Machine Learning Project

Matthew van Bommel

November, 2016

The Task

An unknown training dataset containing 1000 observations of 21 explanatory variables (labeled X_1 - X_{21}) and one response variable (y) was provided. No additional information was revealed. The task was to develop a method to use the X variables to predict y for a prediction set of 8068 observations of X_1 - X_{21}.

Preliminary Exploration

Missing Data

Neither the training nor test dataset contain missing data.

Variable Types

- X_{21} is a categorical variable with 4 categories (A, B, C, and D)
- X_7 and X_{17} are discrete variables
- All other variables (including y) are continuous

Variable Distributions

- The distributions of each input variable in the training set appear to match the corresponding distribution in the test set
- X_5, X_{13}, and X_{14} have highly right skewed distributions. This is noteworthy as some prediction methods struggle when provided highly skewed distributions. However, as shown in Figure 1, after a logarithmic transformation, the distributions of these variables are approximately symmetric. Thus, such a transformation may increase the predictive ability of these variables.

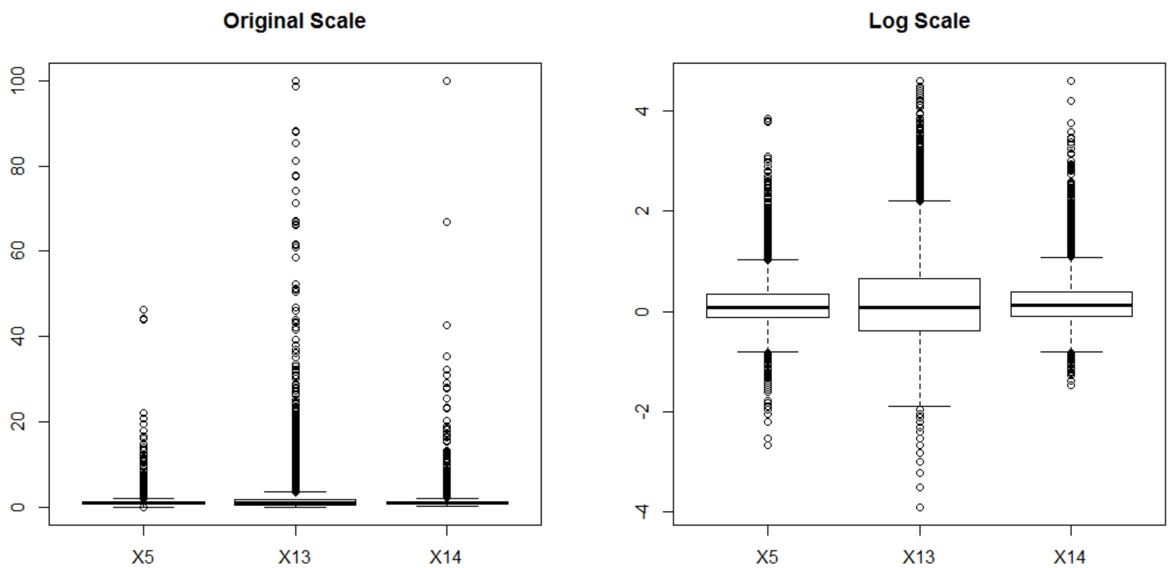

Figure 1: Distribution of variables X_5, X_{13}, and X_{14} in their original scale (left) and after a logarithmic transformation (right).

Collinearity of Explanatory Variables

Many pairs of explanatory variables are highly correlated (as high as 0.93 for X_2 and X_{18}). Thus, the effects of multicollinearity are a concern.

Correlation Between Explanatory Variables and Output Variable

Several input variables are highly correlated with the output variable including X_{12} (-0.95), X_{18} (-0.87), X_2 (-0.81), X_{11} (-0.81), and X_6 (0.78).

Regression Methods Examined

The following list contains the regression methods tested in this analysis:

- Ordinary Least Squares (OLS) Regression
- Stepwise Regression using the Bayesian Information Criterion (BIC)
- Least Absolute Shrinkage and Selection Operator (LASSO) Regression
  - Using the λ value that minimizes cross validation error (LASSO min)
  - Using the λ value that which gives the most regularized model such that cross validation error is within one standard error of the minimum (LASSO 1SE)

- Bayesian Model Averaging (BMA)
- Projection Pursuit Regression (PPR)
  - Using 1 term (PPR-1)
  - Using 2 terms (PPR-2)
  - Using 3 terms (PPR-3)
- Regression Trees
  - With no pruning (Full Tree)
  - Pruning such that the cross validation error is minimized (Min Tree)
  - Pruning such that the smallest tree remains while keeping the cross validation error within one standard error of the minimum (1SE Tree)
- Multivariate Adaptive Regression Splines (MARS)
  - With 1 degree of interaction (MARS-1)
  - With 2 degrees of interaction (MARS-2)
  - With 3 degrees of interaction (MARS-3)
- Random Forest (RF) Regression
- Bayesian Additive Regression Trees (BART)
- Neural Networks (NN)

Initial Test of Regression Methods

After tuning the parameters of the Random Forest, BART, and NN models using a similar procedure, the performance of each regression method was compared by examining the distribution of the root mean square prediction errors (MSPE) across 10-fold cross validation (where the folds were identical for all methods). Additionally, relative root MSPE values were also computed in each fold for all methods by dividing the root MSPE values by the best root MSPE values for that fold (so the best method in a fold has a relative root MSPE of 1, and the values for all other methods are greater than 1). Both sets of results are displayed in Figure 2 (with some of the worse performing methods excluded). The methods which produce the 3 lowest mean root MSPE values are Random Forests (4.16), BART (4.25), and MARS-1 (4.32). From the relative root MSPE plot, Random Forests appears to be the dominant method.

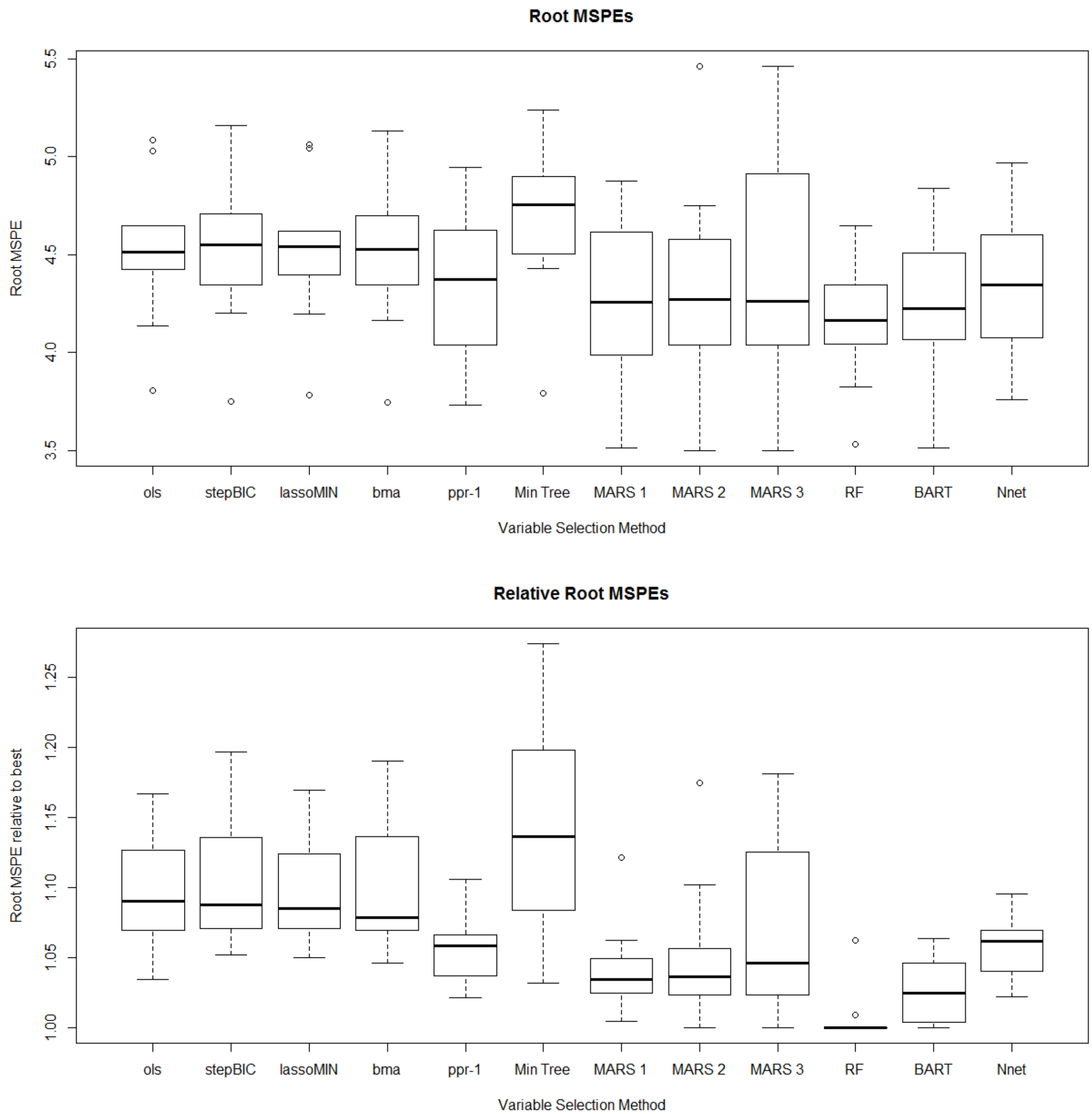

Figure 2: Root MSPE (top) and relative root MSPE (bottom) values across the regression methods for the initial test.

Variable Selection

If some of the explanatory variables have no (or very little) predictive capability, then their inclusion will increase the variance of the predictions but will not sufficiently decrease the bias. Thus, the importance of the variables must be determined. Several of the regression methods used produce measures of variable importance and the results from a subset of these methods are presented in Table 1. Using these variable importance results (giving greater weight to the better preforming methods), along with the correlation value between each explanatory variable and the output variable, the variables are classified into the following 4 groups of decreasing likelihood of their inclusion improving predictions:

- Group 1: X_4, X_{12}, X_{18}, X_{19}
- Group 2: X_2, X_6, X_{11}
- Group 3: X_1, X_5, X_{10}, X_{13}, X_{15}, X_{20}
- Group 4: X_3, X_7, X_8, X_9, X_{14}, X_{16}, X_{17}, X_{21}

Using 5 fold cross validation, the regression methods were tested with all combinations of the four variables in Group 1 (inclusion of any 1, 2, or 3, and of all 4 variables). Of these variable combinations, the models which include all four variables clearly had the best performance. Additionally, all methods had reduced mean 5 fold cross validation MSPE values when using only the Group 1 variables compared to using all variables (see Table 2). Thus, all four variables were concluded to be important. Next, for each of Groups 2, 3, and 4, all methods were tested using all possible combinations of variables in the given group added to the variables in Group 1. However, none of the new combinations clearly outperformed the model with only Group 1 variables. Since including additional variables increases variance, the input variables selected for the final model were only the Group 1 variables: X_4, X_{12}, X_{18}, and X_{19}.

Final Test of Regression Methods

Following the same procedure for the initial test (including re-tuning the Random Forest, BART, and NN models), the predictive performance of the regression methods is tested when the four selected variables are used. The root MSPE and relative root MSPE results are displayed in Figure 3. While both plots indicate Random Forests have the superior prediction ability among the methods, this fact is especially clear when looking at the relative root MSPE values as the Random Forest had the top (or near enough not to matter) performance in 8 of the 10 folds. This test was repeated two additional times using different sets of folds and produced similar results, indicating that the results are consistent and not due to unique characteristics from a single data split. Thus, Random Forests will use to produce the final predicted output values for the test set.

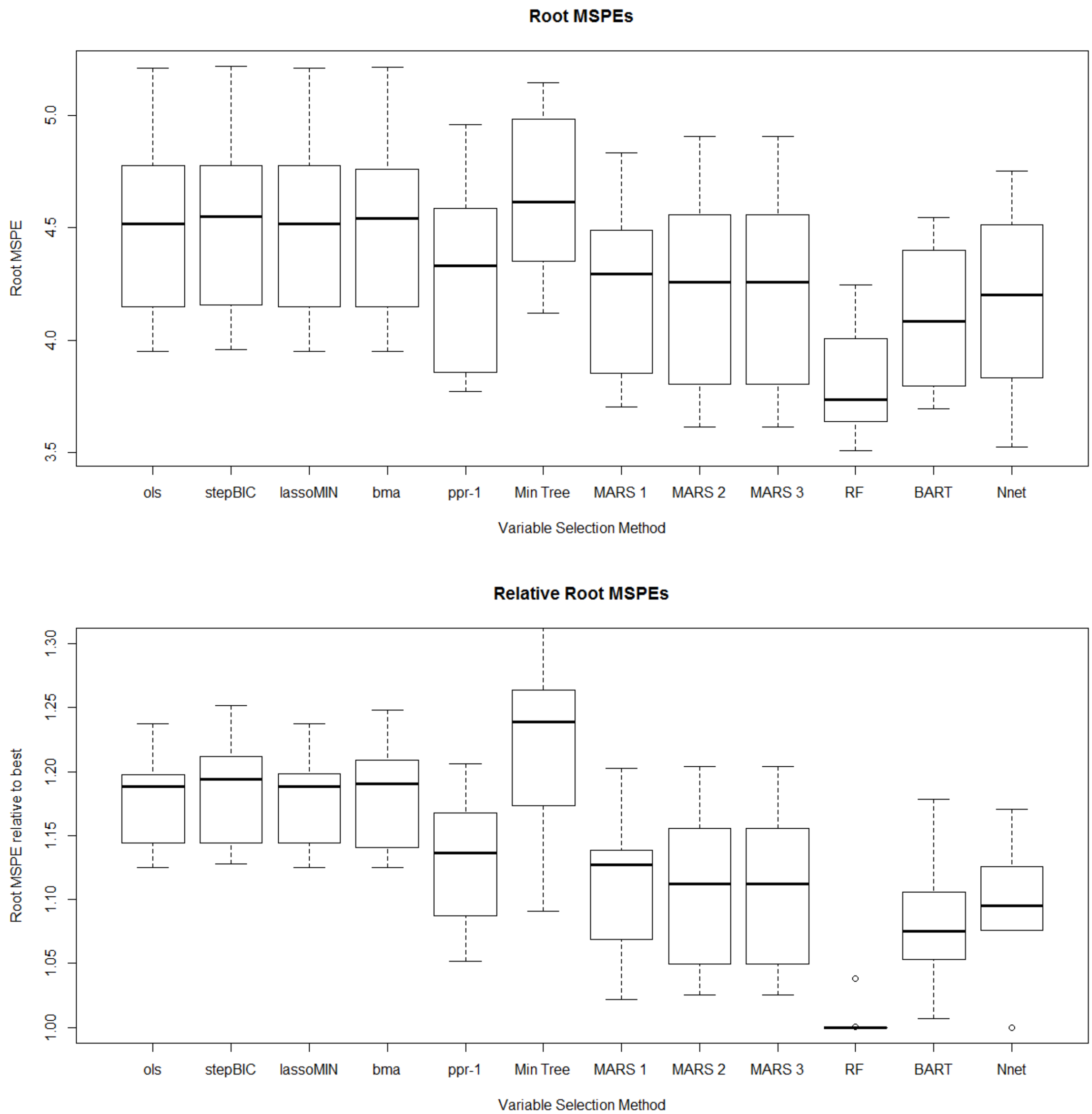

Figure 3: Root MSPE (top) and relative root MSPE (bottom) values across the regression methods for the final test.

Table 1: Variable importance results for a subset of regression methods.

	Stepwise (rank)	BMA (% of models)	PPR-1 (weights)	LASSO Min (coef)	MARS-1 (GCV)	RF (% Inc MSE)	BART (counts)
X1		1.8	0	0		7.42	4107
X2		2.2	0	0		14.50	4131
X3		2.3	0.11	0		1.41	4145
X4	2	100.0	-0.06	-0.15	17.5	32.20	7212
X5		14.5	0.09	0.15		4.08	5952
X6	4	61.7	0.02	0.07		17.31	4311
X7		4.9	-0.05	0		1.99	3706
X8		1.5	0.10	0		1.36	4018
X9		1.7	-0.01	0		0.17	4091
X10		2.2	-0.18	-0.08		1.35	3944
X11		3.6	0.03	0		9.11	5885
X12	1	100.0	-0.85	-1.96	100.0	84.53	20540
X13		11.1	0.01	0.01		5.09	5706
X14		1.5	-0.03	0		-0.03	4548
X15		6.4	0.09	0.02		4.89	3798
X16		2.6	-0.25	0		5.65	3374
X17		3.3	0.17	0		2.71	4344
X18	3	100.0	-0.11	-0.19	10.1	23.96	10104
X19	5	49.9	0.04	0.07	4.4	34.67	7863
X20		1.8	0	0		3.14	5494
X21		NA	NA	NA		4.43	NA
X21A		NA	-0.05	NA		NA	6196
X21B		NA	-0.15	NA		NA	4484
X21C		NA	-0.26	NA		NA	4769
X21D		NA	-0.05	NA		NA	4046

Predict Test Output

Using all 1000 observations of explanatory variables X_4, X_{12}, X_{18}, and X_{19} a Random Forest regression model was produced using 500 trees, a node size of 1, and 2 randomly selected regressors for each potential split. The out of bag (OOB) errors across different numbers of trees were compared and a clear trend demonstrated that 500 trees are a large enough number for the error to have stopped noticeably decreasing. The node size and the number of regressors at each split were selected in the tuning process. The Random Forest model was then used to predict the output values of the test set.

The distributions of the training set output and the predicted test set output were then compared, along with the relationships between the selected variables and output values in both the training and test sets. In all cases, the patterns were similar in both sets of data. 3D scatter plots of the output and all combinations of two of the variables of interest were also examined but no notable trends emerged.

Table 2: Improvement (reduction) in 5-fold cross validation MSPE values from models including all variables to models including only the variables in Group 1.

Method	**MSPE Improvement**
OLS	0.441
Stepwise BIC	0.470
LASSO Min	0.309
LASSO 1SE	0.179
BMA	0.331
PPR-1	0.573
PPR-2	1.252
PPR-3	2.473
Full Tree	2.112
Min Pruned Tree	2.033
1SE Pruned Tree	2.362
MARS-1	0.399
MARS-2	0.816
MARS-3	1.095
Random Forests	1.854
BART	1.428
Neural Nets	0.523

The correlations among the input variables were then examined. Only X_{12} and X_{18} had an absolute correlation value above 0.55 (with a value of 0.85). However, the relationship between the variables did not appear to be linear, and thus there should be unique information contained in each variable. Additionally, Table 3 displays the variable importance results for the final model and indicates that all four included variables made important contributions to the predictive ability of the model.

Table 3: Variable importance results for the final Random Forests regression model.

Variable	X_{12}	X_4	X_{18}	X_{19}
% Increase in MSE	50.22	39.61	34.32	28.54

Thus, overall the selected set of variables appeared reasonable, the predicted values aligned with the given output values, and the root MSPE results indicated that the Random Forests model did a reasonable job of prediction.

Results

The final Random Forest regression model presented in this report had the top predictive performance in the class. Additionally, after submission it was revealed that only 4 variables were truly meaningful to the prediction of the output. Those 4 variables were X_4, X_{12}, X_{18}, and X_{19}, the same 4 variables selected for the final model in this report.

Appendix: Variable Transformations

In the preliminary exploration, explanatory variables X_5, X_{13}, and X_{14} were found to have highly right skewed distributions which became approximately symmetric after a logarithmic transformation. Thus, the initial test of regression methods was repeated after performing the logarithmic transformation on the 3 variables of interest. This transformation led to improved root MSPE results and also increased the relative importance of variables X_5 and X_{13} (though not enough to move either into Group 1). Thus, it appeared as though using the logarithmic transformations would be the logical decision. However, when the variable selection procedure was repeated, the transformation did not affect the inclusion of the variables of interest, and the same 4 variables (X_4, X_{12}, X_{18}, and X_{19}) were selected for the final model. Thus, the transformations had no impact on the final results.