MAC 2233 Spring 2014 Predictive Modeling

Stephen Ippolito

8/21/2014

Executive Summary

The primary purpose of this report is to formulate a classification model for determining whether a student is likely to pass MAC 2233 or if the student is likely to not pass MAC 2233 falling into the DFW category. The data used in the study was taken from the spring 2014 Class of MAC 2233 with a total of 1087 observations. The sources of the data were FAU's Dartboard, FAU Banner and several data requests from FAU's OIT personal. The DFW rate for the entire class was 43% with a passing rate of 57%. For classification purposes a logistic regression was used which consisted of regressing whether a student passed MAC 2233 against whether they passed their midterm, whether they took a prerequisite class at FAU, if they passed their most recent prerequisite class, and how many visits they student spent in the Math Learning Center (MLC). Model construction consisted of using 10-fold cross validation which gives an unbiased estimator of the expected classification error. The estimated model accuracy was 72%, the Negative Predictive value for the model (the probability that a student passed given that the model said they would pass) is 75% while The Positive Predictive Value (the probability that a student failed given that the model said they would fail) was only 66%. The negative predictive value maybe useful for classifying students, after midterm reporting, that are not at risk of failing allowing attention to fall on a smaller group of students. While considering the predictive accuracy it should be noted that 724 of the student had never visited the MLC and only 645 students had recorded Midterm grades. These two variables are highly significant predictors even with these deficiencies, so with improved reporting and MLC attendance we may expect a more accurate report. It should be noted that the selected variables may only be highlighting correlative relationships for student success rather than demonstrating a cause and effect relationship. For the chosen predictors the one which can be dealt with most proactively is the numbers of hours the student has spent in the MLC which has a log odds coefficient of approximately 0.06. The variable INSTRUCTORS was a highly significant variable not used in the model since instructors change from semester to semester. However, what is interesting about this variable is variation between instructors which should be further studied. As a final note a student was considered to pass if they received a "C-"or better. All other grades including withdraws and drops ("W","WM" and "ZR") were included in DFW calculations.

Descriptive Statistics and Variable Definitions.

The variables used in the study and their definitions are given in Table 1. Each of the variables used in the study will be considered for descriptive statistics along with some others that maybe of interest. The preceding subsections will be titled based on the variable they are describing. The names and definitions of the variables are also listed in Table 1.

FINAL	The student's final grade in Spring 2014 MAC 2233
PASS	Did the student Pass MAC 2233? Yes/No
INSTRUCTORS	Instructors who taught MAC 2233 in Spring 2014
MIDTERM_PASS	Did the student receive a passing grade on their midterm report? Yes/No/NA
MIDTERM_GRADE	The student's grade on the midterm report
PREREC_FAU	Did the student take a prerequisite class at FAU? Yes/No
PREREC_OTHER	Did the student take a prerequisite class at another university? Yes/No
OVERRIDES	Did the student receive an override? Yes/No
INSTITUTION	What Institution did the student take the prerequisite at most recently?
PREREC_TAKEN	How many months before the start of the class was the prerequisite taken? At most 5 months. Greater than 5 months and at most 1 year? Greater than 1 year and at most 2 years. Greater than 2 years.
PREREC_PASSED	Did the student pass their most recent prerequisite class? Yes/No/NA
PREREC_GRADE	What Grade did the student receive in their most recent prerequisite class?
VALID_ALEKS	Did the student receive a score of 40 or better on an Aleks test within 5 months of the start of the class? Yes/NO
MLC_HOURS	How many hours did the student spend in the Math Learning Center (MLC) over the course of the Spring 2014 semester?
VISITS	How many visits did the student make to the MLC during the semester?
MLC_HOUR_CATEGORY	Categorical Versions of MLC_HOURS. Lower class limits from 0 to 72 hours with a class width of 6 hours.
Table 1: Variable names (left) and definitions (right) which are used in the study.	

FINAL

The variable FINAL contains the grades for individual students at the end of the course. The distribution of the grades is given in Figure 1 below.

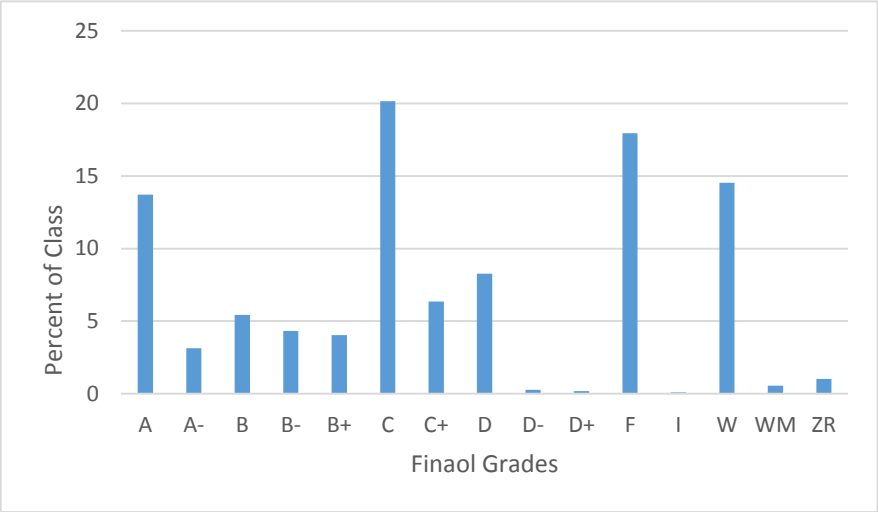


Figure 1: Final grade distribution for MAC 2233 Spring 2014. Grades are on the horizontal axis and the percent of the class that received a grade is on the vertical axis. WM means the student withdrew from the class and ZR means the student dropped before the add drop period.

PASS

The variable PASS indicated whether or not a student received a passing grade in the course or if they were included in DFW statistics. Figure 2 displays the results.

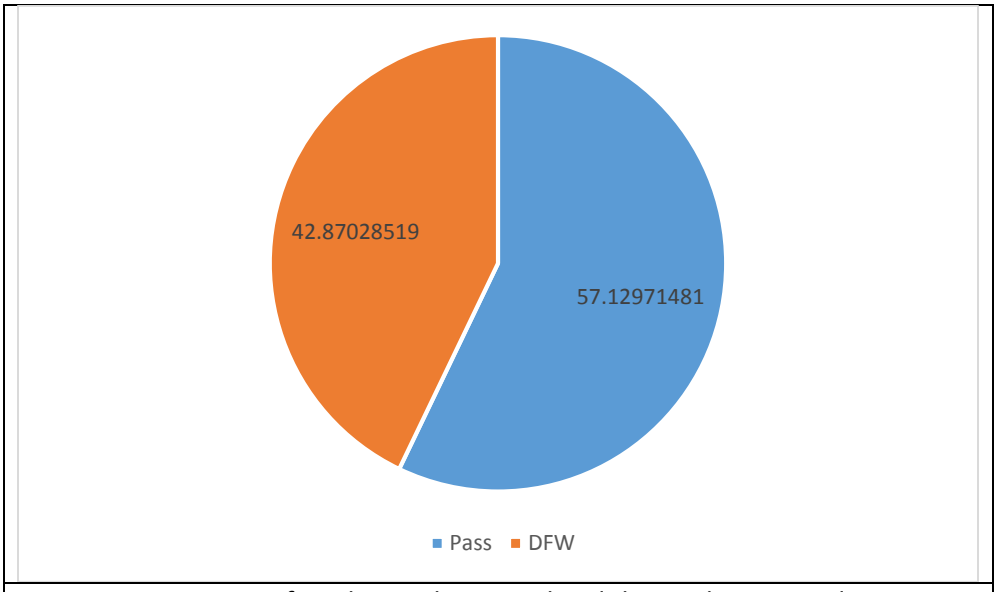


Figure 2: Percentage of Students who passed and those who received a DFW.

INSTRUCTORS

The variable INSTRUCTORS indicates the instructor that a given student had taken Springs 2014 MAC 2233 with. Figure 3 below displays the odds ratio of passing to not passing for each instructor. The

variable INSTRUCTOR was highly significant for prediction however instructors maybe completely different from semester to semester so this variable was not included in the final model.

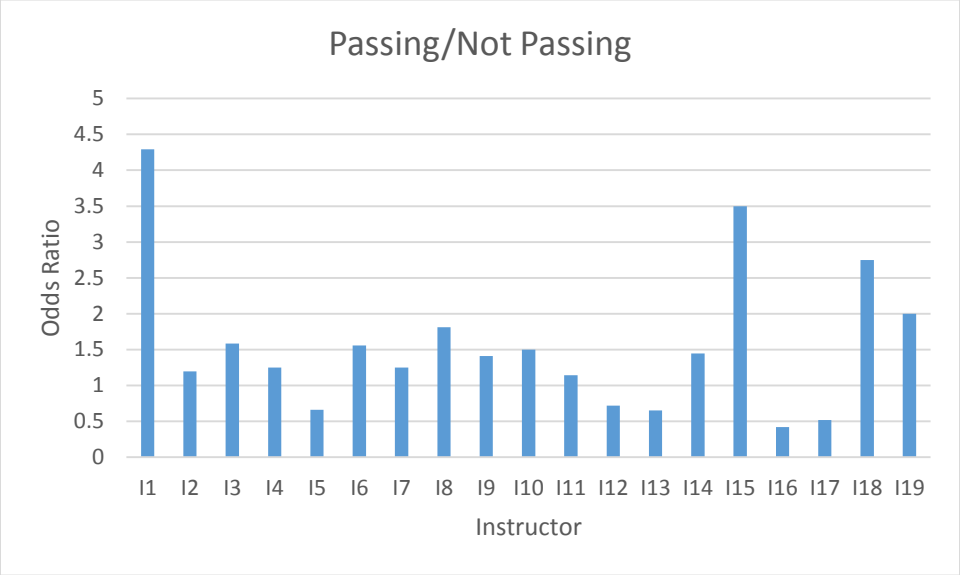


Figure 3: Instructor Names have been replaced by I1-I9 are on the horizontal axis, and the odds ratio of passing to not passing for a given instructor is given on the vertical axis.

.

In the Figure 3 there is a noticeable amount of variation. Consequently this phenomena was investigated further. It was considered that some classes might have better prepared students, for instance in Figure 4 below students in the class for instructor I1 had an odds of passing about 4 times that of instructor I5. We also notice from Figure 4 however, that for students who passed their most resent prerequisite class, students from I1 had an odds of getting a B- or better of about 3 while student from I1 had an odds for getting a B- or better of about 2. This may explain differences for some instructors, however, overall instructors there is no clear pattern between performance in prerequisite and whether the student passed MAC 2233 as seen from Figure 4.

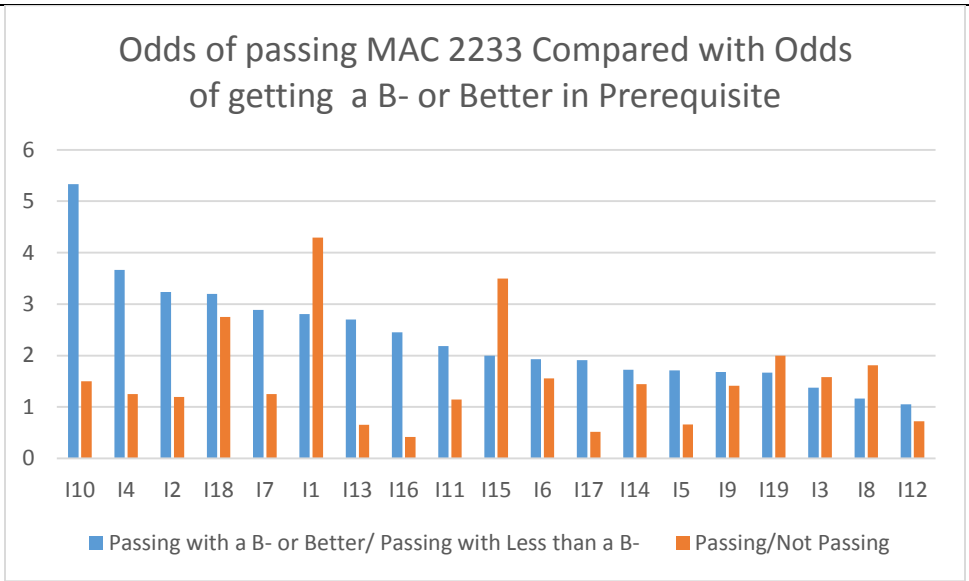


Figure 4: The Blue represents the Odds of Passing with a B- or better in the Prerequisite Class Divided

MIDTERM_PASS

The variable MIDTERM_PASS indicates whether a student received a passing grade, a DFW, or whether a student did not receive a midterm grade.

	Fail Midterm	Pass Midterm	NA
Fail MAC 2233	72	12	44
Pass MAC 2233	28	88	56

Table 2: Conditional probabilities for passing or failing (DFW) MAC 2233 based on midterm grade reporting with the conditionals listed in the columns. NA indicates no reporting for those students.

Concerning group "NA", we would expect that the 442 students whom did receive midterm reporting grades to be a randomly selected group, which seems to be the case as the probability that one of these students receives a DFW is in two percentage points of the class DFW rate, and the probability that one of these students passes is one percentage point of the class passing rate (see Figure 2). It should also be clear from these statistics that eliminating the group with NA's should highly improve the predictability in the final model.

MIDTERM_GRADE

The variable MIDTERM_GRADE is the grade a student received for their midterm report. If we eliminate students that did not receive a midterm grade (about 41% of the class) and those students that received a Final Grade of "I,WM,W" or "ZR" then we can compare the distributions of the final and the midterm grades as in Figure 5. What we notice from the figure is that for final grades 68% of the students passed

while for the midterm 49% passed. Consequently, midterm reporting appears to underestimate passing and overestimate failures which makes it a conservative estimator of the DFW.

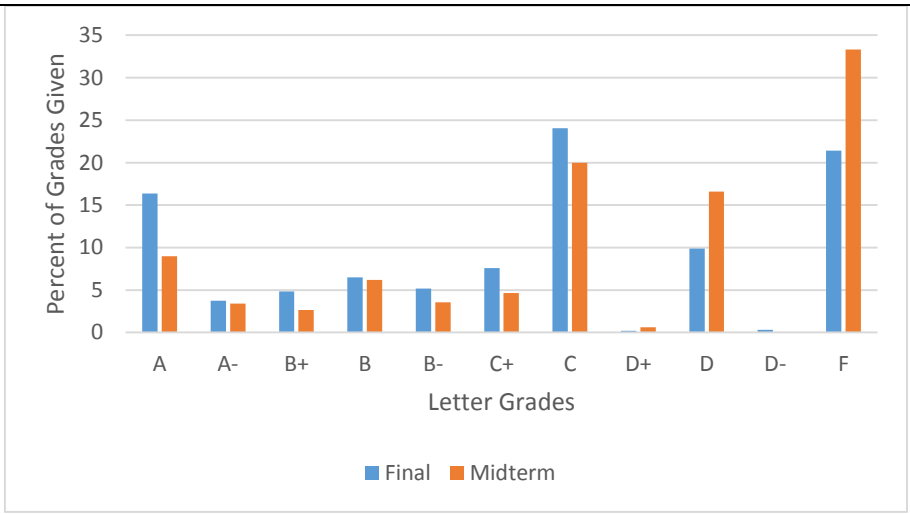


Figure 5: Excluding Students that did not receive a midterm grade, and those student that received a final grade of "W,WM,I,ZR" the percent of total grades given is displayed for both final and midterm reporting.

OVERRIDES

A total of 60 students were given overrides in the class with a passing rate of 67% being slightly better than the class as a whole which had a passing rate of about 57% (see Figure 2).

Over Ride	% Passed
NO	57
YES	67

Table 3: Pass rates, as percentages, for students who had received overrides.

INSTITUTION

The variables INSTITUTION consists of where students took their most recent prerequisite class. Either at FAU or another institution. Most universities were grouped into the category "other" since counts were very low. For model testing an additional category was added called "none" which indicated students that never took the prerequisite. The category "none" had a total of 114 students 71 of whom passed yielding a pass rate of 62%. Table 4 below however extends this variable to include information on Palm Beach State and Broward College. The reason Palm Beach and Broward were not included as levels of this variable for the final model is their presence did not improve model accuracy so it is better to take a simpler model.

University	Number Passed	Total Students	Percent Passed
BROWARD COLLEGE	54	104	52
FAU	364	588	62
other	71	140	51
PALM BEACH STATE COLLEGE	61	141	43
Total	550	973	57

Table 4: Passing Rates for students based on where they took their most recent prerequisite.

To have an idea how many of these students have passed their prerequisites

	Broward	FAU	Palm Beach State	Other
Failed Prerequisite	8	26	3	17
Passed Prerequisite	96	562	138	123
Percent Passed	92	96	97	87

Table 5: Counts of students that passed and failed MAC 2233 along with percent pass rate. The results are separated by the most popular universities.

PREREC_PASSED

PREREC_PASSED is the percentage of students that had passed their prerequisite.

	Failed MAC 2233	Passed MAC 2233	Passing Rate
Failed Prerequisite	37	17	31
Passed Prerequisite	386	533	60
Total	423	550	57

Table 6: Student who had passed a prerequisite course tabulated against whether they passed MAC 2233. The first two columns contain counts of students and the third column “Passing Rate” has the percent of students that passed in that row.

PREREC_FAU

PREREC_FAU indicates if a student had taken a prerequisite at FAU, not necessarily that they passed only at some point that they were enrolled in a prerequisite class at FAU.

	Failed MAC 2233	Passed MAC 2233	Passing Rate
No Prerequisite FAU	235	256	52
Prerequisite at FAU	231	365	61
Total	466	621	57

Table 7: Students who took a Prerequisite class at FAU, though not necessarily passed, cross tabulated against whether they passed MAC 2233. The first two columns contain counts, and the third column labeled "Passing Rate" has the percent of students that passed in certain row.

VISITS

For each student the variable counts the number of visits a student made to the MLC. A plot of passing rate vs time spent in the MLC is given in Figure 6 below. In the figure there seems to be a change in the average passing rate for students that spent more than 10 visits versus those who spent at most 10 visits in the MLC. Investigating this we find that for those who spent more than 10 visits in the MLC the passing rate was 78% while for those who spent at most 10 visits had a passing rate of 55%. It should be noted that the number of students with at least one visit was 364 meaning that 723 students from the class had never been to the MLC.

Total Observations	Min	1st Quartile	Median	Mean	Standard Deviation	3rd Quartile	Max	Passing Rate %
77	11	13	17	21	11	27	74	78

Table 8: Summary Statistics for Students with more than 10 visits.

Total Observations	Min	1st Quartile	Median	Mean	Standard Deviation	3rd Quartile	Max	Passing Rate %
1010	0	0	0	1	1.9	1	10	55

Table 9: Summary Statistics for Students with at most 10 visits.

Total Observations	Min	1st Quartile	Median	Mean	Standard Deviation	3rd Quartile	Max	Passing Rate %
1087	0	0	0	2.278	6.2	1	74	57

Table 10: Summary Statistics for the variable VISITS

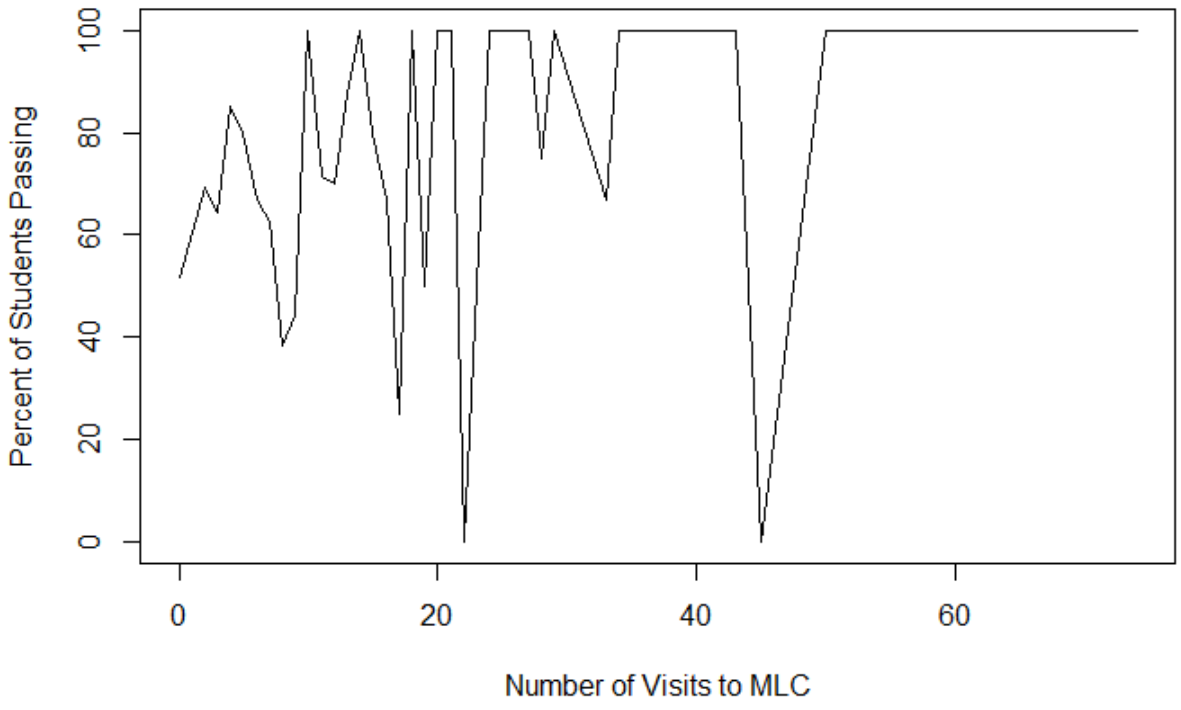


Figure 6: Percent of students passing versus the number of visits to the MLC.

Model Selection

Model selection first proceeded by running a cluster analysis on the set of variables used in this study. This resulted in Figure 7 below. Each of the variables listed in the diagram was regressed against whether or not a student passed the class MAC 2233. Then from each of the lowest level clusters in Figure 7, the predictor most significantly associated with the dependent variable "Pass" was selected. The resulting model was then reduced using stepwise regression to minimize the AIC. The resulting model is given by

$$\log\left(\frac{Passing\ MAC\ 2233}{Not\ Passing\ MAC\ 2233}\right) = \beta_1 MIDTEM\ PASS + \beta_2 PREREC\ FAU + \beta_3 PREREC\ PASSED + \beta_4 VISITS + INTERCEPT$$

Where each of the β_i is a vector with entries corresponding to the levels of that variable. The results are given in Table 11 below.

Vector	Variable	Level	Estimate	Standard Error	p-value
$INTERCEPT$	INTERCEPT	NA	-2.01908	0.35266	0
β_1	MIDTERM PASS	PASS	2.91689	0.21641	0
	MIDTERM PASS	NO REPORTING	1.18367	0.15896	0
β_2	PREREC FAU	YES	0.34714	0.15236	0.022
β_3	PREREC PASS	YES	0.80509	0.33508	0.016
	PREREC PASS	NO PREREC TAKEN	1.03202	0.40345	0.010
β_4	VISITS	NA	0.05781	0.01608	0.000326

Table 11: Coefficient Summary for logistics model.

Using 10-fold cross validation, the estimated model accuracy was 72%, the Negative Predictive value for the model (the probability that a student passed given that the model said they would pass) is 75% while The Positive Predictive Value (the probability that a student failed given that the model said they would fail) was only 66%.

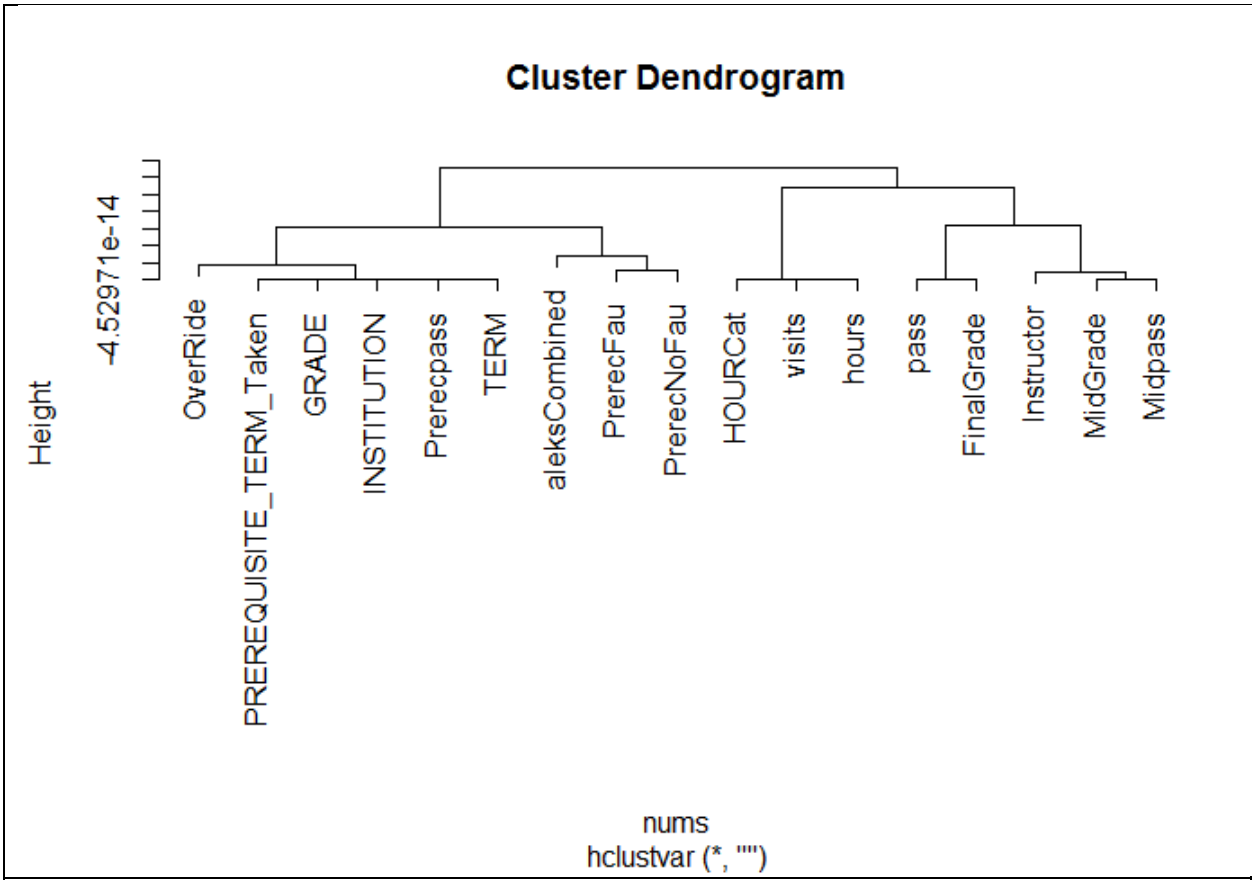


Figure 7: Variable Clustering Diagram

Conclusion

Students that have taken a prerequisite at FAU and have passed their most recent prerequisite tend to do better than those who do not, in. Looking at Table 4 we note that student that have taken their most recent prerequisite at FAU tend to do better than those who took their most recent prerequisite at other universities. Professors giving overrides to students seem to be doing a good job with a pass rate 10% higher than the class as a whole, however, the counts are too low to make a significant impact (see Table 3). Midterm Grades are very significant predictors, and prediction should be greatly improved if the number of students receiving midterm grades is increased, also noting that midterm grades are a conservative estimate of final grades (see Table 2 and Figure 4). The number of visits to the MLC provides a strong indicator of student success as those students taking more than 10 visits per semester (average 20 visits) have a passing rate of 78%.

Using 10-fold Cross Validation the final model was found to be 72% accurate overall with a 75% success rate of predicting passing students and a 66% success rate of predicting failing students. The 10 fold Cross Validation divides the data set into 10 equal parts and tests each part against a model trained on the other 9 parts. The results from each trial are then averaged, yielding an unbiased estimator of the expected classification error. Consequently at midterms this model maybe used to classify students as passing or failing expecting accuracy of 72% or with higher accuracy (75%) if only passing students are identified. It should be noted, however, that this data was trained on spring semester data so it is not clear that this model can be generalized to classify fall or summer students. Also the estimate of the expected accuracy could be greatly improved if the number of students receiving midterm grades is increased (see Table 2) and if the number of students averaging 20 visits per semester is increased (see Table 8).

Concerning the categorical predictors in Table 11, the strongest predictors were those students that have entered the class not taking a prerequisite, (meaning they were admitted by advisers, or took Aleks and never failed the prerequisite either) did better than those who took and passed a prerequisite. This makes sense referring to Table 3 we see that students admitted by an advisor have a passing rate of 67% and although Aleks was not a variable included with descriptive statistics the passing rate for students admitted via Aleks was 62%. We can compare this with students who took and passed a prerequisite at some point having a passing rate of 56%. This is not the same as the variable PREREC_PASSED, since this variable looks at the most recent attempt, but 182 students had multiple attempts and 54 of these students did not pass their most recent attempt. The passing rate for these 54 students was 31%. Students are also better off if they have taken a prerequisite class at FAU and spend more time in the MLC as already mentioned.

Appendix

Since the document has been submitted several questions have been asked via email, which are included below.

First Predictor Comment (Repeating MAC 2233)

"Students who have previously failed MAC 2233 (including those who got a "C-"in their best previous attempt). Do these students have a higher failure rate? Shouldn't they be separated out of the study, since previous indicators such as prerequisite done here or elsewhere may be weaker after the student has tried MAC 2233 once."

There are 211 student who had failed MAC 2233 previously, and 98 of them had passed this attempt, giving a passing rate of 46%. 44 of these students have retaken this class more than once.

Second Predictor Comment (Repeating the Prerequisite class)

"How many times did the student need to take the prerequisite? If a student took MAC 1105 four times before passing it, I can predict that he or she will struggle with MAC 2233. (But, on a case-by-case basis, one might see a student who has turned his/her life around.)"

There were 119 students that had taken the prerequisite twice and they had passing rate of about 37%.

There were 30 students that had taken the prerequisite 3 times. They had a passing rate of 43%. This is higher but the counts are lower so we may expect more variability in the measurement.

There was only one student that had taken the prerequisite 4 times, and they passed.

Looking at students who have taken the prerequisite 2 or more times the passing rate was about 38%. If this variable is included in the final model, it is selected as a predictors using stepwise regression, however, much the information in this variable overlaps with "PREREC_PASSED" which includes a level for failing a prerequisite so there was not much improvement.

Third Predictor Comment (How long are Prerequisite Grades valid indicators?)

I know you've got predictors in their for prerequisite grade and for length of time since prerequisite. Can you deduce from your model how long a students with an "A" in MAC 1105 is "good for" before s/he becomes a bad risk in MAC 2233? How about a "B" student? A "C" student?

I looked at this quite a few times. There should be something there but there is a lot of noise going back more than one semester. For instance it's not only time, but term and university (See the cluster Diagram in the Conclusions section) that matters as well. If I try to account for these, the counts become too low. I believe with historical data this can be studied since I wouldn't expect one year to be that different from previous years during the same term.

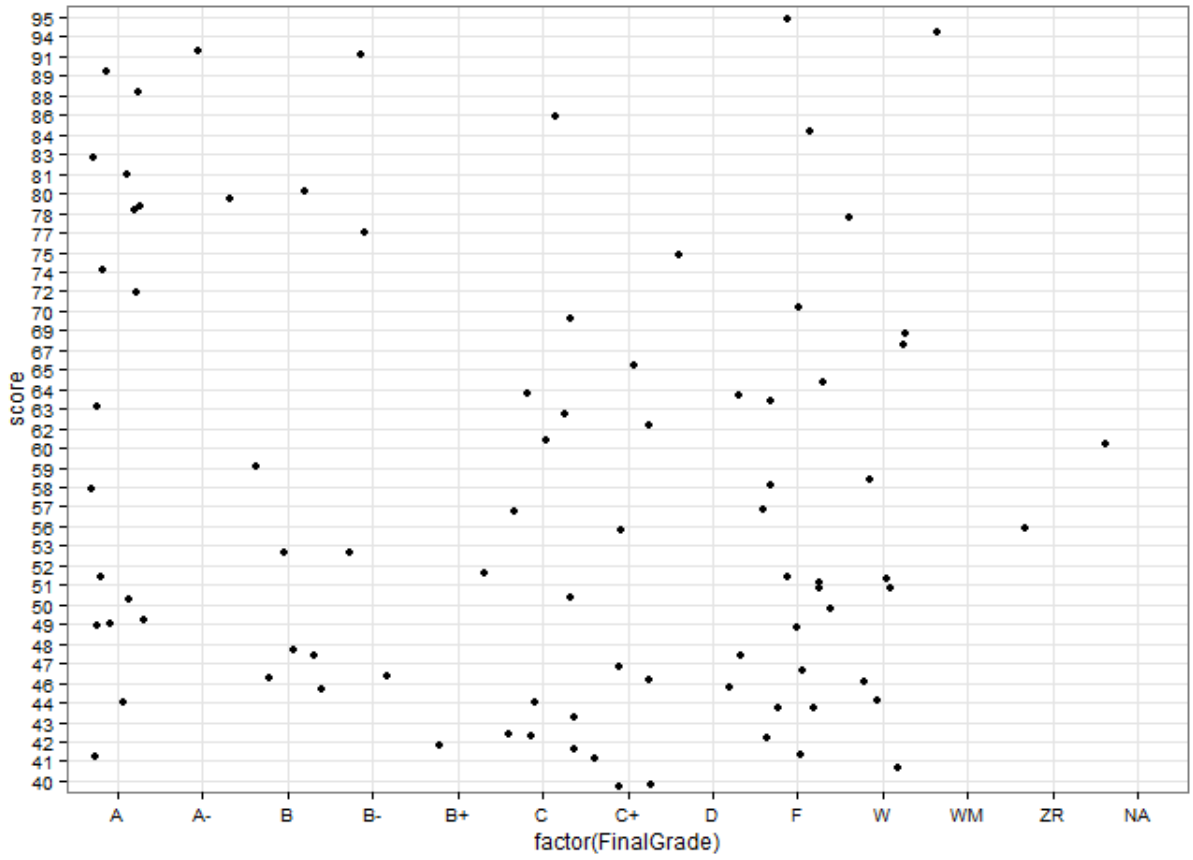
Fourth Predictor Comment (Class Time Slots)

"Did you find a difference in performance between morning and afternoon classes? Afternoon and evening?"

I did not find anything here, to start investigating this I think the best course of action would be to compare different times with the same instructor. Since most instructors only teach one section or two back to back I think this would require historical data as well.

Fifth Predictor Comment (Grades vs Aleks)

Could we see how the ALEKS scores correlated with grades of "D-" or better in MAC 2233?



Fifth Predictor Comment (Grades vs Aleks)

One problem might be that we're thinking one formula fits all. Yes, we can match several indicators against the final grade in MAC 2233, but what if there are two or three distinct groups of students, each with their own characteristics? For example, perhaps Business students have a different type of track record than Biology majors.

I did put splits into the data using indicator variables, and based on the question I decided to look at student majors. The variable looks interesting but will not be done in time for the meeting so I will follow this up with an addendum.