Analyzing US News & World Report Rankings Using Symbolic Regression

Version 2

I' ve been using a relatively new statistical technique called symbolic regression to analyze the most recent US News rankings. I thought the results would be of interest to this group. Here is the bullet point summary

• One can predict the absolute US News score for 2007 using only four factors, a simple polynomial expression, and obtain an r-squared statistic of 97.7 %. In the social sciences, this is considered a very high level of prediction. One does not need to perform any of the "normalizations" employed by US News in order to obtain this result. Thus, one can predict fairly well how a change in a variable used by US News would change one's own US News score without knowing data on any other school or, indeed, without having complete data on one's own school.

• The formula is set forth below. Employed9 means employment 9 months after graduation, LSAT75 means the 75th percentile LSAT score, Peer means the peer rating, and UGPA75 means the 75th percentile UGPA score.

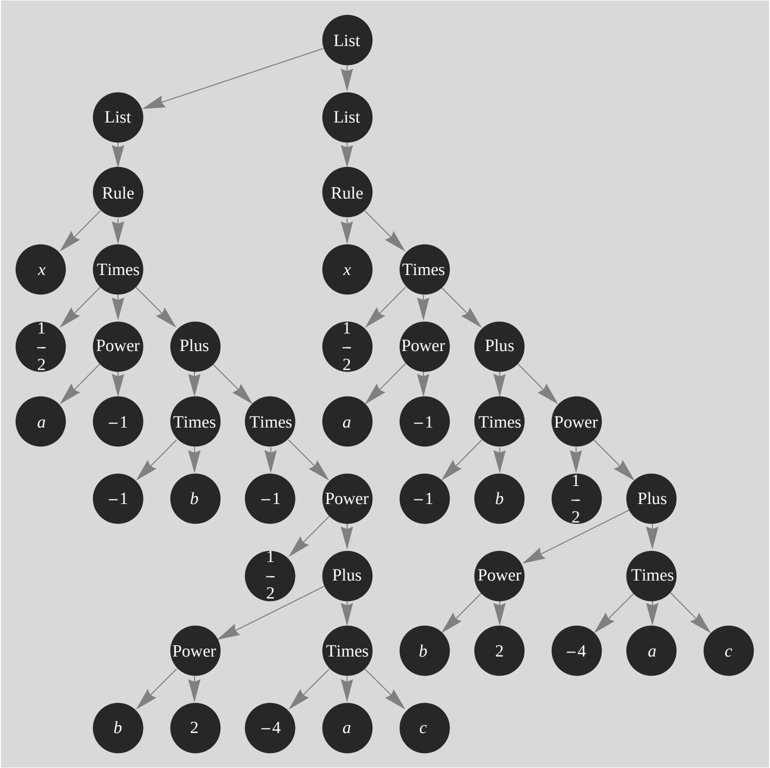
$$\mathbf{8.3214 + 0.04139\,\text{Peer}\,(1.1487\,\text{Employed9} - \text{SFRatio})\,\text{UGPA75}}$$

• If one is allowed a less parsimonious model, one can predict absolute US News score for 2007, with an r-square of over 99% using the following seven factor model.

$$\begin{aligned}&2.71333\times 10^{-6}\,\text{Employed9}\,\text{LSATMedian}\,\text{Peer}\\&\quad(10.573\,\text{Lawyers} + \text{LSAT75} - 16.8212\,\text{Peer} - 0.839826\,\text{SFRatio} - 37.0193)\,\text{UGPA75} + 6.14581\end{aligned}$$

• Of the published US News factors, the ones that most frequently appear along the "Pareto Front" of models that are both good predictors and parsimonious are Peer Rating (100% of Pareto Front models); Employment at 9 months (89.5%), 75th percentile UGPA (79%), student/faculty ratio (79%), 75th percentile LSAT score (74%), lawyer/judge rating (58%), and median LSAT score (53%). No other published factor appears in more than 50% of the models. By Pareto Front, I mean that there is no alternative model that is both more predictive and more parsimonious.

This is not the place for an extensive discussion of symbolic regression but the idea is to construct an evolutionary process that permits algebraic expressions to have sex and figure out which offspring are most fit. Fitness is determined by a mixture of predictive ability (r-squared) and parsimony (incredibly baroque mathematical expressions are disfavored). The technique is arguably (strongly, in my opinion) superior to traditional statistical techniques in the social sciences where there is little basis for imposing some most-likely linear structure on the regression model. Instead, one lets today's powerful personal computers explore a far vaster space of potential models. The technique is possible because mathematical expressions can be represented as mathematical "trees" and these trees can both mutate and interchange branches. By way of example, here's the tree structure of the quadratic formula: $\left\{\left\{x \to \frac{-b-\sqrt{b^2-4ac}}{2a}\right\}, \left\{x \to \frac{\sqrt{b^2-4ac}-b}{2a}\right\}\right\}$.



A major caveat. As you all know, correlation is not the same thing as causation. It may be, for example, that something extrinsic to the model is driving the supposedly causal variables in a certain way that makes it look, absent information on the real cause, that it is the supposedly causal variables doing the work. Still, it is kind of interesting seeing what tends to predict well.

Another interesting factoid. In theory, if one know the marginal cost of increasing employment at nine months, the marginal cost of increasing one' s 75 th percentile LSAT and UGPA scores and the marginal cost of increasing one's Peer rating, one can then create a constrained optimization problem and determine an optimal dollar allocation strategy for improving one's US News score. Also notice that the multiplicative structure of the first formula I set out makes the constrained optimization problem look a lot like classical microeconomic optimization problems, with predictable results.