Modeling Likelihood of Student Success Using Financial Aid Support with Student GPA and Preparedness

Introduction

This two-part study used financial aid support, student high school GPA and student preparedness to predict retention and graduation within six years for University of Alaska first-time freshmen. The study utilized both logistic regression, the typical technique used for this kind of modeling, and gradient-boosting, a machine-learning procedure.

In the first part of the study, 96 logistic regression interaction models, each combining two predictor variables, were fitted to the data and compared to see which combinations of variables best predicted measures of student success related to graduation and retention. Then 960 logistic regression models, each using a single predictor variable, were fitted to the data and examined in aggregate to look for broad patterns in the data.

Student data can be highly variable, with missing data and strongly correlated covariates, which can be challenging for logistic regression. Gradient-boosting, a non-parametric technique, is hypothesized to be more robust to these issues than is logistic regression. In the second part of the study, the predictive accuracy of logistic regression models was compared with that of gradient-boosting models, first with simulated data and then with student data. Data sets used for these comparisons were minimally scrubbed in order to simulate rapid, real-time analyses.

Key Findings

- **Increased average annual scholarship support is associated with increased probability of graduation within six years for all students, but more strongly for students who are more prepared for college. The majority of these students receive little or no scholarship support.**

- **First-year scholarship support is strongly associated with increased graduation and increased retention.**

- **Average annual scholarship support is strongly associated with an increased probability of graduation in six years or in any number of years, and with reduced number of years to graduation.**

- **Gradient-boosting models out-performed logistic regression models in prediction accuracy using simulated data that had significant numbers of missing values.**

- **Gradient-boosting models out-performed logistic regression in predicting graduation of University of Alaska students using minimally scrubbed data.**

Methods

SAS was used to extract the data set from the UA Decision Support Database. The statistical software package R was used for all subsequent data manipulation and for all statistical analysis. Logistic regression models were fitted using the R glm() procedure. Gradient-boosting models were fitted using the gbm() procedure in the R GBM package.

The University of Alaska data set consisted of records for 10,488 students who entered the UA system as first-time freshmen from fall 2000 until fall 2003. This window of entry into the system allowed sufficient time to measure whether students graduated within six years. Each record includes 100-plus variables, including both quantitative and categorical variables. Many variables were manufactured linear combinations of other variables, such as total financial aid, which is the sum of all sources of financial aid. Others were naturally highly correlated, such as high school grade point average and SAT score. Many records had missing values for at least some, often crucial, variables. For example, 26.3% of the sample records were missing values for high school grade point average, which is a powerful predictor of University of Alaska student success. These values were not missing at random: The average age of students with values for high school grade point average was 19.7 years, and the average age of students without was 25.4 years.

Simulated data sets had 2,000 rows each, consisting of two predictor variables with one response variable. The following data issues were simulated: varying numbers of missing values of one predictor variable; varying degrees of correlation between the predictor variables; one nonsense predictor variable, i.e., a variable with no relationship with the response variable.

Detailed Results: Logistic Regression Interaction Models

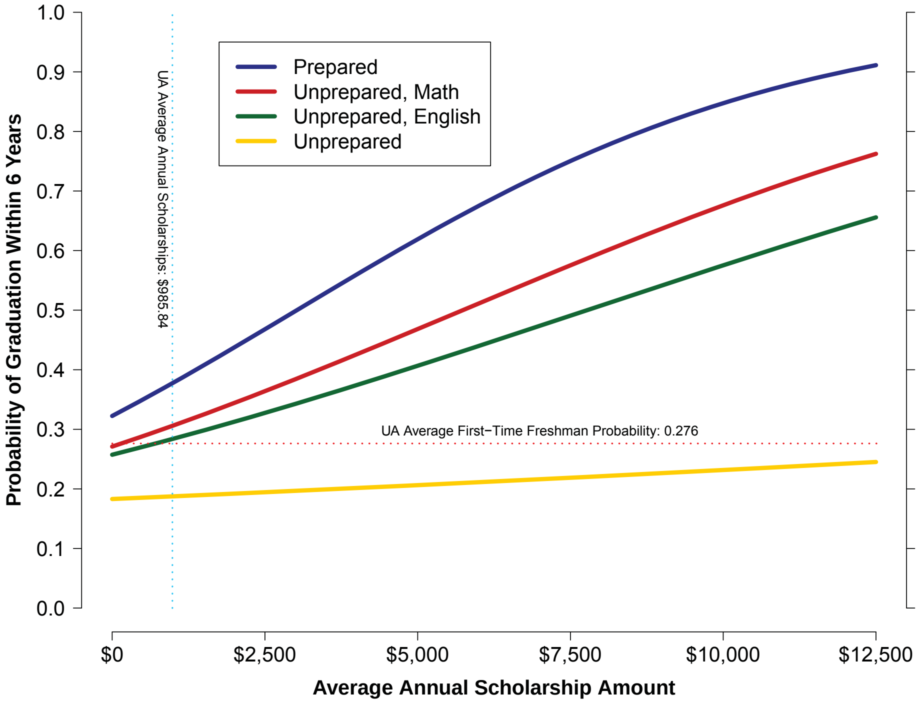

Figure 1. Modeled Probability of Graduation within Six Years With Any Degree or Certificate as a Function of the Interaction Between Annual Scholarship Support and Preparedness. Scholarship support is a continuous variable. Preparedness is a categorical variable based on the number of developmental math and English courses a student takes. The blue dotted vertical line shows mean average annual scholarship support of the 10,488 University of Alaska students in the study group. The red dotted line shows the actual proportion of the students in the study group who graduated within six years with any degree or certificate. All interactions are significant at an α=0.05 level.

Increased average annual scholarship support is associated with increased probability of graduation within six years with any degree or certificate for all students, but more strongly for students who are at least partially prepared for college.

Of the one-way interaction logistic regression models examined that predicted graduation within six years with any degree or certificate, the best (lowest AIC) uses average annual scholarship support and preparedness as predictors.

Figure 1 shows the fitted model. Average scholarship amount is total scholarship amount divided by the number of years a student is in the University of Alaska system. Preparedness is a categorical variable based on the number of developmental courses in math and English a student takes. (Prepared students took no developmental courses. Unprepared students took developmental courses in both math and English.) The dotted lines show the mean probability of graduation and mean average annual scholarship support in the study population.

Values for probability range from 0.0, which in this case is the absolute certainty that a student will not graduate within six years, to 1.0, which is the absolute certainty that a student will. A probability greater than 0.5 means a student more than likely will graduate within six years, while a probability of less than 0.5 means a student more than likely will not graduate within six years.

Note that at the left margin of the plot, where average annual scholarship support is zero, the predicted probability of graduation is well below 0.5 for all categories of preparedness. Students in the highest category of preparedness who receive no scholarship support have a 0.32 probability of graduating within six years.

Increasing average annual support is significantly associated with increased likelihood of graduation within six years, but that effect varies by category. The most prepared students (blue plot) saw an increase in probability of graduation within six years from 0.32 to more than 0.91, or from unlikely to graduate to near certainty of graduation. In contrast, the least prepared students (yellow plot) saw only a modest increase, to approximately 0.25.

Based on where the average student in the study group falls in Figure 1 -- both with respect to scholarship support and probability of graduating within six years -- one can infer that most UA students are less than fully prepared for university study, and that most receive little scholarship support. This is supported by the student data.

Key Findings

- Scholarship support is associated with increased probability of graduation in students who are at least partially prepared for college.
- 60.6% of incoming freshman are at least partially prepared for college.
- 46% of incoming freshmen are at least partially prepared, but receive less than $1,000 per year in scholarship support.

Within the study population, only 24.5% of first-time freshmen fall into the “Prepared” category, 60.6% are at least partially prepared, and 39.4% are unprepared in both math and English.

Almost half (46%) of incoming first-time freshmen are at least partially prepared for college, yet receive less than $1,000 per year in scholarship support. More than a third (38%) are at least partially prepared and receive no scholarship support.

Students within this category, which comprises the top three lines in Figure 1, show a pronounced positive relationship between average annual scholarship support and probability of graduation within six years. This suggests that increased average annual scholarship support could help increase the rate of graduation among these students.

Students who are unprepared for college also showed a significant positive relationship between scholarship support and improved probability of graduation, but this effect was so modest that, for students who fall into this category, improving preparedness may be a much more effective first step in improving graduation rates than simply increasing scholarship support. Students who aren’t prepared in English have lower probability of graduation than students who aren’t prepared in math, which suggests that a focus on English skills might yield the most improvement.

Detailed Results: Simple Logistic Regression Models

There is a broad pattern of increased scholarship support improving graduation and retention.

Figures 2, 3 and 4 show summaries of 48 of the 960 simple logistic regression models examined. Each cell represents a single model using the variable listed at the top of the column to predict the response variable at the left of the row. Green bubbles indicate a model with a significant (p<0.1) relationship. The diameter of a bubble is proportional to the coefficient in that model. This is equivalent to the steepness of the curves in Figure 1. For years to graduation, the bubble is proportional to the negative of the coefficient. Bubble sizes are relative to the models summarized within each table and cannot be compared between tables.

Figure 2 shows a summary of models fitted to the complete University of Alaska student data set. First year financial aid, which includes scholarships, grants and loans, is associated with improvements to all response categories: graduation, years to graduation, and retention through senior year. First-year scholarships have a strongly positive relationship with all categories except years to graduation. Average annual scholarship support, and, to a lesser extent, loans, are strongly associated with improved graduation, as well as reduced years to graduation.

Figure 3 shows models fitted to students in the data set who received Pell grants. This includes those who received other grant support. Because Pell grants are commonly considered to be indicators of low family income in the literature, this category can be seen to represent lower-income students of all abilities: those who qualify for merit-based financial aid, as well as those who don't. In this category, first-year scholarships are associated with improved graduation and retention. Average annual scholarship support is associated with improved graduation. Grant support is associated with reduced years to graduation.

Figure 4 shows models fitted to students who received Pell grants, but no other grant support. This suggests that these students are from lower-income families, but do not qualify for merit-based support. None of the students in this group fell into the "Prepared" category discussed on Page 2, and half fell into the "Unprepared" category. In this group, first-year scholarship support is strongly associated with improvements in graduation, years to graduation and retention. Average annual scholarships, followed by average annual loans, are strongly associated with improvements in graduation and years to graduation. While scholarship support is associated with improved graduation and retention for all students in the study group, it is can be seen that for students who fall into lower income categories, scholarship support becomes a more important indicator of retention and graduation when academic achievement is also low.

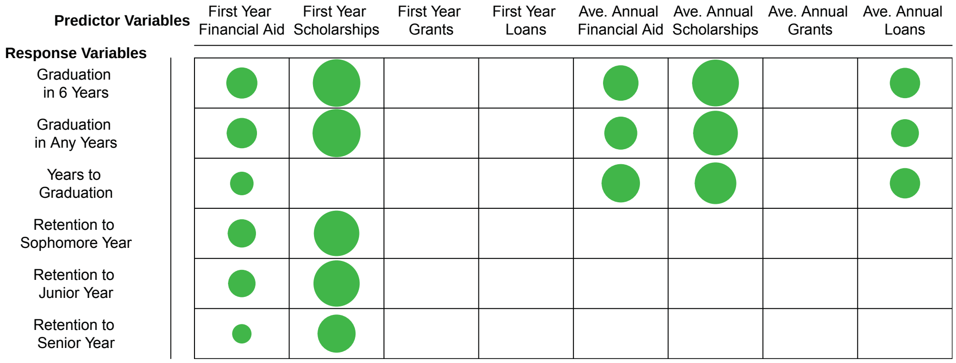

Figure 2. All Students. Summary of results of 48 simple logistic regression models fitted to the entire 10,488-student University of Alaska data set.

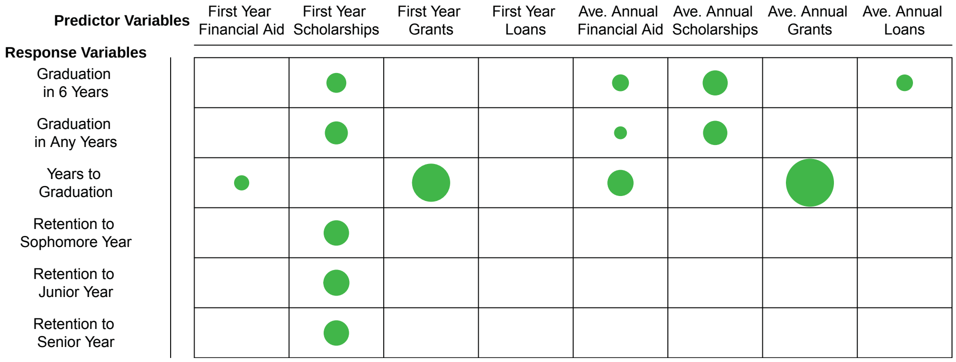

Figure 3. Students who received Pell grants. Summary of results of 48 simple logistic regression models fitted to the data for the 2,114 Pell recipients in the study group. This includes students who received other grants.

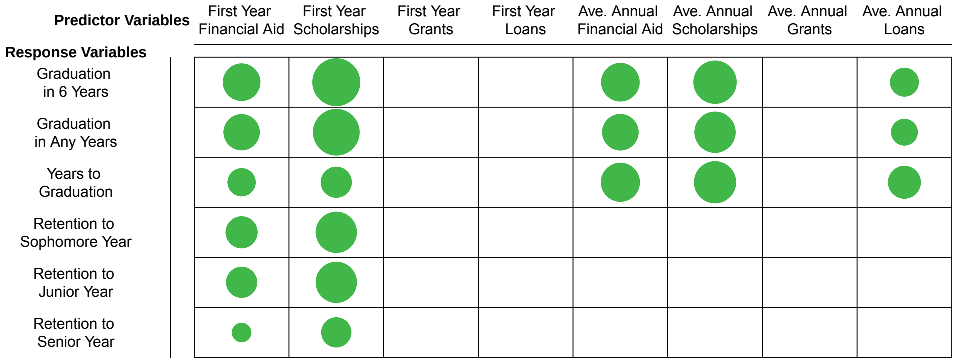

Figure 4. Students who received only Pell grants. Summary of results of 48 simple logistic regression models fitted to data for the 204 University of Alaska Pell grant recipients in the study group who received no other grant support.

Key Findings

- Scholarship support is associated with increased probability of graduation and improved retention in the entire student population.
- For students who receive Pell grants, but no other grants, scholarship support is associated with improvements in all categories of performance examined.

Detailed Results: Gradient-Boosting Models

For data sets with missing data, gradient-boosting models out-performed logistic regression models in prediction accuracy, both with simulated data and with University of Alaska student data.

Figure 5 shows a comparison between gradient-boosting and logistic regression in prediction accuracy using simulated data. Simulated data sets comprised two predictor variables, one binary (0 or 1) response variable and 2,000 rows. Each data set was randomly split into two 1,000-row subsets, one of which was used to train both the gradient-boosting and logistic regression models. The other 1,000-row subset was used to test the models' predictive accuracy. Each comparison point on the horizontal axis represents 100 simulated data sets at that level of values missing from one of the predictor variables.

With no missing values, logistic regression models slightly out-perform gradient-boosting models, though both do well. As the proportion of missing values in one predictor variable increases, the accuracy of both models decreases, but logistic regression drops much more. With 60 percent of values missing, gradient-boosting models maintain 75 percent accuracy, while logistic regression models fall below 50 percent. At that level of missing data, logistic regression is no better at prediction than flipping a coin.

When fitted to relatively unscrubbed University of Alaska student data, gradient-boosting maintains a prediction accuracy of 74.6 percent, while logistic regression manages only 54.5 percent. The most influential predictor variable (see Figure 6) was high school GPA, for which 26 percent of records were missing a value. In conjunction with the results of the simulation study, this explains some of the disparity between the performance of logistic regression and gradient boosting using the real data, but it suggests that other variables are missing values, as well as the possibility that there are other issues with the data that are creating challenges for logistic regression.

Figure 6 shows a relative influence plot of the predictor variables used in fitting the gradient boosting model. The plot can be roughly interpreted as showing percentages of influence contributed by the variables. This should not be interpreted to mean that these are the only important possible predictor variables; it simply describes the weight of their influence relative to one another.

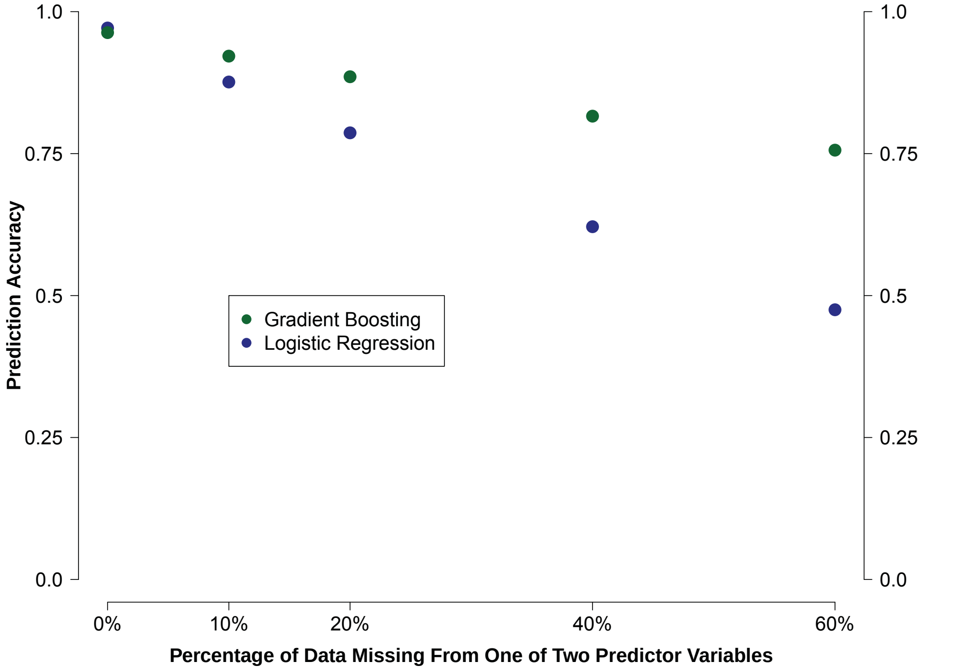

Figure 5. Predictive accuracy of gradient boosting vs. logistic regression using simulated data with missing values. Each simulated data set was 2,000 rows long, consisting of two predictor variables and one response variable. Each data point is the average of 100 simulated sets.

Key Findings

- Gradient-boosting models out-perform logistic regression models in prediction accuracy using simulated data with missing values.
- Gradient-boosting models out-perform logistic regression models in predicting graduation using minimally scrubbed University of Alaska student data.

Not surprisingly, high school GPA is by far the most influential predictor in the model. This is followed at a distance by Preparedness, which is highly correlated with GPA. Average annual scholarship support, the most influential financial aid variable in the model, appears next, followed by three relatively uninfluential variables. Other analyses confirm the importance of GPA, preparedness, scholarships and Pell grants in predicting graduation.

Both modeling techniques are improved by a moderate amount of data preparation and can be brought to very high levels of predictive accuracy using techniques such as imputation -- using predicted values for missing data. However, this test sought to examine how well the techniques performed on relatively raw data.

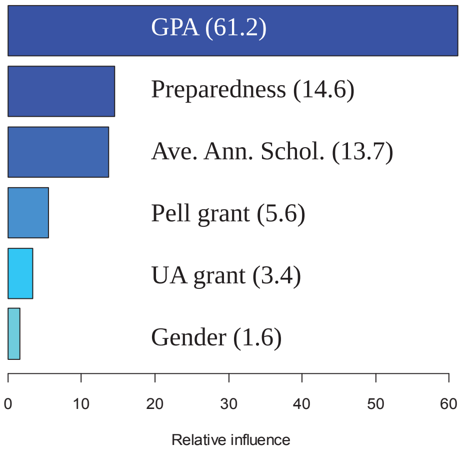

Figure 6. Relative influence plot of predictor variables used in the model. GPA, Preparedness, Annual Scholarships and Pell Grants were selected because of their importance in other models that have been examined. UA Grant and Gender were included arbitrarily.

It is important to keep in mind that while gradient-boosting does provide diagnostic tools, such as the relative influence plot in Figure 6, it is a prediction tool, and it cannot provide inference about the details of a model in the same way that logistic regression can, as in Figure 1, which is a graphical representation of a fitted logistic regression model. There is no equivalent for gradient-boosting.

Source: UA Informations Systems: UA Decision Support Database from Banner SI Closing Extracts 2000-2010. SD546802