Original Article

Replicability of Machine Learning Models in the Social Sciences

A Case Study in Variable Selection

Ranjith Vijayakumar and Mike W.-L. Cheung

Department of Psychology, National University of Singapore, Singapore

Abstract: Machine learning tools are increasingly used in social sciences and policy fields due to their increase in predictive accuracy. However, little research has been done on how well the models of machine learning methods replicate across samples. We compare machine learning methods with regression on the replicability of variable selection, along with predictive accuracy, using an empirical dataset as well as simulated data with additive, interaction, and non-linear squared terms added as predictors. Methods analyzed include support vector machines (SVM), random forests (RF), multivariate adaptive regression splines (MARS), and the regularized regression variants, least absolute shrinkage and selection operator (LASSO), and elastic net. In simulations with additive and linear interactions, machine learning methods performed similarly to regression in replicating predictors; they also performed mostly equal or below regression on measures of predictive accuracy. In simulations with square terms, machine learning methods SVM, RF, and MARS improved predictive accuracy and replicated predictors better than regression. Thus, in simulated datasets, the gap between machine learning methods and regression on predictive measures foreshadowed the gap in variable selection. In replications on the empirical dataset, however, improved prediction by machine learning methods was not accompanied by a visible improvement in replicability in variable selection. This disparity is explained by the overall explanatory power of the models. When predictors have small effects and noise predominates, improved global measures of prediction in a sample by machine learning methods may not lead to the robust selection of predictors; thus, in the presence of weak predictors and noise, regression remains a useful tool for model building and replication.

Keywords: data mining, machine learning, model comparison, variable selection

Psychology has recently suffered from a loss of confidence in its theory-building enterprise. A renewed focus on the lack of replicability in psychology has highlighted concerns with "spurious effects" that achieve significance in one dataset, thus become part of a theory, but are not replicated in future studies (Ferguson & Heene, 2012; Open Science Collaboration, 2015). Concerns about researchers' overemphasis on the *p*-value (Goodman, 1992), *p*-hacking and data-fudging have emboldened cries for changing the methods of analysis in the field (Ioannidis, 2005, 2016). One of the promising venues for this change involves the use of machine learning methods, which are increasingly advocated as supplements to the statistical methods conventionally used in the field of psychology.

Machine learning methods are a set of diverse mathematical techniques that share the explicit aim of helping a system learn the underlying patterns in the data with only a few or no theoretical assumptions about underlying populations; this differentiates them from conventional regression-based inferential methods (Shmueli, 2010). They have become popular in social policy and research fields especially when the accuracy of diagnosis and future prediction is prized over understanding of the underlying causal chains. They are used in financial settings (Liew & Mayster, 2017), criminal policy to predict recidivism among criminals (Zeng, Ustun, & Rudin, 2017), clinical medicine and psychology for accurate pathological diagnosis (Kukar, Kononenko, Grošelj, Kralj, & Fettich, 1999; Lavrač, 1999; Walsh, Ribeiro, & Franklin, 2017), prevention of credit card fraud, email spam detection, and so forth (Bhattacharyya, Jha, Tharakunnel, & Westland, 2011; Goh & Singh, 2015; Ngai, Hu, Wong, Chen, & Sun, 2011). Popular machine learning methods include random forests (RF); several variants of boosting; penalized regression methods such as ridge regression, elastic nets, and least absolute shrinkage and selection operator (LASSO); basis function expansions or *splines* in regression; support vector machines (SVM); and neural nets (Hastie, Tibshirani, & Friedman, 2016).

Apart from the specifics of their separate algorithms, all machine learning practices have the following features that distance them from the regression and have led to their wide spread application. First, their focus on predictive accuracy lets them trade-off between two sources of model uncertainty: *bias* and *variance*. Bias refers to the deviation of the specified model from the true data-generating model; variance refers to the sampling variation in estimates of

the specified model. Conventional regression methods focus on getting an unbiased population estimate, which usually leads to an increase in the variance of the parameter estimates, sacrificing predictive accuracy. Machine learning algorithms choose their parameters to minimize total predictive error and so can trade off an increase in bias for a decrease in variance.

Second, in order to ensure predictive accuracy, machine learning methods explicitly validate the optimal model on a separate dataset from the one used to optimize their parameters. In practice, this is achieved by separating the sample at hand into separate sets: a *training* set, a *validation* set, and sometimes a third *test* set. The model is optimized on the training set. The optimal model's error on the training set is a poor measure of its performance as it will likely overfit to the data; hence, a validation set is used to estimate the out-of-sample performance of the optimal model. When validation error is used to choose the best model among many alternative models, as in our study, it becomes an overly optimistic indicator of this final model's predictive accuracy. Hence, prediction on a separate test is used to estimate the generalization error of the best model. A popular method for validating models is the *k-fold* cross-validation used in our study, where the k-th fold serves as the validation set, training is done on the $k - 1$ sets and validation on the k-th fold. This is repeated for every fold, and the average validation performance is used to choose the best model.

A third feature is the ability in many of these methods to capture non-linear associations without a priori specification by the researcher. In simulated datasets with prescribed multivariate dependencies and non-linear effects, machine learning methods such as forests and boosting performed better than regression in selecting variables important to model fit (Miller, Lubke, McArtor, & Bergeman, 2016). This aspect has also been examined in many applied settings. Rossi, Amaddeo, Sandri, and Tansella (2005) used RF to supplement their logistic regression findings of predictors important to classify patients who come only once for psychiatric consultation. Citing Derksen and Keselman (1992), they found regression coefficients to be less reliable and used RF to elucidate more predictors. Strobl, Malley, and Tutz (2009) also motivate this by claiming that RF might pick interactions not specified in the linear model. In this way, machine learning algorithms could provide evidence of underlying variable patterns, which could then be tested by conventional statistical methods for confirmation (e.g., Gureckis & Markant, 2012). Even though many of the machine learning methods do not have interpretable models, they do output method-specific ranking of predictors which can be used to check the predictors important for prediction.

Researchers advocating the use of machine learning methods are also drawn by this focus on prediction as opposed to explanation. Many methodologists (Breiman, 2001a; Forster, 2002; Varian, 2004) posit that too much focus on explanation has led to stagnant research disciplines in the social sciences. The current "crisis" in psychology (Spellman, 2015) has many observers positing that our predilection for the theory itself might be unjustified, given the current nature of knowledge in psychology. One way out of this "replication crisis" in psychology would be to use machine learning's focus on prediction. Researchers like Yarkoni and Westfall (2017) assert that machine learning practices can help to constrain models by enabling robust replicable results through their focus on predictive accuracy.

Machine learning methods have made inroads in psychology, especially in the fields of clinical psychology and personality research. In clinical psychology, there is grave concern about translatability of research to help diagnosis and treatment of disorders. Here, machine learning methods have led to increased rates of accurate diagnosis of disorders as varied as schizophrenia, depression, anxiety, substance abuse, and eating disorders (for a review, see Dwyer, Falkai, & Koutsouleris, 2018). Such progress in accurate prediction of prognosis (e.g., Whelan et al., 2014) is accompanied by reduction in predictors needed for accurate prediction. With their explicit use of cross-validation, machine learning methods seem a good fit in a field that deals with variables of a varied nature and have an urgent requirement for accurate prediction for applications. Machine learning is also now being used to develop a theoretically derived measures of assessing personality using a large number of records with little or no substantive theory connecting them, focusing solely on prediction (Bleidorn & Hopwood, 2018).

However, there is "no free lunch" with any prediction algorithm (Wolpert, 1996, 2002); no method always performs the best across all situations. Much depends on the nature of the data, the nature of the search algorithms used by the method, the particular error function minimized, and so forth. Machine learning methods do not always outshine conventional methods. For instance, comparisons of logistic regression with neural nets and SVM have shown mixed results, with logistic regression sometimes outperforming machine learning methods (Dreiseitl & Ohno-Machado, 2002; Held, Cape, & Nathan, 2016; Steyerberg, Ploeg, & Calster, 2014).

Sample size can also affect the superiority of machine learning methods. Sanchez-Pinto, Venable, Fahrenbach, and Churpek (2018) compared several machine learning methods with regression methods for variable selection for two differently sized datasets, focusing on *parsimony* in predictor selection, which balances predictive accuracy with the sparsity of predictors. They found that for the smaller of the two datasets (N = 6,564), classical methods

achieved the best parsimony, with machine learning faring better only in the dataset with more than 260,000 participants. Predictive accuracy decreased for machine learning with fewer variables compared to regression models whose predictive accuracy seemed to remain the same or improve by variable selection.

Similarly, Van der Ploeg, Austin, and Steyerberg (2014) compared SVM, neural nets, and RF with logistic regression and trees, focusing on predictive accuracy and stability of predictions. They found that SVM and neural nets needed more data compared to classical methods for achieving similar stability in their predictions. They also found that RF and linear regression methods performed consistently well. On the other hand, Tange, Rasmussen, and Taira (2017) used replications from three separate high dimensional datasets, comparing neural networks (ANN), SVM, and partial least-squares regression, and showed SVMs outperforming ANNs and regression at training sample sizes as small as 143, 20, and 53, respectively. So even though the conventional wisdom is that machine learning methods require large samples, much depends on the characteristics of the data.

Given the increasing use of machine learning algorithms in psychology, and the understanding that different data types prevalent in different fields may modify the effectiveness of these methods, there is a need to systematically examine the ability of these newer tools to enhance replicability of psychology research findings. This paper explores the case for using machine learning by examining the replicability of predictor selection in repeated samples from empirical and simulated data. We focus on whether the increased global predictive accuracy of machine learning tools necessarily lead to enhanced replicability of selected predictors.

Many of the applied studies cited above that compare machine learning tools with regression (e.g., Sanchez-Pinto et al., 2018) focus on variable selection and predictive accuracy in whole datasets, which serve only as a single replication. Simulation studies on variable selection also tend to focus on the predictive accuracy of selected variables, but not on the replicability of predictors. Studies also sometimes use default values of machine learning algorithms (e.g., Sanchez-Pinto et al., 2018; Van der Ploeg et al., 2014), instead of selecting the best tuned machine learning models by cross-validation. In our study, we seek to check replicability of predictor selection through repeated sampling from given datasets, real or simulated. This gives a more reasonable understanding of predictor replicability in research settings. We also systematically tune the parameters of machine learning models and use more than one method of predictor selection to contrast the methods. To our knowledge, no systematic examination of replicability of predictor selection of machine learning algorithms coupled with tracking of predictive accuracy has been conducted in psychology, and we hope this article serves as a case study in this framework.

The tools considered in our study can be categorized loosely into two: (a) *regression-based methods* which lead to interpretable models; these include multiple regression; its regularized variants such as LASSO (Tibshirani, 1996), elastic net (Zou & Hastie, 2005), and MARS (*Multivariate Adaptive Regression Splines*, Friedman, 1991, 1993); and (b) *black-box* methods, referred to as such because the predictor relationships are not easily interpretable. We use two such methods: SVM (Vapnik, 1995) and RF (Breiman, 2001b). Since these techniques are quite standard in machine learning literature, we refer readers to the excellent book-length introductions in Hastie, Tibshirani, and Friedman (2009) and Berk (2016).

Many machine learning statistical packages have built-in functions to rank predictors in terms of importance to prediction; researchers in applied settings tend to use these rankings in their comparison studies. These functions of the different methods, often having the same name in the statistical packages, use different algorithms internally to rank predictors, and this may affect comparisons of predictor selection. Nevertheless, since these methods use the specific machine learning algorithm to select predictors, they give a good measure of the importance of the predictor tailored to that specific algorithm. Our study uses these built-in functions of predictor importance specific to separate machine learning algorithms to make our results compatible with other applied literature. We give relevant details of these importance measures in the Methods section.

Methods

We compare performance measures and replicability of variable selection, contrasting (i) multiple regression with the following methods: (ii) LASSO; (iii) elastic net; (iv) MARS; (v) random forest; (vi) SVM.

Our aim is to compare different machine learning methods to regression on replicability of selected predictors in an empirical dataset (GESIS data, detailed below) and to see (a) whether increase in predictive accuracy for the global model leads to appreciable increase in replicability of selected predictors from samples, and (b) how different methods of variable selection fare at different sample sizes.

Findings of machine learning comparisons seem closely linked to data characteristics; there is likely a complex interplay between the specific algorithm, the magnitude of predictor effects, and sample sizes. Simulated data having identical predictor correlations as the empirical data, with added interactions and squared effects, could suggest

reasons for the pattern of replication and prediction results of the empirical dataset. In our study, we use four sets of replications: replications on the empirical GESIS dataset as well as on three kinds of simulated data: linear additive, interaction, and square terms of predictors. We use sample sizes of 100, 300 and 500 for the studies, each simulation having 500 replications.

GESIS Data

We use a subset of the survey data collected as part of GESIS panel study at GESIS Leibniz Institute of Social Sciences, Germany (Bosnjak et al., 2017). The group sampled German participants aged 18–70 years at recruitment, permanently resident in Germany. Longitudinal data on a variety of psychological and social variables were collected as part of the GESIS Panel Longitudinal Core study. The dataset used in this study is the GESIS Panel Standard Edition that contains information in the years 2014, 2015, and 2016 (GESIS, 2016). We used measurements taken in 2014, denoted "b" in the GESIS dataset, for our analysis.

The independent variables we chose for our study include Personality measures: the Big Five (Rammstedt, Kemper, Klein, Beierlein, & Kovaleva, 2013) and Schwartz Values (Schwartz, 1992) (both collected in GESIS panel study coded "ze"); several work place measures indicating both work place demands and constraints, as well as job satisfaction (collected in GESIS panel study coded "zg"). We combined measures of satisfaction and happiness in GESIS data (OECD, 2013; collected in GESIS panel study coded "zb") to obtain the dependent variable labeled "*well-being.present*." The final GESIS data with complete observations on all variables had 759 participants, 27 predictors measured on a continuous scale, and a single continuous dependent variable. Descriptive statistics of these variables are given in the Results section.

We wanted a set of independent variables from a commonly used survey dataset that refers to both individual and environmental aspects to mirror social studies. We were not interested in these variables for any substantive research questions and hence did not focus on issues such as reliability of measures or theoretical relevance. So the reader should focus not on the interpretation of the results of specific variables, but rather on the replicability and performance aspects of the methods.

Simulated Data

All simulated data were generated from a baseline model of 27 standardized predictors with a mean of zero and variance 1, with inter-predictor correlations taken from the GESIS data. The simulated data were of three types: (a) predictors having an additive linear relationship; (b) models with both additive and interaction effects; (c) model with a non-linear effects specified by square terms of predictors.

Linear Model

Predictors were generated with mean of zero and variance 1, with inter-predictor correlations identical to GESIS data. The dependent variable Y was obtained by specifying the same standardized regression coefficients for predictors as in GESIS data and specifying an error variance of 0.8.

Interaction Model

Predictors were generated with mean of zero and variance 1, with inter-predictor correlations identical to GESIS data. Five interaction terms were formed between variables with varying additive effects; two interaction terms involved the same predictor, while the rest involved separate variables. The coefficients of interaction terms had standardized weights of 0.2 and were added to the baseline coefficient vector of GESIS predictors to form the final weight vector. Error variance of the dependent variable Y was kept at 0.8.

Non-Linear "Square" Model

Predictors were generated with mean of zero and variance 1, with inter-predictor correlations identical to GESIS data. Six predictors had squared terms added to the model, again choosing predictors with varying linear weights. The coefficients of square terms with standardized weights of 0.2 were added to the baseline coefficient vector of GESIS predictors to form the final weight vector. Error variance of the dependent variable Y was kept at 0.8.

Details of Replication

Replication on GESIS Data

We use GESIS data to compare replicated samples from an empirical dataset on performance measures of the methods, as well as a selection of predictors. We randomly split this data into a test set (N = 250) and a sampling set (N = 509). During each replication, samples are drawn from the sampling set without replacement to form the training set. We use 10-fold cross-validation on this training set to get the optimal model for all methods and to get the cross-validation measures of performance. We check the final models of each method for performance on the same test set; this creates a single benchmark for assessing the generalized performance of the different methods at each sample size, which can be compared with their cross-validation performance measures. The variable importance measures are obtained by calling the built-in functions on this final model of each method.

Replications on Simulations

For each simulation, we generate a single test dataset (N = 1,000). Training data are generated in each replication, and 10-fold cross-validation is done to get measures of performance and variable importance. Test performance is measured by model fit on the test set.

Tools for Statistical Analyses

We use the open source R statistical platform (R Core Team, 2017) for our statistical analyses. We used the following packages in R for analysis: *kernlab* (Karatzoglou, Smola, Hornik, & Zeileis, 2004) for radial SVM; *randomForest* (Liaw & Wiener, 2002); *earth* (Milborrow, 2007) for MARS; *glmnet* (Friedman, Hastie, & Tibshirani, 2010) for LASSO and elastic net; and the native *lm* function for regression. Many studies that compare different methods use different packages; the resulting importance and performance measures may suffer from incompatibility in comparisons. The *caret* package (Kuhn, 2008) can use all the specific methods above for its functions, and it eases the workflow of training the different models, the specification of same data folds for cross-validation to the different methods, and also the post-simulation analyses. *Caret* also eases the pre-processing of training data; all predictors are standardized before optimization by caret functions. We use *caret* both for assessing performance measures of cross-validation and test, as well as for examining predictor importance via the ranking functions given in the machine learning packages. We hope that this allows other researchers to better compare our results and serve as an outline for reproducible research.[1]

Optimization of Hyperparameters

Machine learning methods have method-specific parameters (called *hyperparameters*) that are user-determined, and unlike the usual parameters of regression models, not optimized during training of the model. Optimal values for hyperparameters are usually determined by *tuning*: a range of values for each hyperparameter are prescribed; models differing in these values are optimized on the training data, and the model with best cross-validated performance is selected as the final model. The range of values is usually determined either by a grid search through user-determined values, or a random search through an optimal range of values specified by the statistical package. Random search has been shown to be better for certain algorithms, especially at higher dimensions (Bergstra & Bengio, 2012). We use random search, except where grid search is appropriate for the method in question.

LASSO has one parameter λ, which we tune with a random search, sampling 30 values over a uniform distribution between 2^{-10} and 2^{3}, the default range for the random search in the caret package. For the elastic net, we do a random search for two hyperparameters: λ (range same as for LASSO) and α, sampled randomly between values of zero and 1. We tune two hyperparameters for MARS: *nprune*, which specifies the maximum number of possible terms in the final pruned model, and *degree*, the maximum degree of interactions. Random search samples 30 values of nprune over a range from 2 to the number of predictors, examining first and second degree interactions.

SVM was evaluated using radial basis kernels. The hyperparameters we tune include C, which determines a trade-off between training error and model complexity; σ which controls non-linear nature of the decision boundary; and e which determines the tolerance to errors in estimation. The ranges are as follows: (2^{-2} to 2^{5}) for C; (2^{-8} to 2^{0}) for σ; and (2^{-8} to 2^{-1}) for e; these values are commonly used in SVM tuning (e.g., Cortez & Embrechts, 2013; Khondoker, Dobson, Skirrow, Simmons, & Stahl, 2016; Hsu, Chang, & Lin, 2003). For RF, we do a random search sampling of 15 values for the parameter *mtry*, which determines the number of predictors to compare at each split of a tree.

Predictor Importance

Built-in variable importance ranking functions of SVM, RF, and elastic net used by *caret* do not select out predictors; instead, they rank all predictors in order of predictor importance based on internal criteria. For such methods, we choose the 10 best predictors in each round of simulation as important in that iteration. For methods like LASSO and MARS, which select out predictors, we chose the predictors remaining in the optimized model selected through cross-validation as being relevant in that iteration. The overall importance of predictors in the study thus indicates the frequency with which a predictor was selected as important across all replications.

LASSO, MARS, and elastic net have variable selection built-in to their optimization algorithm. RF use a permuted variable importance measure based on the increase in out-of-bag mean square error for permuted predictors (Strobl et al., 2009). For SVM, though many algorithms have been recommended for variable selection (e.g., see Chang & Lin, 2008; Weston et al., 2001; Maldonado, Flores, Verbraken, Baesens, & Weber, 2015; Guyon & Elisseeff, 2003), *caret* does not use built-in functions. Instead, it uses a filter approach using a LOESS smoother between predictor and observed variable (Kuhn, 2008). For regression, *caret* used the absolute value of t-score, which is equivalent to using

[1] The R codes used in our analyses are freely available at https://osf.io/svgrf/?view_only=5e75057a6d6f48f49afb3ecd6bf6cd70

p-values to order the predictors. Although *p*-value does not represent the strength of effect (see e.g., Benjamin et al., 2018; Lakens et al., 2018; Wasserstein & Lazar, 2016), regression methods usually use *p*-values for selection of variables, for instance, in subset-selection algorithms. Hence, this approach helps us compare regression with other tools using the criterion usually used by regression in applied settings.

Performance Estimates

To compare predictive accuracy across methods, we use the two measures of performance: root mean square of error (RMSE) and R^2, giving both validation and test performance. CV prediction accuracy measures are calculated on validation sets during the 10-fold cross-validation process used to select the best model for each method; the validation performance measures are averaged to give the 10-fold CV RMSE and R^2 values. For test performance, we fit the model's predictions to a separate test data. A single test data is used in all replications of each simulation; in each replication, the models' prediction on the test set is used to calculate test RMSE and R^2. These predictions are averaged to give mean test performance measures.

Results and Discussion

The distribution and correlations of variables in the GESIS dataset are shown in Figures 1 and 2 in Electronic Supplementary Material 1 (ESM 1).

Performance Measures

We used two measures of predictive performance, R^2, a measure of relative improvement in prediction over the mean model, and RMSE, a measure of absolute predictive fit. We use both cross-validation as well as a separate tests set to measure predictive performance; the test prediction performances are given in Figures 1 and 2, while the CV performance measures are depicted in Figures 3 and 4 in ESM 1. Since CV error is a biased estimate of prediction error (Cawley & Talbot, 2010; Varma & Simon, 2006),

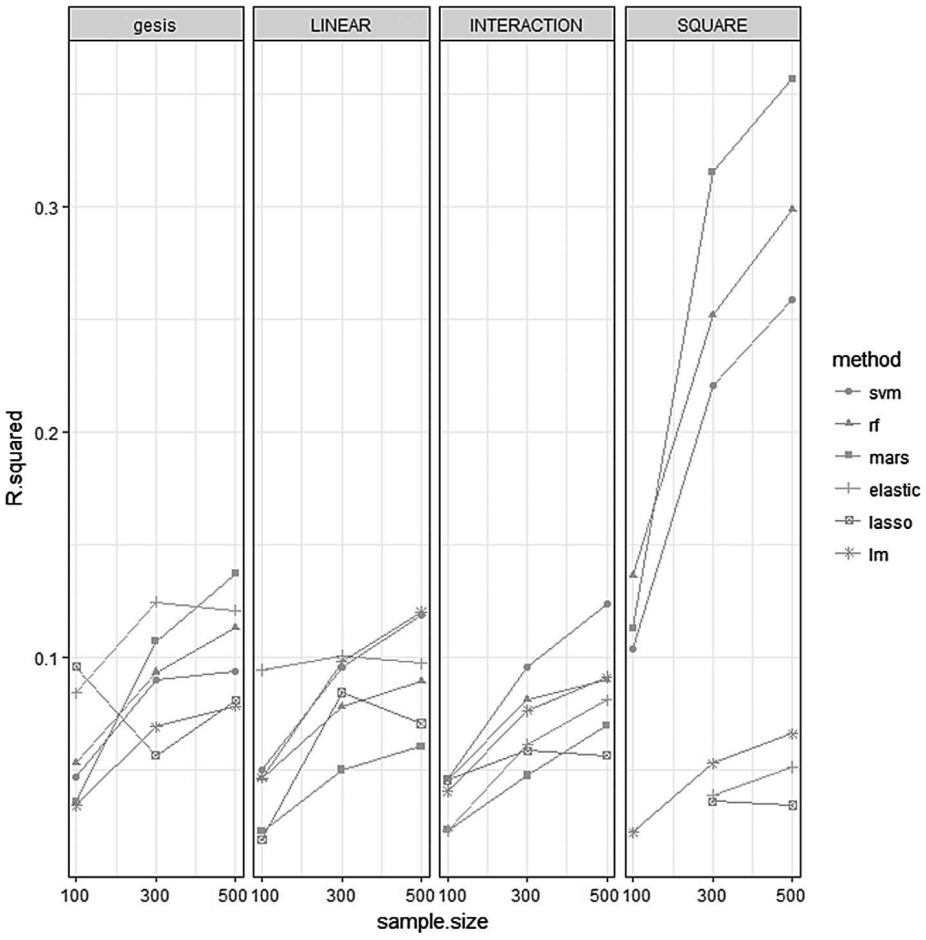

Figure 1. R^2 measure of prediction on test set across sample sizes for empirical (GESIS) and simulated (linear, interaction and square) datasets. Labels: rf = random forest, lm = linear regression model. The performance of SVM, RF, and MARS is superior to regression, both for GESIS data and all simulated scenarios, except linear and interaction models. LASSO and elastic net show performance inferior to regression.

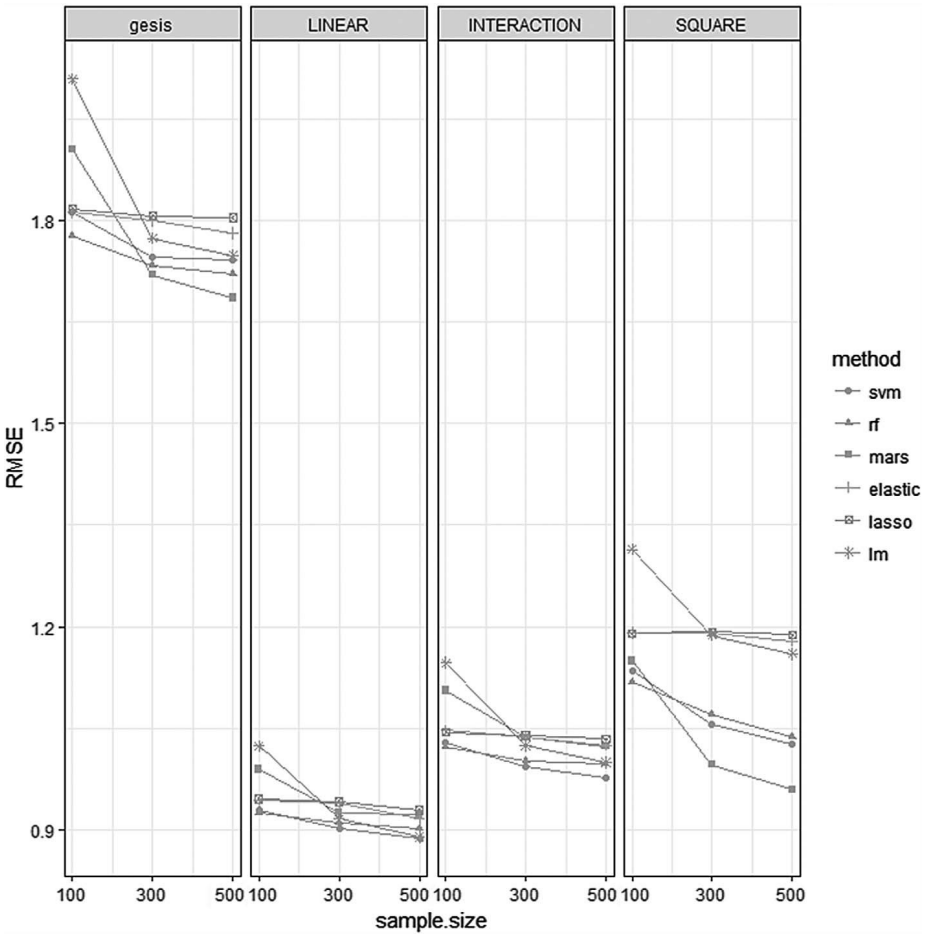

Figure 2. RMSE measure of prediction on test set across sample sizes for empirical (GESIS) and simulated (linear, interaction, and square) datasets. Labels: rf = random forest, lm = linear regression model. The performance of SVM, RF, and MARS is superior to regression, both for GESIS data and all simulated scenarios, except linear and interaction models. LASSO and elastic show performance inferior to regression.

we focus on test error and mention the differences with 10-fold CV prediction results at the end of this section. We compare regression to black-box methods (SVM, RF) and to regression-based methods (LASSO, elastic net, and MARS).

Compared to regression, black-box machine learning tools such as SVM and RF show better test performance in R^2 and RMSE for the replications on both GESIS data as well as simulated data with squared terms; this is true across sample sizes (Figures 1 and 2). For GESIS data, at larger sample sizes of 300 and above, MARS produces better test R^2 and smaller test RMSE than SVM and RF. For square data, MARS, SVM, and RF outperform regression with MARS increasingly becoming the best method. As opposed to the square and GESIS data, in linear and interaction models regression seems to perform as well as complex methods like SVM and RF, with both producing test error less than regression only at the small sample size of 100. As sample sizes increased, regression seems to give comparable test error as SVM and RF, outperforming MARS as well. This lessening of gap between regression and these methods was absent in square models, where regression continued to give higher test error across sample sizes.

LASSO and elastic net show uneven performances for RMSE and R^2; this is because in many replications, these methods produce an optimal model with no predictors, which affects R^2 calculations, so we focus on RMSE to understand their performance. We see that at sample size of 100, both LASSO and elastic net have better RMSE indicators than regression in all datasets. However, from a sample size of 300 on, regression has reduced or similar prediction error compared to its regularized variants. Comparing regression to MARS, regression gives better test error than MARS for linear and interaction data, but MARS outperforms regression for GESIS and square data.

In short, results of the predictive performance show that methods such as SVM, RF, and MARS, which can capture non-linearity without a priori specification, give a better prediction for the empirical GESIS data as well as the square data at all sample sizes. On the other hand, regression does better in presence of additive effects and remains almost as good as complex models in presence of just linear interactions. Regularized variants show better performance than regression at smaller sample sizes for all data, but regression overcomes this performance gap at larger sample sizes.

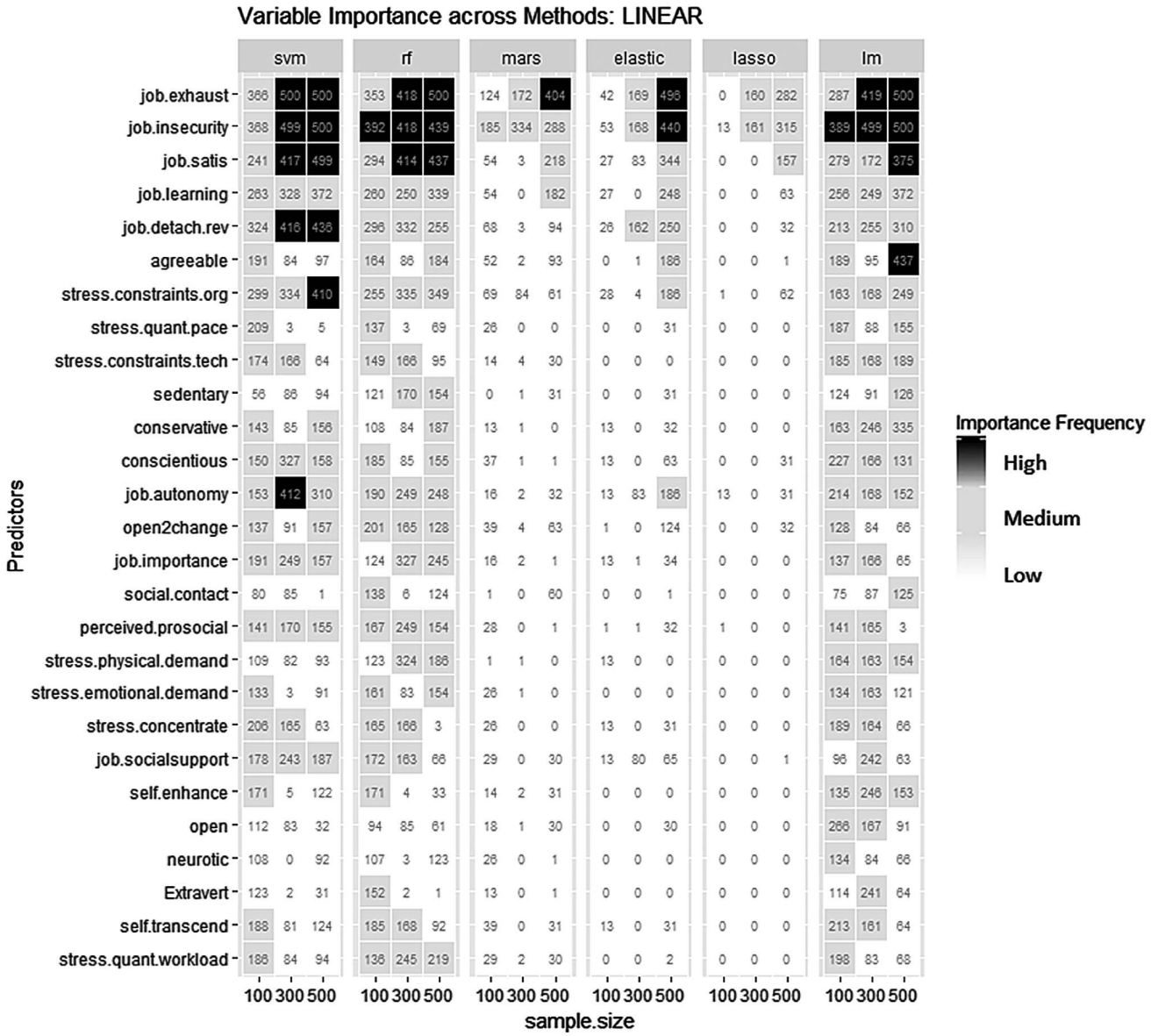

Figure 3. Variable importance across sample sizes in simulated data with linear additive effects. The heat map shows predictors selected as important by the different modeling tools in 500 replications across sample sizes of 100, 300, and 500. Labels: rf = random forest, lm = linear regression model. The numbers in the cells indicate the frequency of variable selection across the 500 replications. The frequency was divided into three categories; "High" and "Low" refer to the 75th and 25th quartile of the importance frequency, respectively. For clarity, predictors are ordered in decreasing order of their standardized coefficients in a linear model.

A note on the CV performance: The 10-fold CV prediction values (Figures 3 and 4 in ESM 1) show optimistic values compared to test measures for machine learning tools. Regression performs very poorly for all sample sizes in all simulations; this is at odds with its test performance. One reason for the dismal CV performance of regression is the small *N* of the validation set; for our training samples of 100 to 500, the validation *N* would range from 10 to 50. We conclude that with sample sizes of 100–500, CV measures are poor indicators of test accuracy for regression and lead to misleading comparisons. We stick to test comparisons for our interpretation.

We now examine whether increased predictive accuracy for GESIS data replications by machine learning methods leads to better replicability of predictors they consider important.

Predictor Importance

Predictor importance is measured by taking the top 10 predictors ranked by a machine learning algorithm in each replication and counting the frequency for each predictor across the 500 rounds of replication. The importance

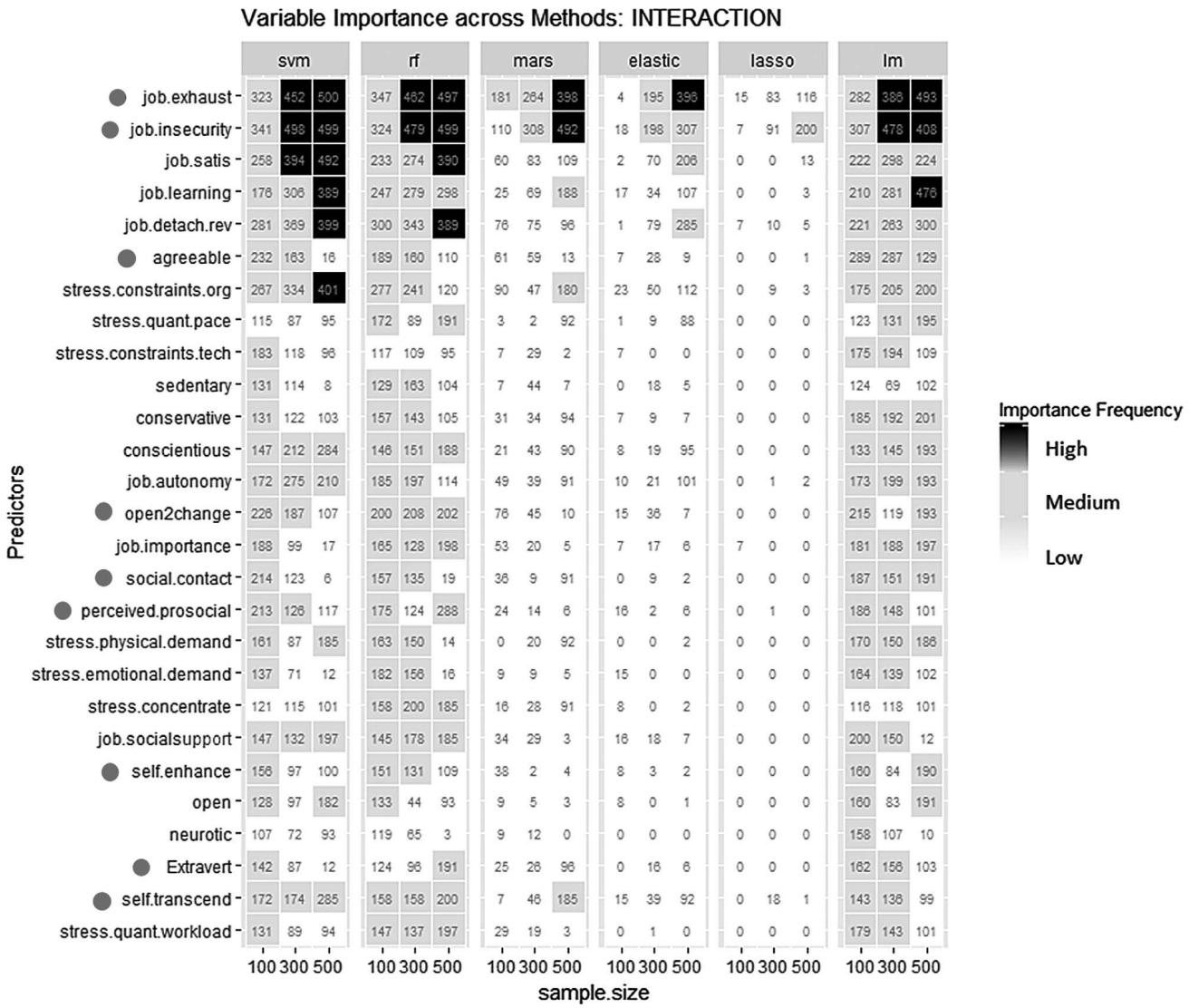

Figure 4. Variable importance across sample sizes in simulated data with linear interaction effects. The heat map shows predictors selected as important by the different modeling tools in 500 replications across sample sizes of 100, 300, and 500. Labels: rf = random forest, lm = linear regression model. The predictors involved in interaction are highlighted with a circle on the left. The numbers in the cells indicate the frequency of variable selection across the 500 replications. The frequency was divided into three categories; "High" and "Low" refer to the 75th and 25th quartile of the importance frequency, respectively. For clarity, predictors are ordered in decreasing order of their standardized coefficients in a linear model.

frequency for all four datasets is depicted in heat maps in Figures 3, 4, 5, and 6. An algorithm that is ideal for replication should be consistent in picking out predictors: predictors should show either very high or very low frequency of selection and not intermediate ranges. For our study, we choose the top- and bottom-quartile of the importance frequency as the "high" and "low" markers; we would like the methods to give predictors importance frequency values in these two ranges. An algorithm with many predictors in the middle half would not be an ideal pick for replicable models. Also when the true model is known, we want a tool that picks out the true effects consistently, but no spurious ones.

Comparisons of the heat maps of simulated linear, interaction and squared effects data with empirical GESIS data are illuminating. First, there is a difference in performance between non-linear square data and linear data with additive or interaction effects. For simulated data with additive or interaction effects (Figures 3 and 4), we find that machine learning methods show almost no improvement over regression. Many predictors are selected as important in the medium range of frequency, suggesting that these methods would lead to similar inconsistent results in studies. Moreover, they do not pick out any predictor with true interaction effects that were not selected by regression. This mediocre performance is anticipated by the perfor-

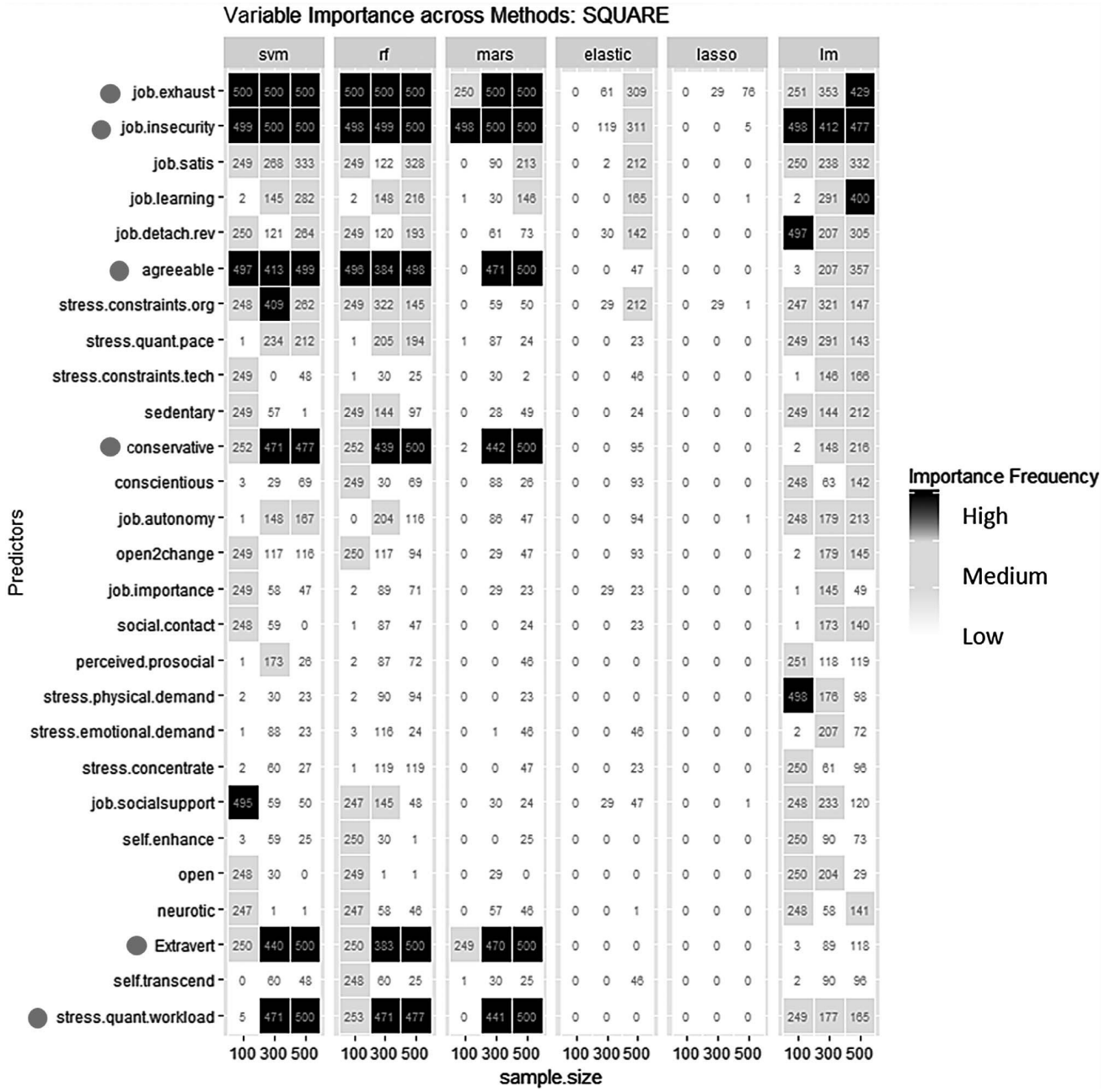

Figure 5. Variable importance across sample sizes in simulated data with non-linear squared effects. The heat map shows predictors selected as important by the different modeling tools in 500 replications across sample sizes of 100, 300, and 500. Labels: rf = random forest, lm = linear regression model. The predictors with squared effects are highlighted by circle on the left. The numbers in the cells indicate the frequency of variable selection across the 500 replications. The frequency was divided into three categories; "High" and "Low" refer to the 75th and 25th quartile of the importance frequency, respectively. For clarity, predictors are ordered in decreasing order of their standardized coefficients in a linear model.

mance in test error by machine learning methods. Recall that for linear and interaction models, machine learning methods fared poorly in test prediction; even black-box methods like SVM and RF showed little or no improvement over regression. SVM did show better test prediction at higher sample sizes for interaction model, but that did not translate into better replicability of models, nor an ability to capture true interaction effects. MARS, LASSO, and elastic net did not capture most predictors; MARS also missed all interactions; this mirrors its poor performance in test prediction in interaction models.

The squared model (Figure 5) tells a different story. Here, the superior performance by machine learning methods in test prediction goes hand in hand with better variable

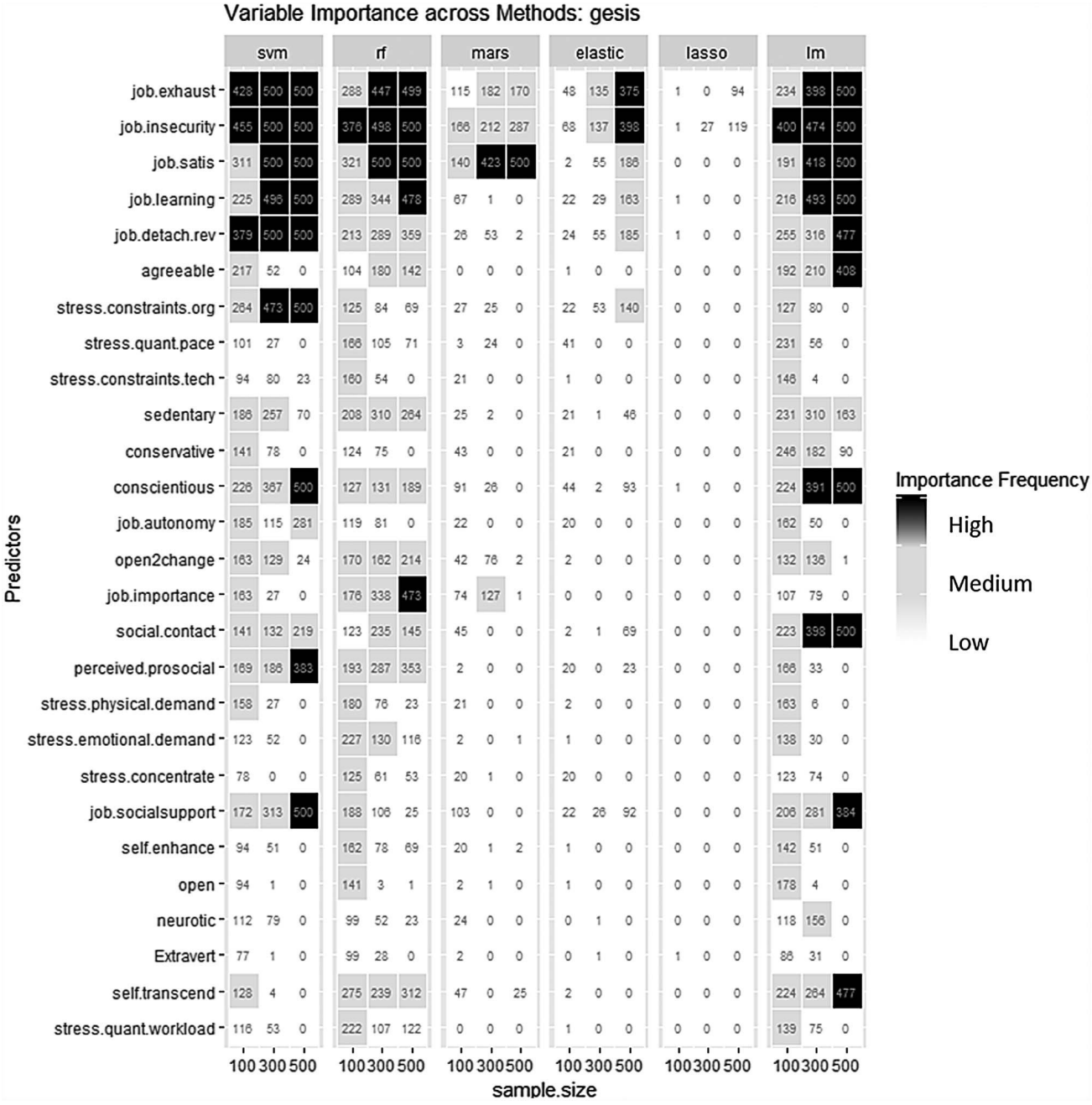

Figure 6. Variable importance across sample sizes in GESIS data. The heat map shows predictors selected as important by the different modeling tools in 500 replications across sample sizes of 100, 300, and 500. Labels: rf = random forest, lm = linear regression model. The numbers in the cells indicate the frequency of variable selection across the 500 replications. The frequency was divided into three categories; "High" and "Low" refer to the 75th and 25th quartile of the importance frequency, respectively. For clarity, predictors are ordered in decreasing order of their standardized coefficients in a linear model.

selection. In terms of replicability, SVM, RF, and MARS show better replicability with many of the selected predictors picked at a high frequency; they have fewer predictors in the medium range compared to regression, and this improves further with sample size. They also pick out the ones with true effects; MARS in particular shows great specificity in selection: it picks all predictors with non-linear squared effects and otherwise very few.

Thus, in simulated data, there seems to be a connection between predictive accuracy and performance in variable selection: in models where machine learning tools predict better, they also lead to better replication of variables.

In models such as additive and interaction models where they show little prediction improvement, it is mirrored in variable selection as well.

The results of variable importance in empirical data replications (Figure 6), however, do not follow the same pattern. Recall that SVM, RF, and MARS produced fewer test errors than regression on GESIS data. However, this is not accompanied by a perceptible improvement in variable selection results. Regression seems to select almost all predictors selected by SVM and RF; both machine learning methods also seem to replicate predictors with a medium frequency, similar to regression. Interestingly, MARS, which also showed better predictive performance compared to regression, obtaining the most impressive prediction R^2 and RMSE values at samples of 300 and 500, performs very poorly in variable selection, similar to the results of LASSO and elastic net. While MARS also had poor variable selection performance in simulations with linear and interaction effects, there it was accompanied by poor predictive accuracy. Here, there is a separation between the performance of machine learning methods in prediction and their performance as variable selection tools: better global prediction does not lead in this empirical dataset to improved replicability in variable selection.

General Discussion

We were interested in tracking the performance of machine learning methods and regression and checking if their performance discrepancies in predictive accuracy are mirrored in the replicability of variable selection in samples. We replicated samples from both an empirical dataset, GESIS data, as well as simulated data having the same covariance structure as the GESIS data, with additive, interaction, or polynomial terms added.

In simulated datasets with linear additive or interaction terms, complex machine learning algorithms such as SVM, RF, and MARS performed mostly at par with regression, reducing test error better than regression only at sample sizes of 100, this gap diminishing remarkably at higher sample sizes. This is different from their performance in squared models, where SVM, RF, and MARS performed substantively better than regression, the gap between them and the regression remaining intact at higher sample sizes.

This global prediction performance in simulated data was closely tracked by their performance in the replicability of variable selection. As expected, in models (linear and interaction) where gains from complex algorithms were meager in predictive accuracy, we found no improvement in replicability as a result of using complex algorithms such as SVM, RF, and MARS. On the other hand, when SVM, RF, and MARS resulted in increased predictive accuracy that persisted across sample sizes, as in the squared terms simulation, it was accompanied by robust replicability in variable selection by these complex algorithms, surpassing the poor performance by regression and its regularized variants, LASSO and elastic net.

However, this consilience between global prediction and variable selection seems to dissipate on examining replications on the empirical dataset at hand, the GESIS data. Here, predictive accuracy is better for RF, SVM, and MARS, with their performance gap with regression in test RMSE remaining even at higher sample sizes for RF and MARS, while reducing to similar levels for SVM only at the largest simulation sample size of 500. But this increase in global predictive accuracy is not accompanied by any visible improvement in variable selection replicability. Regression seems to pick almost all predictors picked by SVM and RF, while MARS, in spite of having best test indices of predictive accuracy, performs very poorly in variable selection, faring the same as LASSO and elastic net. The one or two predictors picked by machine learning tools but not regression are not the same for SVM and RF.

Our study suggests that in empirical datasets having predictors with varied correlations with each other and the dependent variable, machine learning algorithms may not necessarily lead to gains in the replicability of selected variables, despite showing gains in predictive accuracy. This is true even though in simulations with identical covariance structure, they are able to select predictors that have substantive interaction and square effects, aligning their predictive accuracy with variable selection. This might be especially true for spline methods; in our study, MARS showcased best comparative predictive accuracy measures for the GESIS data but could not consistently replicate variables, missing most of predictors picked by all other methods, including regression.

This mismatch might arise from a number of features. First, note that our simulations have "small" to "moderate" sample sizes ranging from 100 to 500. Machine learning algorithms are known to be "data hungry," as per Van der Ploeg et al. (2014) ; their study suggests that these algorithms may need as much as 200 cases per predictor to have stable predictions (but see Tange et al., 2017). Machine learning also excels with an increase in *signal* to *noise* ratio; its performance in social science datasets where noise predominates might be far from optimal. Such conclusions are not surprising; Yarkoni and Westfall (2017) have pointed out that studies taking care not to overfit the data invariably lead to modest results. What is new here is the possibility that replicability in predictor selection seems not to be tied to global measures of predictive accuracy.

Looking at predictive R^2 (Figure 1), we find that in squared effects simulations, the test R^2 for MARS, SVM, and RF were substantive, reaching over 20% in samples of 300 and increasing further to values of 25–35% in samples of 500. This was accompanied by robust replication of selected variables by these algorithms. On the contrary, we find that for linear and interaction effects, neither regression nor machine learning tools managed substantive R^2s of more than 10–15%. It is in these very models that replicability turned out not to improve upon using machine learning models.

So the apparent paradox of predictive performance gaps between machine learning methods and regression not extending to improvements in variable selection might be explained by the lack of explaining power in the models per se. Predictors with true but small effects, such as in simulated interactions, which do not lead to a substantive explanation of the variance in the data, may not be served by using machine learning algorithms.

This is a sobering thought: social science researchers, in fact, *expect* predictors to have small effects that explain a small part of the variation in the data; it is the inability of regression models to consistently discover predictors with small effects that led to calls of "replication crisis" in psychology. Increasing sample sizes from 100 to 500 did not lead to selecting more predictors in simulations or the empirical dataset; it remains to be seen whether further increases in sample size would lead to the better ability of machine learning tools to capture variables missed by regression.

This does not invalidate the use of machine learning in social science data. It must be noted that this study used a group of variables without any prior data cleaning procedures such as dimensional reduction or filtering of predictors to reduce noise; these would be usual first steps in any data-mining study. We decided against this because in many studies in the social sciences, the variables selected have interpretive importance, and researchers have to deal with a set of predictors that cannot be reduced in a mechanistic statistical fashion. In such situations, where researchers try to figure out the appropriate models with small effects in the midst of noise, machine learning algorithms may not be the panacea for all perceived ills of regression modeling.

A few more caveats are in order: this favorable performance of regression may be contingent on peculiarities of the dataset. A better inkling of these methods' performances will need experimentation with diverse datasets used in the social sciences. One should also note that even though we have made an effort to tune parameters of machine learning tools by cross-validation, a case can be made that a different set of tuning considerations may lead to different findings, and hence, comparisons across datasets should be made carefully. Of course, such considerations would affect any machine learning findings across samples, potentially complicating interpretability of their results.

Our study suggests the need to examine more than just comparative improvement in prediction to gauge whether variables selected by the model can be robustly replicated. Improved prediction of a machine learning model over regression, in presence of substantive unexplained variance, may not result in replicable models. Our study also suggests potential divergence between a method's ability to consistently select predictors and its performance in test prediction. MARS was better than all methods in predictions on GESIS data, but performed poorly in variable selection on replications of that very dataset.

In short, replicability in a real-life dataset seems to be as difficult for machine learning tools as for regression. This article tries to examine whether including machine learning tools lead to better replicable models, given their performance superiority in indices of prediction. The answer is unexpected: in a real dataset with varying predictor correlations and substantive noise, regression seems as safe a bet as any in seeking robust replicable models. A more thorough examination with varied datasets usually used in social science fields is required to make judgments on the replicability of predictors selected by machine learning tools.

Electronic Supplementary Material

The electronic supplementary material is available with the online version of the article at https://doi.org/10.1027/2151-2604/a000344

ESM 1. Figures (.pdf)
Figure 1. Correlation matrix for variables in the GESIS dataset.
Figure 2. Histogram of all variables in the GESIS dataset.
Figure 3. 10 fold-CV R2 across sample sizes for empirical (GESIS) and simulated (linear, interaction and square) datasets.
Figure 4. 10 fold-CV RMSE across sample sizes for empirical (GESIS) and simulated (linear, interaction and square) datasets.

References

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., & Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour, 2*, 6–10. https://doi.org/10.1038/s41562-017-0189-z

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research, 13*, 281–305.

Berk, R. A. (2016). *Statistical learning from a regression perspective* (2nd ed.). Basle, Switzerland: Springer International Publishing.

Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems, 50*, 602–613. https://doi.org/10.1016/j.dss.2010.08.008

Bleidorn, W., & Hopwood, C. J. (2018). Using machine learning to advance personality assessment and theory. *Personality and Social Psychology Review*, 1–14. https://doi.org/10.1177/1088868318772990

Bosnjak, M., Dannwolf, T., Enderle, T., Schauer, I., Struminskaya, B., Tanner, A., & Weyandt, K. W. (2017). Establishing an open probability-based mixed-mode panel of the general population in Germany: The GESIS Panel. *Social Science Computer Review, 36*, 103–115. https://doi.orG/10.1177/0894439317697949

Breiman, L. (2001a). Statistical modeling: The two cultures [with comments and a rejoinder by the author]. *Statistical Science, 16*, 199–231. https://doi.org/10.1214/ss/1009213726

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

Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research, 11*, 2079–2107.

Chang, Y.-W., & Lin, C.-J. (2008). Feature ranking using linear SVM. *JMLR: Workshop and Conference Proceedings, 3*, 53–64.

Cortez, P., & Embrechts, M. J. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Sciences, 225*, 1–17. https://doi.org/10.1016/j.ins.2012.10.039

Derksen, S., & Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology, 45*, 265–282. https://doi.org/10.1111/j.2044-8317.1992.tb00992.x

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics, 35*, 352–359. https://doi.org/10.1016/S1532-0464(03)00034-0

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology, 14*, 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science, 7*, 555–561. https://doi.org/10.1177/1745691612459059

Forster, M. (2002). Predictive accuracy as an achievable goal of science. *Philosophy of Science, 69*(Suppl. 3), S124–S134.

Friedman, J. (1991). Multivariate adaptive regression splines. *Annals of Statistics, 19*, 1–67. Retrieved from http://www.jstor.org/stable/2241837

Friedman, J. (1993). *Fast MARS*. Stanford, CA: Stanford University Department of Statistics.

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

GESIS. (2016). *Panel standard edition [ZA5665 Datafile Version 17.0.0]*. Cologne, Germany: GESIS Data Archive.

Goh, K. L., & Singh, A. K. (2015). Comprehensive literature review on machine learning structures for web spam classification. *Procedia Computer Science, 70*, 434–441. https://doi.org/10.1016/j.procs.2015.10.069

Goodman, S. (1992). A comment on replication, p-values and evidence. *Statistics in Medicine, 11*, 875–879. https://doi.org/10.1002/sim.4780110705

Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science, 7*, 464–481. https://doi.org/10.1177/1745691612454304

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

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). New York, NY: Springer New York.

Hastie, T., Tibshirani, R., & Friedman, J. (2016). *Elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York, NY: Springer.

Held, E., Cape, J., & Nathan, T. (2016). Comparing machine learning and logistic regression methods for predicting hypertension using a combination of gene expression and next-generation sequencing data. *BMC Proceedings, 10*, 141–145. https://doi.org/10.1186/s12919-016-0020-2

Hsu, C., Chang, C., & Lin, C. (2003). *A practical guide to support vector classification*. Retrieved from https//www.csie.ntu.edu.tw/ cjlin/papers/guide/guide.pdf

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine, 2*, e124. https://doi.org/10.1371/journal.pmed.0020398

Ioannidis, J. P. (2016). Why most clinical research is not useful. *PLoS Medicine, 13*, e1002049. https://doi.org/10.1371/journal.pmed.1002049

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab – An S4 Package for Kernel Methods in R. *Journal of Statistical Software, 11*, 1–20. Retrieved from http://www.jstatsoft.org/v11/i09/

Khondoker, M., Dobson, R., Skirrow, C., Simmons, A., & Stahl, D. (2016). A comparison of machine learning methods for classification using simulation with multiple real data examples from mental health studies. *Statistical Methods in Medical Research, 25*, 1804–1823. https://doi.org/10.1177/0962280213502437

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software, 28*, . https://doi.org/10.18637/jss.v028.i05

Kukar, M., Kononenko, I., Grošelj, C., Kralj, K., & Fettich, J. (1999). Analysing and improving the diagnosis of ischaemic heart disease with machine learning. *Artificial Intelligence in Medicine, 16*, 25–50. https://doi.org/10.1016/S0933-3657(98)00063-3

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., … Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour, 2*, 168–171. https://doi.org/10.1038/s41562-018-0311-x

Lavrač, N. (1999). Selected techniques for data mining in medicine. *Artificial Intelligence in Medicine, 16*, 3–23. https://doi.org/10.1016/S0933-3657(98)00062-1

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News, 2*, 18–22.

Liew, J. K., & Mayster, B. (2017). *Forecasting ETFs with machine learning*. Retrieved from https://ssrn.com/abstract=2899520

Maldonado, S., Flores, A., Verbraken, T., Baesens, B., & Weber, R. (2015). Profit-based feature selection using support vector machines – General framework and an application for customer retention. *Applied Soft Computing, 35*, 740–748. https://doi.org/10.1016/j.asoc.2015.05.058

Milborrow, S. (2007). *earth: Multivariate adaptive regression spline models*. R package version 2.0-2. http://CRAN.R-project.org/package=earth

Miller, Patrick. J., Lubke, Gitta. H., McArtor, Daniel. B., & Bergeman, C. S. (2016). Finding structure in data using multivariate tree boosting. *Psychological Methods, 21*, 583–602. https://doi.org/10.1037/met0000087

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems, 50*, 559–569. https://doi.org/10.1016/j.dss.2010.08.006

OECD. (2013). *OECD guidelines on measuring subjective well-being*. Paris, France: OECD Publishing.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. DOI: 10.1126/science.aac4716

Rammstedt, B., Kemper, C. J., Klein, M. C., Beierlein, C., & Kovaleva, A. (2013). Eine kurze Skala zur Messung der fünf Dimensionen der Persönlichkeit: 10 Item Big Five Inventory (BFI-10) [A short scale for measuring the five personality dimensions: 10-item Big Five Inventory (BFI-10)]. *Methoden, Daten, Analysen, 7*, 233–249. https://doi.org/10.12758/mda.2013.013

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

Rossi, A., Amaddeo, F., Sandri, M., & Tansella, M. (2005). Determinant of once only contact in a community based psychiatric service. *Social Psychiatry and Psychiatric Epidemiology, 40*, 50–56. https://doi.org/10.1007/s00127-005-0845-x

Sanchez-Pinto, L. N., Venable, L. R., Fahrenbach, J., & Churpek, M. M. (2018). Comparison of variable selection methods for clinical predictive modeling. *International Journal of Medical Informatics, 116*, 10–17. https://doi.org/10.1016/j.ijmedinf.2018.05.006

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advantages and empirical tests in 20 countries. In M. P. Zanna (Ed.), *Advances in experimental and social psychology* (Vol. 25, pp. 1–65). New York, NY: Academic Press.

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

Spellman, B. A. (2015). A short (personal) future history of revolution 2.0. *Perspectives on Psychological Science, 10*, 886–899. https://doi.org/10.1177/1745691615609918

Steyerberg, E. W., Ploeg, T., & Calster, B. (2014). Risk prediction with machine learning and regression methods. *Biometrical Journal, 56*, 601–606. https://doi.org/10.1002/bimj.201300297

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

Tange, R. I., Rasmussen, M. A., & Taira, E. (2017). Benchmarking support vector regression against partial least squares regression and artificial neural network: Effect of sample size on model performance. *Journal of Near Infrared Spectroscopy, 25*, 1–10. https://doi.org/10.1177/0967033517734945

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B, 58*, 267–288.

Van der Ploeg, T., Austin, P. C., & Steyerberg, E. W. (2014). Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology, 14*, 137. https://doi.org/10.1186/1471-2288-14-137

Vapnik, V. (1995). *The nature of statistical learning theory*. New York, NY: Springer.

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics, 7*, 91. https://doi.org/10.1186/1471-2105-7-91

Varian, H. (2004). Big data: New tricks for econometrics. *Journal of Economic Perspectives, 28*, 3–28. https://doi.org/10.1257/jep.28.2.3

Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2017). Predicting risk of suicide attempts over time through machine learning. *Clinical Psychological Science, 5*, 457–469. https://doi.org/10.1177/2167702617691560

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*-values: Context, process, and purpose. *American Statistician, 70*, 129–133. https://doi.org/10.1080/00031305.2016.1154108

Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., & Vapnik, V. (2001). *Feature selection for SVMs*. (Advances in Neural Information Processing Systems 13.). Cambridge, MA: MIT Press.

Whelan, R., Watts, R., Orr, C. A., Althoff, R. R., Artiges, E., Banaschewski, T., … Garavan, H. (2014). Neuropsychosocial profiles of current and future adolescent alcohol misusers. *Nature, 512*, 185–189. https://doi.org/10.1038/nature13402

Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation, 8*, 1341–1390. https://doi.org/10.1162/neco.1996.8.7.1341

Wolpert, D. H. (2002). The supervised learning no-free-lunch theorems. In R. Roy, M. Köppen, S. Ovaska, T. Furuhashi, & F. Hoffmann (Eds.), *Soft Computing and Industry* (pp. 25–42). London, UK: Springer.

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

Zeng, J., Ustun, B., & Rudin, C. (2017). Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 180*, 689–722. https://doi.org/10.1111/rssa.12227

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

Received March 1, 2018
Revision received July 30, 2018
Accepted September 17, 2018
Published online February 22, 2019

Acknowledgments
The action editor for this article was Suzanne Jak.

Funding
Mike W.-L. Cheung was supported by the Academic Research Fund Tier 1 (FY2017-FRC1-008) from the Ministry of Education, Singapore.

Ranjith Vijayakumar
Department of Psychology
Faculty of Arts and Social Sciences
National University of Singapore
Block AS4, Level 2, 9 Arts Link
Singapore 117570
r.v@u.nus.edu