USING DATA MINING TO PREDICT FRESHMEN OUTCOMES

Nora Galambos, PhD
Senior Data Scientist
Office of Institutional Research, Planning & Effectiveness
Stony Brook University

Abstract

Data mining is used to develop models for the early prediction of freshmen GPA. Since student engagement has long been associated with student success, the use of service utilization and transactional data is examined along with more traditional student factors. Factors entered into the data mining models include advising visits, freshmen course-taking activity, interactions with the college learning management system, and college activity participation, along with SAT scores, high school GPA, demographics, and financial aid. In models predicting first semester freshmen GPA, factors associated with students' interactions with the campus environment were stronger predictors than SAT scores.

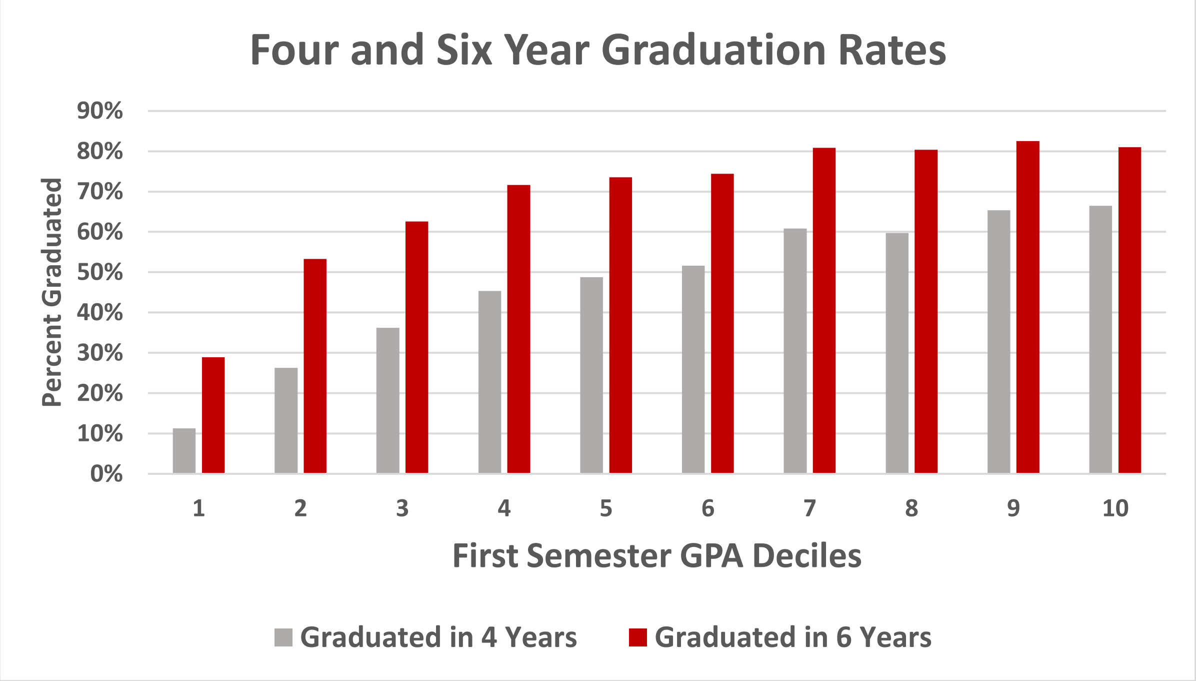
Introduction

The goal is to develop a model to predict at risk first-time full-time freshmen as early as possible in their college careers in order to assist them with interventions. Traditional methods of logistic and linear regression are often good at identifying factors significantly associated with an outcome, but are not always able to make accurate predictions. Linear and logistic regression have one set of predictors to model the outcomes of all of the students in the data and do not assign separate sets of predictors to students having very different characteristics. For example, first-time freshmen entering college with high SAT scores may have very different retention and college GPA predictors than those entering with a low high school GPA and low SAT scores. Inevitably, when using any model, some students will be incorrectly assigned, with some students miss-identified as being at risk or students at risk being not being identified as such by the model. There is an allocation trade-off when resources are expended on students not really in need of interventions or when students who would potentially benefit from interventions do not receive them. Methods capable of more accurate predictions will result in more effective utilization of resources, and higher retention and graduation rates. For that reason the decision was made to explore data mining, because it offers a variety of methods for utilizing different types of data, there are few assumptions to satisfy relative to traditional hypothesis driven methods, and it is able to handle a great volume of data with hundreds of predictors.

At our institution poor academic performance by first-time full-time freshmen in the first semester has a negative impact on graduation and retention outcomes. Figure 1 illustrates that only 11% of students in the lowest GPA decile graduate in four years, and less than 29% of students in that group graduate in six years. For the second decile the four year rate increases to 26% and the six year rate improves to 53%. Those rates, though higher, are still very low relative to the top half of the freshmen class.

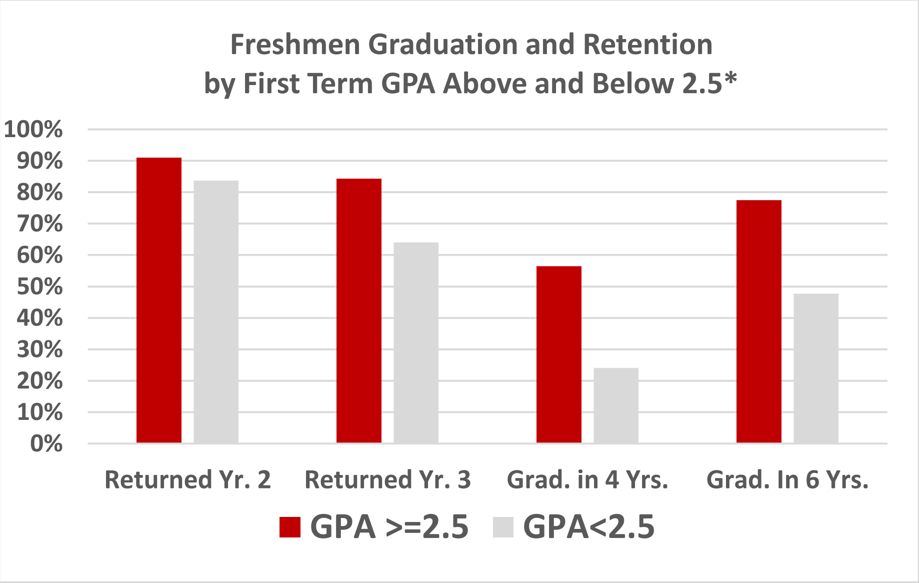
Approximately 30% of first-time full-time freshmen received a GPA below 2.5 in their first semester (Figure 2). Almost 84% of those students returned in year two, however by the next year the retention rate had dropped substantially with only 64% returning for year three and only 48% graduating in six years. In contrast over 77% of students receiving a GPA of 2.5 or greater in their first semester graduated in six years.

Figure 1. Four and Six Year Graduation Rates of First-Time Full-Time Freshmen by GPA Deciles*



*The fall freshmen cohorts of 2006 through 2008 were combined.

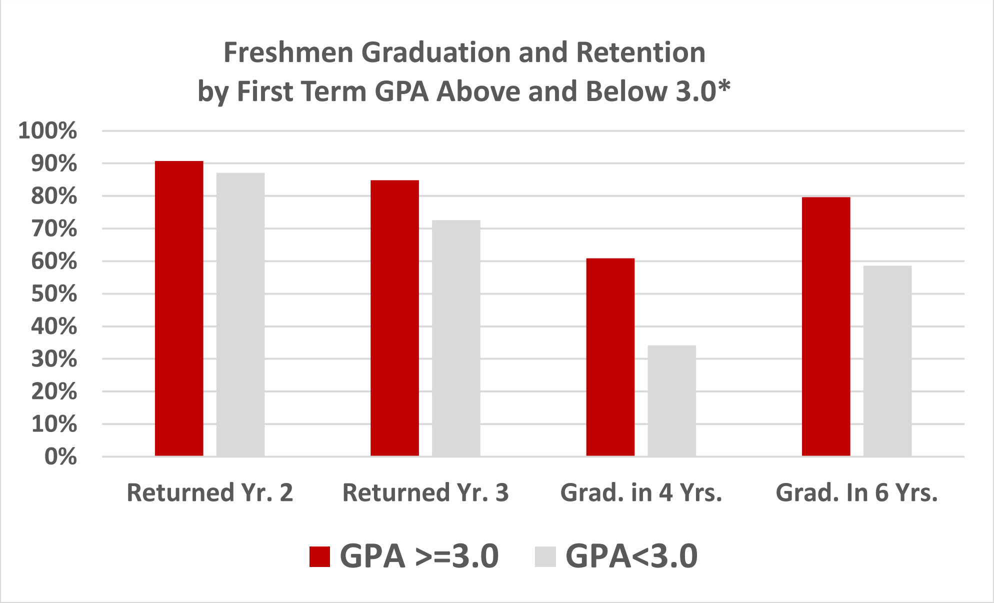
Figure 2. Comparison of Graduation and Retention Rates of First-time Full-time Freshmen by First Semester GPA Above and Below 2.5.



*The fall freshmen cohorts of 2006 through 2008 were combined.

Even when evaluating results for students above and below 3.0 the differences are dramatic (Figure 3). Only 34% of students with a first semester GPA below 3.0 (approximately the median) graduated in four years, which is almost 27 percentage points lower than students above the median.

Figure 3. Comparison of Graduation and Retention Rates of First-time Full-time Freshmen by First Semester GPA Above and Below 3.0.



*The fall freshmen cohorts of 2006 through 2008 were combined.

Given these results we see that it would greatly benefit at risk students if they could be identified as early as possible. In order for the programs to be cost effective and, more importantly, a good match for the needs of the students, the model must be able to make very accurate predictions. The difficulty of this task lies in the fact that there are not many university-level academic measures available on or before the middle of the first semester of the freshmen year. For that reason we have explored the development of a data mining model that combines transactional data such as learning management system (LMS) logins and service utilization such as advising and tutoring center visits with other more traditional measures in an attempt to identify at risk students *before* any grades appear on their transcripts.

Literature Review

The study has cast a wide net in terms of assembling a variety of data for use in studying academic, social, and economic factors to determine elevated risk of a low GPA, which can translate to increased risk of early attrition or longer time to degree. Consistent with the retention study of Tinto (1987), we evaluate many types of data representing students' interactions with their campus environment to determine if higher levels of campus engagement are predictive of improved freshmen outcomes. These measures of engagement include interactions with the learning management system, intramural sports and fitness class participation, and academic advising and tutoring center visits. It appears that students who are identified to be at risk in their first term and remain at the institution, continue to be at risk, with greater numbers leaving in the subsequent term (Singell and Waddell 2010). This is consistent with the results at our institution which are presented in Figures 1, 2, and 3. Methods capable of more accurate predictions will result in more effective utilization of campus resources, and higher retention and graduation rates. Course-taking behavior is also important, particularly math readiness. Herzog (2005) found math readiness to be "more important than aid in explaining freshmen dropout and transfer-out during both first and second semesters." Herzog also focused on both merit and need-based aid and the role that interaction of aid and academic preparedness plays in student retention. Living within a 60 mile radius of the institution, the percent of students at a high school who take the SAT, along with the percentage at the high school receiving free lunches was explored by Johnson (2008) underlining the need to examine the role of the secondary school and socio-economic factors in developing a model. Persistence increases among students closer to the institution and not surprisingly, decreases among those who were from schools having a high percentage of students receiving free school lunches. The role of differing stop-out patterns exhibited by grant, work-study, and loan recipients (Johnson 2010) demonstrated that grants have the highest positive effect on persistence, but its effect decreases more than that of loans after controlling for other factors. Resource utilization was studied (Robbins et al. 2009) using a tracking system. Services and resources were grouped into academic services, recreational resources, social measures and advising sessions, with all but social measures demonstrating positive associations with GPA even after controlling for other demographic and risk factors. These papers have demonstrated that researchers are examining a range of factors in studying and modeling risk. This research underlines that fact that student success is the result of complex interactions between student engagement, academic service utilization, financial metrics, and demographics, which are combined with student academic characteristics that include high school GPA and SAT scores. Data mining is ideal for developing a model with a large diverse number of predictors.

Data Sources

An attempt was made to include as many types of data as possible, so learning management system logins, not previously explored by our institution were included. Building the dataset began with the traditional measures such as demographics (gender, ethnicity, and geographic area of residence when admitted), to which were added high school GPA and SAT scores. In order to control for high school GPA, the average SAT scores of the high schools were incorporated. Because we are modeling the freshmen GPA at the mid-semester point, in terms of college academic characteristics we only have available the fall semester courses in which the students are enrolled, the area the major, whether a major has been declared, and how many college credits were accepted by the institution upon admission. The number of AP credits received was also captured, with those credits separated into STEM and non-STEM totals.

To explore the effect of high failure rate courses on student outcomes, courses with enrollments of 70 or more students having 10% or more D, F, or W grades were identified and categorized as STEM or non-STEM courses. The total number of high DFW-rate courses, and the highest DFW rate for each student (by STEM indicator) was included in the model. The percentage of freshmen in each DFW course was also tabulated and that percentage for the corresponding course was additionally added. The rationale for examining the percentage of freshmen in these difficult courses is that if the courses are populated by large numbers of upper level students, it may make the course even more difficult for freshmen who are less experienced.

Since student engagement has long been associated with student success, the use of service and academic utilization data was included to determine if it resulted in improved models. Student interactions with the university's learning management system, academic advising, tutoring center visits, intramural sports, and fitness classes, have been incorporated in the analysis to evaluate the association of GPA with students' engagement in the university environment.

Much of the data pertaining to interactions with student services and learning management system logins has not been stored long term. In fact the LMS login data was not available for any fall semester prior to fall 2014. As a result, part of the data mining process has included the initial collecting, saving, and storing of the data. Programs are being developed to automate the formatting and aggregation of the transactional data so it can easily be merged with student records and utilized in the data mining process. For modeling use of the LMS logins, only one login per course per hour was counted, so an individual course can have at most 24 logins per day. This eliminated multiple logins that occurred just few minutes and sometimes a few seconds apart. Further, the courses were categorized as STEM or non-STEM. Next the STEM and non-STEM logins were totaled for week 1 and separately for weeks 2 through 6. Finally the STEM and non-STEM logins were divided by their respective STEM and non-STEM course totals to obtain per-course login rates.

Financial aid data was also assembled. The measures that were captured are the expected family contribution, adjusted gross income (AGI), types and amounts of disbursed aid (athletics aid, loans, grants, scholarships, and work-study). Pell Grants and the Tuition Assistance Program (TAP) recipients were also added to the model.

Because the data mining initiative is new and many data sources are being collected and explored for the first time, research and evaluation of the methods for summarizing and using the data in the model is ongoing. The expectation is that additional data sources will be added. A detailed list of the data elements can be found in the appendix.

Methodology

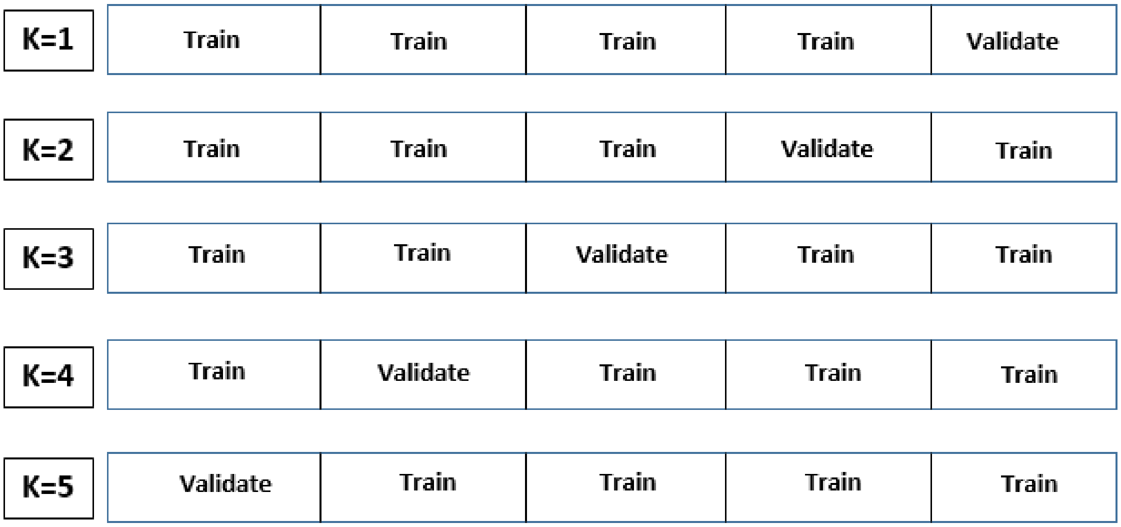
Different models were compared to find the ones that provide the most accurate prediction of the first semester GPA with the lowest average squared errors (ASE)[1] . In developing data mining models it is advisable to partition the data into training and validation

[1] ASE = SSE/N or ASE = (Sum of Squared Errors)/N

sets. The training set is used for model development, then the model is run on the validation set to check its accuracy and calculate the prediction error. It is also important to avoid developing an overly complex model, overfitting. If the model is too complex it can be influenced by random noise, and if there are outliers an overly complex model may be fit to them. Unfortunately, when using such a model on new data its ability to accurately predict the outcomes will be diminished. One way of detecting overfitting is to compare the ASE of the training and validation data. A large increase in the ASE when running the model on the validation data may be a sign of overfitting. However, with less than 3,000 subjects and over 50 variables to predict the GPA's of the bottom 20% of the class, setting aside 40% of the data as is typical for a validation set, is not practical because it would not leave enough of the lower GPA students for building the model. As an alternative, k-fold cross validation was used. It works with limited amounts of data, and its initial steps are similar to traditional analysis. The entire dataset is used to choose the predictors and the error is estimated by averaging the error of the *k* test samples. In subsequent years, when more than one semester of LMS data has been collected, the easier to implement training-validation-partitioning method can be used.

To implement k-fold cross validation, the dataset is divided into *k* equal groups or folds. In this case five folds were used. Four groups are taken together and are used to train the data and one is used for validation. The procedure is repeated five times, each time leaving out a different set for validation as in Figure 4. The model error is estimated by averaging the errors of the five validation samples.

Figure 4: K-fold cross-validation sampling design.



K=1	Train	Train	Train	Train	Validate
K=2	Train	Train	Train	Validate	Train
K=3	Train	Train	Validate	Train	Train
K=4	Train	Validate	Train	Train	Train
K=5	Validate	Train	Train	Train	Train

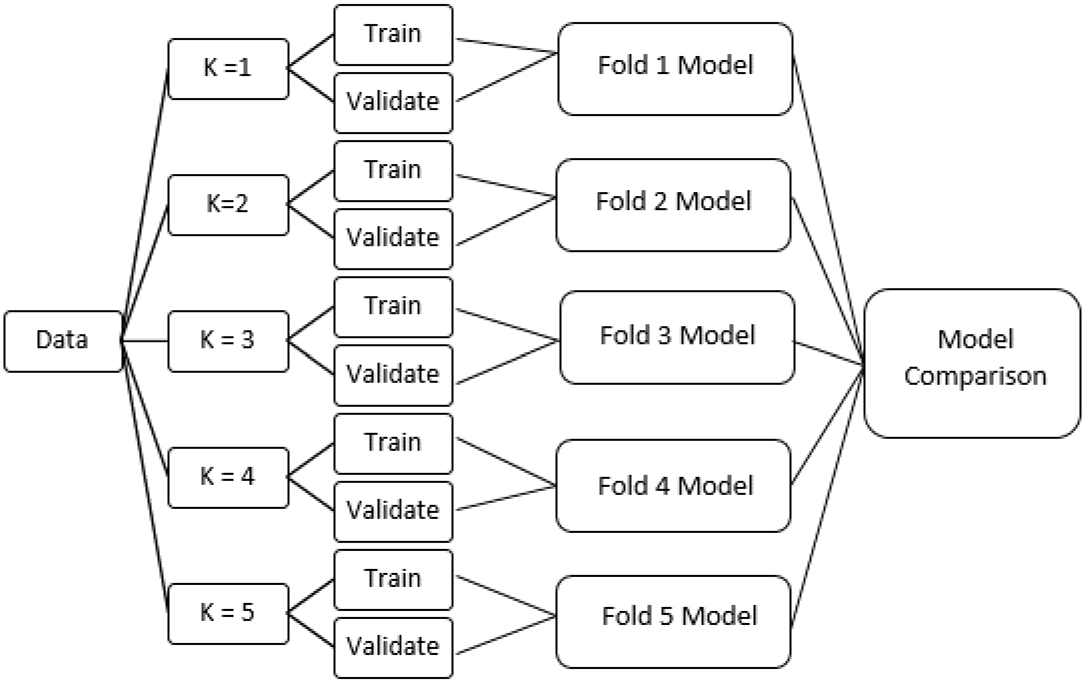
Five different modeling methods were tested and compared using k-fold cross validation. A general data mining diagram for running a modeling method with k-fold cross validation can be seen in Figure 5. Filters can be applied to select the proper groups for the validation and training sets for each fold, then the training and validation sets are sent to the modeling nodes where the same modeling method is run for each of the five training sets. The model is then run on each validation set for calculating the error. A model comparison node provides the relevant model evaluation statistics for each of the five folds.

The five different methods used to develop predictive models were: CHAID[2] (chi-square automatic interaction detection), BFOS-CART (the classification and regression tree method; Breiman, Friedman, Olshen, and Stone, 1984), a general decision tree, gradient boosting, and linear regression.   Each model was developed to predict the first semester GPA of the first-time

[2] The CHAID and CART methods have been closely approximated by using Enterprise Miner settings. SAS Institute Inc. 2014. *SAS® Enterprise Miner™ 13.2: Reference Help*. Cary, NC: SAS Institute Inc. p. 755-758.

full-time fall 2014 freshmen cohort. The average squared errors (ASE) of the five validation samples for each method were averaged and compared with the average errors of the training samples to check for overfitting and to find the method with the smallest error.

Figure 5. A general data-mining diagram for running 5-fold cross-validation to evaluate the accuracy of a model.



With the exception of linear regression, the methods tested were decision tree-based methods. The CART method begins by doing an exhaustive search for the best binary split. It then splits categorical predictors into a smaller number of groups or finds the optimal split in numerical measures. Each successive split is again split in two until no further splits are possible. The result is a tree of maximum possible size, which is then pruned back algorithmically. For interval targets the variance is used to assess the splits; for nominal targets the Gini impurity measure is used. Pruning starts with the split that has the smallest contribution to the model and missing data is assigned to the largest node of a split. This method creates a set of nested binary decision rules to predict an outcome.

Unlike CART with binary splits evaluated by the variance or misclassification measures, the CHAID algorithm uses the chi-square test (or the F test for interval targets) to determine significant splits and finds independent variables with the strongest association with the outcome. A Bonferroni correction to the p-value is applied prior to the split. CHAID may find multiple splits in continuous variables, and allows splitting of categorical data into more than two categories. This may result in very wide trees with numerous nodes at the first level. As with CART, CHAID allows different predictors for different sides of a split. The CHAID algorithm will halt when statistically significant splits are no longer found in the data.

The software was also configured to run a general decision tree that does not conform or approximate mainstream methods found in the literature. To control for the large number of nodes at each level, the model was restricted to up to four-way splits (4 branches), as opposed to CHAID which is finds and utilizes all significant splits and CART which splits each node in two. The F test was used to evaluate the variance of the nodes and the depth of the overall tree was restricted to 6 levels. Missing values were assigned to produce an optimal split with the ASE used to evaluate the subtrees. The software's cross validation option was selected in order to perform the cross validation procedure for each subtree. That results in a sequence of estimates using the cross validation method explained earlier to select the optimal subtree.

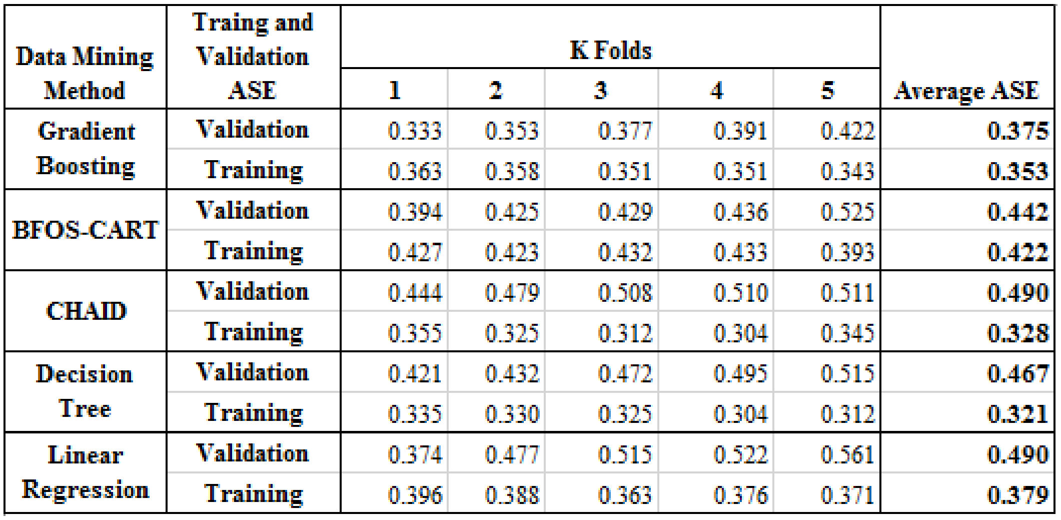
The final tree method was gradient boosting which uses a partitioning algorithm developed by Jerome Friedman. At each level of the tree the data is resampled a number of times without replacement. A random sample is drawn at each iteration from the training data set and the sample is used to update the model. The successive resampling results in a weighted average of the re-sampled data. The weights assigned at each iteration improve the accuracy of the predictions. The result is a series of decision trees, each one adjusted with new weights to improve the accuracy of the estimates or to correct the misclassifications present in the previous tree. Because the results at each stage are weighted and combined into a final model, there is no resulting tree diagram. However, the scoring code that is generated allows the model to be used to score new data for predicting outcomes.

The final method tested was linear regression. The discussion that follows highlights some of the difficulties in implementing linear regression in a data mining environment. Decision tree methods are able to handle missing values by combining them with another category or using surrogate rules to replace them. Linear regression, on the other hand, will listwise delete the missing values. Data in this study was obtained from multiple campus sources, and as such, many students did not have any records for some predictors. For example, students who did not apply for financial aid will have missing data on financial aid measures, a small percentage of the entering freshmen do not have SAT scores, and some students may not have courses utilizing the LMS. These measures result in an excessive amount of data being listwise deleted. The software has an imputation node that can be configured to impute missing data. For this study the distribution method was used whereby replacement values are calculated from random percentiles of the distributions of the predictors. There are many imputation methods and a thorough study of missingness for such a large number of variables is very time consuming. If the linear regression method appeared promising, other imputation methods would be explored and studied in greater detail. Another issue of concern in the linear regression analysis was multicollinearity. That is another issue that can take time to address thoroughly. For this analysis clustering was employed to reduce multicollinearity. With a large volume of predictors, it would be difficult and time consuming to evaluate all of the potential multicollinearity issues, so the software clustering node was used to group highly correlated variables. In each cluster, the variable with the highest correlation coefficient was retained and entered into the modeling process, and the others were eliminated.

Results

Gradient boosting had the smallest average ASE followed by that of CART (Table 1). Additionally, gradient boosting and BFOS-CART, on average, had the smallest differences between the validation and training errors. Those absolute errors were both approximately 0.02, while for the other methods it was greater than 0.1. Gradient boosting had the lowest average validation error, 0.375, while CHAID and linear regression had the highest at 0.49. Though gradient boosting had the lowest average validation ASE, the CART method was chosen for the modeling process. Close inspection of the CART results did not show evidence of any problems with the fit of the model, and it had a relatively low average ASE. The main reason for choosing the CART model is that gradient boosting, without an actual tree diagram, would make the

Table 1. Average Squared Error (ASE) Results for the Five Data Mining Methods

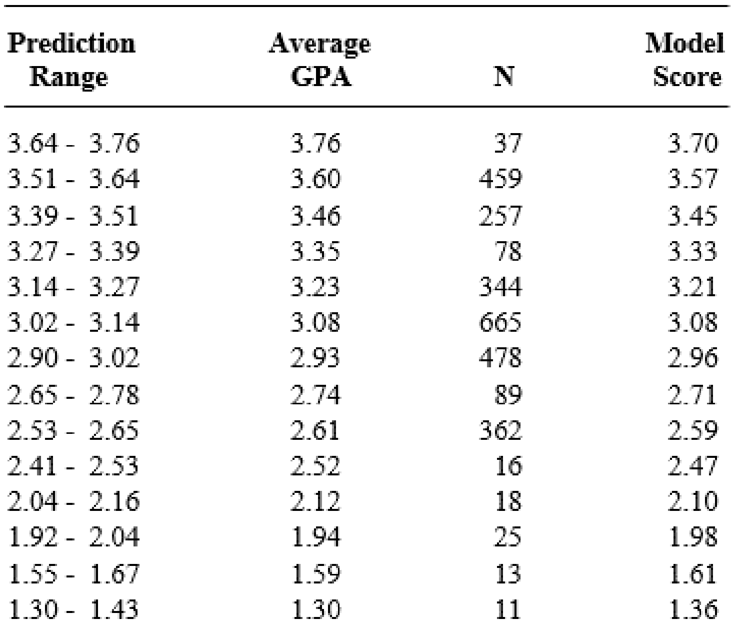


Data Mining Method	Traing and Validation ASE	K Folds					Average ASE
		1	2	3	4	5	
Gradient Boosting	Validation	0.333	0.353	0.377	0.391	0.422	**0.375**
	Training	0.363	0.358	0.351	0.351	0.343	**0.353**
BFOS-CART	Validation	0.394	0.425	0.429	0.436	0.525	**0.442**
	Training	0.427	0.423	0.432	0.433	0.393	**0.422**
CHAID	Validation	0.444	0.479	0.508	0.510	0.511	**0.490**
	Training	0.355	0.325	0.312	0.304	0.345	**0.328**
Decision Tree	Validation	0.421	0.432	0.472	0.495	0.515	**0.467**
	Training	0.335	0.330	0.325	0.304	0.312	**0.321**
Linear Regression	Validation	0.374	0.477	0.515	0.522	0.561	**0.490**
	Training	0.396	0.388	0.363	0.376	0.371	**0.379**

results much more difficult to explain, use, and visualize. Having a set of student characteristics assigned to each node, as well as the ability to graphically display the decision tree adds to the utility of the CART model. Once the CART method was selected, the model was run again using all of the data, and scoring output was created.

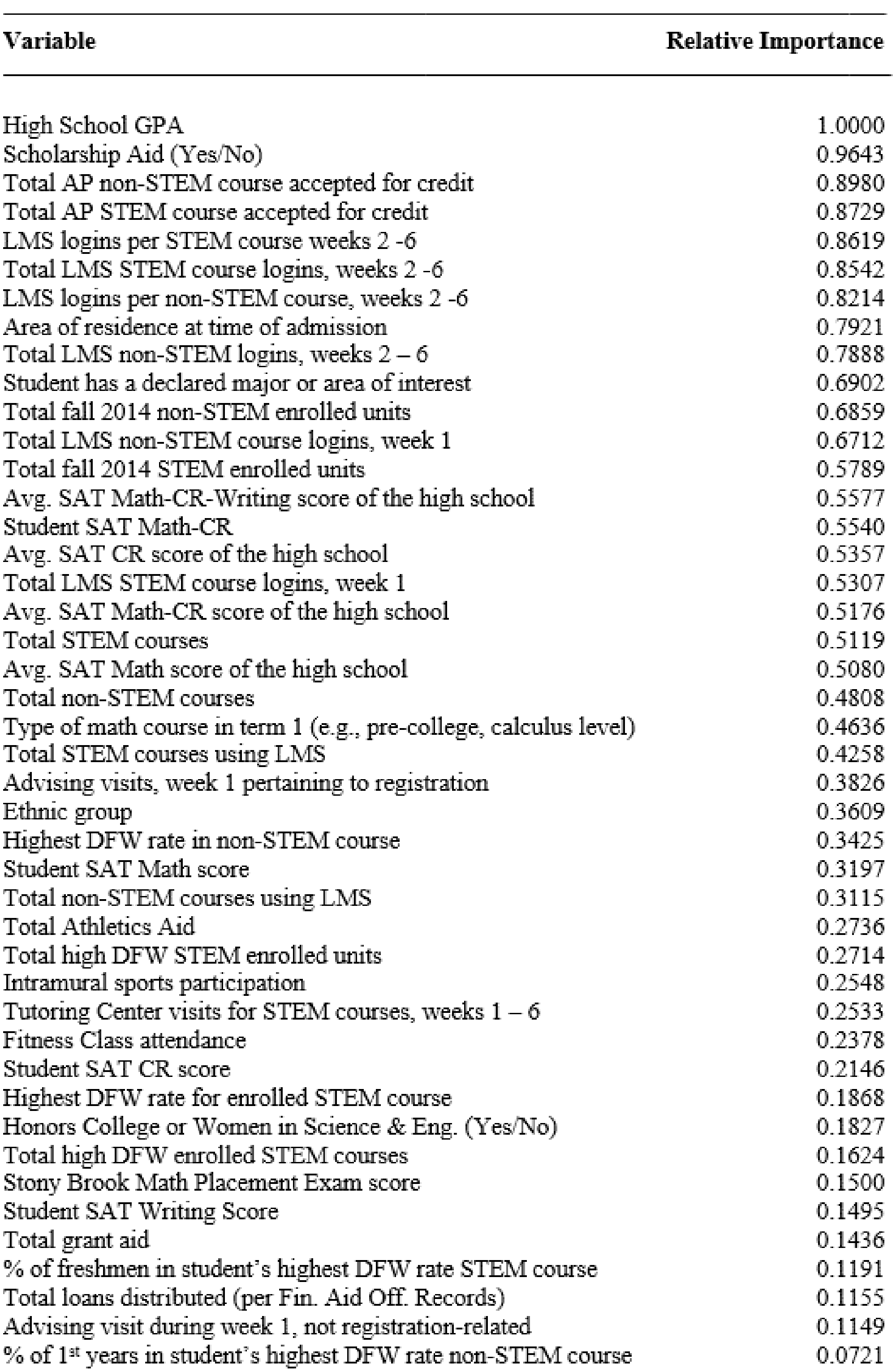
The score distribution table, Figure 2, which is part of the decision tree output allows us to view the frequencies of the model predictions. Twenty bins, the prediction ranges, are created by evenly dividing the interval between the lowest and highest predictions, 1.30 and 3.76. (Intervals without students are not listed.) The model score is calculated by taking the mid-point of the prediction range. The average GPA column contains the average GPA of the N students in the data that fall within the given range. The table can aid us in choosing GPA cut points for different interventions since it shows the number of students at the various prediction levels.

Table 2. Score Distribution Table



Prediction Range	Average GPA	N	Model Score
3.64 - 3.76	3.76	37	3.70
3.51 - 3.64	3.60	459	3.57
3.39 - 3.51	3.46	257	3.45
3.27 - 3.39	3.35	78	3.33
3.14 - 3.27	3.23	344	3.21
3.02 - 3.14	3.08	665	3.08
2.90 - 3.02	2.93	478	2.96
2.65 - 2.78	2.74	89	2.71
2.53 - 2.65	2.61	362	2.59
2.41 - 2.53	2.52	16	2.47
2.04 - 2.16	2.12	18	2.10
1.92 - 2.04	1.94	25	1.98
1.55 - 1.67	1.59	13	1.61
1.30 - 1.43	1.30	11	1.36

Table 3. Variable Importance Table.



Variable	Relative Importance
High School GPA	1.0000
Scholarship Aid (Yes/No)	0.9643
Total AP non-STEM course accepted for credit	0.8980
Total AP STEM course accepted for credit	0.8729
LMS logins per STEM course weeks 2 -6	0.8619
Total LMS STEM course logins, weeks 2 -6	0.8542
LMS logins per non-STEM course, weeks 2 -6	0.8214
Area of residence at time of admission	0.7921
Total LMS non-STEM logins, weeks 2 – 6	0.7888
Student has a declared major or area of interest	0.6902
Total fall 2014 non-STEM enrolled units	0.6859
Total LMS non-STEM course logins, week 1	0.6712
Total fall 2014 STEM enrolled units	0.5789
Avg. SAT Math-CR-Writing score of the high school	0.5577
Student SAT Math-CR	0.5540
Avg. SAT CR score of the high school	0.5357
Total LMS STEM course logins, week 1	0.5307
Avg. SAT Math-CR score of the high school	0.5176
Total STEM courses	0.5119
Avg. SAT Math score of the high school	0.5080
Total non-STEM courses	0.4808
Type of math course in term 1 (e.g., pre-college, calculus level)	0.4636
Total STEM courses using LMS	0.4258
Advising visits, week 1 pertaining to registration	0.3826
Ethnic group	0.3609
Highest DFW rate in non-STEM course	0.3425
Student SAT Math score	0.3197
Total non-STEM courses using LMS	0.3115
Total Athletics Aid	0.2736
Total high DFW STEM enrolled units	0.2714
Intramural sports participation	0.2548
Tutoring Center visits for STEM courses, weeks 1 – 6	0.2533
Fitness Class attendance	0.2378
Student SAT CR score	0.2146
Highest DFW rate for enrolled STEM course	0.1868
Honors College or Women in Science & Eng. (Yes/No)	0.1827
Total high DFW enrolled STEM courses	0.1624
Stony Brook Math Placement Exam score	0.1500
Student SAT Writing Score	0.1495
Total grant aid	0.1436
% of freshmen in student's highest DFW rate STEM course	0.1191
Total loans distributed (per Fin. Aid Off. Records)	0.1155
Advising visit during week 1, not registration-related	0.1149
% of 1st years in student's highest DFW rate non-STEM course	0.0721

Table 3 lists the relative importance measure for variables that were entered into the modeling process. The relative importance measure is evaluated by using the reduction in the sum of squares that results when a node is split, summing over all of the nodes.[3] In the variable importance calculation when variables are highly correlated they will both receive credit for the sum of squares reduction, hence the relative importance of highly correlated variables will be about the same. For that reason some measures may rank high on the variable importance list, but do not appear as a predictors in the decision tree.

On Table 3 high school GPA is highest on the variable importance list for predicting freshmen GPA when modeled mid-semester, followed by whether or not a student received a scholarship. Next are AP STEM and non-STEM courses accepted for credit, and then LMS system logins. A student's combined SAT Math and Critical Reading Exam Score is 15th on the list just behind the high school average score for the combined SAT Math, Critical Reading, and Writing exam. Some other measures that exceeded SAT scores in relative importance are whether a student has a declared major, and the geographic area of residence when admitted.

The decision tree generated by the model is presented in two parts in Figures 6 and 7. The CART method, employing only binary splits as previously discussed, selected high school GPA for the first branch of the tree modeling first semester freshmen GPA. High school GPA was split into two nodes, less than or equal to 92.0, and greater than 92.0 or missing. Figure 6 displays the portion of the decision tree with high school GPA less than or equal to 92.0 and Figure 7 has the portion of the tree with high school GPA greater than 92.0 or missing.

[3] . SAS Institute Inc. 2014. *SAS® Enterprise Miner™ 13.2: Reference Help*. Cary, NC: SAS Institute Inc. p. 794.

Figure 6. Part 1 of the CART Decision Tree Model Predicting Freshmen GPA for Students Having a High School GPA <= 92.0.

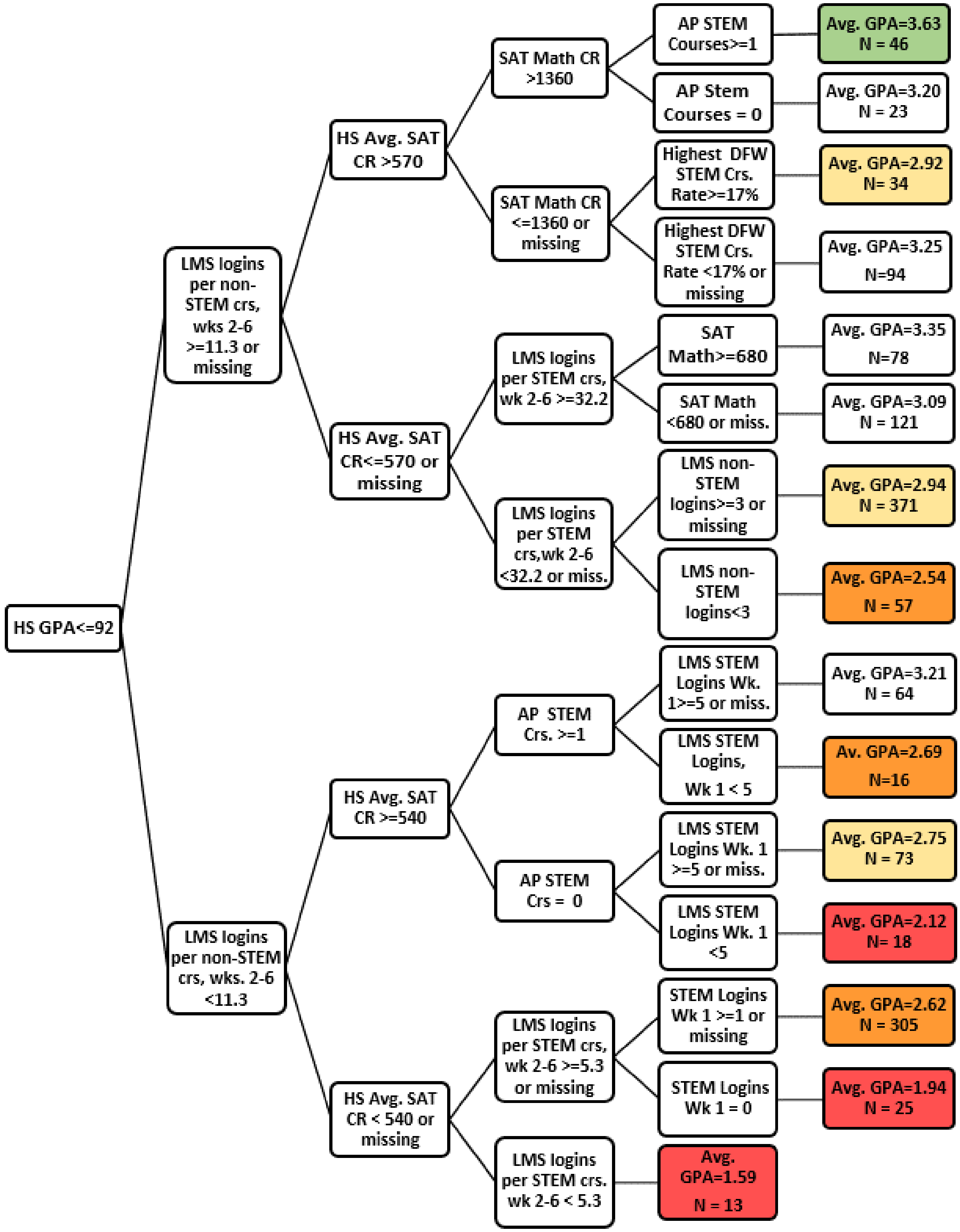
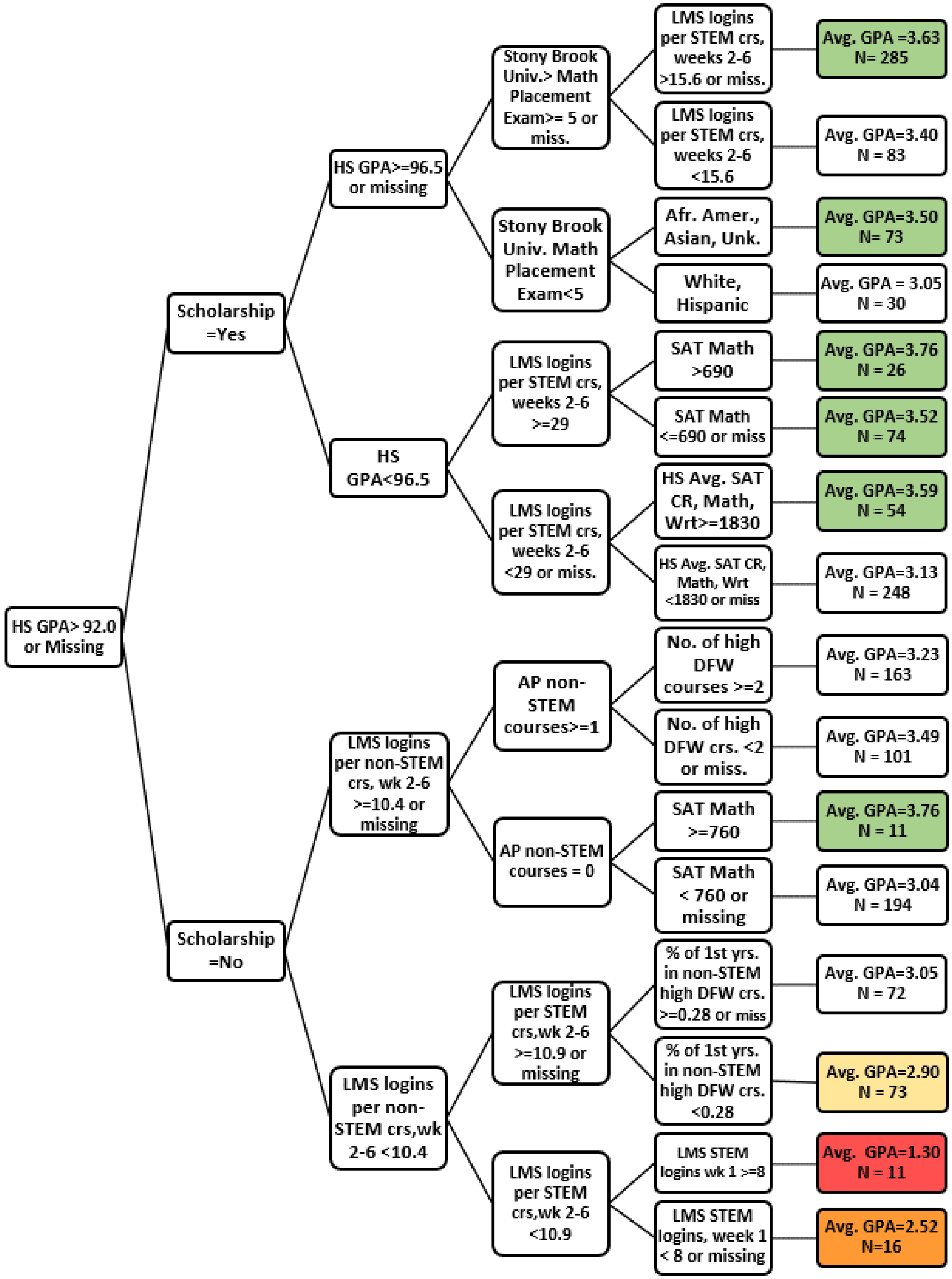

HS GPA<=92
LMS logins per non-STEM crs, wks 2-6 >=11.3 or missing
HS Avg. SAT CR >570
SAT Math CR >1360
AP STEM Courses>=1
Avg. GPA=3.63 N = 46
AP Stem Courses = 0
Avg. GPA=3.20 N = 23
SAT Math CR <=1360 or missing
Highest DFW STEM Crs. Rate>=17%
Avg. GPA=2.92 N= 34
Highest DFW STEM Crs. Rate <17% or missing
Avg. GPA=3.25 N=94
HS Avg. SAT CR<=570 or missing
LMS logins per STEM crs, wk 2-6 >=32.2
SAT Math>=680
Avg. GPA=3.35 N=78
SAT Math <680 or miss.
Avg. GPA=3.09 N = 121
LMS logins per STEM crs,wk 2-6 <32.2 or miss.
LMS non-STEM logins>=3 or missing
Avg. GPA=2.94 N = 371
LMS non-STEM logins<3
Avg. GPA=2.54 N = 57
LMS logins per non-STEM crs, wks. 2-6 <11.3
HS Avg. SAT CR >=540
AP STEM Crs. >=1
LMS STEM Logins Wk. 1>=5 or miss.
Avg. GPA=3.21 N = 64
LMS STEM Logins, Wk 1 < 5
Av. GPA=2.69 N=16
AP STEM Crs = 0
LMS STEM Logins Wk. 1 >=5 or miss.
Avg. GPA=2.75 N = 73
LMS STEM Logins Wk. 1 <5
Avg. GPA=2.12 N= 18
HS Avg. SAT CR < 540 or missing
LMS logins per STEM crs, wk 2-6 >=5.3 or missing
STEM Logins Wk 1 >=1 or missing
Avg. GPA=2.62 N = 305
STEM Logins Wk 1 = 0
Avg. GPA=1.94 N = 25
LMS logins per STEM crs. wk 2-6 < 5.3
Avg. GPA=1.59 N = 13

Figure 7. Part 2 of the CART Decision Tree Model Predicting Freshmen GPA for Students Having a HS GPA > 92.0 or missing.


HS GPA> 92.0 or Missing
Scholarship =Yes
HS GPA>=96.5 or missing
Stony Brook Univ.> Math Placement Exam>= 5 or miss.
LMS logins per STEM crs, weeks 2-6 >15.6 or miss.
Avg. GPA =3.63 N= 285
LMS logins per STEM crs, weeks 2-6 <15.6
Avg. GPA=3.40 N = 83
Stony Brook Univ. Math Placement Exam<5
Afr. Amer., Asian, Unk.
Avg. GPA=3.50 N= 73
White, Hispanic
Avg. GPA = 3.05 N = 30
HS GPA<96.5
LMS logins per STEM crs, weeks 2-6 >=29
SAT Math >690
Avg. GPA=3.76 N = 26
SAT Math <=690 or miss
Avg. GPA=3.52 N = 74
LMS logins per STEM crs, weeks 2-6 <29 or miss.
HS Avg. SAT CR, Math, Wrt>=1830
Avg. GPA=3.59 N = 54
HS Avg. SAT CR, Math, Wrt <1830 or miss
Avg. GPA=3.13 N = 248
Scholarship =No
LMS logins per non-STEM crs, wk 2-6 >=10.4 or missing
AP non-STEM courses>=1
No. of high DFW courses >=2
Avg. GPA=3.23 N = 163
No. of high DFW crs. <2 or miss.
Avg. GPA=3.49 N = 101
AP non-STEM courses = 0
SAT Math >=760
Avg. GPA=3.76 N = 11
SAT Math < 760 or missing
Avg. GPA=3.04 N = 194
LMS logins per non-STEM crs,wk 2-6 <10.4
LMS logins per STEM crs,wk 2-6 >=10.9 or missing
% of 1st yrs. in non-STEM high DFW crs. >=0.28 or miss
Avg. GPA=3.05 N = 72
% of 1st yrs. in non-STEM high DFW crs. <0.28
Avg. GPA=2.90 N = 73
LMS logins per STEM crs,wk 2-6 <10.9
LMS STEM logins wk 1 >=8
Avg. GPA=1.30 N = 11
LMS STEM logins, week 1 < 8 or missing
Avg. GPA=2.52 N=16

The next branch for the lower high school GPA group is the non-STEM course LMS logins during weeks 2 through 6. Average high school SAT scores appear at the next level. Figure 7 displays the section of the tree having the students with a high school GPA greater than 92.0 or missing. A small number of students, some of them international students, do not have a high school GPA in their records. The CART algorithm has combined those observations with the node having high school GPA > 92.0. In that way, those observations remain in the model and are not listwise deleted as they would be in a standard linear regression analysis. The next two levels are different than those for the lower high school GPA students. The next split after high school GPA is whether the students received a scholarship or not. For those who received a scholarship another high school GPA node follows that splits the students into groups above and below 96.5, while for those without a scholarship LMS non-STEM logins during weeks 2 through 6 is most important

Examining both sections of the tree in Figures 6 and 7, we see that LMS logins factored in numerous splits confirming that students' interactions with the college environment plays a role in their academic success. We also observe the differences in the decision rules for students in the higher high school GPA group as compared to the students in the lower high school GPA group.

The actual GPA predictions can be found in the nodes in the right-most column of the tree and are the average GPA's of the students represented by the characteristics of each particular node. The characteristics associated with the GPA predictions can be ascertained by tracing the paths from the high school GPA node on the left to the desired average GPA node on the right. For example, to determine the characteristics for the students represented in the top right average GPA = 3.63 node in figure 6, we have students with high school GPA < =92, LMS logins per non-STEM course in weeks 2 to 6 >= 11.3 or missing, high school average SAT critical reading > 570, SAT Math – Critical Reading combined score > 1360, and finally, receiving credit for 1 or more AP STEM courses. The prediction, 3.63, is the actual average GPA of students in the fall 2014 cohort having the characteristics just listed. Hence, we can say that students with characteristics represented in the final nodes have, *on average*, the GPA that is listed in the node.

The average GPA nodes have been color-coded to assign estimated risk to the GPA levels. The red nodes have average GPA's of 2.20 or less and are at the *highest risk* of receiving a low GPA The orange nodes represent *high risk* students and on average have GPA's of above 2.20 to 2.75. Yellow nodes with average GPA's of above 2.75 to 3.0 represent *moderate risk,* white nodes represent neutral risk with average GPA's ranging from above 3.0 to below 3.5, and the green nodes are *low risk* students who, on average, have GPA's of 3.5 and above. The given risk levels can be adjusted based on university outcomes and the number of students assigned to various planned interventions.

Conclusion

It is clear from studying the decision tree model that weaker students from high schools with lower average SAT scores, who additionally are interacting with the LMS at diminished rates are over-represented in the lower GPA groups. The model can assist in identifying these students before the end of the semester so they can be assigned to interventions that may help to improve their outcomes. Since enrollment in courses with higher failure rates is also a factor appearing in the decision tree, developing a pre-orientation model could assist advisors in steering some students from course loads that may be excessively burdensome. The model results can also be shared with departments to inform their advising and intervention efforts. Automated methods for easily sharing the results are being planned. The goal is to find the students who need assistance in fulfilling their potential, thereby reducing the number who end up leaving due to poor performance.

References

Bahr, P.R. (2008). Does mathematics remediation work?: a comparative analysis of academic attainment among community college students. *Research in Higher Education*. 49:420-450.

Bean J. (1983). The application of model of turnover in work organizations to the student attrition process. *Review of Higher Education*, 6(2), 129-148.

Breiman, L., Friedman, J., Olshen, R., Stone, D. (1984): *Classification and Regression Trees*. Wadsworth Books.

Chen R. (2012). Institutional characteristics and college student dropout risks: a multilevel event history analysis. *Research in Higher Education*. 53:487–505.

Friedman, J., Hastie, T., Rosset, S., Tibshirani, R., Zhu, J. (2003) Discussion of Boosting Papers. Retrieved from http://web.stanford.edu/~hastie/Papers/boost_discussion.pdf

Herzog S. (2005). Measuring determinants of student return vs. dropout/stopout vs. transfer: a first-to-second year analysis of new freshmen. *Research in Higher Education*. 46:883-928.

Johnson I. (2006). Analysis of stop-out behavior at a public research university: the multi-spell discrete-time approach. *Research in Higher Education*. 47:905-93.

Johnson I. (2008). Enrollment, persistence and graduation of in-state students at a public research university: does high school matter? *Research in Higher Education*. 49:776-793.

Parker M. (2005). Placement, retention, and success: a longitudinal study of mathematics and retention. *The Journal of General Education*. 54:22-40.

Robbins S, Allen J, Casillas A, Akamigbo A, Saltonstall M, Campbell R, Mahoney E, Gore P. (2009). Associations of resource and service utilization, risk level, and college outcomes. *Research in Higher Education*. 50: 101-118.

SAS Institute Inc. (2014). *SAS® Enterprise Miner™ 13.2: Reference Help*. Cary, NC: SAS Institute Inc.

Singell L, Waddell, GR. (2010). Modeling retention at a large public university: can at-risk students be identified early enough to treat? *Research in Higher Education*. 51:546-572.

Stater M. (2009). The impact of financial aid on college GPA at three flagship public institutions. *American Educational Research Journal.* 46:782-815.

Stinebrickner R, Stinebrickner T. (2014). Academic performance and college dropout: using longitudinal expectations data to estimate a learning model. *Journal of Labor Economics*. 32:601-644.

Thomas EH, Galambos N. (2004). What satisfies students? mining student-opinion data with regression and decision-tree analysis. *Research in Higher Education*. 45:251-269.

Tinto, V. (1987). Leaving college: rethinking the causes and cures of student attrition. Chicago, IL: The University of Chicago Press.

Zwick R, Sklar JG. (2005). Predicting college grades and degree completion using high school grades and SAT scores: the role of student ethnicity and first language. *American Educational Research Journal*. 42:439-464.

Appendix

Variable List

<u>Demographics</u>
Gender
Ethnicity
Area of residence at time of admission: Suffolk County, Nassau County, New York City,
    other NYS, other US, International

<u>Pre-college Characteristics</u>
High School GPA
College Board SAT Averages by High School
    Average High School Critical Reading
    Average High School SAT Math
    Average High School SAT Critical Reading + Math
SAT: Math, Critical Reading, Writing, Math+Critical Reading

<u>College Characteristics</u>
Number of AP STEM courses accepted for credit
Number of AP non-STEM courses accepted for credit
Total credits accepted at time of admission
Total STEM courses
Total STEM units
Total Non-STEM courses
Total No-STEM units
Class level
Dorm Resident
Intermural Sports Participation
Fitness Class Participation
Honors College
Women in Science and Engineering
Educational Opportunity Program
Stony Brook University Math and Writing Placement Exams
College of student's major or area of interest: Arts and Sciences, Engineering, Health Sciences,
    Marine Science, Journalism, Business
Major Group: business, biological sciences health sciences, humanities and fine arts,
    physical sciences and math, social behavioral science, engineering and applied sciences,
    journalism, marine science, undeclared, other
Major type: declared major, undeclared major, area of interest
High DFW Rate Courses: enrollment >= 70, percent DFW >=10%
    Total high DFW STEM units
    Total high DFW non-STEM units
    Highest DFW rate among the DFW Courses in which the student is enrolled
    Highest DFW rate among the DFW Courses in which the student is enrolled

Proportion of freshmen in a student’s highest DFW rate STEM course
Proportion of freshmen in a student’s highest DFW rate non-STEM course
Type of math course: high school level, beginning calculus, sophomore or higher math

Financial Aid Measures
Aid disbursed in the Fall 2014 – Spring 2015 academic year
Total grant funds received
Total Loans recorded by the Financial Aid Office
Total scholarship funds received
Total work study funds received
Total athletics aid received
Athletic aid, grant, loan, PLIS loan, subsidized/unsubsidized loan, scholarship, work study, TAP, Perkins, Pell indicators
Adjusted Gross Income
Federal Need
Federal Expected Family Contribution
Dependent status

Services/Learning Management System (LMS)
Advising Visits/Tutoring Center Usage
Tutoring center appointment no shows
Number of STEM Course Center Visits, weeks 1 to 6
Number of non-STEM Course tutoring Center visits, weeks 1 to 6
Advising Visits during week 1
Advising visits during weeks 2 – 6
Course Management System Logins
F14_Stem_Login_N
F14_NonStem_Login_Week1_N
Non-STEM course related logins during weeks 2 - 6
Non-STEM Course related logins during week 1
STEM Course related logins during week 1
STEM Course related logins during weeks 2 to 6
Number of STEM course logins per STEM course using the CMS, weeks 2 – 6.
Number of non-STEM course logins per non-STEM courses using the CMS, weeks 2 – 6.