SOCIUS
Special Collection: Fragile Families Challenge

ASA AMERICAN SOCIOLOGICAL ASSOCIATION

Socius: Sociological Research for a Dynamic World
Volume 5: 1–10

DOI: 10.1177/2378023119825886
srd.sagepub.com
SAGE

Variable Selection and Parameter Tuning for BART Modeling in the Fragile Families Challenge

Nicole Bohme Carnegie[1] and James Wu[2]

Abstract

Our goal for the Fragile Families Challenge was to develop a hands-off approach that could be applied in many settings to identify relationships that theory-based models might miss. Data processing was our first and most time-consuming task, particularly handling missing values. Our second task was to reduce the number of variables for modeling, and we compared several techniques for variable selection: least absolute selection and shrinkage operator, regression with a horseshoe prior, Bayesian generalized linear models, and Bayesian additive regression trees (BART). We found minimal differences in final performance based on the choice of variable selection method. We proceeded with BART for modeling because it requires minimal assumptions and permits great flexibility in fitting surfaces and based on previous success using BART in black-box modeling competitions. In addition, BART allows for probabilistic statements about the predictions and other inferences, which is an advantage over most machine learning algorithms. A drawback to BART, however, is that it is often difficult to identify or characterize individual predictors that have strong influences on the outcome variable.

Keywords

regression trees, Bayesian additive regression trees, prediction

Introduction

Background on the Fragile Families Challenge

In this article, we present one approach to participation in the Fragile Families Challenge, a predictive modeling challenge using data from the Fragile Families and Child Wellbeing Study (FFCWS; Reichman et al. 2001). FFCWS is a longitudinal study following more than 4,000 children in large U.S. cities from birth. For purposes of the challenge, we built predictive models for six outcomes (eviction, GPA, grit, material hardship, layoff, and job training) at age 15 using data from the first five waves of data collection (at birth and 1, 3, 5, and 9 years of age).

The goals of the challenge were to (1) build the best possible predictive model for these outcomes and (2) identify children and families who substantially outperformed or underperformed predictions for further study. Both goals will further future avenues of research by the FFCWS team.

Our Approach to the Challenge

We are a team of statisticians with expertise in causal inference but with little to no subject-area expertise. Thus, our approach to the challenge was to develop a procedure that could be readily applied in various settings and that could identify relationships not specified in theory-based models. We were interested in automated variable selection approaches and in minimizing modeling assumptions required for predictive modeling.

We first explored the choice of variable selection procedures. We used the least absolute selection and shrinkage operator (LASSO), the horseshoe prior, Bayesian generalized linear models (BGLM), and Bayesian additive regression trees (BART; Chipman, George, and McCulloch 2007). LASSO and BGLM represent simpler, commonly used methods, while the horseshoe prior is a more recent development with better properties (Carvalho, Polson, and Scott 2010). All of these methods use linear models, how-

[1]Montana State University, Bozeman, MT, USA
[2]New York University, New York, NY, USA

Corresponding Author:
Nicole Bohme Carnegie, Montana State University, P.O. Box 172400, Bozeman, MT 59717-2400, USA.
Email: nicole.carnegie@montana.edu

ever, so we also consider BART, which can detect nonlinear relationships.

The method we used for building the predictive model was BART. BART is a tree-based nonparametric regression-fitting algorithm that can accurately model even complex response surfaces without specifying functional forms or interactions a priori. It is an effective tool in causal inference and predictive modeling (Chipman, George, and McCulloch 2010; Hill 2011). In fact, the authors have been successful using BART in previous black-box causal modeling competitions associated with the Atlantic Causal Inference Conference (Dorie et al. 2017). One major advantage of BART over other machine-learning algorithms is its basis in a probabilistic framework, which permits assessment of uncertainty. The default priors and hyperparameters generally give good predictive performance without a significant amount of tuning. We did, however, explore the effect of the number of trees on predictive performance.

In this article, we first describe the methods used in development of models for submission to the challenge (Methods section), then address questions regarding variable selection methodology and BART tuning parameters (Results section), and finally discuss implications of those results and our experience participating in the Fragile Families Challenge more broadly (Discussion section).

Methods

Creating an entry for the Fragile Families Challenge required three main steps: data processing, variable selection, and model fitting. We describe our approach to all three steps below.

Data Processing

Data processing was by far the most time consuming of the three main steps and the least amenable to automation. The data for the challenge consisted of five waves of data, each containing multiple surveys. Altogether, the data files as provided comprised more than 12,000 variables, with a significant amount of missing data.

We first dropped variables not considered to be of analytic interest. These included computed sampling weights, variables related to the administration of the survey (subject ID numbers, duration of interview, date of survey administration, etc.), and variables taking a single value for all training observations. Variables were then identified as either categorical or quantitative by inspection of the survey documentation. This required perusing more than 15 separate questionnaire files and corresponding codebooks. We evaluated the question and the set of possible responses to decide whether a variable was purely quantitative (e.g., height, age). Rating scales (e.g., *strongly disagree* to *strongly agree*) were considered categorical. Categorical variables were labeled as factors, forcing them to be treated as sets of indicator variables for analysis purposes. Once the set of variables was refined and they were classified, we moved on to missing data.

Missing data were categorized in a number of ways. We spent the most time on values that were missing because of skip patterns in the survey rather than a refusal to respond on the part of the respondent. Since these "missing" values represented something meaningful, it was important to code them sensibly. We evaluated skip patterns in all surveys and filled in logical values where possible. For example, "Number of other biological children" was skipped if the parent responded no to the question "Do you have any biological children other than the focal child?" This could logically be filled in with a 0. When there was no logical value for a categorical response, we allowed the skip value (–6) to be included in the set of categories. In some cases, the same question was asked in different sections of the survey (e.g., depending on whether the mother was married to the child's father, in a nonmarital relationship with him, or no longer in a relationship with him). These questions were merged into a single variable. These two processes allowed us to fill in about 22 percent of missing values.

After handling legitimate skips, we imputed other missing values. For this exercise, we imputed using the median of quantitative variables and the mode of categorical variables. A more nuanced approach, such as hot-deck or regression-based multiple imputation, would likely further improve predictive performance (see, e.g., Little and Rubin 2014).

BART and Predictive Model Fitting

Model fitting was our third step, but since our predictive model is also one of our candidate approaches for variable selection, we describe BART first.

Our method of choice for model fitting was BART (Chipman et al. 2007). BART uses regression trees as its building block. A regression tree is an algorithm that partitions data into nonoverlapping subsets to minimize the variance in the response variable within these subsets. The fit from this partition is expressed simply as the mean of each subset (see Figure 1). In some ways, this is a similar idea to linear regression, which fits a line to minimize overall (squared) deviations from that line. However, regression trees don't assume a linear relationship between subgroups and in fact can support a wide variety of implicit functional relationships between predictors and outcomes.

BART embeds this predictive algorithm into a likelihood framework. This allows for coherent probabilistic statements about the resulting predictions and other inferences. This is an advantage over most machine learning algorithms in that BART is able to produce coherent uncertainty intervals while machine learning methods generally cannot. See the appendix for model details.

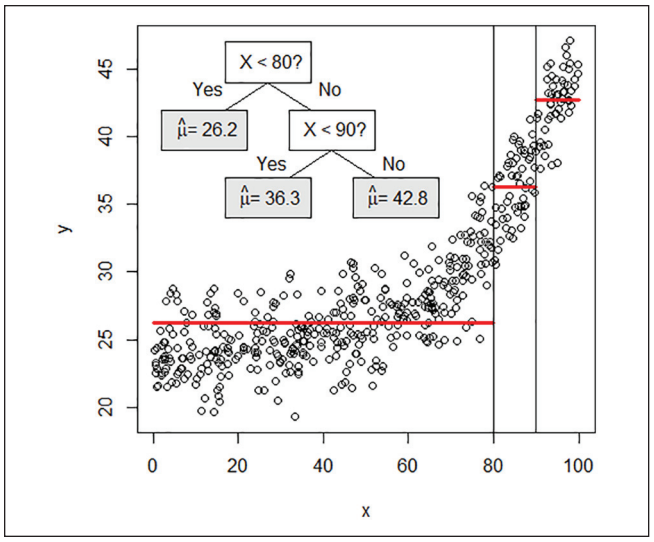

Figure 1. Tree-fitting example.
Note: This is an example of the construction of a single tree. The data are partitioned first at $X = 80$ and then at $X = 90$ among those observations where $X \geq 80$. The fit for the tree is the mean of observations that fall in each terminal node (shaded in gray) and is shown as horizontal line segments on the scatterplot.

Variable Selection

We now return to the second step of the process for creating each entry: variable selection. After preparing the data, it was necessary to reduce the number of variables used in modeling. With approximately 2,000 observations available (depending on the outcome used), it would be easy to overspecify a model when trying to incorporate all 12,000-plus predictors. We opted to compare four approaches to variable selection. These were LASSO, regression with a horseshoe prior, BGLM, and BART (Carvalho et al. 2010; Gelman and Hill 2007; James et al. 2017). See the appendix for details.

Three of the above methods (BGLM, LASSO, horseshoe prior) are based on linear models. This means that they will select variables based on the strength of their linear relationships with the response (conditional on any other variables in the model). We are using a nonparametric approach to response surface modeling (see BART and Predictive Model Fitting section); this opens up the possibility that variables with important nonlinear relationships with the response will be excluded. To combat this, we also considered using BART for variable selection. While BART is not formally a variable selection method, it does inherently include regularization, as with the previous three methods. The advantage of BART is that BART allows for an extremely flexible functional form for the relationship between the response variable and predictors. On the other hand, the properties of using BART for selection have not been established. See the appendix for details about the use of BART for variable selection.

We compare the methods on two dimensions: number of variables selected and out-of-sample prediction error for comparable BART models fit to the selected variables. The first is primarily descriptive—a measure of the stringency of the selection criterion. When comparing prediction error, we assume that methods that are more likely to capture "real" relationships in the data will generate training data sets that result in better predictions. Since we use the same model (BART) for predictive model fitting, differences in prediction error can be attributed at least partially to differences in variables used.

Results

Predictive results for the final holdout data were provided by competition organizers in terms of mean squared error (MSE). However, MSE varied widely across outcomes (see Figure 2), at least partially as a function of the scale and variability of the outcomes (we would expect MSE to be lower for lower-variance outcomes). For this reason, we also compare results using the root mean squared error (RMSE) standardized by the standard deviation of the variable being predicted, estimated using the nonmissing observations in the training data set. This gives a consistent, interpretable scale that allows comparisons across outcomes. A standardized RMSE (sRMSE) of 1 indicates that predictions are off by 1 standard deviation, on average. Once we account for the inherent variability of the outcome using sRMSE, we see that among continuous outcomes, our methods did best predicting GPA and worst predicting grit.

MSE corresponds to the Brier score for binary predictions, with 0.00 corresponding to perfect prediction, 0.25 corresponding to complete uncertainty ($\hat{p} = 0.5$ for all observations), and 1.00 corresponding to perfect error (Brier 1950). We see that predictions for eviction are close to a perfect score while job training and layoff are somewhat better than complete uncertainty. This is at least partially due to the fact that only about 6 percent of families in the training data experienced eviction while 21 percent experienced layoff and 23 percent received job training.

Early in the competition, we made submissions based on a single wave of data—both as a way to test out the submission system while still working through data preparation and to permit statements about whether a single wave could perform as well as the full five waves of data. These single-wave predictions were generally substantially worse than predictions based on all waves. For this reason, subsequent results will consider only submissions using all waves of data. We first consider the prediction error results for the different variable selection methods. To facilitate comparisons across methods, we compute a residual prediction error adjusted for outcome. This makes comparisons between methods clearer by reducing the variability within method that is due to applying all methods to all outcomes. The residual is computed by subtracting the average error (MSE or sRMSE) for a particular outcome from the error of each submission for that outcome. Thus, if a particular method

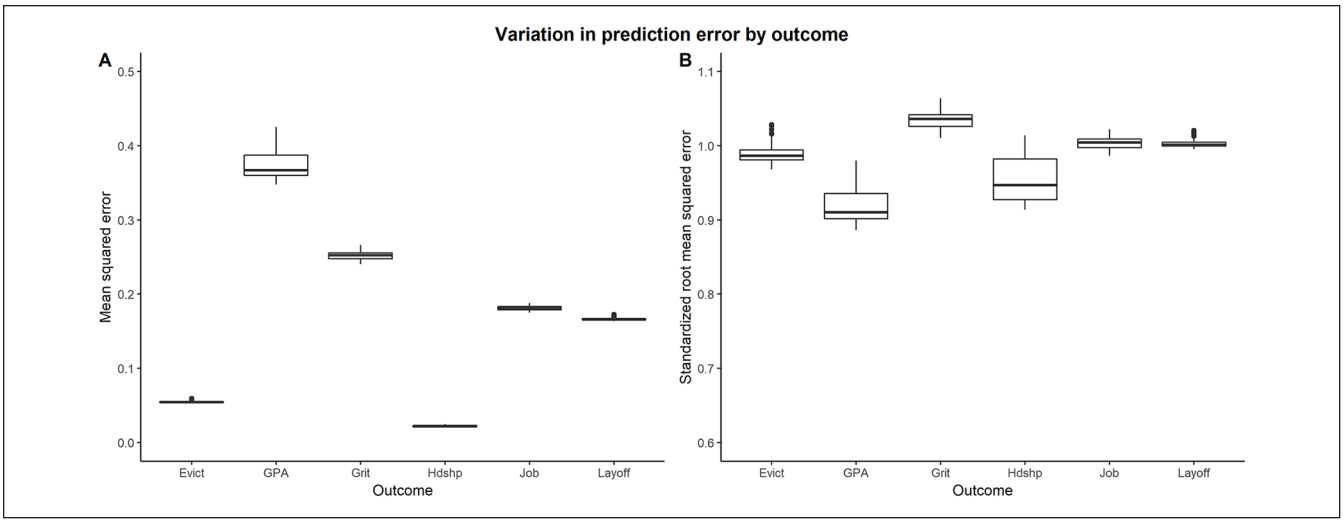

Figure 2. Predictive model performance by outcome.
Note: This is the distribution of prediction error using the metric of mean squared error (MSE, A) and standardized root mean square error (sRMSE, B). Each box in a boxplot represents results across all combinations of variable selection method, number of trees used in Bayesian additive regression trees fit, and subset of data used for the specified outcome: in total, approximately 100 observations per box.

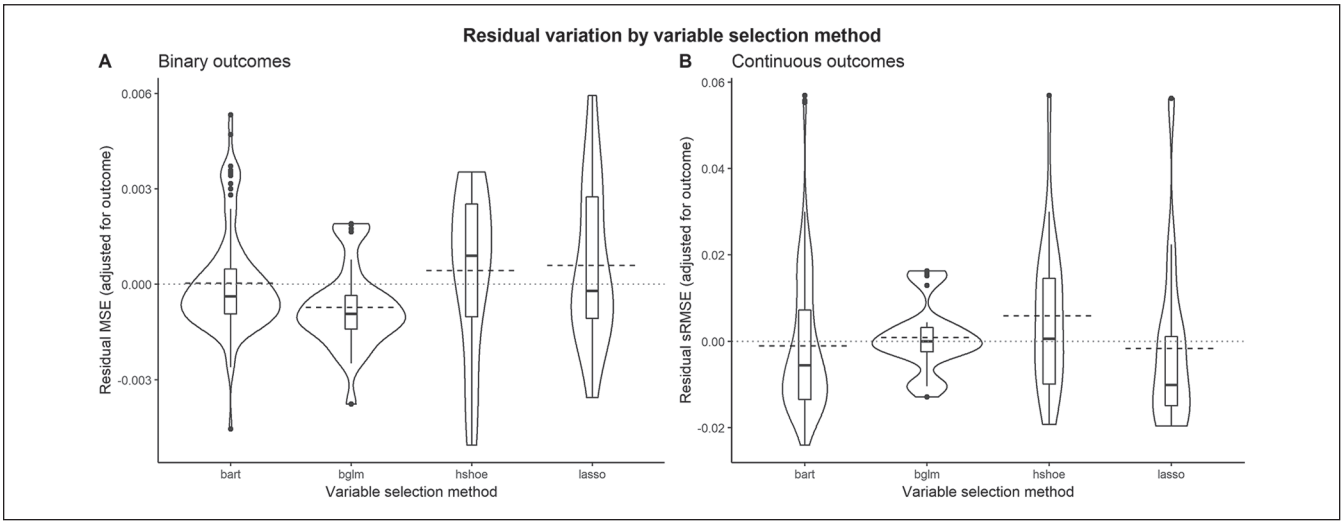

Figure 3. Predictive model performance by variable selection method.
Note: Given variability in overall performance by outcome, we calculated the residual variability in predictive performance by subtracting the average mean squared error (MSE, A) or standardized root mean squared error (sRMSE, B) for the outcome being predicted from MSE or sRMSE for each observation. The plots presented here are combined violin plots and boxplots, with a dashed line representing the mean for each grouping and a dotted line at zero for reference. The violin plots are a form of density plot, indicating the full distribution of results for each variable selection method. The data used for each box/violin plot include all combinations of the outcomes of the appropriate type (binary or continuous) and Bayesian additive regression trees tuning parameters—in total, 45 observations per plot.

outperforms the others, the distribution of residual prediction errors for that method will be shifted down. There was some variability in distribution between methods, but not of large magnitude (see Figure 3). Results for BGLM are far more consistent than results from other methods, which have occasional large values for MSE and sRMSE.

In terms of number of variables selected, LASSO and the horseshoe prior were generally much more stingy with the number of variables chosen: BGLM averaged 597 across all waves for one outcome, LASSO 361, and horseshoe 55 (see Figure 4). The number of variables chosen by BART was controlled by tuning a variable c and ranged from 370 to 932. We also observe that a higher percentage of variables was kept from waves 1 and 5 for most outcomes (see Figure 5), particularly when predicting job training.

When varying the number of trees used to create a BART fit, we hypothesize that the performance will be related to the ratio of trees to predictors. If this ratio is too high, the model will likely

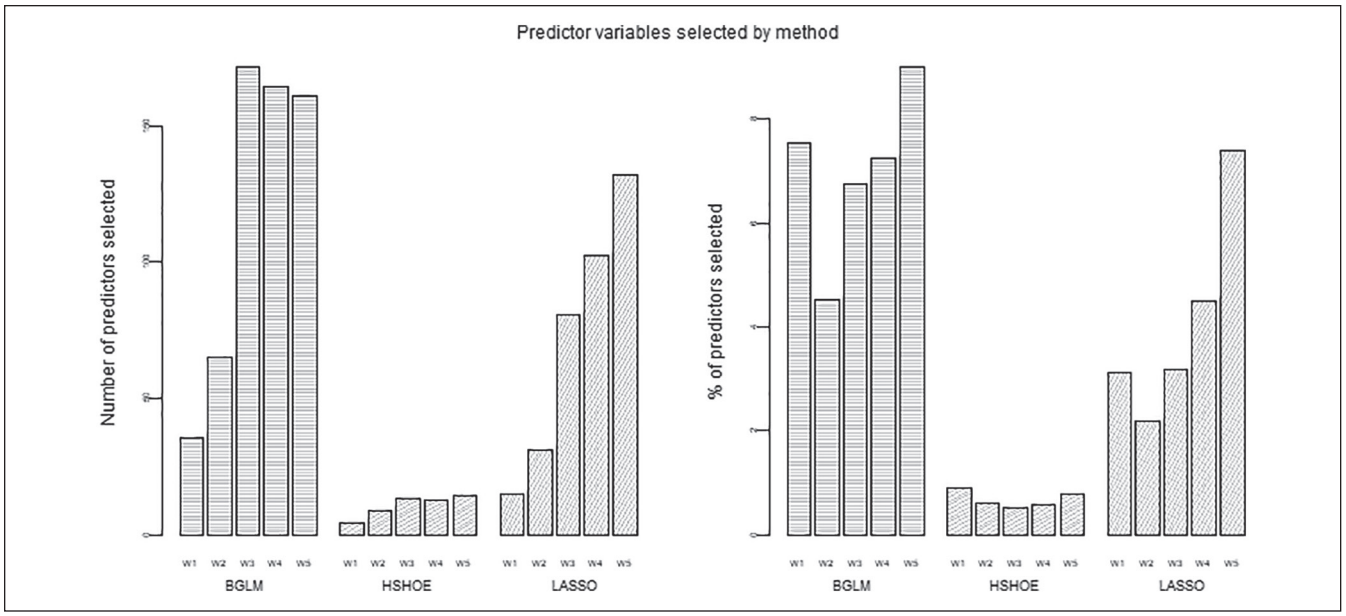

Figure 4. Predictor variables selected by variable selection method.
Note: Bar plots indicate the number of predictors selected by wave for each method (L). Since the number of predictors varied by wave, we also include the percentage of predictors selected (R).

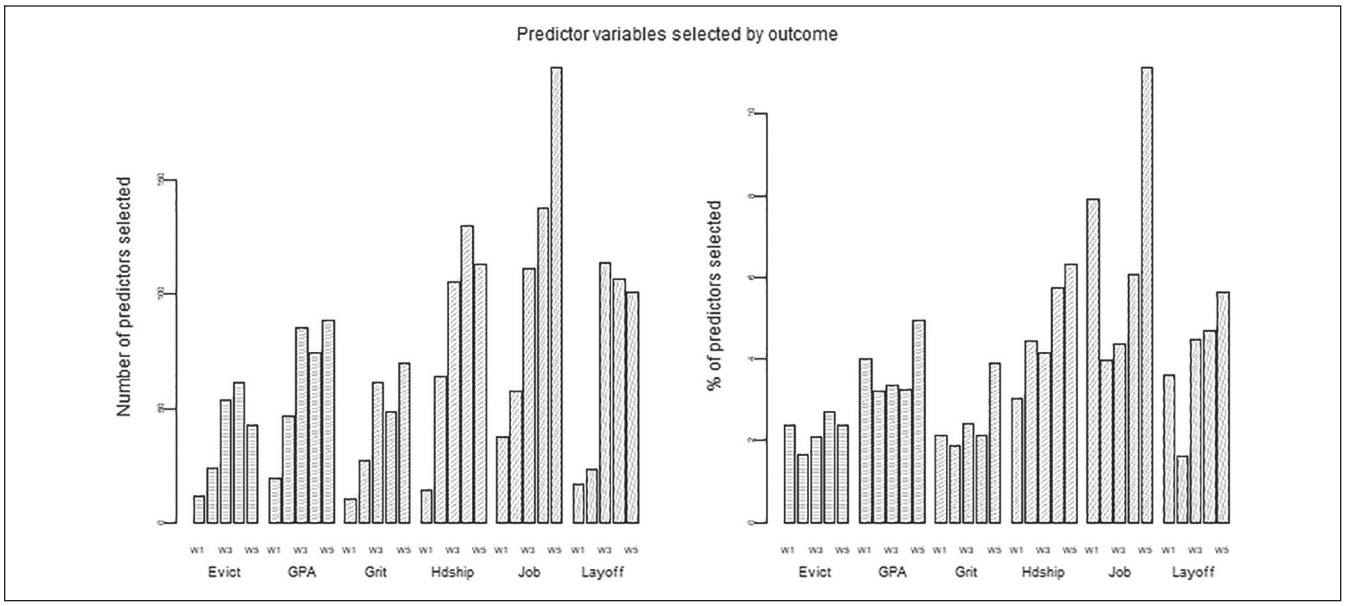

Figure 5. Predictor variables selected by outcome.
Note: Bar plots indicate the number of predictors selected by wave for each outcome (L). Since the number of predictors varied by wave, we also include the percentage of predictors selected (R).

overfit; this leads to improved performance in the training sample but decreased accuracy in predictions in test samples. We use LOESS—a local bivariate polynomial smoother, a descendant of Locally Weighted Scatterplot Smoothing—to generate a smoothed representation of the relationship between tree-to-predictor ratio and residual prediction error, with 95 percent confidence bounds based on a t approximation to the prediction interval. After accounting for differences in performance by outcome, we observe some remaining relationship with the tree-to-predictor ratio (see Figure 6) for continuous outcomes. Primarily, the pattern is quite different for LASSO, with a declining error up to a ratio of about 50:1, which then increases and stabilizes. For other methods, the trend is for slightly increasing or stable error as the ratio increases.

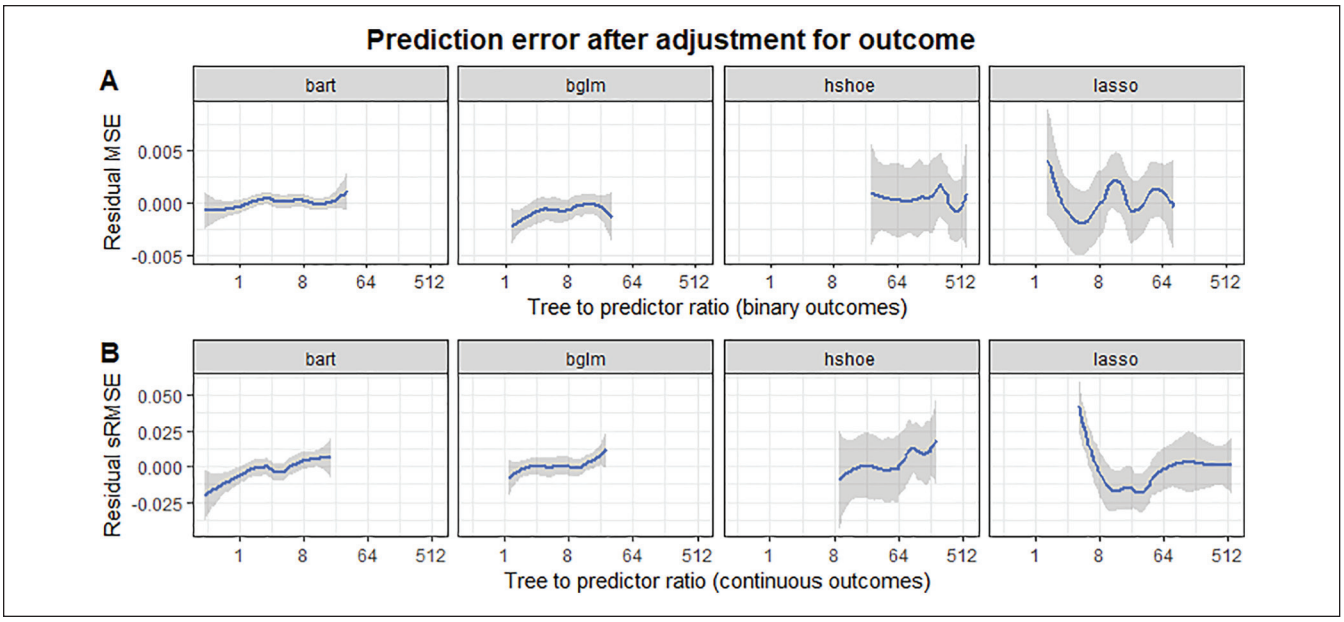

Figure 6. Predictive model performance by tree-to-predictor ratio and variable selection method.
Note: In this plot, we take the residual prediction error after accounting for outcome differences (subtracting the mean prediction error in each outcome group from the corresponding observations) and relate it to the ratio of the number of trees used in the final Bayesian additive regression trees fit to the number of predictor variables available. Since this ratio is associated with the variable selection method used, results are presented separately by method. The plot gives a smoothed curve with 95 percent confidence bounds.

To simultaneously assess the contribution of all components to the variability in predictive performance, we fit a linear regression of MSE or sRMSE on the outcome being predicted (MSE for binary, sRMSE for continuous), variable selection method used, and ratio of number of trees to number of predictors. An interaction between tree-to-predictor ratio and LASSO-based variable selection was included due to the obvious difference in the pattern for this group seen in Figure 6. There are strong differences by outcome and some differences by variable selection (see Table 1). Interesting to note, there is a relationship between only the tree-to-predictor ratio and the final performance for continuous outcomes.

For binary outcomes, the vast majority of the explained variability comes from the outcome. There is a statistically significant improvement in predictive performance when using BGLM for selection, but the effect is three orders of magnitude smaller than the differences between outcomes. In this setting, the tree-to-predictor ratio seems to matter little.

For continuous outcomes, we again see that most of the explained variability comes from the outcome—in fact, outcome alone accounts for 92 percent of the variability in sRMSE (using ANOVA variance decomposition). The variable selection method accounts for only 2 percent of the variability that remains, while the tree-to-predictor ratio accounts for about 4 percent. The differences between outcomes amount to 0.03 to 0.12 units' difference in sRMSE. Three variable selection methods—BART, BGLM, and horseshoe—are effectively equivalent, with differences between variable selection methods on the order of 0.1 percent of a standard deviation. (See the appendix for further unpacking of results.)

Our findings here suggest that 500 trees (or possibly fewer) are adequate for model fitting in most cases, with the exception of the dramatic differences when using LASSO for variable selection with continuous outcomes. It is difficult to generalize based on results from a single data set, but our observations are largely consistent with previous results (Chipman et al. 2010).

Discussion

Given the limited time frame of the challenge, one of the biggest difficulties was getting the data into a usable state in a timely fashion. The greatest portion of our time, by far, was spent on data preparation—making sure qualitative variables were treated as factors, recoding legitimate skips to reduce missingness, and identifying variables to be merged. We suggest that in other challenges, if the primary interest is the predictive modeling, it would benefit both the participants and the organizers to provide a data set with this type of basic cleaning done to allow participants to maximize effort on model development.

Our final results suggest that an automated procedure such as the one we propose here can perform quite well for predictive model building relative to competitors. None of the submissions had stellar predictive performance. We hypothesize that this may be related to the ubiquitous use of median imputation's masking relationships in the data.

Table 1. Estimates of Contributions of Inputs to Predictive Modeling Error Using a Linear Regression Approach.

	Binary			Continuous		
	Estimate	*SE*	*p*	Estimate	*SE*	*p*
Outcome						
Eviction	Reference	Reference	Reference	—	—	—
Job training	.126	.0003	<.001	—	—	—
Layoff	.112	.0003	<.001	—	—	—
GPA	—	—	—	Reference	Reference	Reference
Grit	—	—		.124	.003	<.001
Hardship	—	—	—	.032	.003	<.001
Variable selection						
Bayesian additive regression trees	Reference	Reference	Reference	Reference	Reference	Reference
Bayesian generalized linear models	−.001	.000	.029	−.002	.003	.629
Horseshoe	−.000	.001	.887	−.009	.007	.159
LASSO	.000	.001	.455	.089	.023	<.001
Tree-to-predictor (T:P) ratio						
log(T:P ratio)	.000	.000	.333	.004	.002	.043
log(T:P ratio)2	—	—	—	.000	.001	.767
LASSO by T:P interaction						
LASSO × log(T:P)	—	—	—	−.052	.013	<.001
LASSO × log(T:P)2	—	—	—	.006	.002	<.001
Intercept	.054	.0003	<.001	.899	.003	<.001
Adjusted *R*2		.999			.931	

Note: For binary outcomes, the response is the mean squared error of prediction from the final holdout data. For continuous outcomes, the response is the standardized root mean squared error of prediction (standardization by the standard deviation of the response). The largest differences are between outcomes with nontrivial contributions of tree-to-predictor ratio and some differences among variable selection methods. LASSO = least absolute selection and shrinkage operator. Dashes indicate terms not included in the model.

Our finding that the variable selection method matters little, at least for appropriate choices of tree-to-predictor ratio, is perhaps surprising. This may be because there was an extremely large number of potential predictors, many of which are likely to have little or no association with the specific outcome of interest. (This is reflected in the <10 percent selection rate across waves for all methods.)

It is possible that all approaches are able to identify a core set of primary predictors and only vary in their inclusion or exclusion of predictors that have minimal impact in the final predictive model; however, initial investigation suggests this is not the case. Only a quarter of variables selected by any method appeared in the results for at least one other method, and only 13 were selected by more than half of the variable selection methods. A number of sets of variables measure the same construct; it was beyond the scope of our analysis to determine if the same thematic groups of variables were selected by all methods. Nonetheless, it seems unlikely that that would substantively change the conclusion that the results of the variable selection methods differ widely.

For the scientific purposes of the challenge, the use of BART does present one difficulty. Given the structure of BART, with the response surface approximated by a sum of decision trees, it can be difficult to identify individual predictors that have a strong influence on the outcome. It is straightforward to use predictions from the black-box approach to identify outliers (those that outperform or underperform predictions) and for more general prediction needs such as defining models for propensity scores in other causal models. Hypothesis generation requires more sophisticated exploration to extract relationships in the data that do not come from existing theory. See Green and Kern (2012) for an example of how to extract these relationships from the BART fit.

Were we to propose a single model based on the results of our investigation, we would include first a more nuanced approach to missing data—at the very least some form of hot-deck imputation, if not a regression-based approach. Then we would suggest the use of BGLM for variable selection due to its advantages for binary outcomes and relative equivalence to other methods for continuous outcomes. The final predictive model would be BART, with the number of trees close to the number of predictors included in the model and with cross-validation used for binary outcomes.

Appendix

Data

The individuals surveyed varied by wave; all waves included interviews with the child's mother, father, and primary caregiver (if not the mother/father). Some later waves also

included interviews with a childcare provider, the child's primary teacher, or the child. There was also a home visit in certain waves.

Variable Selection Methods

The least absolute selection and shrinkage operator (LASSO), horseshoe, and Bayesian generalized linear models (BGLM) introduce a penalty in the form of a regularization parameter that controls the trade-off between the penalty and a likelihood-based loss function (e.g., residual sum of squares). The interaction of penalty and the regularization parameter will then have the effect of shrinking the estimates toward zero or even (depending on the penalty type) reducing some estimates to zero (James et al. 2017).

In least squares, the maximum likelihood estimate is given by $\hat{\beta}_{\text{MLE}}=\arg\min \beta \sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2$. The coefficients ($\beta_j$) are chosen to minimize the squared error between the observed values (y_i) and the predicted values ($\hat{y}_i$). Penalized regression methods are based on this likelihood but add a penalty and regularization parameter. LASSO uses an $l-1$ penalty, $\lambda\sum_{j=0}^{p}\left|\beta_j\right|$, where the maximum likelihood would be now given as $\hat{\beta}_{\text{MLE}}=\arg\min \beta \sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2+\lambda\sum_{j=0}^{p}\left|\beta_j\right|$. This adds one parameter, λ, which controls the strength of the shrinkage toward zero. In our submissions, λ was chosen by cross-validation to reduce prediction error (implemented as cv.glmnet in the glmnet package in **R**) (Friedman, Hastie, and Tibshirani 2010). We used the largest λ such that prediction error is within 1 standard error of the minimum to reduce overfitting. Variables with their coefficients shrunk to 0 were dropped from the model.

In a Bayesian framework, choosing a prior distribution with a zero-mean is analogous to choosing a regularization parameter. LASSO's $l-1$ penalty can be approximated by a Laplacian prior (Case and Park 2008). When using BGLM for variable selection, we used a gaussian prior with zero-mean for variable selection, which is equivalent to the $l-2$ penalty or ridge regression. In this case, the variance of the prior controls the strength of shrinkage toward the prior mean (typically zero). We used the bayesglm function in the arm package in **R** (Gelman and Su 2016) to fit the model and selected any variables whose posterior credible interval excluded 0.

The horseshoe prior is a more complex shrinkage prior that proposes a global-local mixture and is claimed to be superior to the Laplacian prior when most coefficients are truly zero (Carvalho, Polson, and Scott 2010). The priors for coefficient estimates can be represented by $\beta_j \sim N(0, \sigma^2\tau^2\lambda_j^2)$, where $\lambda_j \sim$ *Half Cauchy*(0, 1) and $\tau \sim$ *Half Cauchy*(0, 1) (the default prior distributions in the **R** package horseshoe). Here, τ controls global shrinkage toward zero while the local parameters λ_j allow coefficient-specific deviations in the degree of shrinkage. Where LASSO and ridge regression have a single estimated penalty parameter λ, the horseshoe prior has both the global parameter τ and a λ_j for each coefficient. We again selected any variables with posterior credible interval for the regression coefficient excluding 0. This approach was implemented using the horseshoe package in **R** (van der Pas et al. 2016).

When using Bayesian additive regression trees (BART) for variable selection, we used a decision rule based on the frequency that each variable was used in a tree to determine which variables were kept. We average the number of times each training variable was used across trees to get a measure of the informativeness of each variable. For all variable selection runs, we used 1,000 trees. Variable selection was performed within waves of data collection, yielding differing numbers of training variables across waves. We found that selecting all variables whose average usage in trees was greater than $c/$(number of training variables) yielded a roughly constant percentage of variables kept in each iteration. This percentage varied with the cutoff c, which we varied between 200 and 500 to yield predictor sets of differing sizes for comparison. Note that this implies two runs of BART: one to thin the predictor set and a second to fit the final model. When using variable selection prior to BART modeling, there is model uncertainty due to the selection that is not incorporated into the posterior credible intervals.

BART Model

Relying on one regression tree to fit the data has the disadvantage that it overprivileges the role of interactions between variables, making it difficult to find additive relationships. Boosted regression trees combat this issue by fitting many small trees using a back-fitting algorithm. Simply described, the algorithm fits a small tree (perhaps one that partitions the data into four subgroups, or terminal nodes) and gets the fitted values from that tree. The fitted values are subtracted from the observed values of the response, and another small tree is fit. This process is repeated until some number of small trees has been fit. The trick is to not overfit to the data; therefore in practice, the fit from each small tree is multiplied by a small constant. This tuning parameter is often chosen via cross-validation.

BART is akin to a Bayesian form of boosted regression trees that gets around this computationally intensive cross-validation step by shrinking the fit of each tree using an intelligent prior (Hastie, Tibshirani, and Friedman 2003). This regularization prior controls the tendency for trees to overfit by keeping individual trees small. The recommended prior probability that a node at depth $d = 0, 1, 2, \ldots$ is nonterminal is $0.95(1 + d)^{-2}$, which yields $p = .05, .55, .28, .09, .03$ for trees with 1, 2, 3, 4, or ≥5 terminal nodes.

Using BART, we model the response Y as a sum of trees g, plus random error:

$$Y = \sum_{j=1}^{m} g(x; T_j, M_j) + \varepsilon, \varepsilon \sim N\left(0, \sigma^2\right),$$

where T_j is the jth tree with means M_j at its terminal nodes. The prior for σ^2 is an inverse $\chi^2(\nu, \lambda)$, with default degrees of freedom $\nu = 3$ and $\lambda = 3$. The number of trees m is left as a tuning parameter.

The prior for the means associated with terminal nodes is chosen so that the majority of the prior weight is in the interval between the minimum and maximum values of the outcome. This is controlled by the parameter k: $k = 2$ implies that the maximum and minimum are each approximately 2 standard deviations from the mean, or ≈ 95 percent prior probability in the interval (y_{min}, y_{max}) when Y is continuous and symmetrically distributed. $k = 3$ would yield approximately 99 percent prior probability in this interval.

When Y is binary, this definition is more difficult to interpret since all observations are at either $y_{min} = 0$ or $y_{max} = 1$. In addition, results are highly dependent on the choice of k (Dorie et al. 2016). For this reason, we use an extended version of BART that uses cross-validation to choose the $k \in 1, 2, 4, 8$. This extension will soon be available as part of the dbarts package in **R** (Dorie, Chipman, and McCulloch 2014).

BART has been shown to be competitive with other machine learning algorithms while requiring less tuning, less computational time, and also producing uncertainty intervals (Chipman, George, and McCulloch 2010; Dorie et al. 2017). The advantages of BART in causal inference stem from the very precise modeling of the response surface, which permits much more thorough control for confounding than that observed with traditional parametric models.

Results and Discussion

As seen in Table 1, LASSO performs significantly worse than BART when the tree-to-predictor ratio is 1, with an effect of 8.9 percent of a standard deviation, comparable to the differences between outcomes. There appears to be a linearly increasing relationship between log tree-to-predictor ratio and standardized root mean squared error (sRMSE), except in the case of LASSO, where there is evidence of a parabolic relationship. This actually yields lower prediction error using LASSO when the tree-to-predictor ratio is around 50, although the difference from the optimum of the other settings (selection via horseshoe prior with a tree-to-predictor ratio of 1) is minimal. The coefficient on log tree-to-predictor ratio suggests that each doubling of the number of trees relative to the number of predictors is associated with a 0.25 percent standard deviation increase in the prediction error—again, hardly large enough to be of major concern. The intercept here is the expected sRMSE for eviction using BART for variable selection with a tree-to-predictor ratio of 1; all other coefficients are differences from that setting.

Based on our results, we would use variable selection with BGLM and BART with 1,000 trees (among the submissions made) for binary outcomes. This set of results has mean squared error 0.3 to 1.1 percent higher than the highest-ranked final submissions on holdout data. For continuous outcomes, any number of combinations is possible, but the results would suggest that LASSO for variable selection followed by a BART fit using 10,000 trees is marginally better than other settings; these results have sRMSE 0.6 to 4.3 percent higher than the top-ranked final submissions on the holdout data. Given that most predictions were around an sRMSE of 1, this corresponds to a difference of less than 5 percent of a standard deviation in prediction.

Authors' Note

The results in this article were created with software written in R 3.3.3 (R Core Team 2017) using the following packages: ggplot2 2.2.1 (Wickham 2009), gridExtra 2.2.1 (Baptiste 2016), Rmisc 1.5 (Hope 2013), dplyr 0.5.0 (Wickham 2016), dbarts 0.8-7 (Dorie 2016), arm 1.9-3 (Gelman 2016), horseshoe 0.1.0 (van der Pas 2016), glmnet 2.0-5 (Hastie 2016), and methods 3.3.2 (R Core Team 2016).

Acknowledgments

The authors wish to thank Jennifer Hill for her contributions to the development of the approach and for her support in the writing of this manuscript.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Funding for the Fragile Families and Child Wellbeing Study was provided by the Eunice Kennedy Shriver National Institute of Child Health and Human Development through grants R01HD36916, R01HD39135, and R01HD40421 and by a consortium of private foundations, including the Robert Wood Johnson Foundation. Funding for the Fragile Families Challenge was provided by the Russell Sage Foundation.

ORCID iD

Nicole Bohme Carnegie https://orcid.org/0000-0001-7664-6682

Supplemental Material

Supplemental material for this article is available with the manuscript on the *Socius* website.

References

Brier, Glenn W. 1950. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review* 78:1–3.

Carvalho, Carlos M., Nicholas G. Polson, and James G. Scott. 2010. "The Horseshoe Estimator for Sparse Signals." *Biometrika* 97(2):465–80.

Case, George, and Trevor Park. 2008. "The Bayesian Lasso." *Journal of the American Statistical Association* 103:681–86.

Chipman, Hugh A., Edward George, and Robert McCulloch. 2007. "Bayesian Ensemble Learning." Pp. 265–72 in *Advances in Neural Information Processing Systems* 19, edited by B. Schölkopf, J. Platt, and T. Hoffman. Cambridge, MA: MIT Press.

Chipman, Hugh A., Edward I. George, and Robert E. McCulloch. 2010. "BART: Bayesian Additive Regression Trees." *Annals of Applied Statistics* 4(1):266–98.

Dorie, Vincent, Hugh Chipman, and Robert McCulloch. 2014. *dbarts: Discrete Bayesian Additive Regression Trees Sampler*. R package version 0.8-5. Retrieved January 14, 2019 (https://CRAN.R-project.org/package=dbarts).

Dorie, Vincent, Masataka Harada, Nicole B. Carnegie, and Jennifer Hill. 2016. "A Flexible, Interpretable Framework for Assessing Sensitivity to Unmeasured Confounding." *Statistics in Medicine* 35(20):3453–70.

Dorie, Vincent, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. 2017. "Automated versus Do-it-yourself Methods for Causal Inference: Lessons Learned from a Data Analysis Competition." Retrieved January 9, 2019 (https://arxiv.org/abs/1707.02641).

Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33(1):1–22.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.

Gelman, Andrew, and Yu-Sung Su. 2016. *arm: Data Analysis Using Regression and Multi-level/Hierarchical Models*. R package version 1.9-3. Retrieved January 14, 2019 (https://CRAN.R-project.org/package=arm).

Green, Donald P., and Holger L. Kern. 2012. "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees." *Public Opinion Quarterly* 76(3):491–511.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2003. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag.

Hill, Jennifer. 2011. "Bayesian Nonparametric Modeling for Causal Inference." *Journal of Computational and Graphical Statistics* 20(1):217–40.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2017. *An Introduction to Statistical Learning*. New York: Springer.

Little, Roderick J., and Donald B. Rubin. 2014. *Statistical Analysis with Missing Data*. Vol. 333. New York: Wiley & Sons.

Reichman, Nancy E., Julien O. Teitler, Irwin Garfinkel, and Sara S. McLanahan. 2001. "Fragile Families: Sample and Design." *Children and Youth Services Review* 23(4/5):303–26.

van der Pas, Stephanie, James Scott, Antik Chakraborty, and Anirban Bhattacharya. 2016. *horseshoe: Implementation of the Horseshoe Prior*. R package version 0.1.0. Retrieved January 14, 2019 (https://CRAN.R-project.org/package=horseshoe).

Author Biographies

Nicole Bohme Carnegie, PhD, is an assistant professor of statistics at Montana State University. Dr. Carnegie's research focuses on the intersections between causal inference, infectious disease modeling, and networks. This includes network-based infectious disease models to inform strategies for HIV prevention and methodologic work on making causal inferences in infectious disease settings, where observations are inherently not independent. She has a line of research developing methods for analyzing potential sensitivity of causal inferences to unobserved confounding in a variety of settings, including multilevel models and Bayesian additive regression trees (BART). For fun, she occasionally enters predictive modeling or causal inference competitions using BART.

James Wu is currently pursuing a master's degree in applied statistics for social science research at New York University. He has worked in the public sector for more than five years and is looking to apply his statistics knowledge to improving program and policy evaluation. James is interested in doing more research at the intersection of machine learning and casual inference fields.