, correlation networks, topic models from natural language processing) are outside the scope of this editorial.

### Recent shifts in statistical prediction

Recent development in statistical prediction methods have lead to two main shifts:

1. An increased focus on prediction.
2. Increased flexibility, allowing for capturing non-linear associations and/or simultaneously modeling large numbers of potential predictor variables.

We believe that the first shift is highly beneficial for our field, because prediction of behavior is one of the core tasks of psychological assessment. In contrast, the field of psychology at large has traditionally often been interested in *explanation*; that is, developing and testing theories of human behavior. This has sometimes led researchers to overlook *prediction*, perhaps because their main aim was to explain behavior. A theory, however, can only explain real-world phenomena to the extent that it can accurately predict them (Yarkoni & Westfall, 2017). This focus on explanation has lead many researchers to focus on effect sizes (e.g., *R2*, Cohen’s *d*) computed on the *training* data; that is, using observations that were used to fit the model. At the same time, most researchers in psychology are probably familiar with the phenomenon of *R2* automatically and spuriously increasing when explanatory variables are added to the model. More realistic *R2* values can be obtained, for example, through cross validation: by computing effect sizes on a sample of observations *not* used for fitting the model (Rooij & Weeda, 2020). This importance of cross validation is certainly not new to the field of assessment, where it has been discussed for almost a century (Larson, 1931; Mosier, 1951).

We believe that the second shift brings both promises and pitfalls. The great promise of AI/ML techniques lies in their built-in overfitting control, because unrestricted flexibility leads to overfitting. In statistical learning, this has been formalized in the *bias-variance trade-off*. Informally, this trade-off states that the more flexible a model is allowed to approximate any possible patterns of association between predictors and response (i.e., the lower the bias), the worse the model will generalize to new samples from the same population (i.e., the higher the variance). To obtain optimally generalizable results for a given data problem and sample (size), bias and variance should thus be carefully balanced.

Bias can be increased and variance reduced in several ways, including:

* Restricting the complexity of the functional form (e.g., model only linear associations; model only main effects and no interactions);
* Limiting the number of potential predictors used (e.g., include only few predictors; use sum-, component- or factor scores instead of item scores as predictors);
* Using regularized estimation procedures (e.g, lasso, ridge, and elastic net regression; use of Bayesian priors; penalizing wigglyness);
* Ensembling: Aggregating the contributions of many variables, observations and/or learners. For example, in psychometrics we may aggregate individual items into a subscale or factor score to obtain a more reliable measure of the construct of interest. In ML, ensemble methods aggregate the predictions of many simpler or unstable models to obtain a stronger (more stable) final predictive model.

If the bias is well-chosen and realistic, generalizability of the fitted model will be improved. In other words: we can buy predictive power by assumption. If the bias is not well chosen (i.e., our assumptions are too unrealistic), predictive accuracy and generalizability will obviously suffer.

In the next section, we wish to illustrate the flexibility and predictive performance of several statistical learning techniques through an empirical example. The predictive (validity) question is: To what extent do responses on a measure of vocational preferences predict the type of university major completed. However, the prediction question itself is not our prime interest. Instead, we wish to use it to illustrate the more general principles of flexibility, dimensionality, overfitting and interpretability in prediction problems, in psychological assessment.

# Empirical illustration

### Method

#### Dataset

To illustrate a range of statistical learning and prediction methods, we use a dataset from the Open Psychometrics Project (<https://openpsychometrics.org/_rawdata/>). Data were collected through their website from 2015 to 2018. Respondents answered 48 items on vocational preferences, and additional items on personality and sociodemographic characteristics.

We investigate predictive validity of the RIASEC vocational preferences scales (Liao et al., 2008). The RIASEC uses six occupational categories from Holland’s Occupational Themes (Holland, 1959) theory: Realistic (R), Investigative (I), Artistic (A), Social (S), Enterprising (E), and Conventional (C). There are 8 items for each category, each describing a task (e.g., R6: “Fix a broken faucet” or I2: “Study animal behavior”), to which respondents answer on a 1-5 scale, with 1=Dislike, 3=Neutral, 5=Enjoy. The items are included in Appendix A.

From the full dataset, we selected participants who completed at least a university degree, yielding a sample of *N =* 55593 observations. Our criterion of interest is a binary variable, indicating whether respondents majored in Psychology, or in a different topic. We separated the data into 75% training observations and 25% test observations. Our training sample thus consists of 41694 respondents, of which 19.47% majored in psychology. Our test sample consisted of 13899 respondents, of which 19.32% majored in psychology. Univariate distributions and bivariate correlations between predictors are presented in Appendix B.

#### Model fitting and parameter tuning

We fitted (penalized) logistic regression models using package **glmnet** [version 4.1.2; Friedman et al. (2010)]; generalized additive models (GAMs) with smoothing splines using package **mgcv** (version 1.8.35, Wood, 2017); conditional inference trees using package **partykit** (Hothorn et al., 2006; version 1.2.13, Hothorn & Zeileis, 2015); gradient boosted tree ensembles using package **gbm** (version 2.1.8, Greenwell et al., 2020); random forests using package **ranger** (version 0.13.1, Wright & Ziegler, 2017); prediction rule ensembles using package **pre** (version 1.0.1, Fokkema, 2020); nearest neighbours using package **class** (version 7.3.19, Venables & Ripley, 2002).

For tuning the parameters of random forests, boosted tree ensembles and prediction rule ensembles, we used package **caret** (version 6.0.89, Kuhn, 2021). For tuning the parameters of conditional inference trees and nearest neighbours, we wrote custom code; we tuned the penalized regression models using function cv.glmnet from package **glmnet**. We did not tune the parameters of the GAMs with smoothing splines, because we expected the defaults to work well out of the box.

#### Evaluating predictive accuracy

To evaluate predictive accuracy, we used the Brier score. The use of accuracy measures derived from the confusion matrix of actual and predicted classes, like the misclassification error, sensitivity (a.k.a. recall), positive predictive value (a.k.a. precision) or Matthew’s correlation coefficient are pervasive in the machine learning literature. However, these measures disregard the quality of predicted probabilities from a fitted model and we therefore recommend against their use for evaluating predictive accuracy (Gneiting & Raftery, 2007). Methods for predicting a binary outcome should not only provide a predicted class, they should also provide predicted probabilities to quantify the uncertainty of the classification. To evaluate performance, the quality of this probability forecast should be evaluated.

The Brier score quantifies this quality by taking the mean squared error of the predicted probabilities:

Where is the observed outcome for observation , taking a value of 0 or 1; is the model’s predicted probability, taking a value between 0 and 1. Note that we can compute the Brier score on the training and on the test observations; thus can be taken to be the training or the test sample size. A Brier score equal to the variance of indicates performance no better than chance (in the current dataset, the variance was 0.1947 for training and 0.1932 for test data). If we take 1 minus the Brier score divided by the variance of , this yields a pseudo- measure, which takes values between 0 (indicating performance no better than chance) and 1 (indicating perfect accuracy).

# Results

### Prediction using subscale scores

#### (Penalized) logistic regression

Using CV to determine the optimal value of the penalty parameter , the optimal value was found to be 0 for the ridge as well as the lasso penaly. We thus fitted an unpenalized logistic regression, which provides the benefit of standard errors (and values) for the coefficients:

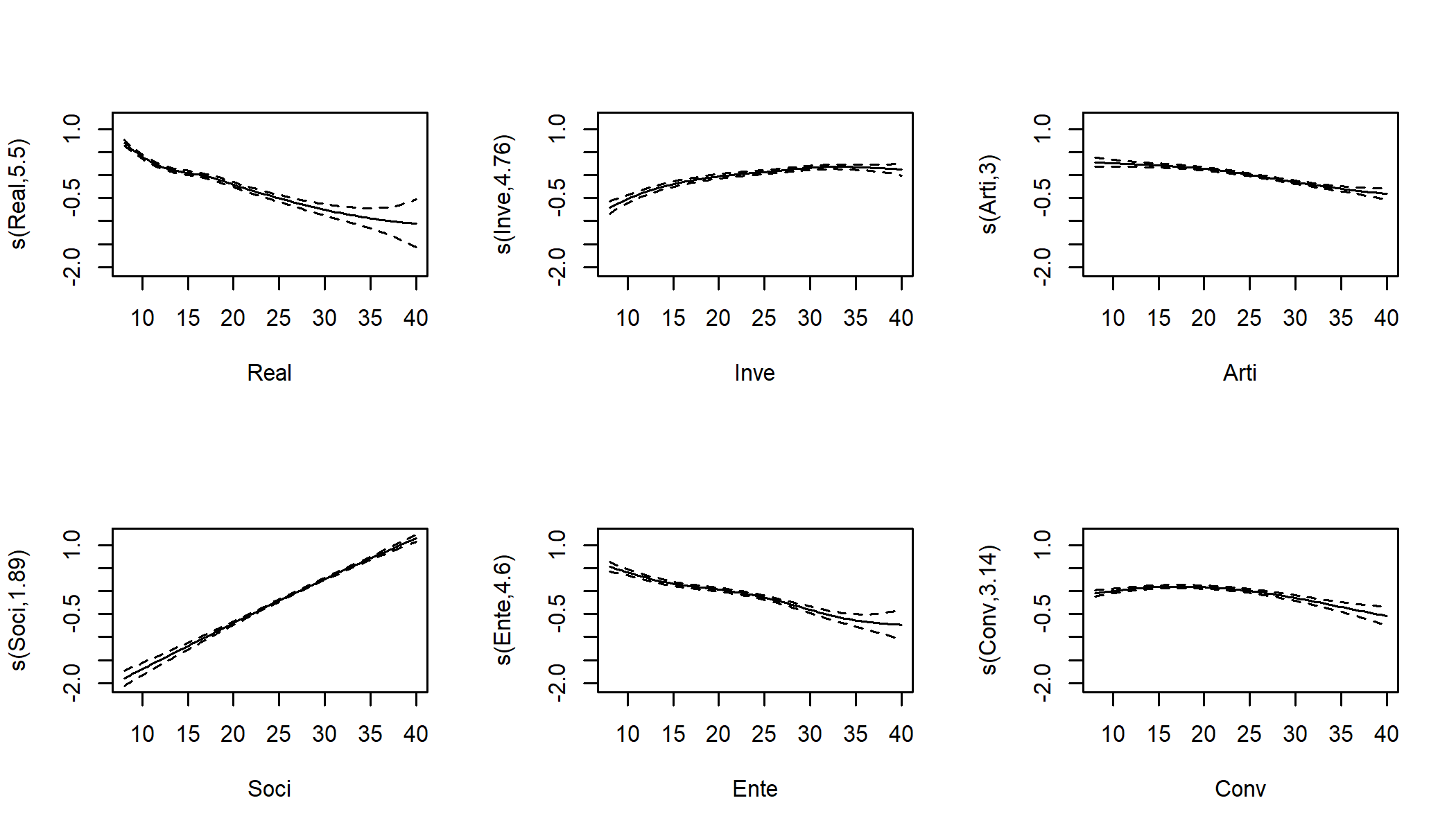
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -2.122 0.076 -27.909 <.001  
## Real -0.062 0.003 -24.734 <.001  
## Inve 0.022 0.002 12.151 <.001  
## Arti -0.023 0.002 -12.194 <.001  
## Soci 0.094 0.002 41.083 <.001  
## Ente -0.039 0.002 -16.114 <.001  
## Conv -0.010 0.002 -4.595 <.001

The results indicate that each vocational preference scale has a significant effect on the choice of major. We see the strongest effect is a positive effect from the Social preferences scale, and the weakest effect is a negative effect from the Conventional preferences scale.

#### Generalized additive model

We fitted a generalized additive model with a smoothing spline for each subscale score. Smoothing splines allow for flexibly approximating non-linear shapes of association between a predictor and response. At the same time, the method prevents overfitting by penalizing the wigglyness of the fitted curves. This explains their name: They provide a flexible but smooth approximation to the data. It is an additive model, resembling the GLM fitted earlier: Both assume interactions to be absent from the data, but GAMs additionally allow for capturing non-linear effects of each predictor.

We employed the defaults of thin-plate regression splines and generalized cross validation method to fit the smoothing splines. This provides built-in regularization of the wigglyness, without the need for choosing an optimal value for penalty or tuning parameters. For smaller samples, restricted maximum likelihood (REML) would be preferred. The fitted curves look as follows:



Like in the logistic regression, we see positive effects of the Social and Investigative subscales, and negative effects of the Realistic, Artistic, Enterprising and Conventional subscales. The Social subscale appears to have a linear effect, but the other subscales’ effects exhibit some non-linearity. Note that the plots show the *conditional* effect of each predictor. The values on the -axis thus reflect the expected increase in log-odds for each value of the subscale on the -axis, assuming all other variables remain constant. We can also test the significance of the effect of each predictor:

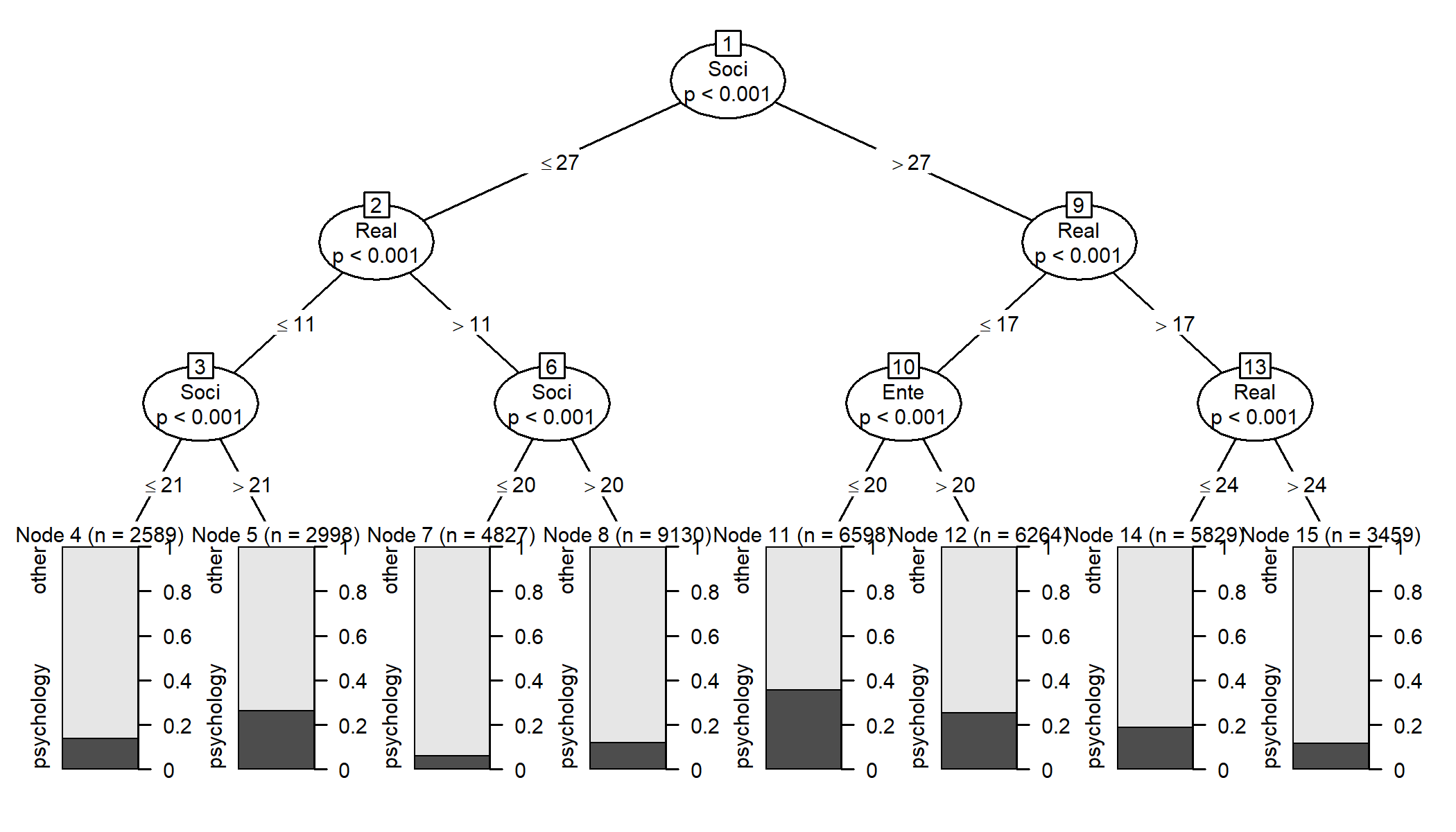
## edf Chi.sq p.value  
## s(Real) 5.504 654.162 <.001  
## s(Inve) 4.764 190.956 <.001  
## s(Arti) 2.997 161.529 <.001  
## s(Soci) 1.888 1625.900 <.001  
## s(Ente) 4.597 252.407 <.001  
## s(Conv) 3.142 60.893 <.001

The EDF values reflect the empirical degrees of freedom. The stronger the non-linearity in a curve, the more degrees of freedom it uses. For a linear slope, the EDF would equal 1; higher values indicate stronger non-linearity. For the social preferences subscale, the EDF value indicates a near linear effect, which we also observed in the plots above. The values quantify the strength of the evidence of each predictor’s effect. Each predictor shows a statistically significant effect on the outcome, and the strength of the effect have the same ordering as in the logistic regression model.

#### Decision tree

We fit a single decision tree using the conditional inference tree algorithm (Hothorn et al., 2006). This algorithm eliminates the variable selection bias present in many other decision-tree algorithms. Decision-tree methods and variable selection bias are discussed in more detail by by Strobl et al. (2009), who provide a comprehensive introduction to decision-tree methods aimed at psychologists. According to the CV results, a tree depth of 6 was optimal, yielding a tree with terminal nodes. Thus, for this data problem, the most accurate tree is surely not the most interpretable.

For illustration, we show the decision tree pruned to a depth of three below:



We see that Social, Realistic and Enterprising preferences occur in the first splits of the tree, which were also identified by the two previous models as the most important predictors. The bars in the terminal nodes show the proportions of observations who (did not) choose psychology as a major. Again, the Social preferences scale shows a positive association with the response, and the Realistic and Enterpristing preferences scales show a negative association. Note also that split number 10, involving the Enterprising subscale, may reflect an interaction effect, because this predictor only appears relevant for higher values of the Social subscale, and lower values of the Realistic subscale. However, such a split may also reflect additive effects combined with multicollinearity. Thus, although decision trees can *capture* interaction effects, they cannot be used to statistically test their significance.

#### Boosted tree ensemble

Although decision trees are easy to interpret, they suffer more strongly from instability than GLMs and GAMs. With instability, we mean that a small change in the training data can lead to large changes in the resulting model. The cause of this instability partly lies in the rather rough cuts made in the tree. Every next split that is implemented, is conditional on previous splits. If the first split in the tree would have been made using a different predictor or different split point, the following splits may also turn out different, yielding quite a different tree structure.

Decision tree ensembling method take advantage of this instability. Ensemble methods derive a large number of learners (e.g., trees), and each of these learners is fit on different versions of the training dataset. For example, different versions of the training data can be created by drawing different random samples from the training set for fitting each tree, a method also known as bagging. More powerful tree ensembling methods are random forests and boosting.

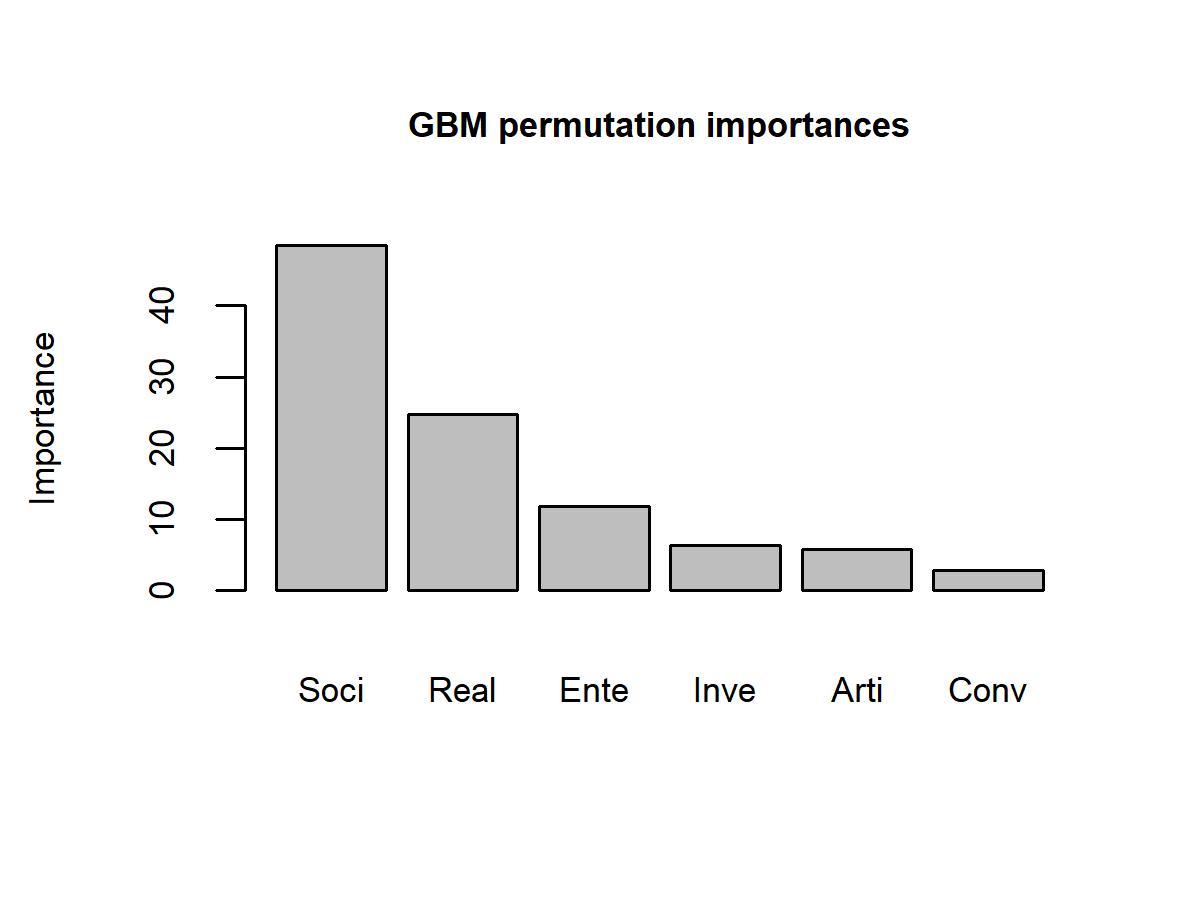
Tree ensembles are powerful prediction methods, because averaging over the predictions of many unstable learners yields a final model that is more accurate than any of the individual learners. This is comparable to averaging or summing items into subscales: Individual items may have low reliability (high error), but a subscale score composed of a sufficient number of items with at least some reliability will have high reliability.

The first tree ensemble method we apply to the data is a gradient boosted ensemble. Boosting (Bühlmann & Hothorn, 2007) uses sequential fitting of so-called weak learners to create a strong learner. Weak learners are simple models, that provide predictive accuracy (slightly) better than chance. When boosting trees, we use weak learners in the form of small trees, e.g., trees with only 1-7 splits. Sequential learning means that each consecutive tree is adjusted for the predictions of previous trees. In effect, observations that were well (badly) predicted by previous trees receive less (more) weight when fitting the current tree. An introduction and tutorial to tree boosting methods is provided in Elith et al. (2008).

Gradient boosting is one of the top prediction approaches. Many forecasting competitions have been won by using gradient boosting. To obtain good performance with boosting, careful tuning of the model-fitting parameters is necessary. The most important parameters are tree size, the number of trees in the ensemble and the shrinkage or learning rate. The latter parameter reflects the weight that is attributed to the predictions of each previous tree, when fitting the current tree. In the current application, the optimal number of trees found by 10-fold CV was 1100, with an learning rate of 0.01 and a tree depth of three. The latter means that interaction effects of at most three variables can be captured, and the individual trees contain at most seven splits and eight terminal nodes.

A disadvantage of decision tree ensembles is their black box nature: While individual trees are generally easy to interpret, an ensemble of trees is impossible to grasp. Therefore, so-called variable importance measures have been developed for interpretation of tree ensembles, which aim to quantify the effect of predictor variable on the predictions of the ensemble. In this paper, we use the permutation importances proposed by Breiman (Breiman, 2001). These quantify how much an ensemble’s predictive accuracy would be reduced, if the values of each of the predictor variables were randomly shuffled.

Importance measures provide a useful ranking of the contributions of each predictor to the ensemble’s predictions, but should be interpreted with care. They should not be used to judge the significance of the effect of predictors; tree ensembles can easily include predictors in the model that in fact have no effect on the outcome. Furthermore, there are many ways to compute variable importance measures [TODO: ADD REFS], which each may yield different conclusions, especially when predictors are correlated. Especially with correlated predictors, permuting the values of predictor variables may lead to unrealistic data patterns.



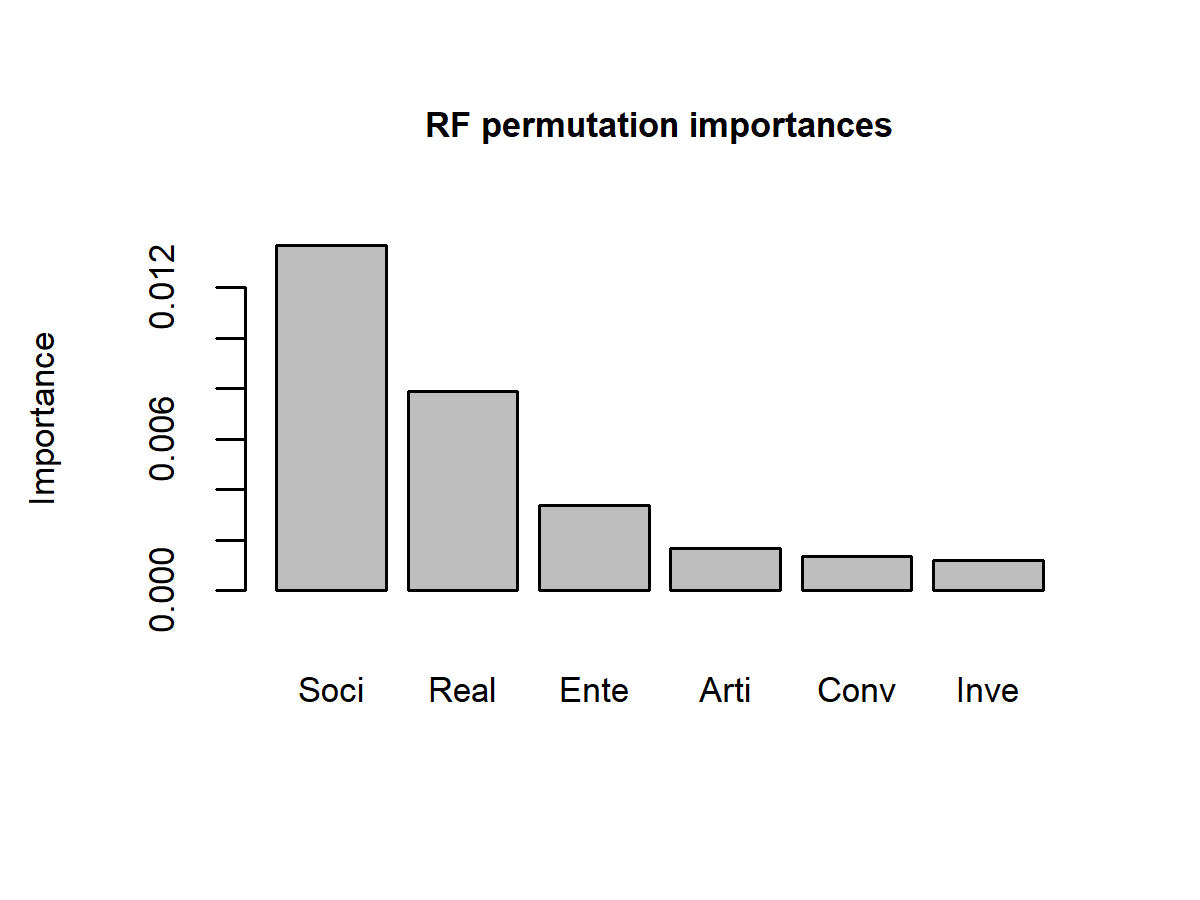
The variable importances of the gradient boosted ensemble indicate a similar pattern as the logistic regression and the GAM.

#### Random forest

Another popular decision-tree ensembling method are random forests (Breiman, 2001). Like boosted tree ensembles, random forests fit a large number of decision trees, and the ensemble’s predictions average over the predictions of the individual trees. Random forests do not employ sequential learning: each tree is fitted, without adjusting for predictions of the other trees in the ensemble. Unlike boosting, random forests employ trees with many splits: in the original algorithm (Breiman, 2001), the training observations are split into subgroups until no further splitting is possible. This yields very complex trees, that have very low bias, but very high variance. This very high variance is reduced by averaging over the predictions of the many trees in the ensemble. Later studies, however, have shown that large trees can lead to unstable results when there are many correlated predictors that are at best weakly correlated to the response (Segal, 2004). It is thus beneficial to grow large, but not too large trees; the optimal tree size is best determined through CV.

The most characteristic feature of random forests is how it selects variables for splitting. Random forests select a random sample of *mtry* candidate predictor variables for every split in every tree. From this set of predictor variables, the best splitting variable and value is then selected. This adds a level of randomization to the algorithm, from which the method also received its name. This randomization might appear counter intuitive: In fact, it only increases the variance of each individual tree, without reducing bias. However, the predictive performance of the ensemble is actually improved by the randomization. Without the randomization, each tree of the ensemble would likely use the same set of relatively strong predictors, and thus be very similar. Averaging over many very similar trees is unlikely to improve predictive accuracy. Because the randomization makes the trees more dissimilar, it strengthens the performance of the ensemble.

Using 10-fold CV, we found the optimal value of the *mtry* parameter to be 3, and the optimal minimum sample size in each terminal node to be 1,000. With a training sample size of 41694, this yields trees with at most 41 terminal nodes.



The variable importances of the random forest indicate a similar pattern as the logistic regression, GAM and boosted tree ensemble.

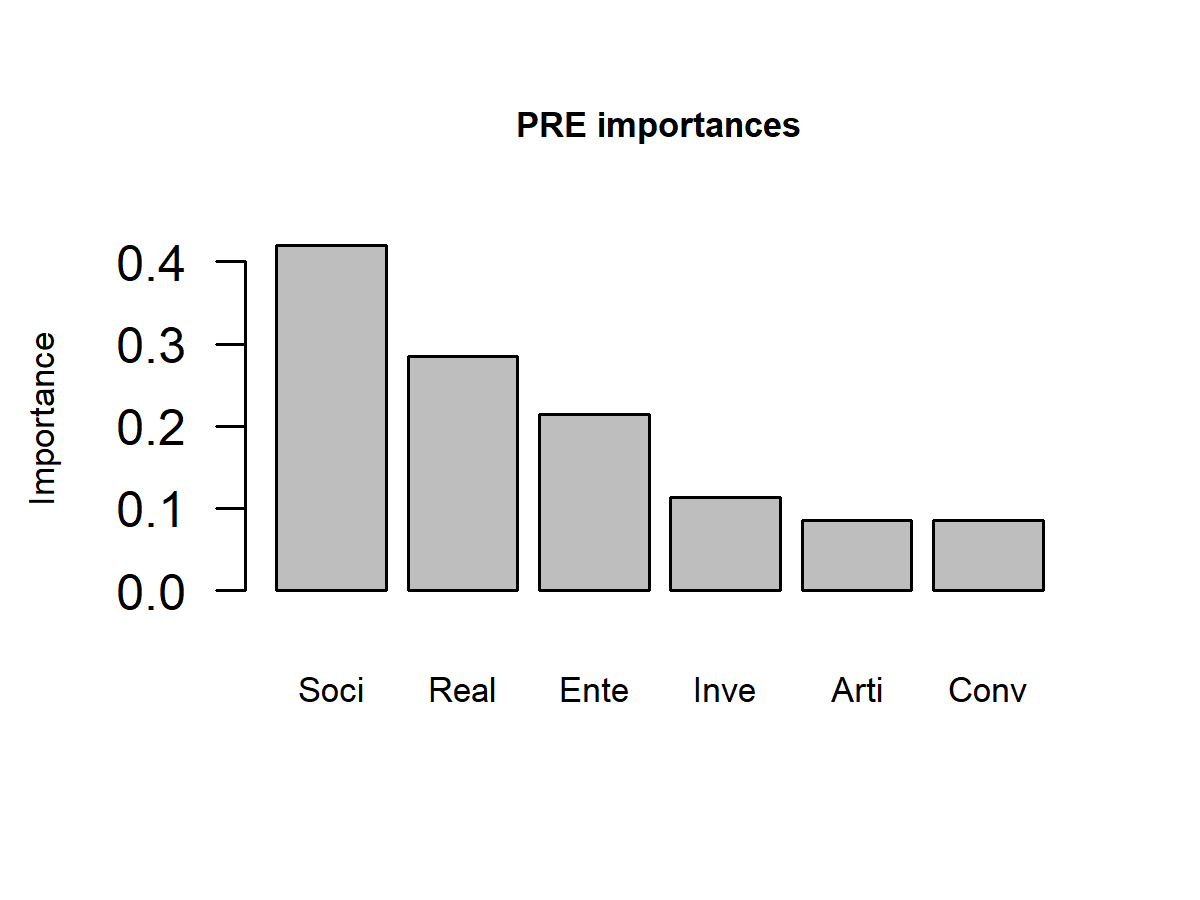
#### Prediction rule ensemble

Prediction rule ensembles aim to strike a balance between the high predictive accuracy of decision tree ensembles, and the ease of interpretability of single decision trees and GLMs (Fokkema, 2020; Fokkema & Strobl, 2020). The method fits a boosted decision tree ensemble to the training dataset, takes every node from every tree in the ensemble and codes it as a rule of the form *if [conditions] then [prediction]*. For example, we can code membership of node 2 in the decision tree above as the condition: *Social 27*. Similarly, membership of node 14 can be coded as a set of conditions: *Social > 27 & Realistic > 17 & Realistic 24*. The rules are then dummy variables, indicating membership of a node: A rule takes a value of 1 if its conditions apply (i.e., the observation is a member of the node), and takes a value of 0 if its conditions do not apply (i.e., the observation is not a member of the node).

Through applying a penalized regression on a dataset containing both these rules as well as the original predictor variables, we obtain a prediction rule ensemble (PRE). This technique combines the strengths of penalized regression, decision tree ensembles and single decision trees. When a sparse penalized regression method, like lasso regression is used, we can obtain a small set of rules as the final ensemble. Although the boosted decision tree ensemble will initially contribute a very large number of nodes rules), use of lasso regression will give many of these rules a weight of zero, thus eliminating them from the ensemble.

The fitted PRE consists of 48 rules, providing a great simplification compared to the trees of the boosted ensemble and random forest. Note that the current dataset is exceptionally large, which tends to result in longer rule lists if only predictive accuracy is optimized, because very large samples allow for capturing even very nuanced effects.

Below, the eight most important rules are provided, and the variable importances are depicted. Each rule has obtained an estimated coefficient, these are logistic regression coefficients. They reflect the expected increase in log-odds if the conditions of the rule apply. The importances show a pattern similar to that of the previous methods. Note that variable importances for PRE are computed differently for decision-tree ensembles, and do not require permutation of predictor variables. More detail about the method can be found in Fokkema & Strobl (2020), which provides an introduction to PRE aimed at researchers in psychology.



## description coefficient  
## Soci > 27 & Ente <= 31 & Conv <= 30 0.182  
## Soci > 23 & Ente <= 29 & Real <= 24 0.181  
## Real > 10 & Soci <= 35 -0.175  
## Real <= 22 & Soci > 19 & Inve > 18 0.138  
## Inve > 10 & Real <= 13 0.120  
## Conv <= 23 & Arti <= 29 & Soci > 21 0.112  
## Real <= 26 & Soci > 18 & Ente <= 27 0.113  
## Soci <= 33 & Real > 9 -0.111

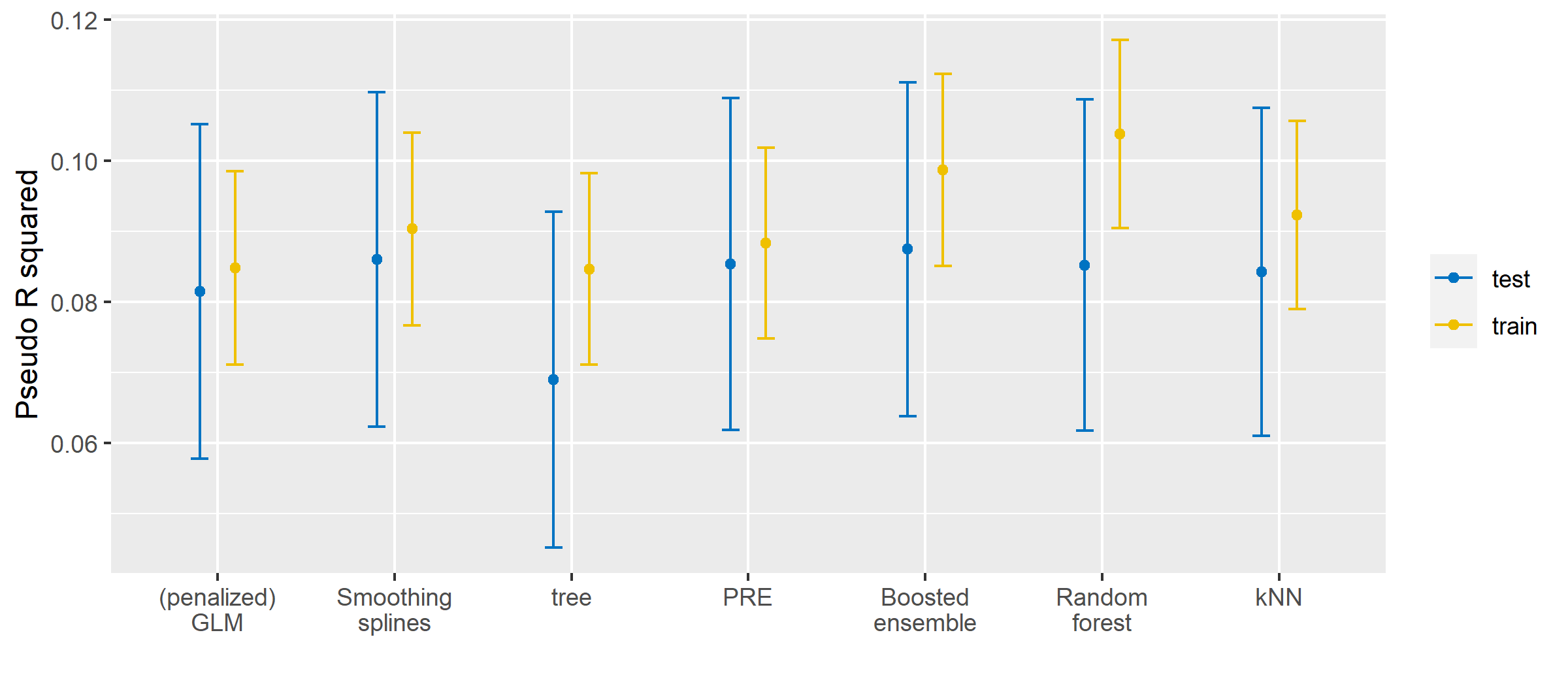
#### nearest neighbours

A prime example of a highly flexible method, perhaps the most non-parametric method of all, is the method of *k*-nearest neighbours (kNN). In fact, kNN does not even fit a model; it merely remembers the training observations. To compute predictions for new observations, kNN computes the distance of a new observation to all the training observations in order to find the *k* nearest ones (the neighbours). It then takes the response variable values of these neighbours and takes their mean as the predicted value for the new observation. This provides the greatest possible flexibility of all prediction methods, as it does not impose *any* a-priori restriction on the shape of association between predictors and response. This flexibility is both the strength and weakness of the method: when the number of predictor variables increases, the performance of kNN worsens fast. Only in lower dimensions is the great flexibility of kNN a beneficial property.

The method has only a single tuning parameter: *k*. With larger values of *k*, the predicted value for a new observation averages over a larger number of observations (neighbours). Thus, higher values of yield lower variance, but higher bias. Using 10-fold CV, we found the optimal value for to be 300.

#### Comparing predictive accuracy

For each of the fitted models, we computed predictive accuracy on test observations. Below, pseudo- values are depicted; the error bars depict 95% confidence intervals.



The results indicate that best performance on the test data was obtained with the boosted tree ensemble, followed by the generalized additive model, followed by the prediction rule ensemble, followed by the random forest, followed by nearest neighbours, followed by logistic regression, and finally the decision tree. This latter result is rather unsurprising: a single decision tree is generally expected to have both higher variance as well as higher bias than other prediction methods (but they often ‘win’ in terms of interpretability). The boosted tree ensemble performing best is also not very surprising, giving its high competitive performance in forecasting competitions.

From the graph, we obtain three key take-aways:

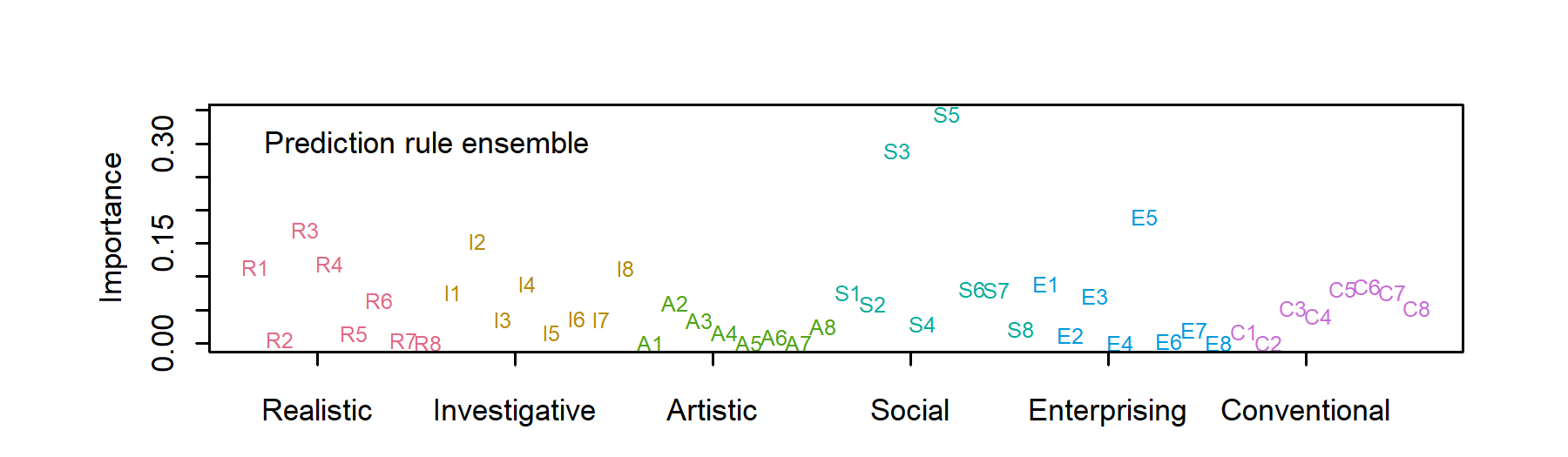
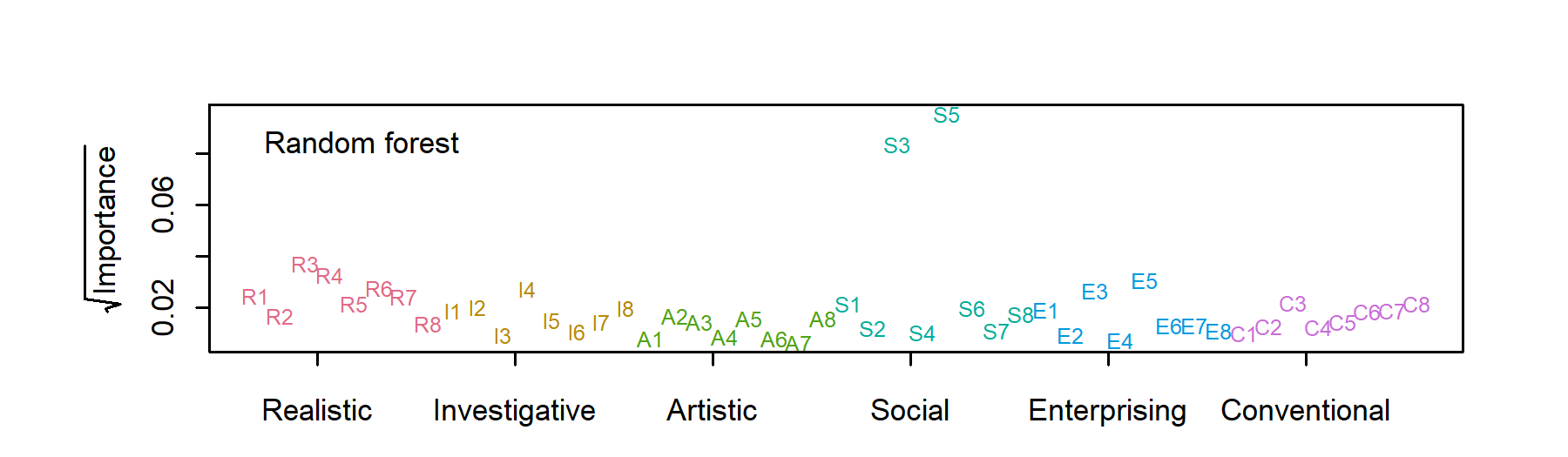
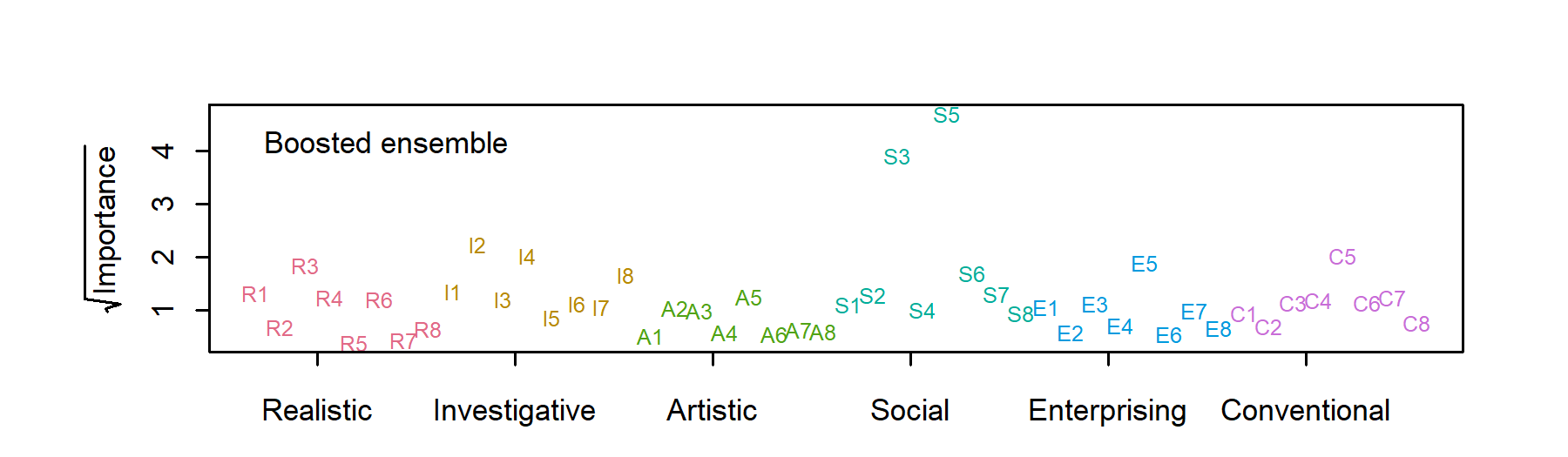
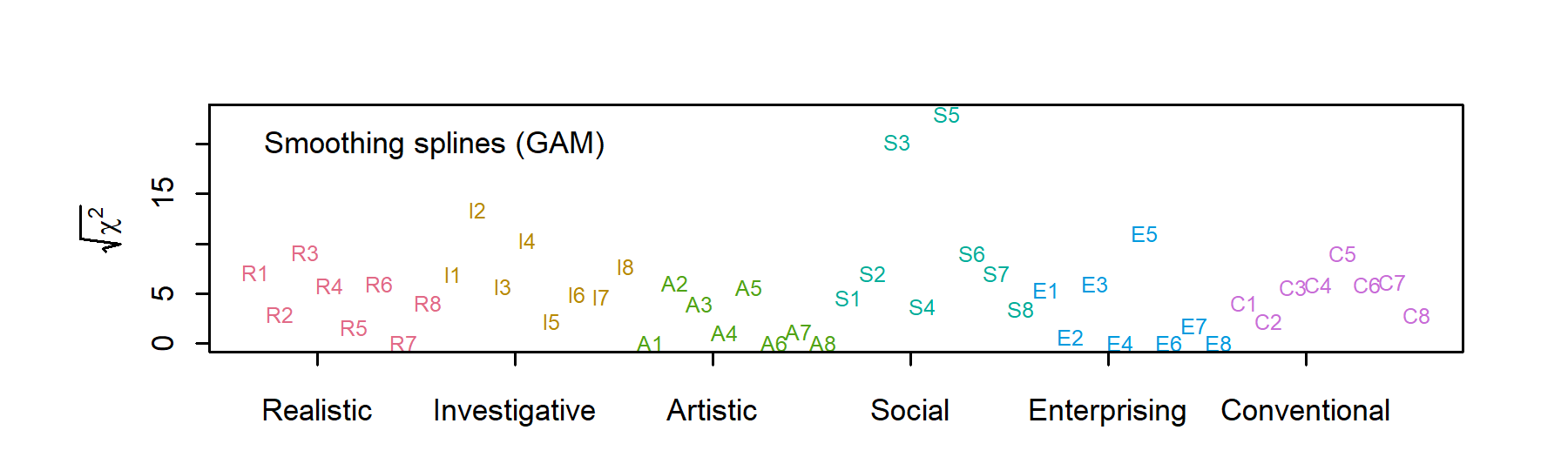
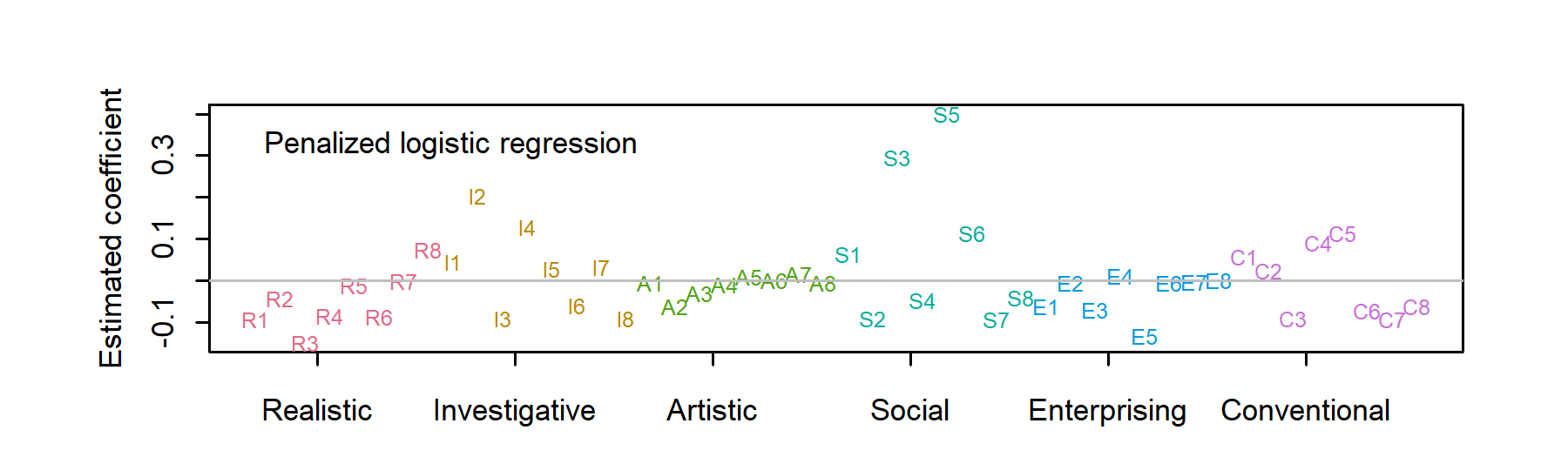
1. None of the methods performs significantly worse or better than any of the other methods.
2. The difference between training and test performance increases with increasing flexibility. The methods that can incorporate linear main effects (logistic regression, smoothing splines, prediction rule ensembles) show the smallest difference in performance on training and test data. These methods thus appear less likely to overfit. Especially the generalized additive model appears to provide good interpretability as well as accuracy, and good overfitting control.
3. The more flexible methods (single decision tree, nearest neighbours, boosted ensemble and the random forest) show greater susceptibility to overfitting.

The subscale scores did not provide strong predictive power, with indicative of a moderate effect. We therefore evaluate whether use of 48 RIASEC item scores for prediction improves performance:

### Prediction using item scores

For the logistic regression, we found that regularization with the lasso penalty would improve predictive accuracy, according to the 10-fold CV results. For the generalized additive model, we set the number of basis functions of the thin-plate regression splines to 4, instead of the default of 8, because item responses can only take 5 possible values; increasing the number of basis functions would yield an unidentified model. Furthermore, because of the larger number of predictors, and the benefit of penalization observed in the logistic regression, we also applied a penalty to the linear part of each smoothing spline function, so that some items can be completely eliminated from the model. For all other methods, we again applied 10-fold CV on the training data to determine the optimal value of the model-fitting parameters.

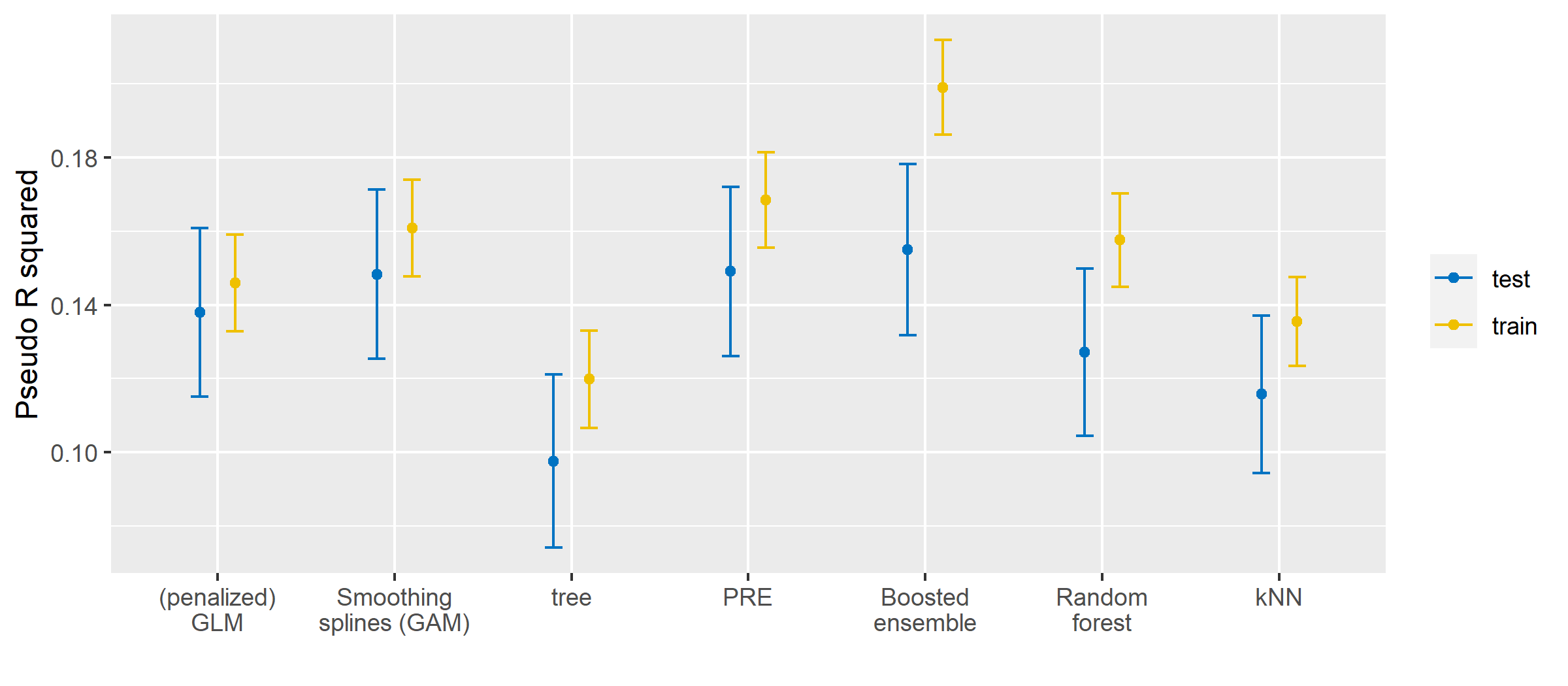
For each method, except the single decision tree and NN, we again computed measures of variable importance, which are plotted below. Note that the coefficients of the logistic regression reflect both direction and strength of the effect. For the other models, the measures of variable importance only reflect the strength, not the direction, of the effects of the included variables. Furthermore, the coefficients in the logistic regression and the importance measures of the prediction rule ensemble are on the scale of standard deviations, not variances. The importance measures for the generalized additive model, boosted ensemble and random forest are on the scale of variances, and we thus plotted their square roots to allow for easier comparison between methods.



Compared to the analyses using subscale scores, the variable contributions yield similar conclusions: the Social preferences scale contains the strongest predictors, while the items in the Conventional and Artistic subscales contribute least. All methods found items S3 (“I would like to help people who have problems with drugs or alcohol”) and S5 (“I would like to help people with family-related problems”) to contribute most. For most methods, this was followed by item R3 (“I would like to work on an offshore oil-drilling rig”), I2 (“I would like to study animal behavior”) and E5 (“I would like to manage a department within a large company”). The main difference with the analyses using subscale scores is that the item score analyses suggest greater relevance of the Investigative subscale items.

#### Comparing predictive accuracy

Again, we computed pseudo- values for each of the fitted models:



Using item instead of scale scores yielded a substantial (about 50%) increase in variance explained. Again, best performance on the test data was obtained with the boosted tree ensemble. This time, it was followed by the prediction rule ensemble, then the generalized additive model, logistic regression, random forest, nearest neighbours, and finally the decision tree.

The graph adds to our earlier take-aways as follows:

1. With a larger number of predictors, differences in performance between the methods become more pronounced, but the penalized logistic regression is not substantially, nor significantly outperformed by more sophisticated methods.
2. With a larger number of predictors, the difference in performance on training and test data becomes more pronounced. The higher dimensionality seems to have created more room for overfitting, even for methods with powerful built-in overfitting control.
3. Like in the subscale analyses, we see that methods incorporating linear main effects (logistic regression, smoothing splines, prediction rule ensembles) appear less susceptible to overfitting. The more flexible methods (single decision tree, nearest neighbours, boosted ensemble and random forest) appear more susceptible to overfitting.

# Discussion

Our conclusions can be succinctly summarized as: It is difficult to outperform logistic regression. Linear main effects models (i.e., GLMs) tend to capture most of the explainable variance. This finding corresponds to a range of previous studies noting a lack of (substantial or significant) benefit of sophisticated machine learning methods over (penalized) regression, in prediction problems from psychology (Elleman et al., 2020; Littlefield et al., 2021; Perlis, 2013; Webb et al., 2020) and medicine (Christodoulou et al., 2019; Desai et al., 2020; Gravesteijn et al., 2020; Lynam et al., 2020; Nusinovici et al., 2020).

Sophisticated methods can only improve upon linear main-effects models by capturing more nuanced non-linearities and interactions. Almost by definition, these effects are of smaller size. Capturing these smaller, more nuanced effects comes at the price of an increased tendency to overfit. To reliably approximate small effects, much larger sample sizes are needed. Even if sophisticated methods outperform simpler methods like logistic regression in terms of predictive accuracy on test data, their tendency to overfit and their black-box nature may make them less suited for increasing scientific understanding, and/or making influential decisions about individuals (e.g., clinical, selection settings).

Perhaps GAMs and PREs may provide the most steady improvement on (penalized) GLMs. They are essentially GLMs with added flexibility for capturing non-linearities, but provide robust overfitting control and also retain interpretability. Especially GAMs may provide the ‘best of both worlds’: They provide the flexibility of modern statistical learning, robust overfitting control and allow for performing statistical inference. Flexible machine-learning methods especially fall short in terms of the latter, which limits their use for increasing scientific understanding.

Our finding that item scores can provide better predictive accuracy than subscale scores corresponds to previous studies (e.g., Seeboth & Mõttus, 2018; Stewart et al., 2021). As also noted by Yarkoni (2020), a (very) large number of item scores will outperform any predictive model fitted on subscale scores, given a large enough sample size. At the same time, a handful of subscale scores is easier to interpret and use than hundreds of personality items. Also, with smaller samples (e.g., or 500), including prior knowledge about the subscale structure, through the use of subscale or factor scores, may likely improve predictive accuracy (Rooij et al., under review).

Big-data applications involving, for example, image-, video- and text-based analytics may likely exhibit (much) stronger patterns of non-linearity and interaction than the analytic example presented here. More sophisticated methods like deep neural networks may even be called for in such applications. However, similar rules of sampling and statistics apply in such applications: The more complex the patterns that we want to capture, the larger the sample sizes required. Sample size requirements for artificial neural networks by far exceed the sample sizes common in our field (e.g., Alwosheel et al., 2018).

Furthermore, in image classification, rather arbitrary image characteristics can easily lead neural nets to misclassify images (Szegedy et al., 2014, pp. AlcoyLi19, HendyZhao21). This has been attributed to the extreme non-linearity of deep neural networks, but it is likely largely due to the high dimensionality of image data (Goodfellow et al., 2015). High dimensionality yields great flexibility for any algorithms to adapt to chance or irrelevant characteristics of images, and this may be worse for highly flexible and non-linear algorithms. For example, with MRI data, scanner effects are substantial (Glocker et al., 2019). That is, it is relatively easy for an algorithm to identify subjects that were put in the same scanner. This will lead to inflated estimates of predictive accuracy and invalid conclusion when, for example, we are interested in predicting who suffers from a neuro-degenerative disease, but the patients and controls in a study were scanned with different machines. While evaluating predictive accuracy in the study data, the model may appear to be able to discriminate between patients and controls, while in fact it can discriminate well between subjects put in different scanners. The number of irrelevant characteristics that an algorithm may be using for prediction increases fast with increasing dimension.

There is no doubt that image, text, audio, video and sensor-based data (will) provide novel ways of assessing psychological traits (Boyd et al., 2020; Gillan & Rutledge, 2021). Their relatively unobtrusiveness also opens up new avenues for assessment. At the same time, the unobtrusiveness brings ethical risks. Furthermore, the focus of machine learning on predictive accuracy on unseen observations is beneficial for the field of assessment. We should, however, guard against a blind focus on maximizing predictive accuracy on test observations, as this disregards two important issues:

* Data points analysed in, e.g., research settings or forecasting competitions, will generally differ from the data points that the predictive model will be applied to in practice. These differences may be subtle in relatively closed, low-stakes systems, like online recommender systems. But much of psychological assessment is focused on offline, out-of-lab human behavior, often with high stakes. Generalizing research findings to the real world remains difficult; external validity has not become irrelevant all of a sudden. Gains in predictive accuracy in controlled research settings may be swamped by practical aspects of data problems, like population drift, measurement error, ethics (interpretability) and costs (of data collection) (Efron, 2020; Fokkema et al., 2015; Hand, 2006; Luijken et al., 2019; Rauthmann, 2020).
* From both an ethical and scientific perspective, validity has become more (not less!) important with new and bigger sources of data. A blind focus on predictive validity leads to black-box assessment procedures with limited content, internal and construct validity. For opening the black box, there is an important role for the field of psychological assessment and psychometrics. Not only by applying our existing theory, evidence and methods, but also by continually improving, adopting and developing them (Alexander III et al., 2020; Bleidorn & Hopwood, 2019; Iliescu & Greiff, 2019; Tay et al., 2020).

Finally, although modern statistical prediction methods have certainly improved our ability to predict, attribute significance and interpret has not become better. Attribution (assigning significance to individual predictors) requires strong individual predictors and large sample sizes (Efron, 2020). This task only becomes more difficult when datasets contain increasing numbers of predictors with modest effects. The task also becomes more difficult with sophisticated prediction methods, that can capture increasingly nuanced non-linear and interaction effects. A range of interpretation tools for black box models have been proposed (variable importances, LIME, Shapley values, SHAP). However, the accuracy of their explanations cannot be quantified (Carvalho et al., 2019; Ross et al., 2017), and their inner workings pose another black box to most users, resulting in misinterpretation and misuse (Kaur et al., 2020; Kenny et al., 2021; Kumar et al., 2020; Rudin, 2019; Waa et al., 2021). With large numbers of predictors, fitted models become inherently difficult to interpret and black-box interpretation tools are unlikely to help with this.

# References

Alexander III, L., Mulfinger, E., & Oswald, F. L. (2020). Using big data and machine learning in personality measurement: Opportunities and challenges. *European Journal of Personality*, *34*(5), 632–648.

Alwosheel, A., Cranenburgh, S. van, & Chorus, C. G. (2018). Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. *Journal of Choice Modelling*, *28*, 167–182.

Ammannato, G., & Chiesi, F. (2020). Playing with networks: Using video games as a psychological assessment tool. *European Journal of Psychological Assessment*, *36*(6), 973. <https://doi.org/10.1027/1015-5759/a000608>

Bleidorn, W., & Hopwood, C. J. (2019). Using machine learning to advance personality assessment and theory. *Personality and Social Psychology Review*, *23*(2), 190–203.

Boyd, R. L., Pasca, P., & Lanning, K. (2020). The personality panorama: Conceptualizing personality through big behavioural data. *European Journal of Personality*, *34*(5), 599–612.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Bühlmann, P., & Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, *22*(4), 477–505.

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8), 832.

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, *110*, 12–22.

Desai, R. J., Wang, S. V., Vaduganathan, M., Evers, T., & Schneeweiss, S. (2020). Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. *JAMA Network Open*, *3*(1), e1918962–e1918962.

Efron, B. (2020). Prediction, estimation, and attribution. *International Statistical Review*, *88*, S28–S59.

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802–813.

Elleman, L. G., McDougald, S. K., Condon, D. M., & Revelle, W. (2020). That takes the BISCUIT: Predictive accuracy and parsimony of four statistical learning techniques in personality data, with data missingness conditions. *European Journal of Psychological Assessment*, *36*(6), 948.

Finkelman, M. D., Jamison, R. N., Kulich, R. J., Butler, S. F., Smits, N., & Weiner, S. G. (2020). A comparison of short forms of the Screener and Opioid Assessment for Patients With Pain–Revised (SOAPP-R). *European Journal of Psychological Assessment*, *36*(2), 387.

Fokkema, M. (2020). Fitting prediction rule ensembles with R package pre. *Journal of Statistical Software*, *92*(1), 1–30.

Fokkema, M., & Greiff, S. (2018). Would you prefer your coefficients with a little bias, or rather with a lot of variance? *European Journal of Psychological Assessment*, *34*(6). <https://doi.org/10.1027/1015-5759/a000514>

Fokkema, M., Smits, N., Kelderman, H., & Penninx, B. W. (2015). Connecting clinical and actuarial prediction with rule-based methods. *Psychological Assessment*, *27*(2), 636.

Fokkema, M., & Strobl, C. (2020). Fitting prediction rule ensembles to psychological research data: An introduction and tutorial. *Psychological Methods*.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1.

Gillan, C. M., & Rutledge, R. B. (2021). Smartphones and the neuroscience of mental health. *Annual Review of Neuroscience*, *44*.

Glocker, B., Robinson, R., Castro, D. C., Dou, Q., & Konukoglu, E. (2019). Machine learning with multi-site imaging data: An empirical study on the impact of scanner effects. *arXiv Preprint arXiv:1910.04597*.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *3nd International Conference on Learning Representations, ICLR 2015*. <https://arxiv.org/abs/1412.6572>

Gravesteijn, B. Y., Nieboer, D., Ercole, A., Lingsma, H. F., Nelson, D., Van Calster, B., Steyerberg, E. W., Åkerlund, C., Amrein, K., Andelic, N., & others. (2020). Machine learning algorithms performed no better than regression models for prognostication in traumatic brain injury. *Journal of Clinical Epidemiology*, *122*, 95–107.

Greenwell, B., Boehmke, B., Cunningham, J., & Developers, G. (2020). *Gbm: Generalized boosted regression models*. <https://CRAN.R-project.org/package=gbm>

Hand, D. J. (2006). Classifier technology and the illusion of progress. *Statistical Science*, *21*(1), 1–14.

Holland, J. L. (1959). A theory of vocational choice. *Journal of Counseling Psychology*, *6*(1), 35.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674. <https://doi.org/10.1198/106186006X133933>

Hothorn, T., & Zeileis, A. (2015). partykit: A modular toolkit for recursive partytioning in R. *Journal of Machine Learning Research*, *16*, 3905–3909. <https://jmlr.org/papers/v16/hothorn15a.html>

Iliescu, D., & Greiff, S. (2019). The impact of technology on psychological testing in practice and policy. *European Journal of Psychological Assessment*, *35*(2), 151–155.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020). Interpreting interpretability: Understanding data scientists’ use of interpretability tools for machine learning. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.

Kenny, E. M., Ford, C., Quinn, M., & Keane, M. T. (2021). Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. *Artificial Intelligence*, *294*, 103459.

Kuhn, M. (2021). *Caret: Classification and regression training*. <https://CRAN.R-project.org/package=caret>

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., & Friedler, S. (2020). Problems with Shapley-value-based explanations as feature importance measures. *Proceedings of the 37th International Conference on Machine Learning*, 5491–5500.

Larson, S. C. (1931). The shrinkage of the coefficient of multiple correlation. *Journal of Educational Psychology*, *22*(1), 45.

Liao, H.-Y., Armstrong, P. I., & Rounds, J. (2008). Development and initial validation of public domain basic interest markers. *Journal of Vocational Behavior*, *73*(1), 159–183.

Littlefield, A. K., Cooke, J. T., Bagge, C. L., Glenn, C. R., Kleiman, E. M., Jacobucci, R., Millner, A. J., & Steinley, D. (2021). Machine learning to classify suicidal thoughts and behaviors: Implementation within the common data elements used by the military suicide research consortium. *Clinical Psychological Science*, *9*(3), 467–481.

Luijken, K., Groenwold, R. H., Van Calster, B., Steyerberg, E. W., & Smeden, M. van. (2019). Impact of predictor measurement heterogeneity across settings on the performance of prediction models: A measurement error perspective. *Statistics in Medicine*, *38*(18), 3444–3459.

Lynam, A. L., Dennis, J. M., Owen, K. R., Oram, R. A., Jones, A. G., Shields, B. M., & Ferrat, L. A. (2020). Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: Application to the discrimination between type 1 and type 2 diabetes in young adults. *Diagnostic and Prognostic Research*, *4*, 1–10.

Mosier, C. I. (1951). I. Problems and designs of cross-validation 1. *Educational and Psychological Measurement*, *11*(1), 5–11.

Nusinovici, S., Tham, Y. C., Yan, M. Y. C., Ting, D. S. W., Li, J., Sabanayagam, C., Wong, T. Y., & Cheng, C.-Y. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of Clinical Epidemiology*, *122*, 56–69.

Opitz, A., Heene, M., & Fischer, F. (2021). Using differential item functioning to analyze the domain generality of a common scientific reasoning test. *European Journal of Psychological Assessment*. <https://doi.org/10.1027/1015-5759/a000662>

Park, J. J., Chow, S.-M., Fisher, Z. F., & Molenaar, P. (2020). Affect and personality: Ramifications of modeling (non-) directionality in dynamic network models. *European Journal of Psychological Assessment*, *36*(6), 1009. <https://doi.org/10.1027/1015-5759/a000612>

Perlis, R. H. (2013). A clinical risk stratification tool for predicting treatment resistance in major depressive disorder. *Biological Psychiatry*, *74*(1), 7–14.

Rauthmann, J. F. (2020). A (more) behavioural science of personality in the age of multi–modal sensing, big data, machine learning, and artificial intelligence. In *European Journal of Personality* (Vol. 34, pp. 593–598). SAGE Publications Sage UK: London, England. <https://doi.org/10.1002/per.2310>

Rooij, M. de, Karch, J. D., Fokkema, M., Bakk, Z., Pratiwi, B. C., & Kelderman, H. (under review). *SEM-based out-of-sample predictions*.

Rooij, M. de, & Weeda, W. (2020). Cross-validation: A method every psychologist should know. *Advances in Methods and Practices in Psychological Science*, *3*(2), 248–263.

Ross, A. S., Hughes, M. C., & Doshi-Velez, F. (2017). Right for the right reasons: Training differentiable models by constraining their explanations. *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2662–2670.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.

Sauer, S., Buettner, R., Heidenreich, T., Lemke, J., Berg, C., & Kurz, C. (2016). Mindful machine learning. *European Journal of Psychological Assessment*.

Seeboth, A., & Mõttus, R. (2018). Successful explanations start with accurate descriptions: Questionnaire items as personality markers for more accurate predictions. *European Journal of Personality*, *32*(3), 186–201.

Segal, M. R. (2004). Machine learning benchmarks and random forest regression. *eScholarship Repository, University of California*. <http://repositories.cdlib.org/cbmb/bench_rf_regn>

Stewart, R. D., Mõttus, R., Seeboth, A., Soto, C. J., & Johnson, W. (2021). The finer details? The predictability of life outcomes from big five domains, facets, and nuances. *Journal of Personality*.

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.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. *2nd International Conference on Learning Representations, ICLR 2014*. <https://arxiv.org/abs/1312.6199>

Tay, L., Woo, S. E., Hickman, L., & Saef, R. M. (2020). Psychometric and validity issues in machine learning approaches to personality assessment: A focus on social media text mining. *European Journal of Personality*, *34*(5), 826–844.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th Edition). Springer. <https://www.stats.ox.ac.uk/pub/MASS4/>

Waa, J. van der, Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, *291*, 103404.

Webb, C. A., Cohen, Z. D., Beard, C., Forgeard, M., Peckham, A. D., & Björgvinsson, T. (2020). Personalized prognostic prediction of treatment outcome for depressed patients in a naturalistic psychiatric hospital setting: A comparison of machine learning approaches. *Journal of Consulting and Clinical Psychology*, *88*(1), 25.

Wood, S. N. (2017). *Generalized additive models: An introduction with R*. CRC press.

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), 1–17. <https://doi.org/10.18637/jss.v077.i01>

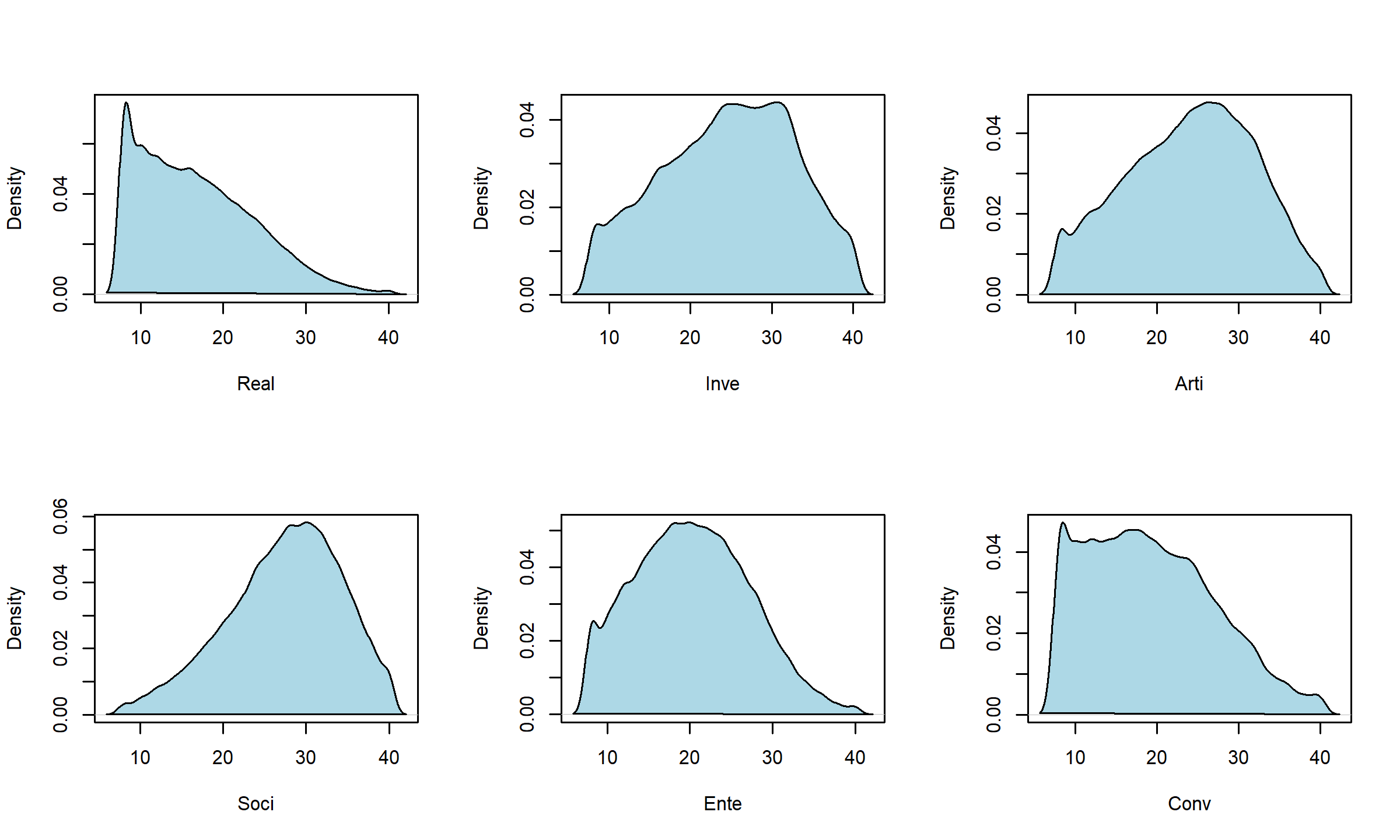
Yarkoni, T. (2020). Implicit realism impedes progress in psychology: Comment on fried (2020). *Psychological Inquiry*, *31*(4), 326–333. <https://doi.org/10.1080/1047840X.2020.1853478>

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.

# Appendix A: RIASEC items

|  |
| --- |
| **The following items were rated on a 1-5 scale of how much they would like to perform that task, with the labels 1=Dislike, 3=Neutral, 5=Enjoy:** |
| R1 Test the quality of parts before shipment |
| R2 Lay brick or tile |
| R3 Work on an offshore oil-drilling rig |
| R4 Assemble electronic parts |
| R5 Operate a grinding machine in a factory |
| R6 Fix a broken faucet |
| R7 Assemble products in a factory |
| R8 Install flooring in houses |
| I1 Study the structure of the human body |
| I2 Study animal behavior |
| I3 Do research on plants or animals |
| I4 Develop a new medical treatment or procedure |
| I5 Conduct biological research |
| I6 Study whales and other types of marine life |
| I7 Work in a biology lab |
| I8 Make a map of the bottom of an ocean |
| A1 Conduct a musical choir |
| A2 Direct a play |
| A3 Design artwork for magazines |
| A4 Write a song |
| A5 Write books or plays |
| A6 Play a musical instrument |
| A7 Perform stunts for a movie or television show |
| A8 Design sets for plays |
| S1 Give career guidance to people |
| S2 Do volunteer work at a non-profit organization |
| S3 Help people who have problems with drugs or alcohol |
| S4 Teach an individual an exercise routine |
| S5 Help people with family-related problems |
| S6 Supervise the activities of children at a camp |
| S7 Teach children how to read |
| S8 Help elderly people with their daily activities |
| E1 Sell restaurant franchises to individuals |
| E2 Sell merchandise at a department store |
| E3 Manage the operations of a hotel |
| E4 Operate a beauty salon or barber shop |
| E5 Manage a department within a large company |
| E6 Manage a clothing store |
| E7 Sell houses |
| E8 Run a toy store |
| C1 Generate the monthly payroll checks for an office |
| C2 Inventory supplies using a hand-held computer |
| C3 Use a computer program to generate customer bills |
| C4 Maintain employee records |
| C5 Compute and record statistical and other numerical data |
| C6 Operate a calculator |
| C7 Handle customers’ bank transactions |
| C8 Keep shipping and receiving records |

# Appendix B: univariate distributions and correlations between subscales



## Real Inve Arti Soci Ente Conv  
## Real 1.000 0.332 0.182 0.049 0.305 0.460  
## Inve 0.332 1.000 0.329 0.141 0.016 0.065  
## Arti 0.182 0.329 1.000 0.290 0.253 -0.056  
## Soci 0.049 0.141 0.290 1.000 0.356 0.124  
## Ente 0.305 0.016 0.253 0.356 1.000 0.464  
## Conv 0.460 0.065 -0.056 0.124 0.464 1.000