Final Project

(student name withheld)

DAT 520

8/24/14

A decision tree was conducted using R statistical software using the Rattle package in R. The results show that four years of high school math are a significant predictor of placement in college level math upon entry into college.

Moran (2008), conducted a causal-comparative study using chi-squared analysis and logistic regression. Moran found that students taking four or more years of HS math and who also took advanced/honors courses within those 4 years were five times more likely to place in non-developmental math than their peers who did not take such courses. Using a decision analysis approach, we can take a more direct examination of first year college students’ progression from high school into their first college math course.

The data utilized in the present study are from 82 students who transitioned from a local high school into a small local community college. The data were part of a pilot study designed to take a first look at the possible impact of moving to a 4-year math requirement for students in high school in the state of Arizona. Although the original data set had demographic data, no demographic information is provided with the present data and the data set has been de-identified. The high school and college names are also withheld. All computations were done in R (The R Project for Statistical Computing, 2014), with the addition of the Rattle package (Williams, 2009).

The variables in the present study are:

* HSyrs: categorical coded 1-4 for the number of years of high school math; “dependent”variable in the model
* HSgpa: uncoded continuous variable; the student’s high school GPA, used as a weight
* CCgrade: categorical, coded A-F; the student’s grade on their first college math course (no D grades in this sample)
* CCmath: categorical, the student’s first assigned undergraduate college math course coded to remedial level, college level, or advanced

From the top down (starting with node 1, the root node), the decision tree shows that of the students in the sample (at the root node on top of **Figure 1**., also see **Tables 1 and 2**): 49% had 4 years of math, 24% had 2 years, another 24% had 3 years, and 3% had 1 year of math. The root node was split on the students’ placement in either remedial level math in college or regular to advanced college math with 32% going into college math and 68% placed in remedial math. Node 3 which branches to the right of the top/root node is a terminal node and shows that the probability of a student with 4 years of math placing at college level is .92, whereas those with less than 4 years of math have a .04 or less probability of placing at a college level. Node 3 holds 16 students.

Node two is not a terminal node but splits the students on their grades in their first college math course, which can only be a remedial math course since this branch contains only students placed in remedial math. There were five grades possible A, B, C, D, and F, but no students received a D grade so the grades were effectively A,B,C,F. Students with 3 years of high school math (node 5, terminal node) have a .49 probability of obtaining an A in remedial math and those with 2 years of math have a .21 probability of an A. Interestingly, those with 4 years of math in high school who ended up in remedial math had a lower probability of an A than those with 3 years of math. This could be due to the fact that some students took 4 years of low level math, not because they like math, but rather because they struggled with math. Those students with 2 or fewer years of high school math tended to earn a lower grades. Students with 2 years of high school math had a probability of .45 of earning a B or lower in remedial math at the college.

The outcome is a bit complex as the probabilities seem to indicate. Students with 3 years of math actually have a higher probability of an A grade in remedial math in node 5 than students with 4 years of math. This may be due to the inconsistent nature of years of math in high school to actual level of math achieved. Three years of higher level math for a student with high math aptitude can allow the student to outpace a student with 4 years of math who has struggled through those three years and largely taken remedial high school math courses. In the present study, no attempt was made to censor or filter out students who had 4 years of math but who failed to achieve college preparatory levels of high school math courses.

What is plain to see is that 4 years of math does make a difference. However the model is far from perfect. There is a high root node error (1.0974), which means that there is a strong likelihood of misclassification. Looking at the error matrices (**Tables 3 and 4**), the model tends to under-predict the number of years of high school required, meaning that some students had more years of high school math than the model would have predicted. This again, is likely due to those students taking lower level courses for more years being mixed in with students taking 4 years of progressively more difficult math terminating at a higher level. Given more time, and a larger sample it would be interesting to further delineate the population by collapsing categories to look at differences in students who have both 4 years of higher level math versus those who have 4 years of lower level math and the different combinations of years and achievement level in order to obtain a less convoluted picture.

**TABLE 1.** **Summary of the Decision Tree model for Classification** (built using 'rpart'):

n= 57

node), split, n, loss, yval, (yprob)

\* denotes terminal node

1) root 57 62.55 4 (0.02559548 0.23980622 0.23964473 0.49495357)

2) CCMath=Remedial 41 56.15 3 (0.03764846 0.33289786 0.33313539 0.29631829)

4) CCgrade=B,C,F 25 24.33 2 (0.07227542 0.44528044 0.18832649 0.29411765) \*

5) CCgrade=A 16 20.55 3 (0.00000000 0.21070897 0.49058007 0.29871096) \*

3) CCMath=Advanced,College\_lvl 16 3.30 4 (0.00000000 0.04211854 0.04110971 0.91677175) \*

Classification tree:

rpart(formula = HSyrs ~ ., data = crs$dataset[crs$train, c(crs$input,

crs$target)], weights = (crs$dataset$HSgpa)[crs$train], method = "class",

parms = list(split = "information"), control = rpart.control(usesurrogate = 0,

maxsurrogate = 0))

Variables actually used in tree construction:

[1] CCgrade CCMath

Root node error: 62.55/57 = 1.0974

n= 57

CP nsplit rel error xerror xstd

1 0.11487 0 1.00000 1.0000 0.088955

2 0.01000 2 0.77026 1.0056 0.088948

Time taken: 0.01 secs

**Table 2.** **Tree as rules**:

Rule number: 3 [HSyrs=4 cover=16 (28%) prob=36.35]

CCMath=Advanced,College\_lvl

Rule number: 4 [HSyrs=2 cover=25 (44%) prob=12.90]

CCMath=Remedial

CCgrade=B,C,F

Rule number: 5 [HSyrs=3 cover=16 (28%) prob=12.05]

CCMath=Remedial

CCgrade=A

**TABLE 3**. Error matrix for the Decision Tree model on SEM.2 (counts):

Predicted

Actual 1 2 3 4

1 0 4 0 0

2 0 14 4 1

3 0 6 10 3

4 0 12 9 19

**TABLE 4.** Error matrix for the Decision Tree model on SEM.2 (%):

Predicted

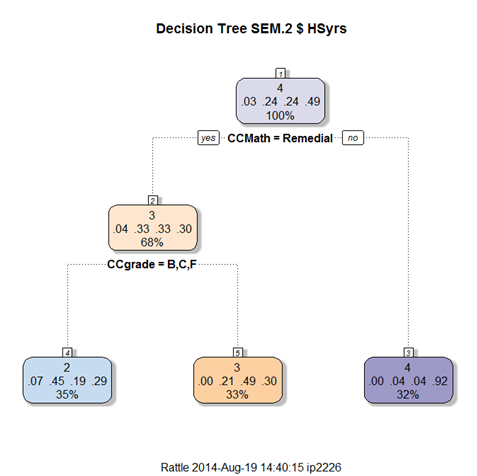
Actual 1 2 3 4

1 0 5 0 0

2 0 17 5 1

3 0 7 12 4

4 0 15 11 23



**Figure 1.** The model output from R using Rattle data mining package which implements the Rpart package with additional graphics from the Lattice package. Within each box is the number of years of math (at the top), the probabilities of each different number of years of high school math at that node (middle row), and the proportion of the total prior sample that the node contains (bottom percentage).

**References**:

Williams, G. (2009). Rattle: A Data Mining GUI for R, Graham J Williams, The R Journal, 1(2):45-55.

Moran, K. (2008). Examining the mathematical preparedness of first year college students entering college directly from traditional high schools in Connecticut. Research & Teaching in Developmental Education, 25(1), 30-44.

The R Project for Statistical Computing (2014). What is R? Retrieved from: <http://www.r-project.org/index.html>.