Masters Statistical Professional Writing Project

Analyzing Student Data before Enrollment to Determine Academic Success at Montana State University

Ben Sharp May 2007

The aim of the following analyses is to use first-time Freshmen as a sample for evaluating potential academic success before enrolling at Montana State University. Measuring academic achievement is difficult to generalize for all new Freshmen, so an initial goal is to define a couple of useful metrics. Once a response has been determined, the next task is to identify meaningful variables for predicting an optimal outcome.

First, consider remaining enrolled at MSU until achieving a degree as the overall measure of success. It seems reasonable to assume that a student embarking on a college career does so with the ultimate goal of acquiring a Bachelor's degree. Graduating within six years is a standard and sensible choice for the time to degree. Measuring this response is straightforward, but it has a major draw back: waiting six years to analyze the outcome of students is inconvenient, especially when assessing newly implemented programs designed to influence graduation rates.

After testing a diverse array of variables available during the first year of enrollment, none was as significant to acquiring a Bachelor' degree as the variable of retention to the student's second Fall. Typically, about 45% of first-time Freshmen graduate within six years. For this analysis, a retained student is one who is enrolled in at least one course during their second Fall after beginning as a first-time freshmen the prior Fall. The retention rate is normally around 70%.

Looking at the Fall 2000 cohort of Freshmen, there are several other relatively strong predictors of graduating within six years. For example, Freshmen starting with part-time enrollment in Fall 2000 graduated within six years at a rate of almost 13% versus almost 50% for those who began at full time. Freshmen starting college with a declared major graduated at rate of 51% versus 36% for those who did not initially declare a major. Students who are retained to the subsequent Spring semester are also much more likely to graduate within six years. There are many other potential variables that could serve as a substitute for six-year graduation rate.

Clearly, it would be ideal to measure only the six-year graduation rates for academic success of new Freshmen at MSU; however, this project will focus largely on retention rates to the second Fall. In the end, retention seems to be a sound indicator for graduation rate. In the Fall of 2000, there were 2032 new Freshmen enrolled. Of the 599 Freshmen that failed to enroll in Fall '01, 32 of them graduated, a rate of 5.3%.

To get a general idea of first-time Freshmen preparedness and persistence, examine the following table. The grids of data represent information that is systematically tracked on the progress of first-time **full-time** Freshmen. The graphic is a web page posted on the Office of Planning and Analysis web site where other Freshmen subcohort data are also posted.
See: http://www.montana.edu/opa/facts/gradrate.html

Table 1:

MONTANA STATE UNIVERSITY — Mountains & Minds — Search MSU GO

ACADEMICS | ADMINISTRATION | ADMISSIONS | A-Z INDEX | DIRECTORIES

> Office of Planning & Analysis

Retention and Graduation Rates

Montana State University - Bozeman

Profile of First-time, Full-time, Degree-Seeking Freshmen

High School Background

First Fall	Class Size	Final GPA		Percentile		ACT Comp		SAT Total	
		Average	Num	Average	Num	Average	Num	Average	Num
1997	1824	3.28	1583	66%	1499	23.1	1349	1090	808
1998	1889	3.34	1608	68%	1541	23.6	1363	1095	809
1999	1894	3.33	1682	68%	1558	23.4	1379	1120	688
2000	1854	3.35	1682	67%	1582	23.4	1407	1120	801
2001	1722	3.33	1535	66%	1410	23.3	1261	1097	734
2002	1924	3.34	1758	67%	1587	23.4	1402	1103	898
2003	2011	3.35	1850	66%	1610	23.3	1497	1107	896
2004	2000	3.35	1808	66%	1575	23.5	1449	1119	981
2005	1985	3.36	1821	66%	1564	23.7	1487	1131	987
2006	1942	3.36	1783	67%	1532	24.0	1348	1128	950

College Persistence

First Fall	Class Size	Percent Enrolled Each Subsequent Fall									Cumulative Percent Graduated						
		2nd	3rd	4th	5th	6th	7th	8th	9th	10th	4yr	5yr	6yr	7yr	8yr	9yr	10yr
1997	1824	70.4	57.7	52.9	36.2	13.2	7.7	4.1	1.9	1.3	14.3	35.1	44.3	46.6	47.8	48.2	
1998	1889	70.2	57.6	54.2	36.3	15.2	8.0	2.7	1.4		15.8	40.1	47.2	50.0	51.5		
1999	1894	70.8	59.8	54.5	37.2	13.7	4.6	2.3			18.6	40.4	46.8	49.1			
2000	1854	72.8	61.5	57.7	38.8	13.1	5.1				19.3	41.4	49.6				
2001	1722	72.2	60.0	55.1	35.7	12.5					17.0	39.5					
2002	1924	70.3	58.7	55.7	36.0						17.3						
2003	2011	71.6	61.2	56.3													
2004	2000	70.5	60.2														
2005	1985	70.6															
2006	1942																

The top panel show indicators of high school preparedness based on high school GPA, ACT score and SAT score. The rows represent individual Freshmen cohorts by fall. The number in the cohort and the number with available preparedness data are also shown. The bottom panel shows persistence from Fall to Fall at Montana State University. Persistence is the retention rate at which students continue to be enrolled at MSU from one Fall semester to the next. When students begin graduating after the 4th Fall, the retention rate drops. On the right side of that panel, the cumulative graduation rate grid includes those students graduating after the 4th Fall. For the Fall 2002 Freshmen, one might add the 4th year graduation rate of 17.3 % to 5th Fall persistence rate of 36% to arrive at a graduation rate plus persistence rate 5th year success rate of 53%.

Before settling on retention to the second Fall as the best measure of academic success at MSU, briefly consider Freshmen first-term GPA a response variable. To some, simply performing well during the semester at MSU might be the ultimate measure of success. This response in some circumstances may prove better than graduation rates and retention rates. For example, a student may achieve good marks in her first semester, then transfer to another reputable institution where she achieves exactly her academic goals. Maybe after her Freshmen year at MSU she transfers to University of Washington's medical school. Additionally, first-term GPA can be shown to be a strong indicator of graduating in six years. The first analysis takes a look at first semester academic performance at MSU.

Use first-term GPA as potential continuous variable scaled from 0.00 to 4.00. Now, consider the two obvious predictors: college preparatory test scores and high school GPA. This first example helps to determine which predictor is best at showing academic success where first-term GPA is the metric. Define maximum test score as the best score between comprehensive scores on the ACT and the equivalent ACT scores based on SAT results. Montana State University's Office of Admissions employs a crosswalk of scores from the SAT total to ACT comprehensive score. An 1180 on the SAT equates to a 26 on the ACT. College-bound students may take both or each exam more than once, or they may take none at all. The few incoming students without a high school GPA or test score are not included in the analyses.

Also, Freshman residency status will be considered separately. From a students perspective, tuition and distance from home are important distinctions. From an administrative perspective, differences in tuition and policy merit analyzing the groups independently. Students are either Resident, Nonresident, or WUE. Resident Freshmen pay resident tuition rates and are generally those who graduated high school within the state of Montana. Nonresident Freshmen are generally those who graduate from high school outside of the state, and nonresident tuition and fees are about three times that of resident students. WUE students are part of the Western Undergraduate Exchange agreement. Students from nearby western states meeting certain academic standards and requirements are charged 1.5 times the resident tuition rate.

Resident 2005 Freshmen:

Fitting a linear model will be done using the R statistical package (R 2.2.1). The resident student's first-term GPA will be a linear function of the maximum ACT score. A positive increase in ACT scores should contribute to a positive increase in GPA for the first semester.

Below is R output for the linear model.

```
lm(formula = TERM_GPA ~ ACT_Max, data = resGPA_ACT)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1345 -0.3652  0.2066  0.5855  2.0357

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.950096   0.140692   6.753 2.16e-11
ACT_Max     0.078014   0.005941  13.131  < 2e-16

Residual standard error: 0.8849 on 1320 degrees of freedom
Multiple R-Squared: 0.1155,     Adjusted R-squared: 0.1149
F-statistic: 172.4 on 1 and 1320 DF,  p-value: < 2.2e-16
```

ACT_Max is clearly a significant predictor of first-term GPA, due to a tiny p-value (2.2 e – 16). Also, this model produces an R-squared of 0.1155. That is, ACT scores can explain approximately 11% of the variability in MSU GPA. This leaves 89% of variability left to unexplained phenomenon.

The next model shows first-term GPA as a linear model of HS GPA. One would expect that increasing high school GPA would contribute to increasing first-term GPA at MSU.

```
lm(formula = TERM_GPA ~ HS_GPA, data = resGPA_ACT)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3303 -0.3272  0.1502  0.5100  2.8040

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.90537    0.15567  -5.816 7.56e-09
HS_GPA       1.09449    0.04591  23.839  < 2e-16

Residual standard error: 0.7867 on 1320 degrees of freedom
Multiple R-Squared: 0.301,      Adjusted R-squared: 0.3004
F-statistic: 568.3 on 1 and 1320 DF,  p-value: < 2.2e-16
```

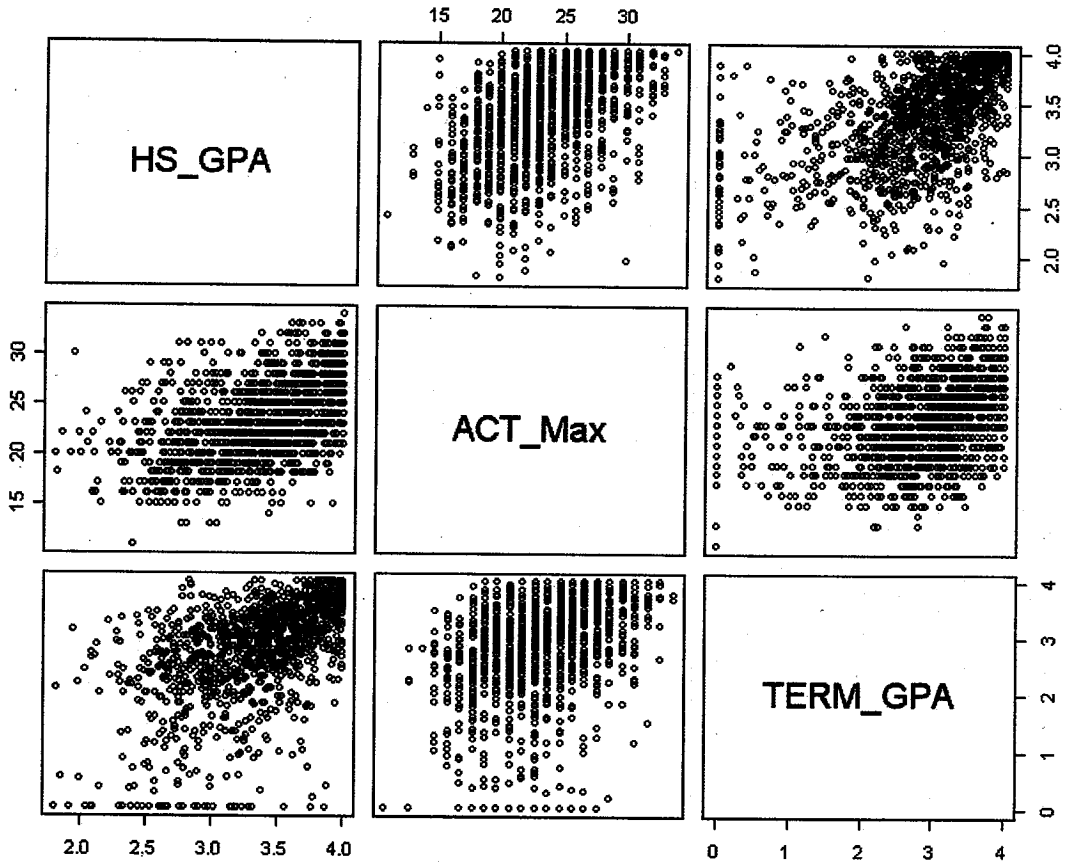
High school GPA is again a significant predictor of MSU first-term GPA for resident students, but now the model produces an R-squared of .301, almost three times larger. Now, there is about 70% left unexplained for first-term GPA results.

Does considering both HS GPA and test scores in the model noticeably improve the R-squared? First, there may be concerns with multicollinearity. The output below shows the correlation among the two predictors and the response.

	HS_GPA	ACT_Max	TERM_GPA
HS_GPA	1.000	0.503	0.549
ACT_Max	0.503	1.000	0.340
TERM_GPA	0.549	0.340	1.000

The next set of plots reveal more of the correlation problems between HS GPA and maximum ACT. Note the narrower linear pattern for HS GPA to Term GPA as compared to maximum ACT to Term GPA.

Pairs Plot for Resident Student
High School GPA, Maximum ACT and First-term MSU GPA.



The following R output shows the results of modeling with both predictors in the model.

```
lm(formula = TERM_GPA ~ ACT_Max + HS_GPA, data = resGPA_ACT)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2798 -0.3252  0.1410  0.4798  2.8837

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.076075   0.163824  -6.568  7.3e-11
ACT_Max      0.019720   0.006087   3.239  0.00123
HS_GPA       1.008345   0.052915  19.056  < 2e-16

Residual standard error: 0.7839 on 1319 degrees of freedom
Multiple R-Squared: 0.3065,     Adjusted R-squared: 0.3054
F-statistic: 291.4 on 2 and 1319 DF,  p-value: < 2.2e-16
```

The effectiveness of the model including both ACT_Max and HS_GPA is dampened by their correlation. Given a choice of one predictor over the other, for resident Freshmen, one should choose high school GPA.

Non-Resident 2005 Freshmen:

Below are results for nonresident new Freshmen from the Fall 2005 cohort. Again, shown is output of first-term GPA according to test score.

```
lm(formula = TERM_GPA ~ ACT_Max, data = nonresGPA_ACT)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8009 -0.4866  0.2029  0.6744  1.4939

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.20858    0.28776   4.200 3.18e-05
ACT_Max      0.05898    0.01217   4.845 1.71e-06

Residual standard error: 0.947 on 485 degrees of freedom
Multiple R-Squared: 0.04617,    Adjusted R-squared: 0.0442
F-statistic: 23.47 on 1 and 485 DF,  p-value: 1.706e-06
```

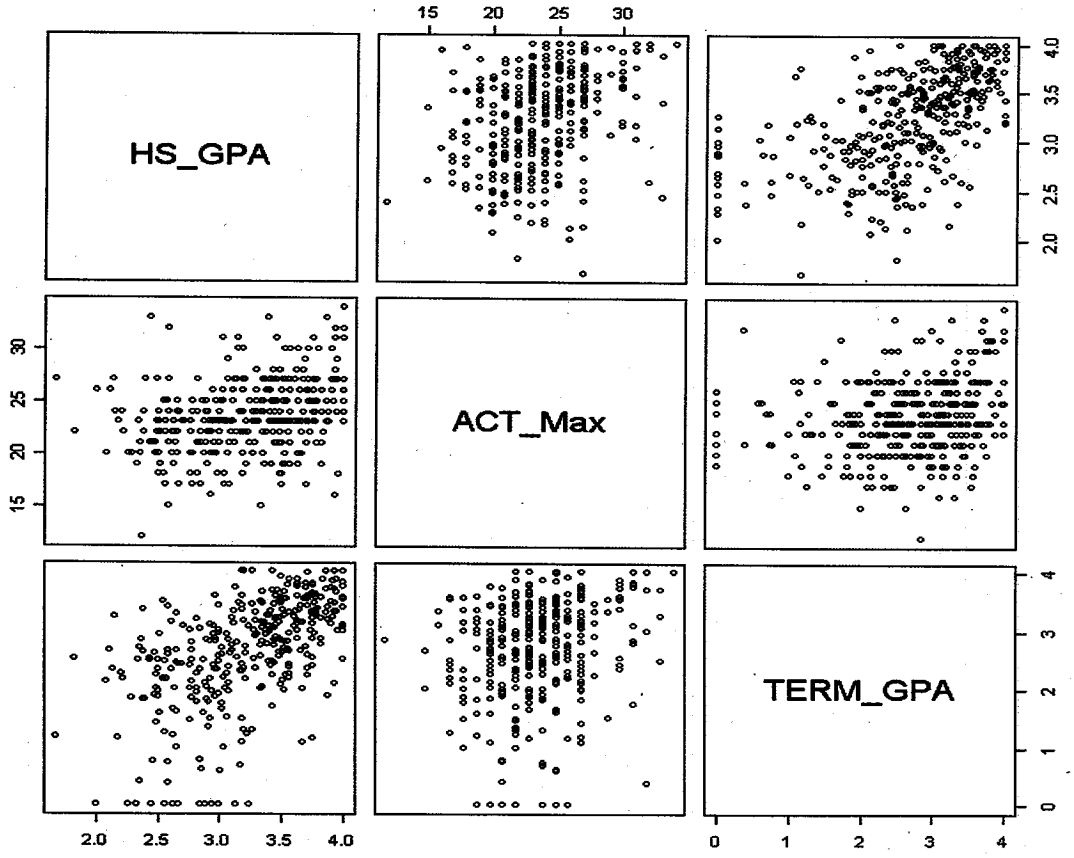
Next is output for first-term GPA according to high school GPA.

```
lm(formula = TERM_GPA ~ HS_GPA, data = nonresGPA_ACT)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6881 -0.3800  0.1445  0.5610  1.7641

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.60678    0.24186  -2.509   0.0124
HS_GPA       1.01069    0.07561  13.368   <2e-16

Residual standard error: 0.8289 on 485 degrees of freedom
Multiple R-Squared: 0.2692,     Adjusted R-squared: 0.2677
F-statistic: 178.7 on 1 and 485 DF,  p-value: < 2.2e-16
```

Notice the R-square when modeling by test score is .0462 versus an R-squared of .2692 when predicting with high school GPA. For nonresident Freshmen, the relative strength of using high school GPA over test scores is superior.

The pairs plot of ACT_Max and HS GPA with term GPA as the response for nonresident is given below.

Pairs Plot for NON-Resident Student
High School GPA, Maximum ACT and First-term MSU GPA.



The corresponding correlations among the predictors and response for nonresidents are shown below.

	HS_GPA	ACT_Max	TERM_GPA
HS_GPA	1.000	0.301	0.519
ACT_Max	0.301	1.000	0.215
TERM_GPA	0.519	0.215	1.000

The regression results with both predictors are also shown. Again there is negligible improvement in the R-squared value over using HS GPA alone due to multicollinearity.

```
lm(formula = TERM_GPA ~ ACT_Max + HS_GPA, data = nonresGPA_ACT)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7307 -0.3757  0.1261  0.5400  1.7066

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.90105    0.30448  -2.959  0.00323
ACT_Max      0.01770    0.01116   1.587  0.11320
HS_GPA       0.97287    0.07916  12.290  < 2e-16

Residual standard error: 0.8276 on 484 degrees of freedom
Multiple R-Squared: 0.273,      Adjusted R-squared:  0.27
F-statistic: 90.89 on 2 and 484 DF,  p-value: < 2.2e-16
```

WUE 2005 Freshmen:

The similar set of analyses are shown for the WUE Freshmen.

```
lm(formula = TERM_GPA ~ ACT_Max, data = wueGPA_ACT)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4139 -0.1856  0.1416  0.3891  0.7782

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.42903    1.16619   1.225    0.222
ACT_Max      0.06403    0.03941   1.625    0.106

Residual standard error: 0.7138 on 144 degrees of freedom
Multiple R-Squared: 0.018,      Adjusted R-squared: 0.01118

F-statistic:  2.64 on 1 and 144 DF,  p-value: 0.1064
```

```
lm(formula = TERM_GPA ~ HS_GPA, data = wueGPA_ACT)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7289 -0.2163  0.0619  0.2921  1.1283

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.5832     0.3886  -1.501    0.136
HS_GPA        1.0824     0.1070  10.118   <2e-16

Residual standard error: 0.5507 on 144 degrees of freedom
Multiple R-Squared: 0.4155,     Adjusted R-squared: 0.4114
F-statistic: 102.4 on 1 and 144 DF,  p-value: < 2.2e-16
```

In this case it does not appear that ACT scores are significant at the .10 alpha level mainly due to the narrow range of values (27-34). WUE students for Fall 2005 are generally required to have a 28 ACT or SAT equivalent scores. Again, high school GPA is the better option when predicting first-term success for WUE Freshmen.

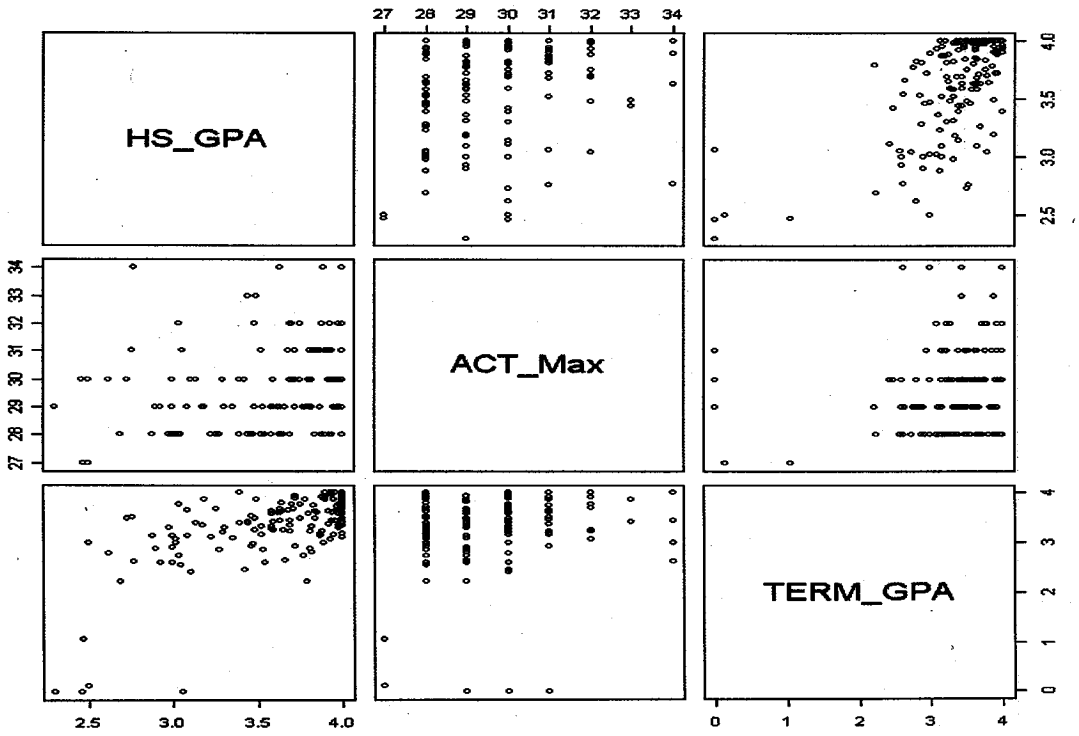
Below is the correlation of the two predictors and first-term GPA for WUE Freshmen.

	HS_GPA	ACT_Max	TERM_GPA
HS_GPA	1.000	0.146	0.645
ACT_Max	0.146	1.000	0.134
TERM_GPA	0.645	0.134	1.000

correlation between HS_GPA and Term GPA.

Again, the pairs plot show the ~~multicollinearity for those variables~~.

Pairs Plot for WUE Student
High School GPA, Maximum ACT and First-term MSU GPA.



Up to this point, first term GPA at MSU has been the response variable. Certainly the R-squared values indicate that high school GPA is a much better predictor over test scores in the preceding analysis. Now, reconsider the response as one of two possibilities. A student is either retained to the second Fall or they are not retained. As concluded at the beginning of the paper, retention to the second Fall may be regarded as a better measure of academic success at MSU.

Continuing with the same predictors, HS_GPA and ACT_Max, but now the binary response of retention to the following Fall will be used. In order to increase the number of observations, consider also using the cohorts of Fall 2002 – Fall 2004. One way to view the difference among retention rates according to the two variables is to look at rates by different test score categories and then by high school GPA categories. Essentially, increasing levels of test scores will likely show an increase in retention rate. An increase in a high school GPA category should also result in an increase in retention rate. The object is to see if test score groups or HS GPA groups reveal trends in retention rates across cohort years. Also, within a single cohort, observing patterns between categories can help determine which variable best differentiates retention rate .

Basically, this analysis displays the numbers of Freshmen under different levels of preparedness over time. It is meaningful to see head count figures for administrative reasons and to quickly get an idea of which college preparatory categories are larger or smaller and then which of those categories the better retention rates.

A similar spread in distribution among the categories would be ideal. Comparing retention rates from one category to another, one would like to be comparing averages of the same number of Freshmen. For example, comparing the retention rate of 25 Freshmen scoring between 15 and 18 on the ACT to the retention rate of 200 scoring between 25 and 28 may not be insightful. More importantly, comparing the retention rate of 25 Freshmen scoring between 15 and 18 on the ACT to 200 Freshmen with high school GPAs between 2.00 and 2.60 may not be useful.

Unfortunately, the discrete values of test scores and GPA do not allow for equal categorization. The method employed in the following analysis was to first create a distribution according to test scores. That same distribution is then matched with an assortment of ranges of high school GPA. Table 2 has four sets of grids. Each grid represents headcounts and the corresponding retention rates of resident Freshmen for Fall 2002 through Fall 2005. The set of grids on the left show headcounts and rates by ACT score (or equivalent SAT score) and on the right are headcounts and rates by GPA category. Initially, look at the cohort sizes by test category and compare it to the number in the same row across the page for GPA category. Generally that distribution is fairly similar throughout each cohort year. Note that WUE cohorts are excluded due to different test score requirements and smaller numbers in single categories.

Next, notice the last set of grids at the bottom of Table 2. These cells represent total headcounts for the listed resident Freshmen cohorts and show a significant number of observations by category. The larger numbers help to stabilize retention rates in each group. For resident students, the lowest retention rate for test score begins at 56% for scores under 19 and goes up to 88% for scores above 30. On the other hand, the GPA categories push out that spread indicating more strength in predicting retention. It ranges from 42% to 91%. Interestingly, students with GPA's from about 3.91 to 3.99 out of high school are equally likely to be retained as those with 3.99 and 4.0. Table 3 shows similar results for the nonresident subcohorts. Many conclusions can be made from these tables. Primarily, as test scores and high school GPA increase, retention rates increase reinforcing observations when first-term GPA was the metric for academic success. Another point is that high school GPA seems to give more information about retention rate.

A graphical summary of the data in Table 2 and 3 are found on the subsequent pages. Chart 1 and 2 represent resident Freshmen by ACT score and high school GPA, while Chart 3 and 4 display same information for the nonresidents.

Table 2: Resident Freshmen Retention by ACT and High School GPA

Freshmen Fall 2002

ACT	Cohort	Retain	%
Under 19	156	84	53.85%
19-20	181	114	62.98%
21-22	232	169	72.84%
23-24	219	146	66.67%
25-26	173	139	80.35%
27-28	151	115	76.16%
29-30	78	69	88.46%
Over 30	59	51	86.44%
Total	**1249**	**887**	**71.02%**

GPA	Cohort	Retain	%
Under 2.68.	150	65	43.33%
2.68-3.03	181	108	59.67%
3.03-3.30	221	144	65.16%
3.30-3.57	214	163	76.17%
3.57-3.76	168	133	79.17%
3.76-3.91	154	130	84.42%
3.91-3.99	82	74	90.24%
3.99-4.00	79	70	88.61%
Total	**1249**	**887**	**71.02%**

Freshmen Fall 2003

ACT	Cohort	Retain	%
Under 19	167	98	58.68%
19-20	205	130	63.41%
21-22	230	167	72.61%
23-24	216	166	76.85%
25-26	191	146	76.44%
27-28	154	124	80.52%
29-30	87	72	82.76%
Over 30	40	37	92.50%
Total	**1290**	**940**	**72.87%**

GPA	Cohort	Retain	%
Under 2.68.	132	58	43.94%
2.68-3.03	196	121	61.73%
3.03-3.30	219	142	64.84%
3.30-3.57	228	172	75.44%
3.57-3.76	187	154	82.35%
3.76-3.91	161	139	86.34%
3.91-3.99	92	85	92.39%
3.99-4.00	75	69	92.00%
Total	**1290**	**940**	**72.87%**

Freshmen Fall 2004

ACT	Cohort	Retain	%
Under 19	160	94	58.75%
19-20	188	129	68.62%
21-22	228	165	72.37%
23-24	197	128	64.97%
25-26	185	144	77.84%
27-28	152	118	77.63%
29-30	101	83	82.18%
Over 30	68	57	83.82%
Total	**1279**	**918**	**71.77%**

GPA	Cohort	Retain	%
Under 2.71	144	66	45.83%
2.71-3.03	205	110	53.66%
3.03-3.35	233	174	74.68%
3.35-3.57	199	149	74.87%
3.57-3.76	185	141	76.22%
3.76-3.91	136	120	88.24%
3.91-3.99	101	91	90.10%
3.99-4.00	76	67	88.16%
Total	**1279**	**918**	**71.77%**

Freshmen Fall 2005

ACT	Cohort	Retain	%
Under 19	180	96	53.33%
19-20	187	127	67.91%
21-22	231	148	64.07%
23-24	241	166	68.88%
25-26	200	154	77.00%
27-28	160	125	78.13%
29-30	87	76	87.36%
Over 30	64	59	92.19%
Total	**1350**	**951**	**70.44%**

GPA	Cohort	Retain	%
Under 2.77	184	67	36.41%
2.75-3.08	191	125	65.45%
3.03-3.35	227	149	65.64%
3.35-3.57	238	188	78.99%
3.57-3.76	189	143	75.66%
3.76-3.92	153	125	81.70%
3.92-3.99	91	82	90.11%
3.99-4.00	77	72	93.51%
Total	**1350**	**951**	**70.44%**

Total of Four Cohorts

ACT	Total	Retain	%
Under 19	663	372	56.11%
19-20	761	500	65.70%
21-22	921	649	70.47%
23-24	873	606	69.42%
25-26	749	583	77.84%
27-28	617	482	78.12%
29-30	353	300	84.99%
Over 30	231	204	88.31%
Total	**5168**	**3696**	**71.52%**

GPA	Total	Retain	%
Range 1	610	256	41.97%
Range 2	773	464	60.03%
Range 3	900	609	67.67%
Range 4	879	672	76.45%
Range 5	729	571	78.33%
Range 6	604	514	85.10%
Range 7	366	332	90.71%
Range 8	307	278	90.55%
Total	**5168**	**3696**	**71.52%**

Freshmen Fall 2002

ACT	Cohort	Retain	%
Under 19	50	32	64.00%
19-20	60	35	58.33%
21-22	109	69	63.30%
23-24	103	69	66.99%
25-26	56	33	58.93%
27-28	26	18	69.23%
29-30	15	9	60.00%
Over 30	14	13	92.86%
Total	**433**	**278**	**64.20%**

GPA	Cohort	Retain	%
Under 2.45	49	23	46.94%
2.45-2.75	62	34	54.84%
2.78-3.12	108	63	58.33%
3.12-3.53	108	74	68.52%
3.53-3.77	53	42	79.25%
3.77-3.91	27	20	74.07%
3.91-3.99	10	9	90.00%
3.99-4.00	16	13	81.25%
Total	**433**	**278**	**64.20%**

Freshmen Fall 2003

ACT	Cohort	Retain	%
Under 19	51	29	56.86%
19-20	70	41	58.57%
21-22	88	52	59.09%
23-24	125	76	60.80%
25-26	44	29	65.91%
27-28	26	18	69.23%
29-30	12	9	75.00%
Over 30	17	15	88.24%
Total	**433**	**269**	**62.12%**

GPA	Cohort	Retain	%
Under 2.45	47	21	44.68%
2.45-2.85	75	41	54.67%
2.85-3.15	88	49	55.68%
3.15-3.60	128	87	67.97%
3.60-3.77	41	27	65.85%
3.77-3.91	26	19	73.08%
3.91-3.99	9	8	88.89%
3.99-4.00	19	17	89.47%
Total	**433**	**269**	**62.12%**

Freshmen Fall 2004

ACT	Cohort	Retain	%
Under 19	52	33	63.46%
19-20	58	34	58.62%
21-22	110	63	57.27%
23-24	126	80	63.49%
25-26	80	60	75.00%
27-28	76	54	71.05%
29-30	20	13	65.00%
Over 30	14	10	71.43%
Total	**536**	**347**	**64.74%**

GPA	Cohort	Retain	%
Under 2.45	50	24	48.00%
2.45-2.70	54	29	53.70%
2.85-3.02	116	73	62.93%
3.15-3.43	123	86	69.92%
3.40-3.67	81	53	65.43%
3.67-3.91	70	51	72.86%
3.91-3.99	17	13	76.47%
3.99-4.00	25	18	72.00%
Total	**536**	**347**	**64.74%**

Freshmen Fall 2005

ACT	Cohort	Retain	%
Under 19	50	28	56.00%
19-20	58	34	58.62%
21-22	95	59	62.11%
23-24	120	86	71.67%
25-26	85	49	57.65%
27-28	56	39	69.64%
29-30	14	11	78.57%
Over 30	20	15	75.00%
Total	**498**	**321**	**64.46%**

GPA	Cohort	Retain	%
Under 2.50	52	19	36.54%
2.50-2.70	57	32	56.14%
2.85-3.02	93	59	63.44%
3.15-3.43	121	80	66.12%
3.43-3.67	88	60	68.18%
3.67-3.90	56	46	82.14%
3.91-3.95	12	9	75.00%
3.95-4.00	19	16	84.21%
Total	**498**	**321**	**64.46%**

Total of Four Cohorts

ACT	Cohort	Retain	%
Under 19	203	122	60.10%
19-20	246	144	58.54%
21-22	402	243	60.45%
23-24	474	311	65.61%
25-26	265	171	64.53%
27-28	184	129	70.11%
29-30	61	42	68.85%
Over 30	65	53	81.54%
Total	**1900**	**1215**	**63.95%**

GPA	Cohort	Retain	%
Range 1	198	87	43.94%
Range 2	248	136	54.84%
Range 3	405	244	60.25%
Range 4	480	327	68.13%
Range 5	263	182	69.20%
Range 6	179	136	75.98%
Range 7	48	39	81.25%
Range 8	79	64	81.01%
Total	**1900**	**1215**	**63.95%**

Notice that the general gradient for the lines in Chart 2 using high school GPA is much steeper than those lines found in Chart 1 using ACT scores. For resident Freshmen, it appears that high school GPA better differentiates Freshmen retention. Focusing on the bold blue line may simplify these charts. It is the overall average rate for the combined cohorts.

Chart 1:

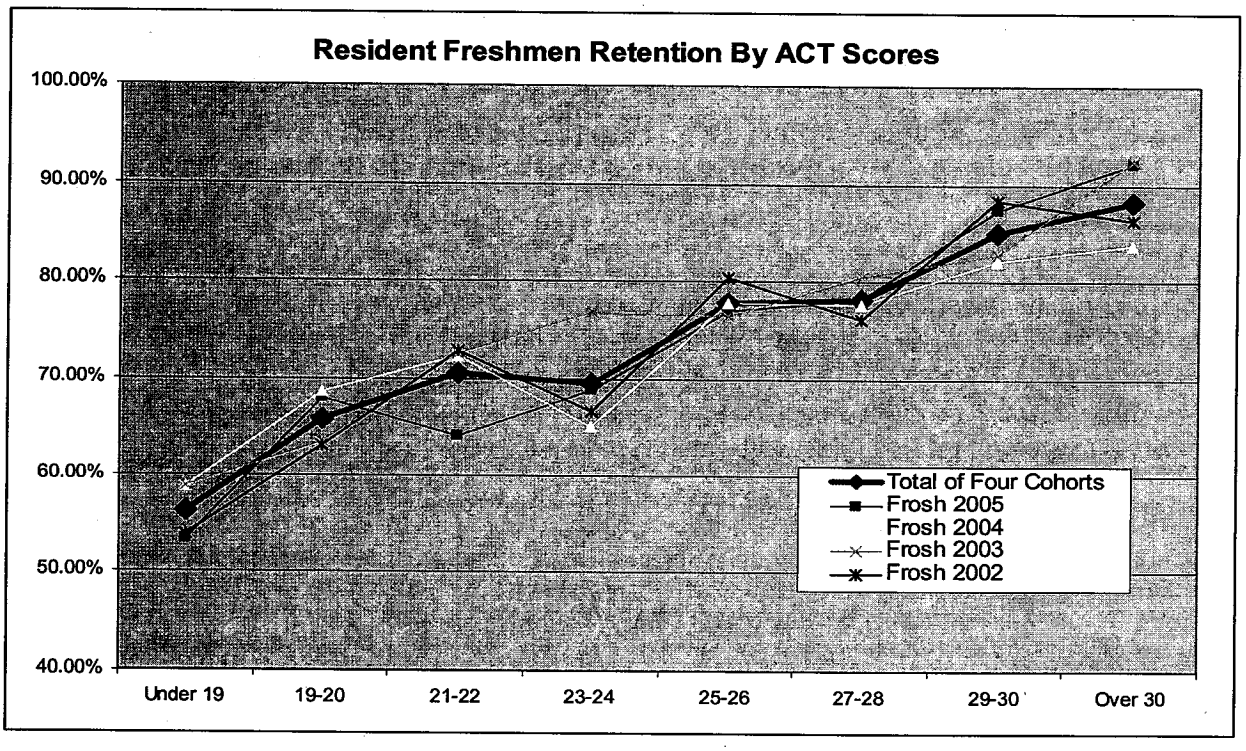
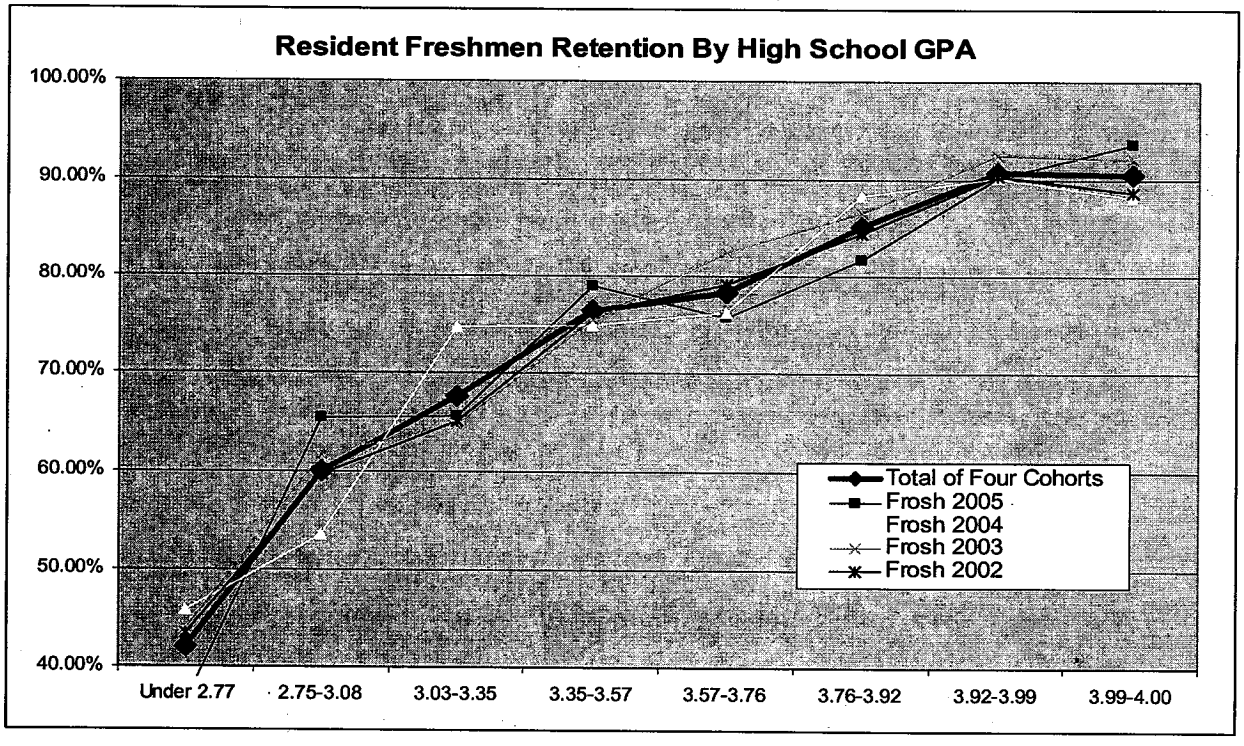


Chart 2:



As with resident Freshmen, high school GPA for nonresidents is also graphically shown to better differentiate retention rates in the charts below.

Chart 3:

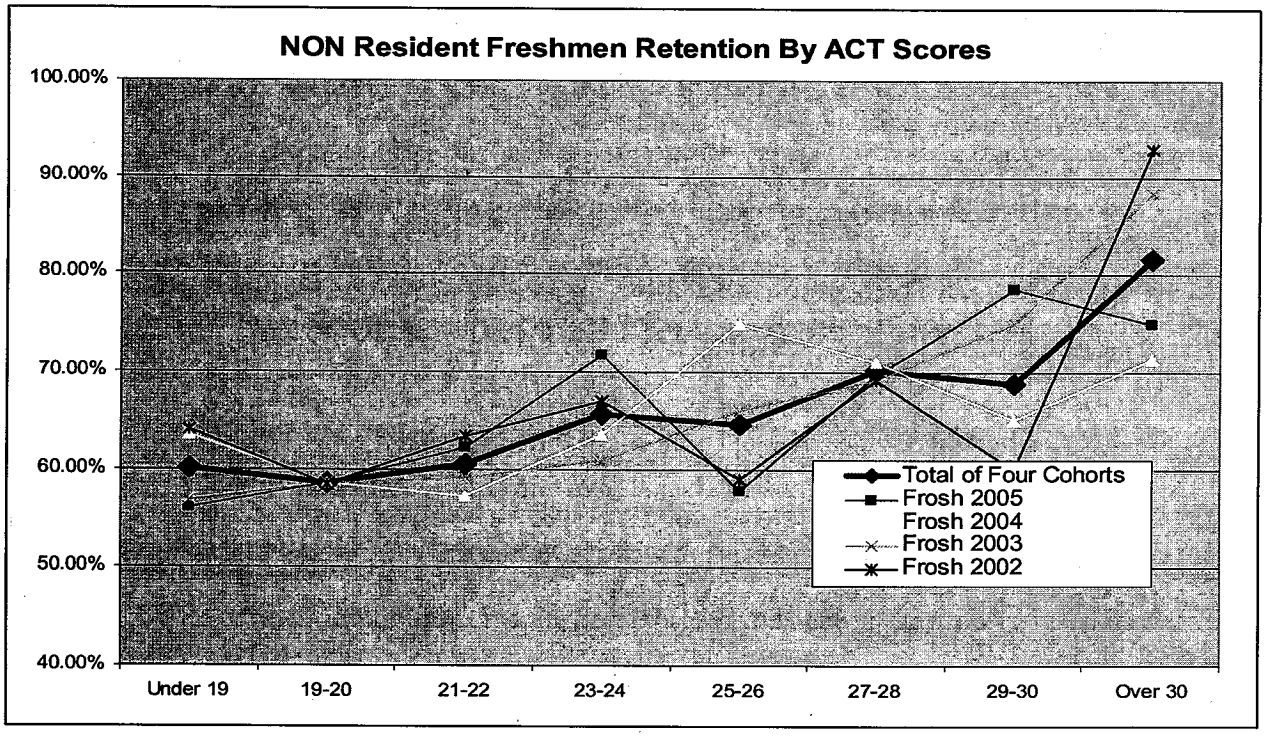
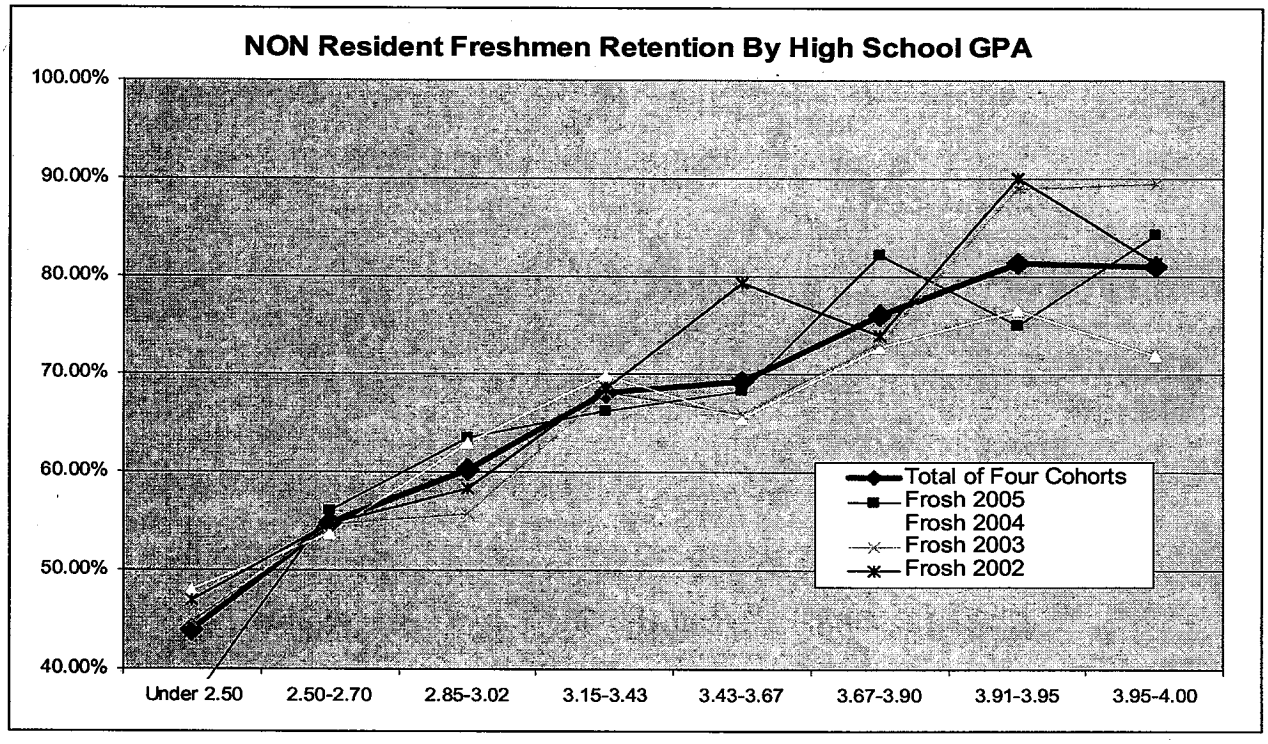


Chart 4:



Again, the goal is to determine a set of characteristics that identify which prospective students are likely to enroll at MSU for a second year. Once a student applies to MSU, there are several potential variables of interest including:

Test Scores
HS GPA
Gender
Ethnicity
Age
Major of Interest
Number of College Preparatory Exams taken
Geographic Data
High school size
Expected Family Contribution
Recruiting contact data
First-generation college student
Orientation Sign-Up

Determining which variables are both useful and have predictive power is a challenge. The above variables were tested for significance by using them in a series of logistic regression models with retention being the binary response. Many were found to be of little significance in predicting retention, especially when high school GPA was included in the model. For example, beyond high-school GPA, gender does not appear to be a strong predictor of retention on its own.

The next set of tables summarize retention rates by Gender. Also included are averages of some preparedness indicators. The preparedness variables are ACT English/verbal scores and the maximum ACT Math/Quantitative scores, as well as, high school GPA. Of lesser importance for these tables is the Average Test Count. This predictor may be thought of as how many college-preparatory exams were taken. The Test Count variable counts each exam subscore separately. This variable will be addressed further in the next section.

Gender:

Resident

Gender	Frosh #	Retention Rate	Avg ACT Verbal	Avg ACT Math	Avg HS_GPA	Avg Test Count
F	598	73.58%	22.34	22.38	3.4407	15.17
M	658	68.09%	22.41	24.32	3.2644	14.73
Total	**1256**	**70.70%**	**22.38**	**23.40**	**3.3483**	**14.94**

NonResident

Gender	Frosh #	Retention Rate	Avg ACT Verbal	Avg ACT Math	Avg HS_GPA	Avg Test Count
F	108	71.30%	22.76	22.62	3.3516	15.91
M	232	62.07%	22.64	23.94	3.1479	13.72
Total	**340**	**65.00%**	**22.68**	**23.52**	**3.2126**	**14.42**

WUE

Gender	Frosh #	Retention Rate	Avg ACT Verbal	Avg ACT Math	Avg HS_GPA	Avg Test Count
F	32	84.38%	30.00	28.63	3.8163	14.88
M	61	70.49%	28.00	29.54	3.6495	15.84
Total	**93**	**75.27%**	**28.69**	**29.23**	**3.7069**	**15.51**

Clearly, women are retained at higher rates for each residency status. Also note the average test score and average high school GPA for women. In each case high school GPA is higher.

Minority students are those who self-reported a category other than 'white' as their ethnicity. The result is similar to that seen with gender. WUE is missing in this category due to smaller numbers.

Minority:

Resident

Minority	Frosh #	Retention Rate	Avg ACT Verbal	Avg ACT Math	Avg HS_GPA	Avg Test Count
Not Minority	1192	71.06%	22.45	23.52	3.3518	14.86
Minority	64	64.06%	20.91	21.13	3.2838	16.42
Total	**1256**	**70.70%**	**22.38**	**23.40**	**3.3483**	**14.94**

NonResident

Minority	Frosh #	Retention Rate	Avg ACT Verbal	Avg ACT Math	Avg HS_GPA	Avg Test Count
Not Minority	318	65.72%	22.70	23.63	3.2165	14.64
Minority	22	54.55%	22.32	21.91	3.1559	11.27
Total	**340**	**65.00%**	**22.68**	**23.52**	**3.2126**	**14.42**

Another variable of interest is a student's chosen major of study. This variable is more difficult to analyze and interpret because there are over 50 categories. Also, often students change majors within a short time of starting class. Grouping by college is one method to reduce the categories to seven and increases the likelihood of capturing a changed major within the same college. In this tabulation, WUE is again not included. With only seven groups, the numbers per cell are still quite small.

College of Initial Major:

Resident

College of Major	Frosh #	Retention Rate	Avg ACT Verbal	Avg ACT Math	Avg HS_GPA	Avg Test Count
Art & Arch	159	62.89%	23.18	23.35	3.3307	14.23
Agriculture	42	88.10%	22.21	23.50	3.5571	16.50
Business	128	67.19%	22.31	23.72	3.3853	14.60
Education	97	76.29%	21.12	21.56	3.3805	15.38
Engineering	211	78.67%	23.75	26.69	3.4636	15.41
Letters & Science	185	77.84%	23.88	24.33	3.4303	15.54
Nursing	66	78.79%	21.76	21.88	3.4442	13.61
University College	368	62.23%	20.96	21.70	3.1863	14.75
Total	**1256**	**70.70%**	**22.38**	**23.40**	**3.3483**	**14.94**

NonResident

College of Major	Frosh #	Retention Rate	Avg ACT Verbal	Avg ACT Math	Avg HS_GPA	Avg Test Count
Art & Arch	98	63.27%	23.04	23.70	3.1584	17.02
Agriculture	10	60.00%	23.10	24.50	3.3940	12.00
Business	38	57.89%	20.84	22.00	3.0216	14.87
Education	10	80.00%	21.60	22.30	3.5170	14.60
Engineering	45	68.89%	24.22	26.89	3.3813	14.07
Letters & Science	56	75.00%	23.39	23.68	3.3043	12.98
Nursing	14	57.14%	24.00	22.86	3.4771	13.36
University College	69	60.87%	21.41	21.91	3.0862	12.41
Total	**340**	**65.00%**	**22.68**	**23.51**	**3.2126**	**14.42**

For each college, generally the preparedness variables, ACT Verbal scores, ACT Math scores, and high school GPA allow for a relative prediction of the retention rate. In particular, in nearly all cases, a decrease in high school GPA realizes a decrease in retention rate.

When coupled with high school GPA and test scores, the categorical variables above generally seem not to offer new information. In the regression models, they were found not to be significant in predicting retention and only made the model more complicated without increasing the R-squared remarkably. Other variables tested and found not to be significant are high school size, the number of contacts during recruitment, age, and the high school percentile in which a student graduated. Certainly more careful analyses of each variable and possibly others should be explored. Greatly increasing the number of Freshmen studied may help to find some differences among groups. For example, combining six or seven cohorts could increase the number in the model by six times the 2005 Freshmen cohort. Then variables such as year could be tested in a regression which may produce significance in some of the listed predictors.

For the final analysis, four predictors will be reviewed according to residency status for the Fall 2005 Freshmen retention to Fall 2006. First, the student's Pell amount paid during their first term will be used to represent financial need. The larger Pell grant amount indicates that the student's effective family contribution is lower. The maximum amount for Pell is $2,025. About 7% of WUE students received some Pell dollars. 12% of nonresident Freshmen received Pell grant money. And, roughly 28% of resident students took advantage of the Pell grant.

The number of times a student appeared in the database for exams taken will also be considered. This variable may represent a rough numeric measure of "college-bound determination." It is intended to roughly capture or index the relative seriousness that a future college student is about college preparedness. So while a student may have relatively poor maximum test scores and high school GPA, he may still be very eager to succeed in college. This phenomenon may be apparent by having 35 records in the test score database. If a student takes the ACT exam ~~and then~~ followed by an ACT writing exam, then he will have at minimum four counts. One for each ACT subscore and then one for the writing score. Other subject area scores are also included in those counts. Occasionally, students may have up to 60 appearances in that table. This variable was used in the preceding summary tables and will be called Test Count .

The final two variables are high school GPA and the maximum ACT Math/ Quantitative scores. There is a strong correlation between maximum ACT Math scores and maximum ACT Verbal scores when including both predictors in the model. For resident Freshmen, it is 0.64. Choosing the best predictor of the two resolves the problem and may produce better a predictor than just considering the total score. To determine which variable to use, observe the p-value for significance when each variable is included in the model independently. For resident Freshmen, the ACT Verbal produces a p-value of 0.0634. When ACT Verbal is replaced by ACT Math, the respective p-value is 0.0195. The ACT Math variable had a smaller p-value for resident Freshmen ,but neither test score is significant for nonresident or WUE Freshmen when including them separately in the model. However, for consistency, the maximum ACT Math score is used instead of ACT Verbal scores for each residency status.

<u>Resident 2005 Freshmen:</u>

Below is a grid of the correlation between those variables.

	Retain	HS_GPA	TestCnt	ACT_MaxVerb	ACT_MaxMath	Pell
Retain	1.00	0.32	0.14	0.20	0.23	-0.06
HS_GPA	0.32	1.00	0.13	0.48	0.55	-0.01
TestCnt	0.14	0.13	1.00	0.09	0.06	-0.02
ACT_MaxVerb	0.20	0.48	0.09	1.00	0.64	-0.10
ACT_MaxMath	0.23	0.55	0.06	0.64	1.00	-0.15
Pell	-0.06	-0.01	-0.02	-0.10	-0.15	1.00

There is also moderate correlation among the test scores and high school GPA (r = 0.48, 0.55).

Below is output from fitting the binary response of retention to the second year for resident first-time Freshmen according to the four predictors for resident Freshmen. The predictors are high school GPA, ACT math score, Pell grant amount, and Test count.

```
glm(formula = Retain ~ HS_GPA + ACT_MaxMath + Pell + TestCnt,
    data = res)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0116  -0.4647   0.1723   0.3123   0.7521

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.850e-01  8.367e-02  -4.601 4.58e-06
HS_GPA       2.535e-01  2.909e-02   8.714  < 2e-16
ACT_MaxMath  6.978e-03  2.985e-03   2.338   0.0195
Pell        -2.954e-05  1.673e-05  -1.765   0.0777
TestCnt      5.729e-03  1.389e-03   4.125 3.92e-05

(Dispersion parameter for gaussian family taken to be 0.1878215)

    Null deviance: 303.92  on 1434  degrees of freedom
Residual deviance: 268.58  on 1430  degrees of freedom
AIC: 1679.6
```

This model seems to be a reasonable mix of intuitive variables that should be in the model and their respective statistical significance. Notice that verbal ACT test scores are not significant and ACT math score p-value is above a 0.10 alpha test level. Previously, a total or single comprehensive test score was shown to be significant. With the math and verbal score separated, the correlation between the two are likely influencing their independent significance in the model.

Non-Resident 2005 Freshmen:

A similar result for nonresident students is shown in the following output.

	Retain	HS_GPA	TestCnt	ACT_MaxVerb	ACT_MaxMath	Pell
Retain	1.00	0.22	0.12	0.09	0.12	-0.09
HS_GPA	0.22	1.00	0.15	0.31	0.36	0.01
TestCnt	0.12	0.15	1.00	0.15	0.07	-0.06
ACT_MaxVerb	0.09	0.31	0.15	1.00	0.59	-0.16
ACT_MaxMath	0.12	0.36	0.07	0.59	1.00	-0.16
Pell	-0.09	0.01	-0.06	-0.16	-0.16	1.00

In the next set of output, notice that as for resident students, larger Pell amounts contribute to a smaller retention rate. The negative value of the Pell coefficient may indicate that students with greater financial need are less likely to be enrolled for a second year. The R output below summarizes the same model that was applied to the resident Freshmen.

```
glm(formula = Retain ~ HS_GPA + ACT_MaxMath + Pell + TestCnt,
    data = nonres)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8763  -0.5338   0.2410   0.3835   0.6593

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.202e-01  1.465e-01  -0.820   0.4123
HS_GPA       2.039e-01  4.429e-02   4.604 5.13e-06
ACT_MaxMath  3.234e-03  5.015e-03   0.645   0.5193
Pell        -8.967e-05  4.326e-05  -2.073   0.0386 *
TestCnt      3.878e-03  2.010e-03   1.929   0.0542 .

(Dispersion parameter for gaussian family taken to be 0.2184612)

    Null deviance: 131.96  on 568  degrees of freedom
Residual deviance: 123.21  on 564  degrees of freedom
AIC: 756.2
```

WUE 2005 Freshmen:

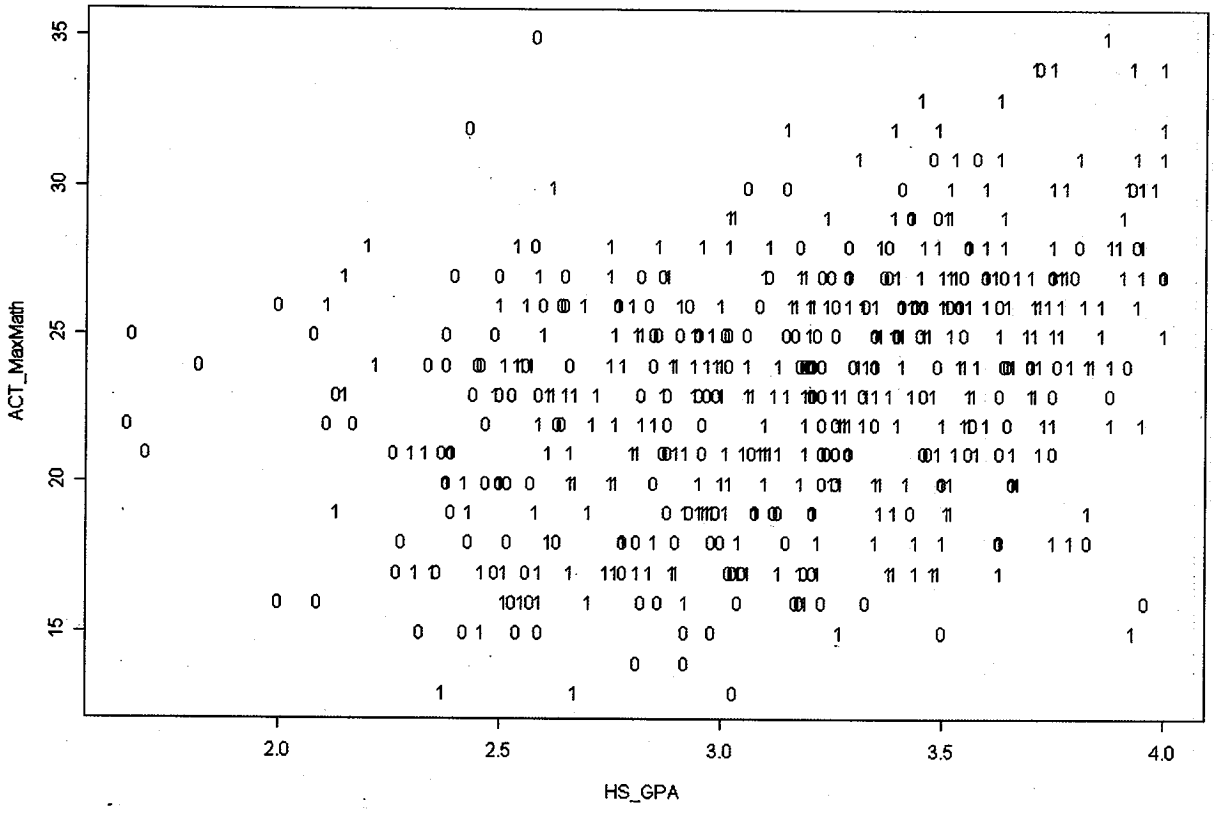
The analysis for WUE retention using high school GPA, ACT math score, Pell grant amount, and Test count as the predictors is as follows.

```
             Retain HS_GPA TestCnt ACT_MaxVerb ACT_MaxMath  Pell
Retain         1.00    0.22    0.05       -0.06        0.08 -0.06
HS_GPA         0.22    1.00    0.29        0.12        0.20  0.03
TestCnt        0.05    0.29    1.00        0.08       -0.11 -0.01
ACT_MaxVerb   -0.06    0.12    0.08        1.00       -0.09  0.03
ACT_MaxMath    0.08    0.20   -0.11       -0.09        1.00 -0.09
Pell          -0.06    0.03   -0.01        0.03       -0.09  1.00

glm(formula = Retain ~ HS_GPA + ACT_MaxMath + Pell + TestCnt,
    data = wue)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8740   0.1232   0.1724   0.2449   0.4738

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.671e-01  4.511e-01  -0.370   0.7116
HS_GPA       2.259e-01  8.814e-02   2.563   0.0114
ACT_MaxMath  4.362e-03  1.382e-02   0.316   0.7526
Pell        -7.415e-05  9.649e-05  -0.769   0.4434
TestCnt     -5.985e-04  3.443e-03  -0.174   0.8622

(Dispersion parameter for gaussian family taken to be 0.1778650)

    Null deviance: 28.168  on 154  degrees of freedom
Residual deviance: 26.680  on 150  degrees of freedom
AIC: 179.15
```

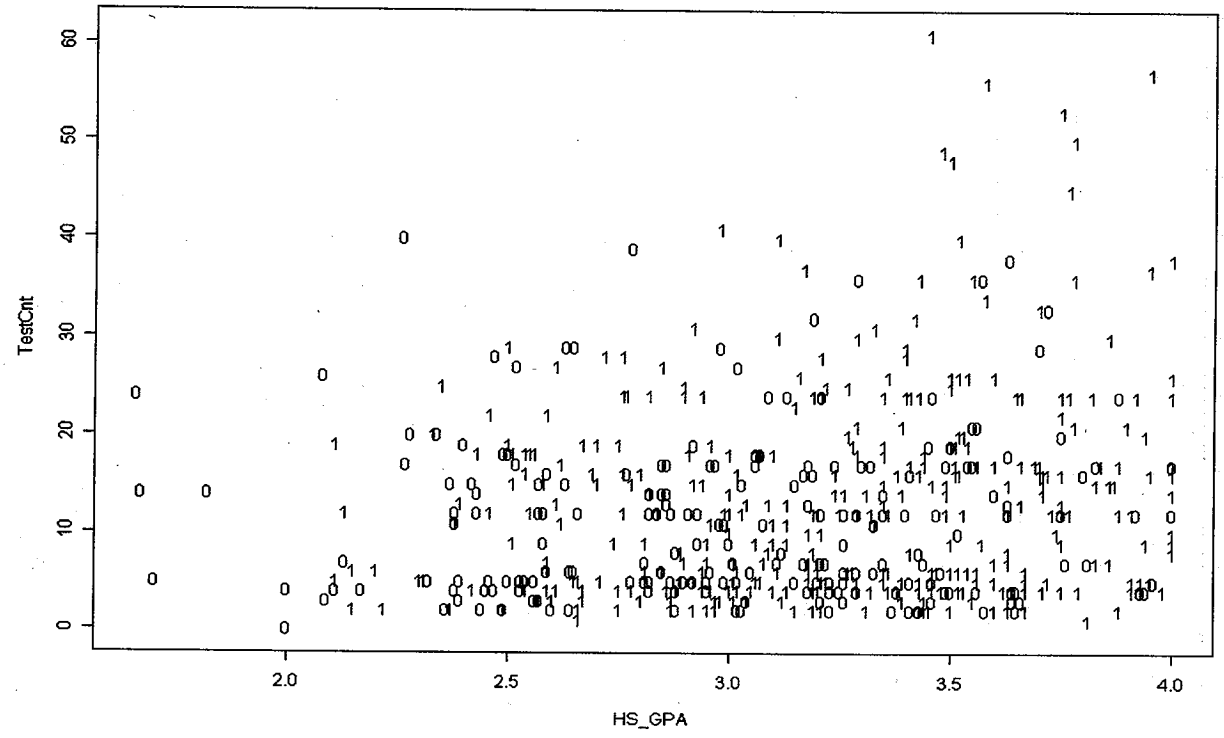
Looking at plots may show more insight into the above models. Focus is now on nonresident Freshmen only. The idea is to see which patterns may influence a students likelihood of being retained. Below are three plots. A "1" represents a retained nonresident student. While it is difficult to discern in detail, the areas where students are retained are somewhat visible.

Plot 1 shows ACT Math scores versus high school GPA. Obviously, those students with high Math scores and high GPA are more likely to remain enrolled. Plot 2 shows the test counts versus high school GPA. Notice that those who have taken a lot of exams and have high GPA's are very likely to be retained. Plot 3 displays test counts by ACT Math scores. This plot is less insightful because both values are discrete creating some overlap among data points, but it still may have some good information to offer.

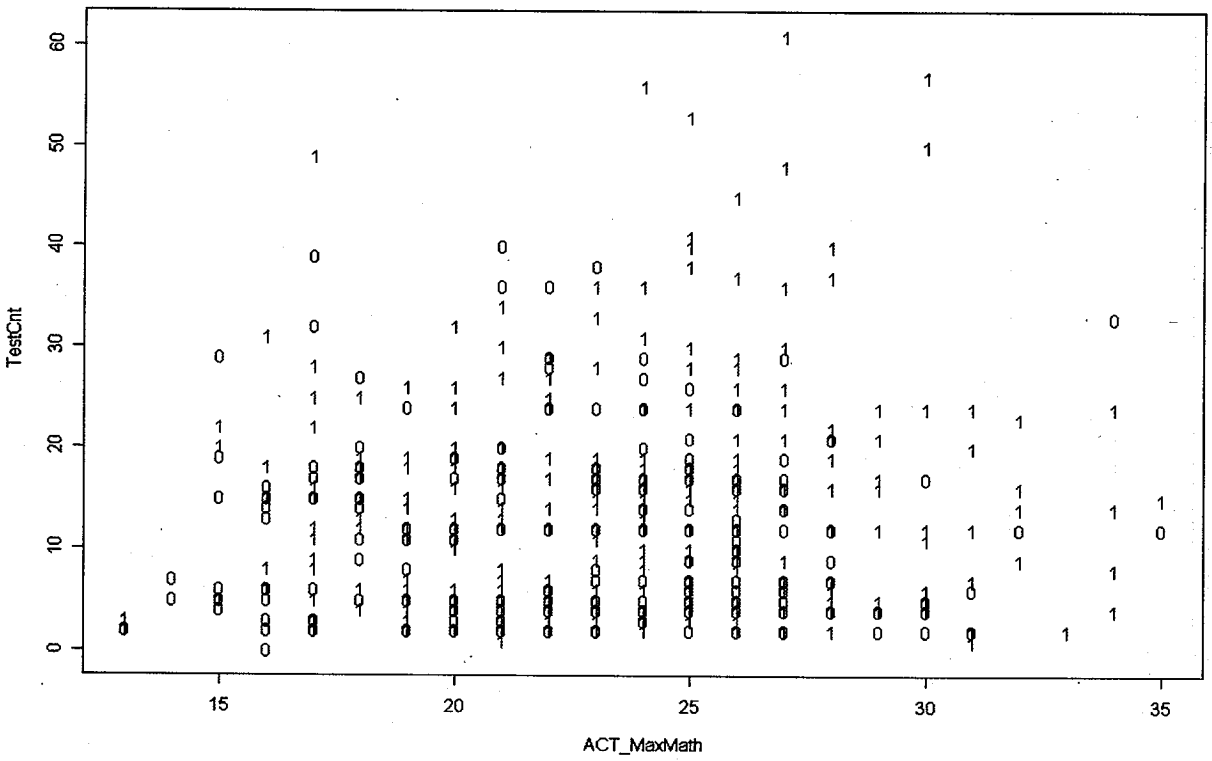
Plot 1:



Plot 2:


TestCnt
HS_GPA
0
10
20
30
40
50
60
2.0
2.5
3.0
3.5
4.0

Plot 3:


TestCnt
ACT_MaxMath
0
10
20
30
40
50
60
15
20
25
30
35

In order to deal with the collinearity concerns a classification tree would make apparent natural breaks in the level of predictors when classifying a student as one who is or is not likely to be retained. Again beginning with resident students, R's rpart() function finds those cut-points in the data where students are retained. In the following R output, the variables used in the classification tree are shown. Next, R displays the error rate using that particular tree. Then the graphic of the tree is shown. Follow the tree according to the decision rules and at the terminus the predicted probability of retention is printed.

For example, notice for resident Freshmen the error for classification is 0.21. This rate shows that using this classification tree with its particular number of branches incorrectly determines whether a students is retained or not with a predicted probability of 0.21. In other words, it produces correct results for almost 80 of 100 new resident Freshmen.

Consider a particular potential student with a high school GPA of 3.25, ACT Verbal of 25, an ACT Math of 29, no Pell amount, and who appears in the test database 45 times. Starting at the top of a tree, determine if high school GPA is above or below 3.305. Move to the left to the next cut-point. The GPA is above 2.655, so go right. Move to the right again because this student is not eligible for Pell grant dollars. Right once again to determine that this student is likely to be retained. The predicted probability is about 0.65.

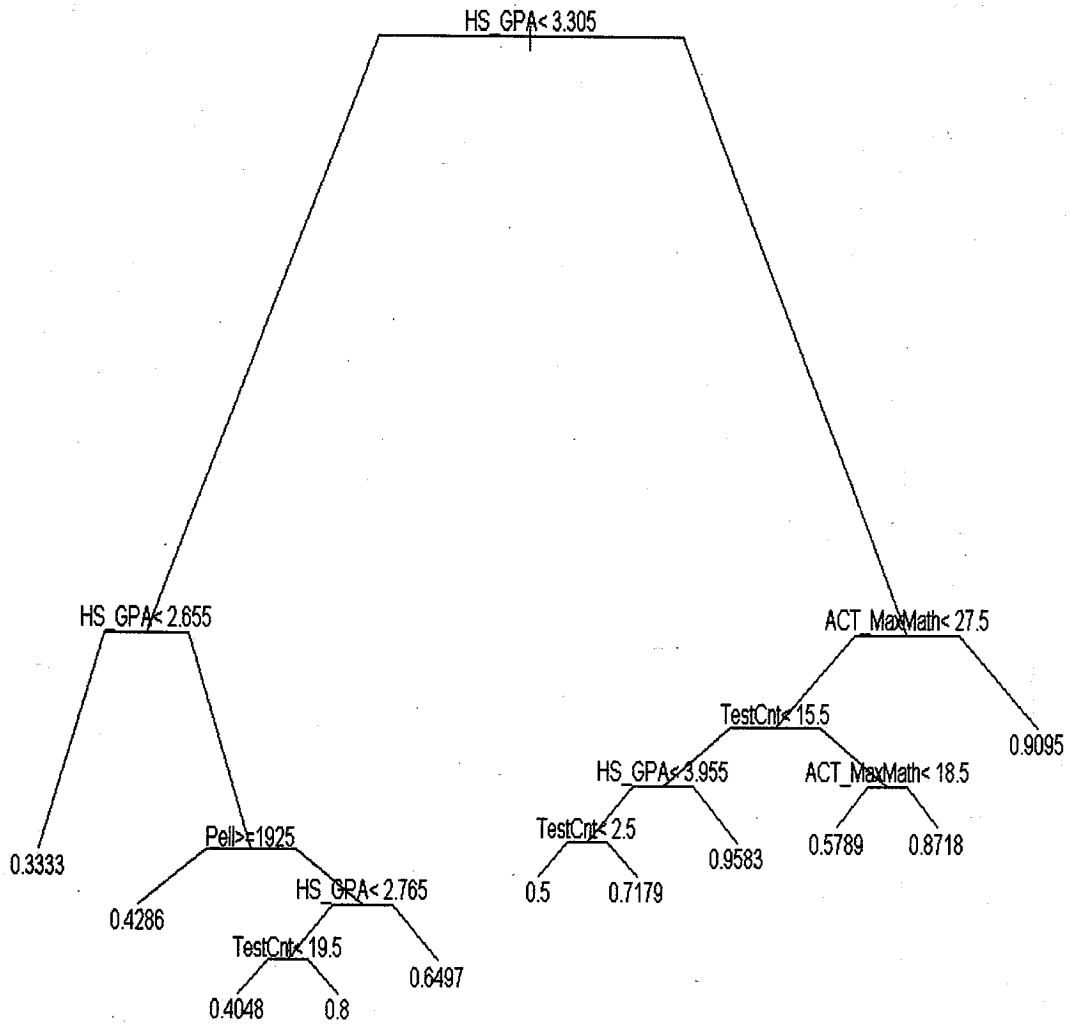
Output for each residency status is available on subsequent pages.

Resident 2005 Freshmen:

```
Regression tree:
rpart(formula = Retain ~ ., data = res, cp = 0.004)

Variables actually used in tree construction:
[1] ACT_MaxMath HS_GPA      Pell        TestCnt

Root node error: 303.92/1435 = 0.21179

n= 1435
```

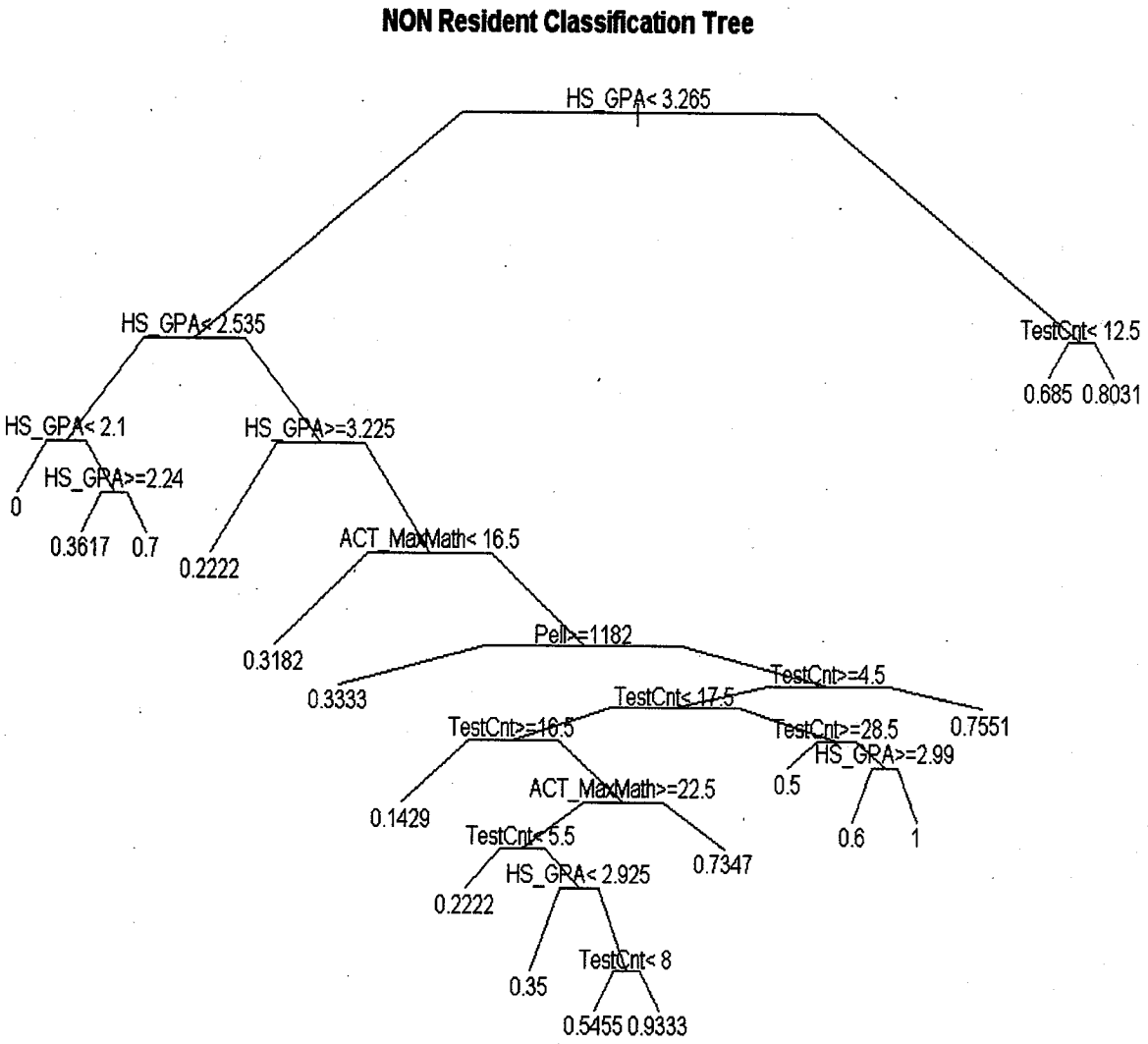
The next figure represents the resultant classification tree. The values at the nodes represent the predicted probability of retention for that classification.

Resident Classification Tree



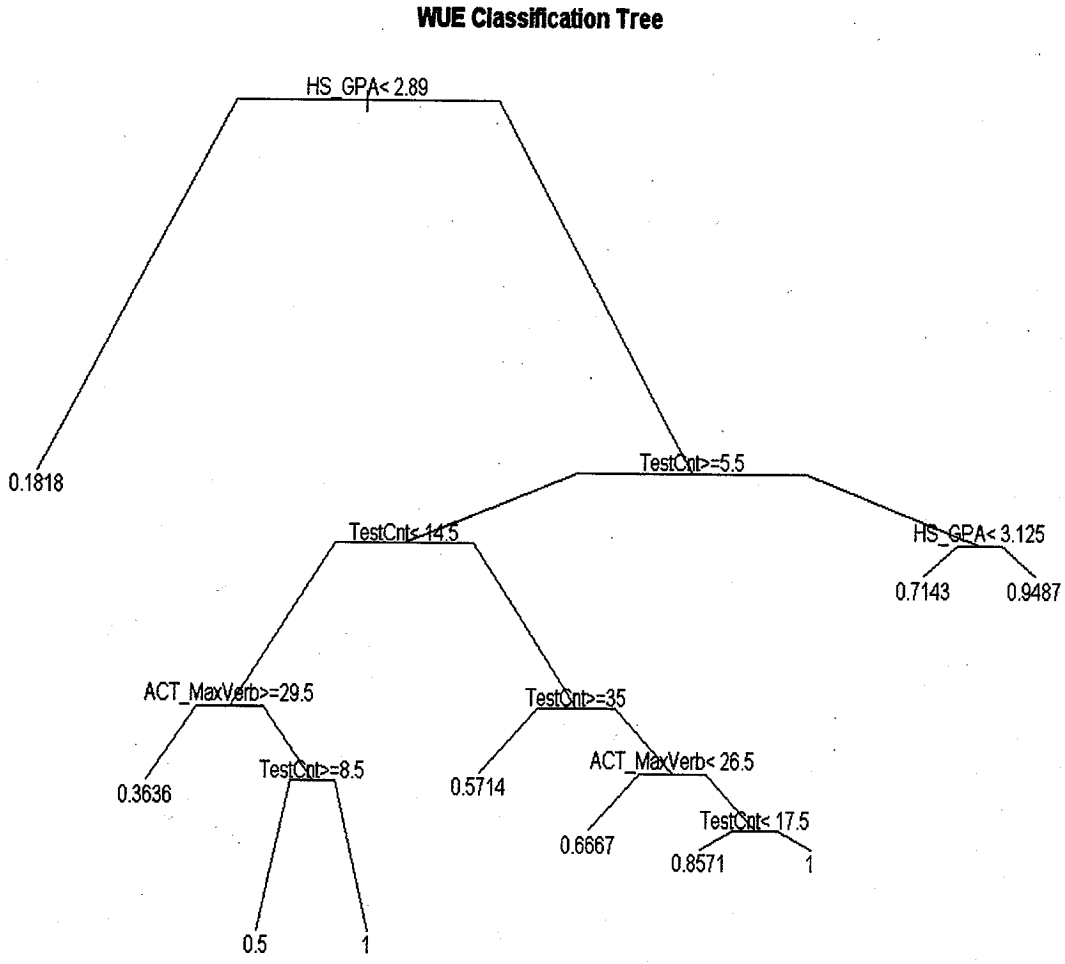
Nonresidents:

```
Regression tree:
rpart(formula = Retain ~ ., data = nonres, cp = 0.004)

Variables actually used in tree construction:
[1] ACT_MaxMath ACT_MaxVerb HS_GPA      Pell        TestCnt

Root node error: 131.96/569 = 0.23192

n= 569
```



WUE:

```
Regression tree:
rpart(formula = Retain ~ ., data = wue, cp = 0.006)

Variables actually used in tree construction:
[1] ACT_MaxVerb HS_GPA      TestCnt

Root node error: 28.168/155 = 0.18173

n= 155
```



Throughout the logistic regression and continuous response modeling and when developing classification trees, one consistent theme appears. Test scores are relatively poor predictors of academic success at MSU measured by retention and by first-term GPA. High school GPA seemed to out perform test scores for the regression setting in each case. Also, there are very few occurrences of cut points involving test scores in the classification trees. For the Freshmen of 2005, it appears that it is more important to examine the number of tests a student takes than to look at the best score the student achieved on those exams. Certainly more analysis is necessary to make more certain conclusions. In particular, models to investigate a curvature effect should be considered. For example, as seen on the graphs, once high school GPA is above 3.90, the likelihood of being retained seems to level off. This effect indicates that better model might be achieved by accounting for that response. Also, there still may be other very strong predictors yet to be considered for success at MSU. But the likely end result, even after extensive analysis of more variables of the Fall 2005 cohort, is that there is a quickly approached limit to predicting academic success. Having substantially more observations still may come up short because certainty in predicting behavior of 18 and 19 year old students may not greatly improve simply by including more of them.

References

R 2.2.1 – The R Development Core Team, Language and Environment Copyright, 2005.

Faraway, Julian J. (2005). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*, Chapman & Hall/CRC.

Advanced Regression STAT 506 Course Notes. Robison-Cox, James F.

Data Available from Montana State University Office of Planning and Analysis. Information was loaded in SCT Banner by the excellent staff in the MSU Office of Admissions and Office of Financial Aid, as well as MSU's Registrar's Office.