Machine learning and prediction in psychological assessment:

Some promises and pittfalls.

"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 prediction methods from machine learning (ML) and artificial intelligence (AI) are becoming increasingly popular, also in the field of psychological assessment. These methods provide unprecedented flexibility for modeling large numbers of predictor variables, and non-linear associations between predictors and response. In this paper, we aim to take a look at what these methods may contribute for the assessment of criterion validity, and what their possible drawbacks are. We apply a range of modern statistical prediction methods to a dataset for predicting the university major completed, based on the subscales and items of a scale for vocational preferences. The results indicate that a traditional logistic regression model performs strikingly well already in terms of predictive accuracy. More recent techniques for regularization and incorporating non-linearities, however, can further contribute predictive accuracy and validity.

Introduction

Machine learning (ML) and artificial intelligence (AI) are by now familiar buzzwords in many fields of empirical research, including psychology. In the field of psychological assessment, interest and application of these methods is also increasing. We believe ML and AI have the potential to contribute to our field, but the buzz around these methods can be 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 they are often not so new. Many of the techniques presented as machine learning (e.g., cross validation, regularization, ensembling) have long been known and fruitfully applied in statistics, psychometrics and psychological assessment.

In the current paper, we look at how several modern methods from statistics, ML and AI may contribute to our field, and what their limitations are. Note, we will use the term statistical learning to refer to both traditional and more recent (sometimes referred to as ML or AI) tools for data analysis. As already suggested by the motto at the start of this paper, there is no consensus on whether specific methods are statistical, AI or ML, so we avoid making the distinction altogether. We focus instead on the aim shared by all these methods: Learning from data. We focus on methods for prediction of a response (dependent, criterion) variable, often referred to as 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 the current paper.

Recent shifts in statistical learning

Modern developments in statistical learning methodology have yielded two main shifts:

1. Increased focus on prediction.
2. Increased flexibility: Modern methods allow for capturing non-linear associations and/or modeling large numbers of predictors.

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. Accurate evaluation of predictive accuracy is needed to provide evidence for the validity of test score interpretations, but also when more complex decision systems are developed for data-driven decision making. Traditionally, the field of psychology at large has been mostly interested in *explanation*, or 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).

The traditional focus on explanation may have motivated researchers to compute effect sizes (e.g., R^2, Cohen's *d*) on *training* data; that is, using observations that were also used to fit the model. This leads to overly optimistic effect size estimates. More realistic effect sizes 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). It is interesting to note that cross validation has been discussed in the field of assessment for almost a century (Larson, 1931; Mosier, 1951), but its use has become more common only in recent years.

We believe that the second shift, towards flexibility, brings both promises and pitfalls for our field. Promises, because few if any real-world phenomena behave in a purely linear and additive fashion. Pitfalls, because assumptions of linearity and additivity (i.e., no interactions) are very powerful when it comes to inference and interpretation, even if they are known to be

only partialy true. This means that the often one-sided focus on maximizing predictive accuracy in AI and ML are of limited value when it comes to understanding and explaining behavior, and the role of these methods is, at best, in hypotheses generation.

Of note, unrestricted flexibility leads to overfitting and poorly generalizable results. 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 shape 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 through choosing an appropriate model-fitting procedure. Bias can be increased and variance reduced in various ways, including:

- Limiting the complexity of the functional form (e.g., model only linear associations; model only main effects);
- Limiting the number of potential predictors used (e.g., include only few predictors; use sum or factor scores instead of item scores as predictors);
- Regularized estimation procedures (e.g, lasso, ridge, or elastic net regression; use of Bayesian priors);
- Ensembling (e.g., in psychometrics, multiple items are often aggregatec into subscale or factor scores; in ML predictions of so-called base learners are often aggregated into the predictions of an ensemble).

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 making realistic assumptions. If the bias is not well chosen, predictive accuracy and generalizability will obviously suffer.

Empirical example

We aim to illustrate and compare the use of a range of statistical learning techniques through a data-analytic example. We focus on a predictive validity question: To what extent do the item and subscale scores on a measure of vocational preferences predict the type of university major completed? Note, we will not focus on substantive aspects of this prediction problem, but we use it to illustrate more general principles of flexibility, overfitting and interpretability in predictive modeling in assessment. In test development, providing evidence for criterion validity of the scores is vital as it often is used by practitioners to choose between existing tests. Thus, establishing test-criterion related evidence is a fundamental part of test construction. Therefore, it seems obvious that the potential of statistical learning procedures should come to bear here.

Readers interested in replicating our analyses will find our annotated code and results in the ESM.

Method

Dataset

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 items on vocational preferences, personality and

sociodemographic characteristics. The sample likely does not represent a random sample from a well-defined population, which would normally be required for evaluating a test's validity.

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 presented in Appendix A. The research question from an assessment perspective is whether the RIASEC scores can be used to predict the university major completed. Such evidence could support the use of the scale in applied settings; moreover, the results could inform decision rules.

From the full dataset, we selected participants who completed at least a university degree, yielding a sample of N = 55,593 observations. As the criterion we take a binary variable, indicating whether respondents majored in Psychology (19.42%), or in a different topic (80.58%). Further descriptive statistics of the sample are presented in Appendix B.

Model fitting and evaluation

We fitted a range of traditional and more recent (ML/AI) methods to model the relation between the RIASEC scores and the criterion. This will show the magnitude of differences in performance such algorithms typically yield. Also, it exemplifies the researcher degrees of freedom in such cases and it is thus important to use separate data for fitting and evaluation of the models.

We separated the data into 75% training observations and 25% test observations. Our training sample thus consists of 41,694 respondents, of which 19.46% majored in psychology. Our test sample consisted of 13,899 respondents, of which 19.3% majored in psychology. Other train and test sample sizes may sometimes be preferred, or k-fold CV. Considering the current sample size, however, we do not expect the results to be very sensitive to this choice.

All analyses were performed in **R** (version 4.1.0, R Core Team, 2021). We tuned the model-fitting parameters for all models using resampling and cross validation (CV) on the training observations. We did *not* tune the parameters of the generalized additive models (GAMs), because we expected the defaults to work well out of the box. The specific packages used, as well as the code and results of tuning and fitting the models are provided in the ESM.

We evaluated predictive accuracy of the fitted models by computing the Brier score on test observations. The use of accuracy measures derived from the confusion matrix of actual and predicted classes, like the misclassification error, sensitivity (or recall), positive predictive value (or precision) 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. Methods for predicting a binary outcome should not only provide a predicted class, but also a predicted probability to quantify the uncertainty of the classification. To evaluate performance, the quality of this probability forecast should thus be evaluated (Gneiting & Raftery, 2007).

The Brier score is the mean squared error of the predicted probabilities:

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{p}_i)^2$$

Where y_i is the observed outcome for observation i, taking a value of 0 or 1; $\hat{p}_i$ is the model's predicted probability. We computed Brier scores on training as well as on test observations; thus N can be taken to be the training or the test sample size. A Brier score equal to the variance of y indicates performance no better than chance (in the current dataset, the variance was 0.1946 for training and 0.193 for test data). To obtain a pseudo-R^2 measure, we take 1 minus the Brier score divided by the variance of y, which takes values between 0 (indicating performance no better than chance) and 1 (indicating perfect accuracy).

Results

Considering the two shifts in predictive modeling discussed in the Introduction, we fitted all models twice: Once using subscale scores, once using item scores. This allows us to evaluate whether our conclusions generalize between the two approaches, and to gauge the effect of having a larger pool of predictor variables (which are likely more noisy but possibly more informative of the criterion).

(penalized) Logistic regression

Our benchmark traditional method is an additive generalized linear model (GLM): Logistic regression. If CV results indicated predictive accuracy could be improved by application of a lasso or ridge penalty, we applied it. For prediction with subscale scores, no penalization was found to be optimal. The estimated coefficients for the subscale scores are presented in Figure 3; as expected with the currently large sample size, all subscale scores obtained p-values $< .001$. The strongest effect was a positive effect from the Social preferences scale, and the weakest effect was a negative effect from the Conventional preferences scale.

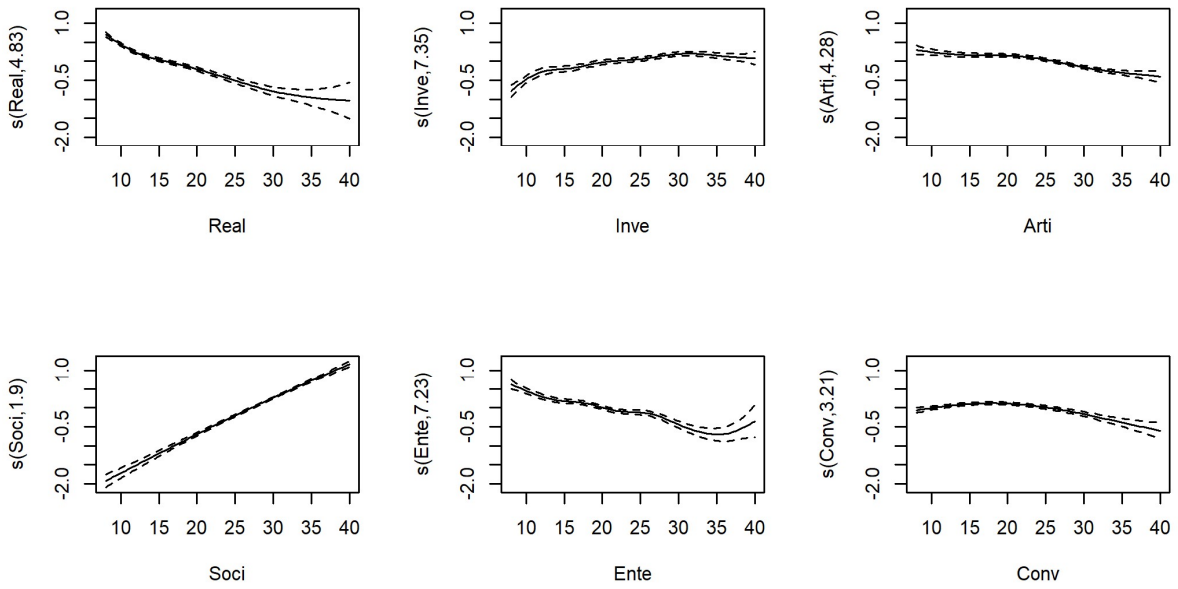
For prediction with item scores, CV indicated optimal performance for a small value of the lasso penalty. The resulting coefficients are depicted in Figure 6 (Appendix C, ESM). The item coefficients indicate similar relevance of the subscales as the previous analysis, but provide a more finegrained view of individual item's contributions.

Generalized additive model

Next we fitted generalized additive models (GAMs) with smoothing splines. Smoothing splines allow for flexibly approximating non-linear shapes of association between predictor and response. At the same time, overfitting is prevented by penalizing the wigglyness of the fitted curves. The splines provide a flexible but smooth approximation to the observed datapoints, while the additive structure provides ease of interpretability because the estimated effects are *conditional* (i.e., keeping the values of all remaining predictors fixed).

Figure 1

Fitted smoothing spline curves for each of the RIASEC subscales



Note. Values on the y-axis reflect the effect on the log-odds of having completed a university major in psychology.

The splines fitted to the subscale scores are presented in Figure 1. Similar to the GLM, we see positive effects of the Social and Investigative subscales, and negative effects of the Realistic, Artistic, Enterprising and Conventional subscales. The Social preferences subscale shows a near-linear effect, while the other subscales' effects clearly exhibit some stronger non-linearity. An advantage of GAMs is that they allow for inference: they provide χ^2 tests to evaluate the significance of the effect of each predictor variable. As expected with the current large sample size, all subscale scores obtained p-values $< .001$.

For the GAM fitted using item scores, we do not depict the curves for space considerations, but Figures 3 and 6 (Appendix C; ESM) show the χ^2 values per subscale and per item, respectively. The figures indicate very similar effects between the GLM and GAM.

We now leave the realm of additive models, and set about fitting models that allow for capturing interaction effects:

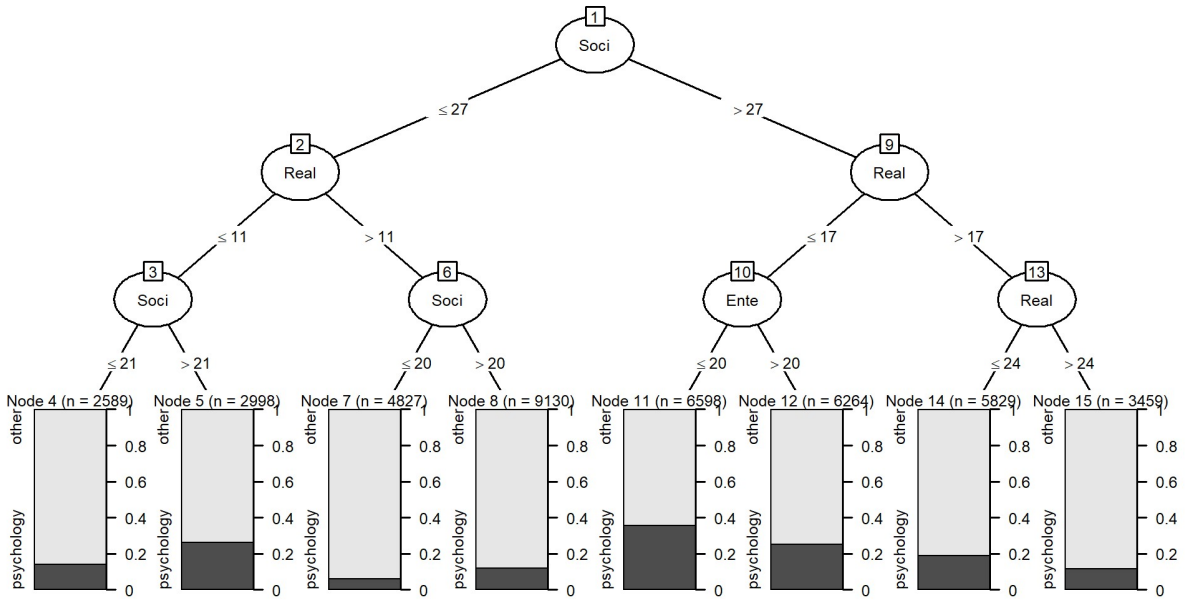
Decision tree

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 Strobl et al. (2009), who provide a comprehensive introduction aimed at psychologists. According to our CV results with the subscales as predictors, a tree depth of seven was optimal, yielding a tree with $2^6 = 128$ terminal nodes. Thus, for this data problem, the most accurate tree is surely not the most interpretable. The predictor variables selected by the trees are depicted in Figures 3 and 6 (Appendix C, ESM).

For illustration, Figure 2 shows the decision tree fitted to the subscale scores, pruned to a depth of three. The Social, Realistic and Enterprising preferences subscales were used in the first splits of the tree. The bars in the terminal nodes depict the proportion of participants within each node that majored in psychology. Thus, the Social subscale shows a positive effect, and the Realistic and Enterprising subscales show a negative effect. With regards to possible interactions, note that split number 10 suggests that the Enterprising subscale appears relevant only for higher values of the Social, and lower values of the Realistic subscales. However, such a split may also reflect additive effects combined with multicollinearity. Although decision trees can *capture* interaction effects, they cannot be straightforwardly used to statistically test their significance.

Figure 2

Conditional inference tree pruned to a depth of three



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. Tree ensembling methods capitalize on this instability. They derive a large number of learners (e.g., trees), each fitted on different versions of the training dataset. Different versions of the training data can be generated, for example, by taking bootstrap samples from the training data, a method also known as bagging. More powerful tree ensembling methods are random forests and boosting. Introductions aimed at psychologists can be found in Strobl et al. (2009) and Miller et al. (2016).

Gradient boosted tree ensemble

The first tree ensemble method we apply to the data is a gradient boosted ensemble. Boosting 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, with only a few 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 next tree.

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 for humans 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 (2001). These quantify how much an ensemble's predictive accuracy would be reduced, if the values of each of the predictor variables are randomly shuffled. The variable importances of the fitted gradient boosting ensembles are depicted in Figures 3 and 4 (Appendix C, ESM).

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 which in fact have no effect on the outcome. Furthermore, there are many ways to compute variable importance measures, which each may yield different conclusions, especially when predictors are correlated (Nicodemus et al., 2010; Nicodemus, 2011; Strobl et al., 2007, 2008). Especially with correlated predictors, permuting the values of predictor variables may

lead to unrealistic data patterns. These issues illustrate the interpretability problems which come along with complex prediction methods such as tree ensembles, support vector machines and (deep) neural networks.

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. The ensemble's predictions are simply the 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 of (Breiman, 2001), trees were grown as large as possible. 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 most characteristic feature of random forests is how it selects variables for splitting: A random sample of *mtry* candidate predictor variables is considered for every split in every tree. From this set of predictor variables, the best splitting variable and value is selected. Without random selection of variables, 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. Thus, the randomization makes the trees more dissimilar, which likely improves performance of the ensemble.

The variable importances of the fitted random forests are depicted in Figures 3 and 4 (Appendix C, ESM).

Prediction rule ensembling

Prediction rule ensembles (PRE) 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, and takes every node from every tree in the ensemble as a rule. For example, node 2 in the tree in Figure 2 contributes the rule *Social* $\leq$ *27*, while node 14 contributes the rule *Social* > *27 & Realistic* > *17 & Realistic* $\leq$ *24*. We can code these rules as a dummy variable, which take a value of 1 if the conditions apply, and 0 if not.

PRE applies lasso regression on a dataset consisting of both these rules and the original predictor variables. It combines the strengths of penalized regression, decision tree ensembles and single decision trees. Although the boosted decision tree ensemble will initially contribute a large number of nodes (rules), use of lasso regression will give many of these rules a weight of zero, which removes them from the final ensemble. As such, PRE provides a sparse and interpretable final model.

The PRE we fitted using the subscale scores consisted of 48 rules, providing a great simplification compared to the > 500 trees of the boosted ensemble and random forest. Note that the current dataset is exceptionally large, which tends to result in longer rule lists when only predictive accuracy is optimized, because very large samples allow for capturing highly nuanced effects. In Table 1, the six most important rules are shown.

Table 1

Six most important rules in the prediction rule ensemble

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

Note that each rule has obtained an estimated coefficient, which are simply logistic regression coefficients: They reflect the expected increase in log-odds if the conditions of the rule apply. PRE also provides variable importance measures, which are presented for the fitted ensembles in Figures 3 and 4 (Appendix C; ESM). An introduction to PRE aimed at psychologists is provided in Fokkema & Strobl (2020).

k 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 training observations, in order to find the *k* nearest ones (the neighbours). It then takes the mean of the response variable over these *k* neighbours as the predicted value. 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 kNN: with

increasing numbers of of predictor variables, the performance of kNN worsens fast. Only in lower dimensions is the great flexibility of kNN beneficial.

kNN 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 k yield lower variance, but higher bias. Furthermore, because kNN is a fully distance-based method, in which all variables obtain the same weight of 1, the method does not provide *any* measure of effect of individual variables, and we thus do not plot variable contributions for kNN here.

Model comparisons

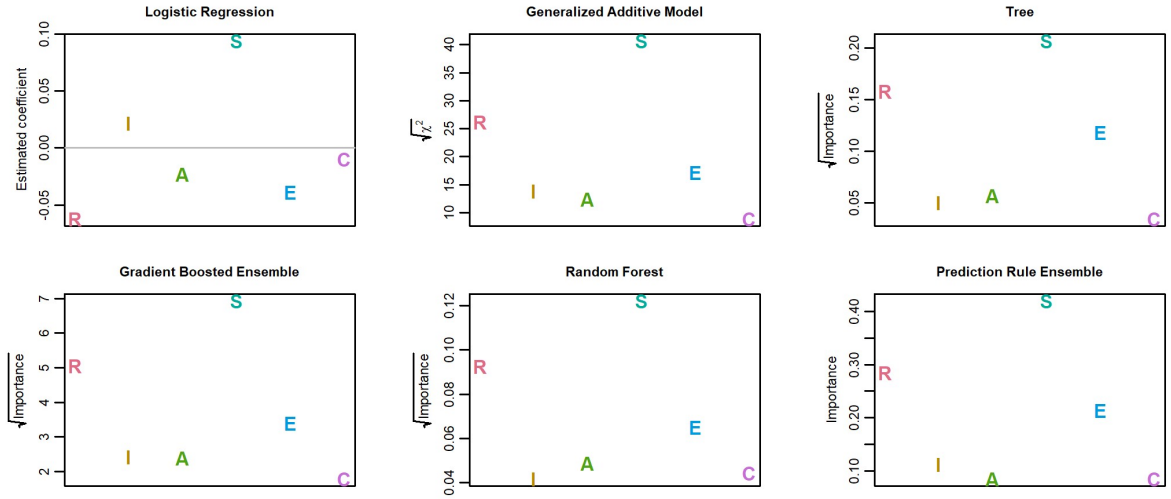
Variable contributions

Figure 3 depicts the variable contributions in the models fitted using RIASEC subscale scores. Note that the coefficients of the logistic regression reflect both direction and strength of the effects. For the other models, the variable contributions only reflect the strength of the variables' effects. Figure 3 shows similar variable contributions for all methods: The Social preferences are most important for predicting university major completed, followed by Realistic, followed by Enterprising preferences, while the Conventional and Artistic subscales contribute least. The variable contributions for models fitted using the item scores as predictors yielded similar conclusions and are provided and discussed in Figure 6 (ESM).

In Figure 4, pseudo-R^2 values on train and test data are depicted with confidence intervals. Confidence intervals for test data are systematically wider than for training data, due to the larger number of training observations.

Figure 3

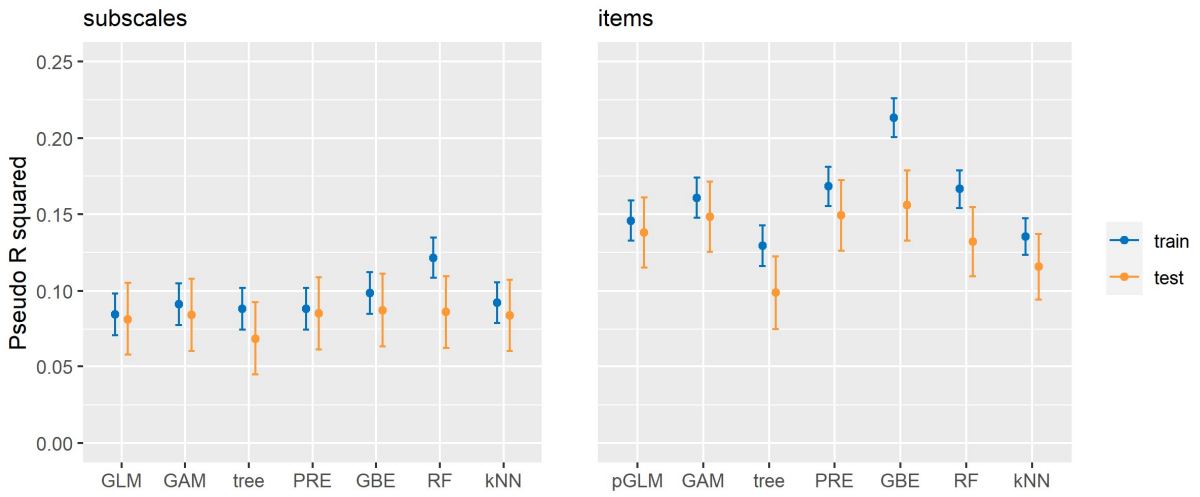
Variable contributions for each of the models fitted using RIASEC subscale scores as predictors



Note. Coefficients in the logistic regression and importance measures of the prediction rule ensemble are on the scale of standard deviations. Importance measures for the other methods are on the scale of variances; for those methods, the square roots are plotted.

Figure 4

Predictive accuracy on train and test observations for each of the models fitted on subscale scores (left panel) and items scores (right panel)



Note. (p)GLM = (penalized) logistic regression; GAM = generalized additive model with smoothing splines; PRE = prediction rule ensemble; GBE = gradient boosted tree ensemble; RF = random forest; kNN = k nearest neighbours.

The left panel of Figure 4 shows that with the subscale score, the best test set performance was obtained with the boosted tree ensemble, closely followed by the generalized additive model, prediction rule ensemble, random forest, *k* nearest neighbours, logistic regression, and finally the decision tree. This latter result is rather unsurprising: a single decision tree is generally expected to have somewhat lower predictive accuracy, but they often 'win' in terms of interpretability, which can be observed in Figure 4, which shows that the decision tree uses only about half of the items for prediction. The boosted tree ensemble performing best is also not very surprising, giving it's high competitive performance in forecasting competitions.

From the left panel in Figure 4, we obtain the following take-aways:

1. On the test data, 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 incorporate linear main effects (logistic regression, GAM, PRE) show the smallest difference in performance between training and test data. These methods thus appear least likely to overfit.

3. The more flexible methods (single tree, kNN, boosted ensemble, random forest) show greater susceptibility to overfitting.

The subscale scores did not provide strong predictive power, with R^2 indicative of a moderate effect. Using item scores as predictors 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, k nearest neighbours, and finally the decision tree.

From the right panel in Figure 4, we can add to our earlier take-aways:

4. With a larger number of predictors, differences in performance between the methods become more pronounced, but none of the more sophisticated methods significantly (or substantially) outperforms the GLM with lasso penalty.

5. With a larger number of predictors, the difference in performance between training and test data becomes more pronounced. Higher dimensionality creates more opportunity for overfitting, even though all methods feature powerful built-in overfitting control.

Discussion

Our conclusions can be succinctly summarized as: Logistic regression is hard to beat. 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 and medicine [Elleman et al. (2020); Littlefield et al. (2021); ChriyJie19; Gravesteijn et al. (2020); Nusinovici et al. (2020); Lynam 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 or 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. Most flexible machine-learning methods especially fall short in terms of the latter, which limits their use for increasing scientific understanding and theory development.

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 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., $N = 300$ 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 exhibit 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 nuanced 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). 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 opens up new avenues for assessment, but the black-box nature of algorithms that can capture complex non-linear effects also brings ethical risks (Boyd et al., 2020; Rudin, 2019).

The focus on predictive accuracy brought about by recent statistical, ML and AI methods 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 may 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 data-collection costs (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 newer and bigger data sources. 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, attribution and interpretation have not become easier. 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 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; 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. Thus, while flexible models might help to inform theory building, their use for making decisions in assessment procedures aimed at individuals is currently limited.

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.

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.

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

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

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.

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

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*.

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

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

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.

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

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.

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.

Miller, P. J., Lubke, G. H., McArtor, D. B., & Bergeman, C. (2016). Finding structure in data using multivariate tree boosting. *Psychological Methods*, *21*(4), 583.

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

Nicodemus, K. K. (2011). On the stability and ranking of predictors from random forest variable importance measures. *Briefings in Bioinformatics*, *12*(4), 369–373. https://doi.org/10.1093/bib/bbr016

Nicodemus, K. K., Malley, J. D., Strobl, C., & Ziegler, A. (2010). The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics*, *11*(1), 1–13. https://doi.org/10.1186/1471-2105-11-110

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.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

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.

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., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(1), 1–11. 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*(1), 1–21. https://doi.org/10.1186/1471-2105-8-25

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.

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.

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.

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.