Competitive Modeling of Educational Achievement

Thomas Ball

August 2011

OBJECTIVE

Preliminary

A methodical and agnostic evaluation of the extent to which randomization impacts five approaches to variable selection and model building on different measures of predictive power in a holdout as well as the stability, quality and parsimony of the resulting models based on a database of high school student performance in their first year of college.

EXECUTIVE SUMMARY

Preliminary

Overall four of the five different approaches to data mining, variable selection and model building were roughly at parity with one another

• While there may be a statistically significant difference between a Brier Score of 0.167 and 0.170 especially based on the large sample size used in this study, the practical difference is not significant

• On the more qualitative factors reflected in metrics specific to the quality, stability and parsimony of the models developed some sharp distinctions did emerge

  • Stepwise regression produced models where 1 in 4 predictors were collinear and, relative to its competing benchmarks, were largely overfitted

In addition some commonly held myths from the marketing literature were debunked:

• Introducing collinearity into a model, while frequently boosting R^2, does nothing to improve predictive power in holdout samples

• This study provides no empirical evidence that adding polynomials, two-way or even three-way interactions to model specification boosts predictive power in a holdout above and beyond what had already been identified based on models using “main effects” only

BACKGROUND

Preliminary

Building on an urban secondary school system's database of characteristics of a high school graduating class and their subsequent college performance

- 16,480 graduating seniors who attended their 1st year of college at a local university
    - Leveraging a unique relationship and unique data
    - College success defined as the accumulation of 20+ credits (Y/N) in year one
- Comparisons based on two datasets: small and large numbers of predictors
    - Small: 57 "main effects" – high school performance and background factors, e.g., middle and high school test scores, student demographics, etc.
    - Large: 1,751 predictors --- all of the above plus their two-way interactions and 2nd degree polynomials
- Each approach will be evaluated using five randomly selected subsets of information
    - Note however that these are overlapping samples and not 5 completely independent samples of roughly 3,300 students each
        - A check of RELR's performance using 5 completely independent samples suggested that the issue was largely of academic concern

APPROACHES TO BE COMPARED

Preliminary

Approach	Description
RELR	Reduced Error Logistic Regression. A proprietary, patented SAS macro and approach developed by Rice Analytics that explicitly models error in its approach to variable selection in logistic regression, identifying the most probable model solution (PARSED).
RFLR	"Random Forests" logistic regression. An extension of Breiman's random forest for decision trees to logistic regression. Uses a two-stage methodology: stage 1, random jacknifing of variables and students to develop a "scorecard" or ranking of variable importance. Stage 2 is a hand-driven approach to drilling down to a final model by identifying the variables that are the most stable and significant.
LASSO	Least Absolute Shrinkage and Selection Operator. A constrained form of ordinary least squares regression where the sum of the absolute values of the regression coefficients is constrained to be smaller than a specified parameter.
LAR	Least Angle Regression. Like forward selection, the algorithm produces a sequence of regression models where one parameter is added at each step, terminating at the full least squares solution when all parameters have entered the model.
Stepwise	Classic stepwise logistic regression.

All analyses executed using SAS

KEY METRICS TO BE COMPARED

Preliminary

	Metric	Description
Predictive Power	Brier Score	A function that measures the holdout accuracy of a set of probabilistic assessments. Calculated as the average squared deviation between actual and predicted outcomes. A lower score is better.
	% Correct Predictions	A function that measures the holdout precision of a probabilistic model. Calculated as the sum of the percent of true negative and true positive predictions. A higher value is better.
Model Stability, Quality and Parsimony	% Variables that are Collinear	A measure of the extent to which a variable-selection method is able to condition the selection process for collinearity, defined as consistency between the sign of the parameter and the sign of the pair-wise correlation between the parameter and the DV. Zero collinear variables is ideal.
	Stability of Coefficients	A measure of the stability of parameter estimates across iterations. Parameters not in a model are plugged with zeros. Calculated as the average correlation between all possible pair-wise combinations of parameter estimates within the same method. A higher value is better.
	Stability of Variables Selected	An information-theoretic approach* to evaluating the extent to which an approach tends to select the same variables across the five random iterations. Calculated as the average (for all pair-wise combinations) of the sum of the entropy for solution A, the entropy for solution B minus two times the mutual information for the combination of A and B, where entropy is defined as $-\sum P(k)\log_2 P(k)$ and mutual information is defined as $\sum\sum P(k,k')\log_2 \frac{P(k,k')}{P(k)P'(k')}$ And 0 means complete instability and 1 is complete stability.
	Average # of Variables	A measure of model parsimony: Average number of variables selected per iteration.

* Marina Meila, Comparing Clusterings, University of Washington Statistics Technical Report 418 and COLT 03

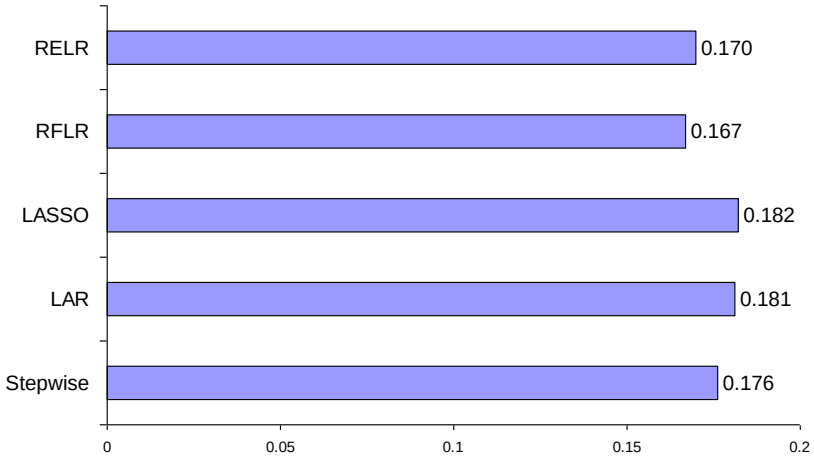
Preliminary

BRIER SCORE COMPARISONS

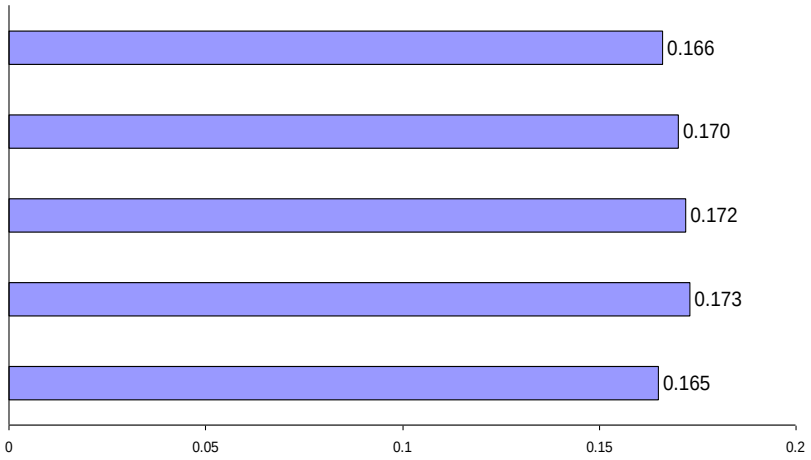
Holdout Brier scores do not reliably differentiate between methodologies as scores averaged about 0.170 overall with a tight distribution

- Directionally, LAR and LASSO performed best on both small and large variable sets

Brier Scores for Small Dataset



Brier Scores for Large Dataset

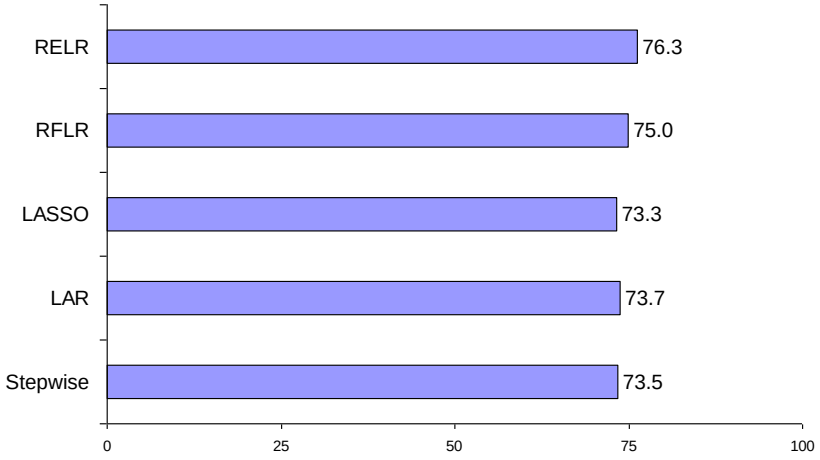


Preliminary

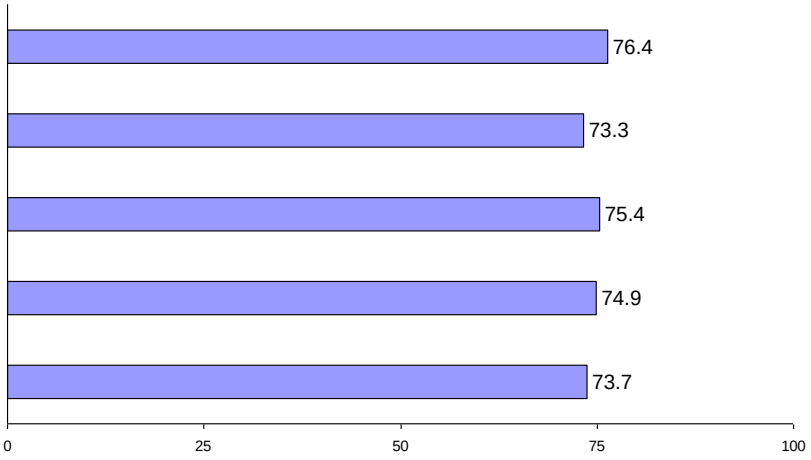
% CORRECT PREDICTION COMPARISONS

In terms of predictive accuracy, RELR holds the top ranking but, from a practical point of view, RELR does not dominate or “own” this metric as other approaches achieve near equivalence

% Correct Predictions for Small Dataset



% Correct Predictions for Large Dataset



% OF VARIABLES THAT ARE COLLINEAR COMPARISONS

Preliminary

Only one technique allowed any collinearity in the variable selection process (defined by a comparison between the pair-wise correlation of a predictor with the dependent variable and the sign of its estimate in the model), and that was Stepwise regression

- Stepwise allowed roughly 1 in 4 variables as collinear

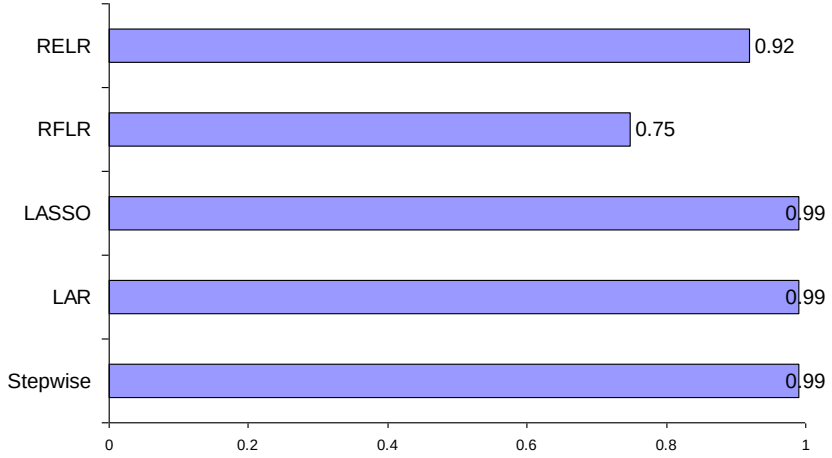
Preliminary

STABILITY OF COEFFICIENT COMPARISONS

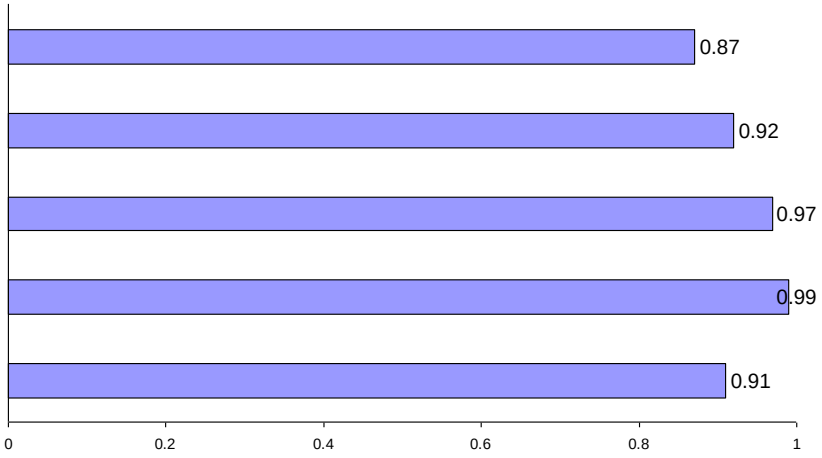
In general, all methodologies delivered equivalent levels of stability in terms of parameter estimates based on their average pairwise correlations across the 5 iterations.

• RFLR (“Random Forests” Logistic Regression) had a shortfall in its stability when dealing with a small dataset but more than compensated for this when fielding a much larger number of candidate predictors

Stability of Coefficients for Small Dataset



Stability of Coefficients for Large Dataset



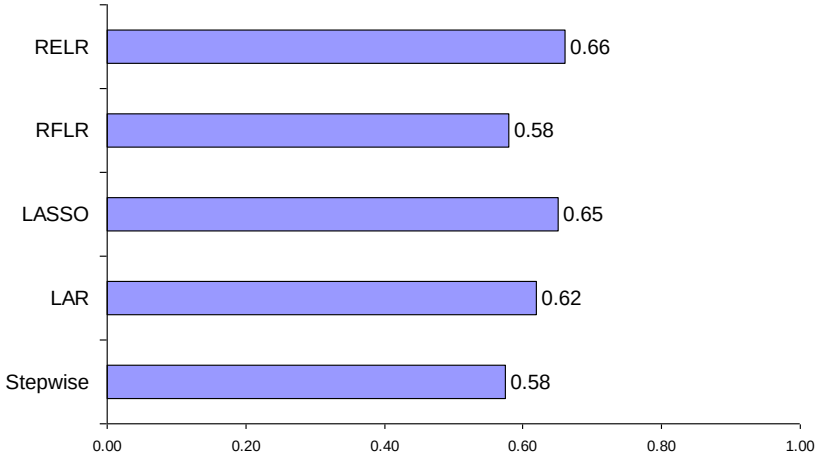
STABILITY OF VARIABLES SELECTED COMPARISONS

Preliminary

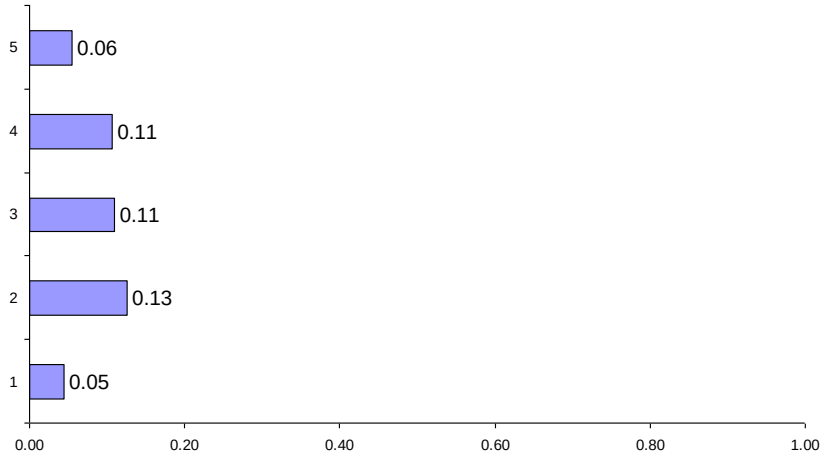
By far the most controversial metric in this analysis is the information-theoretic measure of stability in variables selected. A wide divergence is to be noted in the values between the small and large number of variable samples. This is almost entirely due to the inclusion of a huge percentage of variables (in the large sample) that are never selected in one of the 5 iterations of the models

- Treating each dataset as norm-referenced, no striking findings emerge

Stability of Variables Selected for Small Dataset



Stability of Variables Selected for Large Dataset



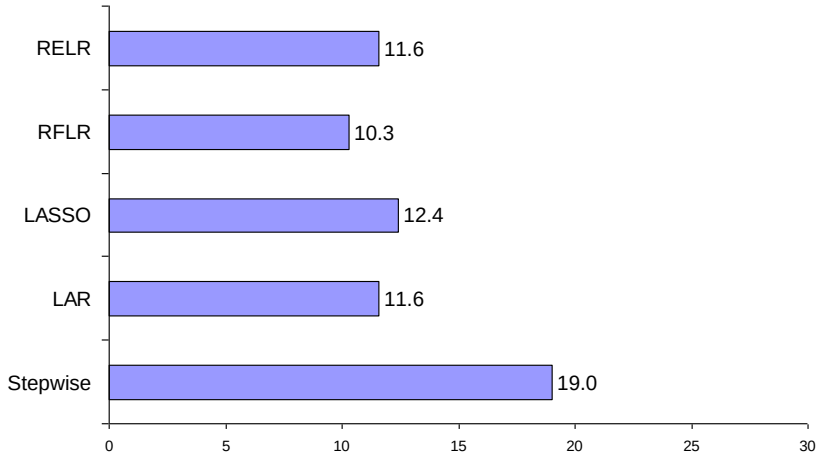
AVERAGE NUMBER OF VARIABLES SELECTED COMPARISONS

Preliminary

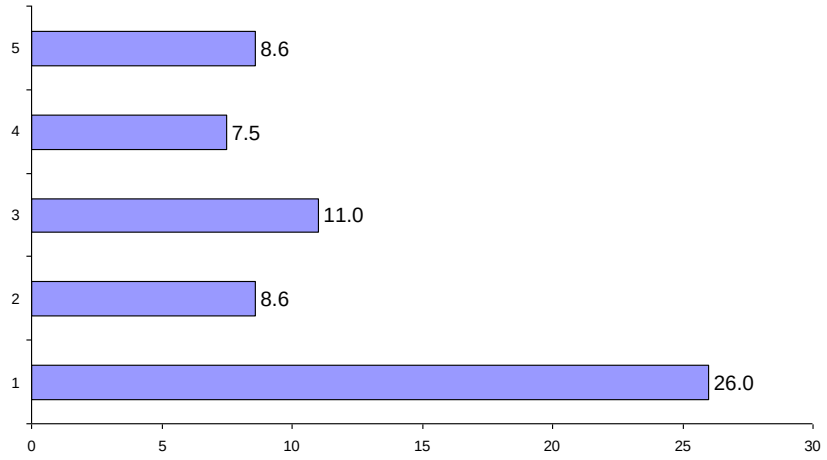
Model parsimony is an important construct and, here again, the only contrast is between stepwise regression and all other approaches.

• Stepwise regression overfits, relative to the other methodologies, by nearly double the number of variables required

Average Number of Variables Selected for Small Dataset



Average Number of Variables Selected for Large Dataset



APPENDIX

Preliminary

Model Comparisons

Averages Across 5 Iterations per Model

Small Data Models, # Vars=57	Brier Score	% Correct Predictions	% of Variables That Are Collinear	Stability of Coefficients	Stability of Variables Selected	Average Number of Variables Per Model
Stepwise	0.176	73.5	26	0.99	0.58	19.0
LAR	0.181	73.7	0	0.99	0.62	11.6
LASSO	0.182	73.3	0	0.99	0.65	12.4
RFLR	0.167	75.0	0	0.75	0.58	10.3
RELR	0.170	76.3	0	0.92	0.66	11.6

Large Data Models, # Vars=1,751	Brier Score	% Correct Predictions	% of Variables That Are Collinear	Stability of Coefficients	Stability of Variables Selected	Average Number of Variables Per Model
Stepwise	0.165	73.7	28	0.91	0.05	26.0
LAR	0.173	74.9	0	0.99	0.13	8.6
LASSO	0.172	75.4	0	0.97	0.11	11.0
RFLR	0.170	73.3	0	0.92	0.11	7.5
RELR	0.166	76.4	0	0.87	0.06	8.6