



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.**
- 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 fresh-

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 in-

fer that most UA students are less than fully

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 an-

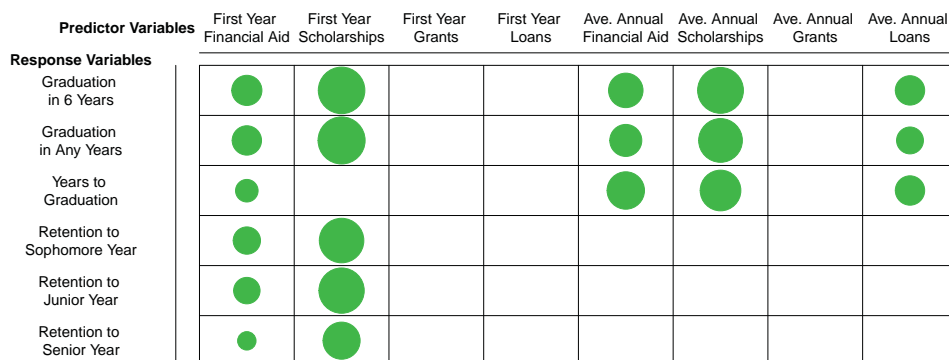


Figure 2. All Students. Summary of results of 48 simple logistic regression models fitted to the entire 10,488-student University of Alaska data set.

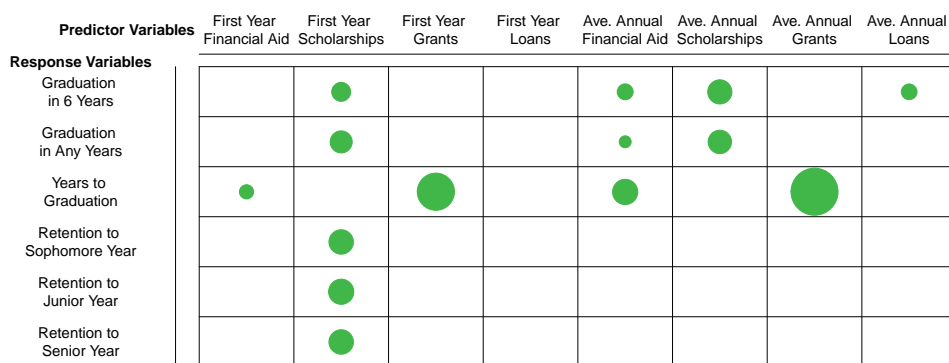


Figure 3. Students who received Pell grants. Summary of results of 48 simple logistic regression models fitted to the data for the 2,114 Pell recipients in the study group. This includes students who received other grants.

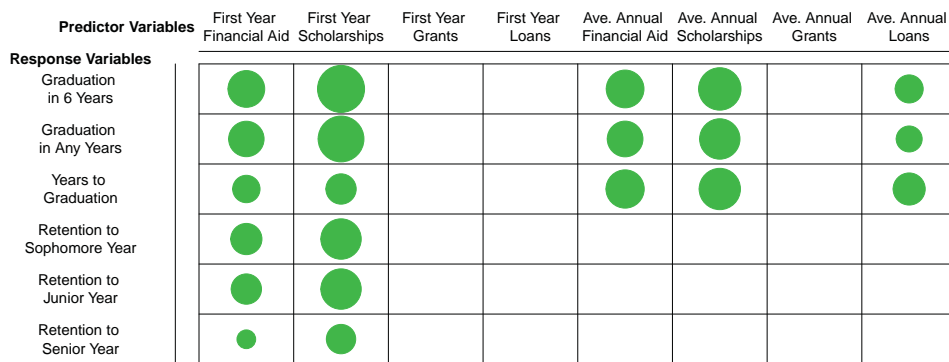


Figure 4. Students who received only Pell grants. Summary of results of 48 simple logistic regression models fitted to data for the 204 University of Alaska Pell grant recipients in the study group who received no other grant support.

nual 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 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.

Key Findings

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.