Using Financial Aid Support to Model Probability of Student Success, Incorporating a Comparison of Predictive Accuracy of Logistic Regression and Gradient Boosting

Adam Watson
University of Alaska Fairbanks

April 19, 2011

Abstract

In the current climate of tightening education budgets, an effort to assess the efficacy of financial aid support relies on tools that allow the identification of students most in need of assistance, as well as the methods of assistance that are most helpful to those students. Preliminary logistic regression model fitting found a positive relationship between financial aid support and probability of success. Student data is 'dirty,' however, with problems such as missing data and collinearity in the predictors, which can interfere with the ability of logistic regression to make accurate predictions. Logistic regression and gradient boosting were compared in their predictive accuracy using simulated data with the problems mentioned, revealing that gradient boosting outperforms logistic regression when data are missing. The two methods were then compared using real University of Alaska student data, where gradient boosting outperformed logistic regression in predicting student success with real data.

1. INTRODUCTION

This study is a preliminary examination of how financial aid and other factors predict success, as measured by retention and graduation for students at the University of Alaska. The drive to understand the relationship between financial aid and student success is motivated by a need to examine the effectiveness of the need-based support that the university distributes to students. In the current climate of tightening budgets, there is a need to justify funding by demonstrating measurably increased student performance. It is also important to be able to identify students who are most at risk of not succeeding in the university system, regardless of their financial aid status, to direct assistance to them in order to improve their likelihood of success.

This paper is organized into two sections: In the first part, I report notable results of exploratory analysis using simple and one-way interaction logistic regression models examining the effect of financial aid on student success. In the second part, I compare the accuracy of multiple logistic regression and gradient-boosting models in predicting outcomes for simulated data. I then compare their ability to predict success or failure using University of Alaska student data.

The student data involved is largely 'dirty', including missing values and highly correlated predictor variables, and a convenient predictive model would be able to digest the data with minimal preparation. I chose to examine gradient boosting, a form of non-parametric decision-tree learning, because I wanted to explore a method that was likely to be more robust to these issues than logistic regression.

2. METHODS

2.1 STUDENT DATA

The University of Alaska data set consists of records for 10,488 students who entered the University of Alaska system as first-time freshmen from fall 2000 until fall 2003. This window of entry into the system allowed time to measure whether students graduated within six years. The data were extracted from the UA Decision Support Database using a SAS SQL query. Each record includes 100-plus variables, which include both quantitative and categorical variables. Many variables are manufactured linear combinations of other variables, such as total financial aid, which is the sum of all sources of financial aid – grants, scholarships, loans, etc. Others are naturally highly correlated, such as high school grade-point average and SAT scores. Not surprisingly in a data set with this many variables, many records are missing at least some values.

2.2 LOGISTIC REGRESSION

2.2.1 Simple Logistic Regression

The simple logistic regression model for estimating the parameters of the logistic response function is described by Kutner et al. (2005). We have Y_1 ,Y_2 … Y_n responses to a predictor X_1 where Y_i can take the values 1 and 0, with probabilities p_i and 1-p_i, respectively. Therefore the probability function of Y_i, a Bernoulli random variable with parameter p_i=P(Y_i=1) is

$$f\left(Y_i\right)=p_i^{Y_i}\left(1-p_i\right)^{1-Y_i}\;,\; Y_i=0,1; i=1,\ldots,n. \tag{1}$$

The simple logistic model states that these responses depend on the predictor variable as follows,

$$P\left(Y_i=1\right)=p_i=\frac{e^{\beta_0+\beta_1 X_1}}{1+e^{\beta_0+\beta_1 X_1}} \tag{2}$$

and

$$P\left(Y_i=0\right)=1-p_i=1-\frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}=\left(1+e^{\beta_0+\beta_1 X_i}\right)^{-1} \tag{3}$$

The model in (1) is non-linear in the parameters. The logit transformation linearizes the regression function and is defined as the log of the odds that Y takes the value 1. This turns the regression function into a linear function of the parameters,

$$\log\left(\frac{p_i}{1-p_i}\right)=\beta_0+\beta_1 X_i \tag{4}$$

Model parameters β_0 and β_1 are typically estimated by the method of maximum likelihood. From (1), the joint probability of the sample $Y_1 \ldots Y_n$ is

$$\prod_{i=1}^{n} f_i\left(Y_i\right)=\prod_{i=1}^{n} p_i^{Y_i}\left(1-p_i\right)^{1-Y_i}=L(p)$$

Thus, the log of the likelihood is

$$\log L(p)=\sum_{i=1}^{n}\left[Y_i \log(p_i)+\left(1-Y_i\right)\log\left(1-p_i\right)\right]=\sum_{i=1}^{n}\left[Y_i \log\left(\frac{p_i}{1-p_i}\right)\right]+\sum_{i=1}^{n}\left[\log\left(1-p_i\right)\right]$$

Using (3) and (4), the log likelihood of β_0 and β_1 is

$$\log L\left(\beta_0,\beta_1\right)=\sum_{i=1}^{n}\left[Y_i\left(\beta_0+\beta_1 X_i\right)\right]-\sum_{i=1}^{n}\left[\log\left(1+e^{\beta_0+\beta_1 X_i}\right)\right] \tag{5}$$

This is maximized with respect to β_0 and β_1 to obtain maximum-likelihood estimates b_0 and b_1. Because no closed form solution exists that maximizes the likelihood function for β_0

and β_1, numerical search procedures are required to carry this out. The resulting estimates are used in the fitted logit response function:

$$\log it\left(\hat{p}_i\right) = \log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = b_0 + b_1 X_i \tag{6}$$

This can also be written as the fitted logistic response function, which yields estimated probabilities:

$$\hat{P}(Y_i = 1) = \hat{p}_i = \frac{e^{b_0 + b_1 X_i}}{1 + e^{b_0 + b_1 X_i}} \tag{7}$$

Predicting a response can be done by examining the estimated probability that $Y_1=1$. A commonly used rule for predicting the value of Y_1 is

$$\hat{y}_i = \left\{\begin{matrix} 1, \hat{p}_i \geq 0.5 \\ 0, \hat{p}_i < 0.5 \end{matrix}\right\} \tag{8}$$

2.2.2 Multiple Logistic Regression

Expanding simple logistic regression to multiple logistic regression requires substituting $\beta_0+\beta_1 X_i$ with $\beta_0 + \beta_1 X_{1,i} + \ldots + \beta_{p-1} X_{p-1,i} = X^T\beta$ in (5) to arrive at the log likelihood

$$\log L(\beta) = \sum_{i=1}^{n} Y_i\left(X_i^T \beta\right) - \sum_{i=1}^{n} \log\left[1 + e^{X_i^t \beta}\right] \tag{9}$$

From this, maximum likelihood estimates for the parameters, fitted probabilities and predicted responses are obtained as described above.

2.2.3 Model Selection

Where logistic regression models were compared, Akaike's information criterion (AIC) was used as part of a stepwise model selection procedure to determine the better model. Models with small values for AIC are preferred. In general, AIC is given by

$$AIC_p = 2p - 2\ln(L) \tag{10}$$

where p is the number of parameters.

AIC will return lower values for models with larger likelihoods, as long as the term, 2p, which penalizes non-parsimonious models, is not too large.

For logistic regression, AIC is adapted as follows:

$$AIC_p = -2\log L(b) + 2p = -2\left[\sum_{i=1}^{n} Y_i\left(X_i^T\beta\right) - \sum_{i=1}^{n}\log\left(1 + e^{X_i^T\beta}\right)\right] + 2p \qquad (11)$$

The R step() procedure was used to compare models in a backward and forward stepwise fashion, which combines forward selection and backward elimination to decide which variables to include or remove from the model, based on their effects on AIC, as defined in (11).

2.3 GRADIENT BOOSTING

2.3.1 Regression Trees, overview

A description of the gradient boosting machine procedure in R, gbm(), should begin with a discussion of regression trees, which are described by Hastie et al. (2008).

Consider a model with a response variable Y, and two predictor variables X_1 and X_2. The feature space of X_1 and X_2, i.e., all combinations of values of each that can be used to predict the response, is first split into two partitions at a point X_1=t_1 (Figure 1a), and the response is modeled by the mean of each partition, with the variable X_1 and split point t_1 chosen to achieve the best fit. (This process will be explained in greater detail below.) The resulting partitions, or nodes, are split at similarly chosen points until some stopping criterion is met, typically the minimum number of elements remaining within a partition, which is commonly referred to as minimum node size. In Figure 1a, the region of $X_1 < t_1$ is split next, at $X_2 = t_2$, using the same procedure to achieve the best fit within the region $X_1 < t_1$. Two more splits, t_3 and t_4 are made in the same manner before the procedure is stopped. The decision tree in Figure 1b represents the same information.

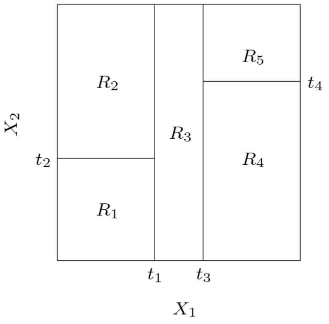

Figure 1a.

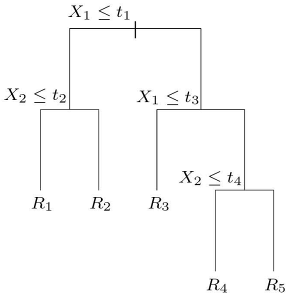

Figure 1b.

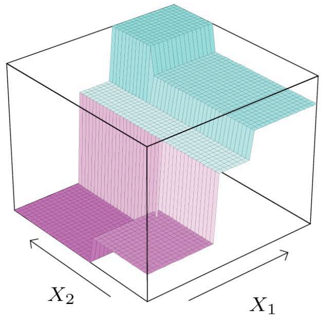

Figure 1c.

Figures from Hastie et al. (2008)

Figure 1c shows the partitioned feature space, with the prediction surface fitted by the regression model

$$\hat{f}(X) = \sum_{m=1}^{5} c_m I\{(X_1, X_2) \in R_m\} \qquad (12)$$

Here m is the partition index, and c_m is the estimated response for the partition (or node) R_m.

2.3.1 Regression Trees, detail

The estimated responses, c_m, are obtained using sum of squares

$$\sum_{i=1}^{n}\left(y_i - f\left(x_i\right)\right)^2 \tag{13}$$

thus the “best” estimate for c_m will be the average of y_i for each region, because the average minimizes the sum of squares:

$$\hat{c}_m = ave\left(y_i \middle| x_i \in R_m\right)$$

Split points are chosen with an algorithm using a splitting variable x_j and split point s. Starting with the complete data, define

$$R_1\left(j,s\right) = \left\{X \middle| X_j \leq s\right\} \text{ and } R_s\left(j,s\right) = \left\{X \middle| X_j > s\right\}$$

Then seek j and s to solve

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i \in R_{1(j,s)}}\left(y_i - c_1\right)^2 + \min_{c_2}\sum_{x_i \in R_{2(j,s)}}\left(y_i - c_2\right)^2\right]$$

For any j,s, the inner terms are solved by

$$\hat{c}_1 = ave\left(y_i \middle| x_i \in R_1(j,s)\right), \text{ and } \hat{c}_2 = ave\left(y_i \middle| x_i \in R_2(j,s)\right)$$

The best splitting variable and split point at each stage can then be determined by scanning through all possible values of j and s. The feature space is split, at X_j=s, into two new partitions, each of which is then subjected to the same treatment. This process continues until a user-defined minimum node size is reached for every partition.

The resulting tree is then "pruned" using the cost-complexity criterion

$$C_\alpha\left(T\right) = \sum_{m=1}^{|T|} N_m Q_m\left(T\right) + \alpha\left|T\right| \tag{14}$$

where $\left|T\right|$ is the number of terminal nodes, or partitions, in the tree T, N_m is the number of y_i in node m, $\hat{c}_m$ is the mean of y_i in node m, α is the tuning parameter, large values of which result in smaller trees, and

$$Q_\alpha\left(T\right) = \frac{1}{N_m}\sum_{x_i \in R_m}\left(y_i - \hat{c}_m\right)^2$$

To find T_α, weakest-link pruning successively collapses the internal node that produces the smallest increase in

$$\sum_{m=1}^{|T|} N_m Q_m(T)$$

This continues until a single node is left, resulting in a sequence of subtrees of T. Among those subtrees will be T_α, the one that minimizes (14). A value for α is estimated using cross-validation, choosing $\hat{\alpha}$ to minimize the cross-validated sum of squares, which results in the final tree, $T_{\hat{\alpha}}$.

2.3.3 Gradient Boosting, overview

Boosting uses the idea that finding and averaging many rough prediction rules is easier than finding a single, highly accurate one. For regression trees, boosting makes use of a loss function that measures the loss in predictive power that results from a suboptimal model. An example is the minimized squared-error loss function (13) from the previous section.

At each step, boosting adds to the model the regression tree that most reduces the loss function. Thus the model moves down the gradient of the loss function. (Elith et al. 2008)

The first fitted regression tree maximally reduces the loss function, given constraints, for the data. The second regression tree maximally reduces the loss function for the residuals of the first tree. The third regression tree maximally reduces the loss function for the residuals of a model that contains both the first and second regression trees. Note that each tree is added to the model without changing the previous trees in the model. Slowing the rate of movement down the gradient toward a model with best predictive performance guards against overfitting the model, so the contribution of each tree is multiplied by the learning rate, a constant between 0 and 1 (usually close to zero, e.g., 0.001). The ultimate result can be imagined as a regression model that includes thousands of terms, each of which is a regression tree. Fitted values of the final model are the sum of the component trees, multiplied by the learning rate. (Elith et al. 2008)

2.3.4 Gradient Boosting, as implemented by gbm()

We begin by selecting the following (from Ridgeway, 2007):

a. A loss function, $\Psi(y, f)$
b. The number of trees to examine, or iterations, N
c. The interaction depth of each tree, K
d. The learning rate parameter, λ
e. The proportion of the data to sample as a training set for the next proposed tree, p

Define

$$\hat{f}(x) = \arg\min_p \Psi(y, p)$$

For t = 1 to N:

1. Calculate the negative gradient:

$$z_i = -\frac{\partial}{\partial f(x_i)} \Psi(y_i, f(x_i)) \Big|_{f(x_i)=\hat{f}(x_i)}$$

2. Randomly select pxN cases, the subset to be used for the training set, from the dataset.

3. Fit a regression tree with K terminal nodes, using only the observations randomly selected for this iteration.

4. Compute predictions, $\rho_1, \ldots, \rho_k$:

$$\rho_k = \underset{\rho}{\arg\min} \sum_{x_i \in S_k} \Psi\left(y_i, \hat{f}(x_i) + \rho\right)$$

where S_k is the set of x's that define terminal node k.

5. Update $\hat{f}(x)$:

$$\hat{f}(x) = \hat{f}(x) + \lambda \rho_{k(x)}$$

3. RESULTS/DISCUSSION

This section will be organized into two parts. First, I will report selected results from exploratory analysis of the student data using simple and two-variable logistic regression models. I describe in detail one model that demonstrates significant interaction between financial aid and student high school grade point average in predicting probability of graduation within six years. I also report on a group of simple logistic regression models. In the second part, I will compare the predictive accuracy of logistic regression and gradient boosting. I do this using both simulated data sets and the real University of Alaska student data.

3.1 EXPLORATORY LOGISTIC REGRESSION ANALYSIS

3.1.1 Interaction Model Fitting

Exploratory examination of simple logistic regression models suggested that the most important non-financial aid variables were high school GPA, a quantitative variable, and preparedness, a four-level categorical variable based on the number developmental courses a student enrolls in at the University of Alaska: “Prepared,” which indicates sufficient preparation in both math and English; “Unprepared, English” – prepared in math, but not in English; “Unprepared, math” – prepared in English, but not math; and “Unprepared” – unprepared in both English and math. I examined one-way interactions between each of these variables and each of eight quantitative measures of financial aid support as predictors of six different measures of graduation and retention. This amounted to 96 one-way interaction models of potential interest.

Of models that used interactions between preparedness or GPA and different measures of financial aid to predict graduation within six years, the model that I report in Table 1 had the lowest AIC. Average annual scholarship amount is a quantitative variable with values ranging from nothing to well above $20,000.

Coefficient	Estimate	P-value
Prepped(Prepared)	-0.743	< 2*10^-16
Prepped(Unprepared_English)	-1.060	< 2*10^-16
Prepped(Unprepared_math)	-0.989	< 2*10^-16
Prepped(Unprepared)	-1.496	< 2*10^-16
Ave. Scholarship	0.0002456	< 2*10^-16
Prepped(Unprepared_English)*Ave. Scholarship	-0.000109	0.0086
Prepped(Unprepared_math)*Ave. Scholarship	-0.0000733	0.0164
Prepped(Unprepared)*Ave. Scholarship	-0.000216	1.03*10^-15

Table 1. Significant parameter estimates with P-values for the lowest-AIC model predicting graduation within six years.

A number of interesting inferences can be drawn from this fitted model. For example, the model predicts that when scholarship support is zero, all categories of preparedness are associated with predicted probabilities of graduation that fall below $\hat{p}_i = 0.5$. These can be found by using (2) with each individual estimate of the categories of the prepared variable.

While increasing average annual scholarship support is associated with increased probability of graduation, that effect varied for the different categories, hence the interaction terms. Table 1 shows that the most negative interaction term estimate is for students who are unprepared in both math and English. The next most negative is for students unprepared in English, followed by students unprepared in math. This is illustrated in figure 2, where we see that for the least prepared students, increasing levels of scholarship support are associated with a modest increase in probability of graduation, followed by a larger increase for those unprepared in English, then those unprepared in math. The 'prepared' category, the baseline against which the other categories are compared, is associated with the highest intercept, as well as the greatest increase in probability of graduation with increasing levels of scholarship support.

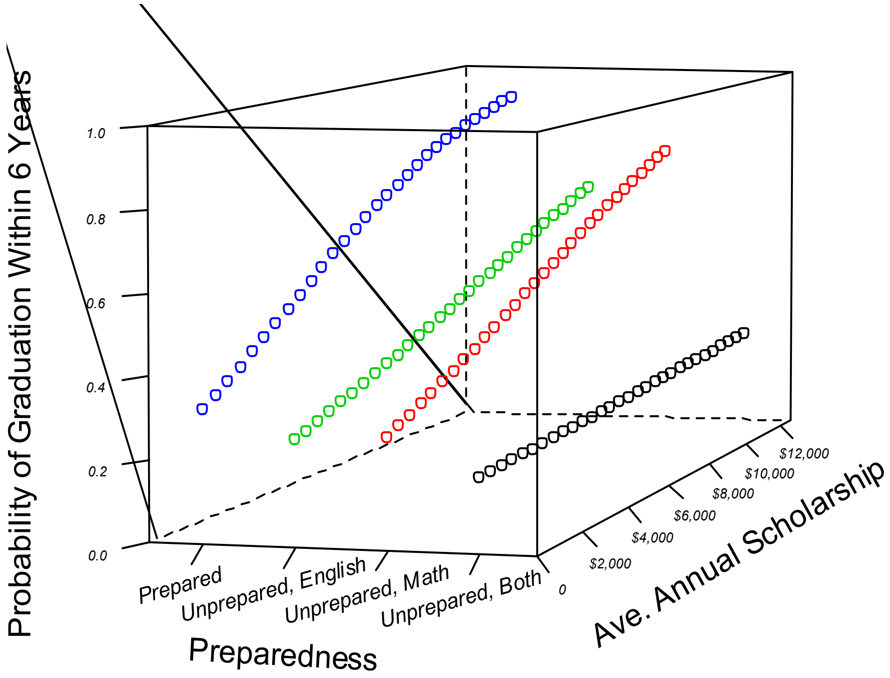

Figure 2. Predicted probability of a UA first-time freshman (fall 2000 to fall 2003 cohorts) graduating within 6 years as a function of average annual scholarship support and preparedness.

3.1.2 Exploratory Simple Logistic Model Fitting

I examined 960 simple models, each considering the relationship between a single predictor variable and measures of retention or graduation. I summarize the results from 48 of those models in Figure 3. All predictor variables were continuous measurements of financial aid support, while all but one response variable described binary measures of success and were fitted with logistic regression models. The exception was "years to graduation," which I fitted with simple linear regression models.

Each cell in Figure 3 represents a single, simple model of the relationship between the predictor variable at the top of the column and the response variable at the left of the row. Cells with green circles represent models with a positive coefficient that was significant at the 0.1 level. The exception to this is the "Yrs. to Grad." row, which represents the negative of the coefficients; in that row, circles indicate a model whose predictor was associated with reduced number of years to graduation. The diameter of each bubble is proportional to the magnitude of the coefficient. Predictor variables with the "Yr1" prefix are measures of first-year support; "Ann." indicates average annual support over the course of a student's academic career. "FA" refers to all financial aid received, which includes scholarships grants and loans. "Ret." refers to retention after a given year. Thus "Yr1 Ret." refers to retention to sophomore year.

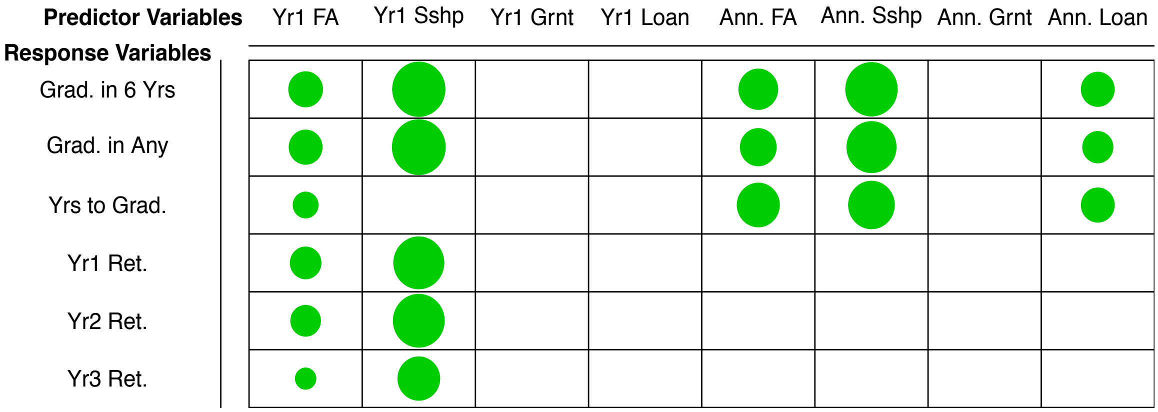

Figure 3. Positive predictors of success for full-time, first-time UA freshmen, fall 2000 - fall 2003 cohorts

There are several notable conclusions that can be drawn. For full-time, first-time freshmen, overall financial aid support, primarily driven by scholarship support, is significantly positively associated with increased likelihood of success. First year scholarship support is significantly positively associated with graduation and retention. Average annual scholarship support across academic career is associated with improved likelihood of graduation and reduced number of years to graduation.

3.2 PREDICTIVE MODEL COMPARISON

3.2.1 Simulation Study

The Alaska student data set contains several problems that could interfere with logistic regression's ability to accurately model and predict student success, particularly for models that contain larger numbers of predictor variables than the models considered above. Specifically, the data set has more than 100 predictor variables to select from, many of which are highly correlated, and numerous observations in the data set contain missing values.

It has been suggested that gradient boosting models are superior for making predictions from data sets that contain these sorts of problems.

I performed a simulation study to compare the predictive ability of logistic regression to gradient boosting. Data were simulated to investigate the effects of multicollinearity, missing values, and model mis-specification on the predictive accuracy of the models.

I used the R function glm() to fit logistic regression models. Because of the simplicity of the models, I did not improve them with techniques such as stepwise model selection. To fit gradient boosting models, I used the R gbm() function in the gbm package with standard settings (Ridgeway 2007) and allowing up to one-way interactions. As with the logistic regression models, the gradient boosting models were simple and many, and I took no action to optimize them.

For each level (i.e., of percent missing, multicollinearity) of each head-to-head comparison, I generated 100 distinct 2,000-record data sets consisting of a response variable and two predictor variables. Each data set was split into two equal parts, one of which became the training set used to fit each logistic regression and gradient boosting model. The other part became the test set on which the predictive powers of the models were assessed. Predictive accuracy and its variance for logistic regression and gradient boosting were recorded for comparison. Predictions, i.e., values of Y, were assigned according to (8).

3.2.2 Simulated Data

I used R to generate data sets with two predictor variables, a binary response variable and known parameters (See Appendix 5.1). Each data set was then split into equally sized training and testing sets.

To compare the effect of missing values on predictive ability, I replaced randomly selected values of one predictor variable of the training set and test set with NAs. Missing values cause the glm() function in R to return NAs, rather than predictions, so I removed rows with missing values for the data used to fit the logistic regression model. I tested the following levels of missingness in the data: 0%, 10%, 20%, 40% and 60%.

To compare the effects of correlation within predictor variables on predictive ability, I used R to generate a data set with a defined correlation between the predictor variables (See Appendix 5.2). I tested the following levels of correlation between the predictor variables: 0.0, 0.1, 0.2, 0.4, 0.75 and 0.90.

To compare the effects of mis-specification of the model, I created a normal data set with two predictor variables, then threw out one of the predictors used to generate the responses and replaced it with a variable generated in the same way, but which was not used to generate responses, for the training and testing steps.

Of the three 'dirty' characteristics that I tested with simulated data, only missing values brought about a marked difference in predictive performance between logistic regression and gradient boosting (Table 2). Figure 4 shows the results of that simulation. Each data point represents the average of the percent correctly predicted by 100 models fitted to 100 distinct data sets. Both models predicted correctly more than 95 percent of the time when no data were missing, and both declined in accuracy as the proportion of missing data increased. With 60 percent missingness, however, gradient boosting was still predicting correctly three-quarters of the time, while logistic regression had fallen below 50 percent. Variance of percentages correctly estimated for both techniques grew as the percent missing grew, but remained small enough (<< 0.1) that error bars are subsumed in the data point markers in the chart.

With no missing values, logistic regression predicted outcomes slightly more accurately than did gradient boosting (97.2% vs. 96.4%), with slightly smaller variance (0.0031 vs. 0.0039). This pattern remained true for tests of varying degrees of correlation and of uselessness of one predictor variable. Predictive accuracy of both techniques dropped as correlation increased, but logistic regression continued to predict slightly (~1%) better than gradient boosting.

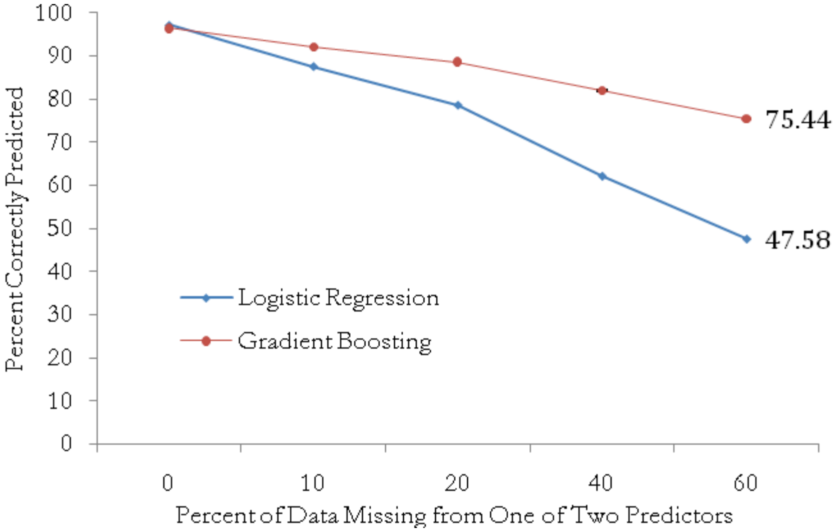

Figure 4. Percentage correct predictions vs. percentage missing data, logistic regression vs. gradient boosting, simulated data.

That logistic regression was able to slightly outperform gradient boosting when no data were missing is likely due in part to the data set being simulated using the logistic function with three parameters, which is likely the ideal data for fitting with a three-parameter logistic regression. The benefit of gradient boosting as a non-parametric technique would be likely to occur when the process generating the data was more cryptic, as in the real-world data set used in the analysis below.

Data Issue	Ave. Percent Correct, Logistic Regression	Ave. Percent Correct, Gradient Boosting
Missing Data (%)		
0	**97.19** (0.0557)	**96.44** (0.0629)
10	**87.64** (0.0446)	**92.20** (0.0849)
20	**78.67** (0.0501)	**88.67** (0.0969)
40	**62.16** (0.0621)	**82.00** (0.1040)
60	**47.58** (0.0733)	**75.44** (0.1478)
Correlation between predictors		
0	**94.24** (0.0776, 0.8342)	**93.56** (0.0834, 0.8078)
0.10	**94.22** (0.0766, 0.8331)	**93.44** (0.0828, 0.8059)
0.20	**94.24** (0.0735, 0.8332)	**93.30** (0.0855, 0.8021)
0.40	**93.77** (0.0698, 0.8199)	**93.89** (0.0810, 0.7902)
0.75	**91.36** (0.0883, 0.7414)	**90.12** (0.1096, 0.7105)
0.90	**87.55** (0.1094, 0.6443)	**85.74** (0.1379, 0.5903)
Extraneous predictor variable	**87.53** (0.0968)	**87.31** (0.1036)

Table 2. Results of simulated data tests. Each level of each data issue represents 100 logistic regression and gradient boosting models fitted to 100 distinct 2,000-record data sets, each consisting of a response variable and two predictor variables. Standard deviation and R-squared of percents correct in parentheses.

3.3 LOGISTIC REGRESSION VS. GRADIENT BOOSTING, REAL DATA

The result of the comparison of performance on data with missing values is of particular concern for fitting predictive models for UA student data. The UA student body has an average age of approximately 30 (UA in Review 2011), and older students are more likely to be missing data values. The average age of students in the data set with 'NA' for high school gpa was 25.4. The average age of students who had a non-'NA' value for gpa was 19.7. It therefore should not be surprising that 26.3 percent of the sample records are missing values for high school GPA, one of the most important predictors of graduation.

I carried out a preliminary comparison using gender, Pell grant status, preparedness, UA Grant status, high school gpa, average annual scholarship support, the interaction between gpa and scholarship support and the interaction between preparedness and scholarship support to predict graduation within six years. I used forward and backward stepwise model selection based on Akaike Information Criterion (R function step()) to improve the logistic regression model, and allowed the gradient boosting machine to use an interaction depth of 3. Other settings for the gradient boosting machine (R function gbm()), were standard, recommended starting settings (Ridgeway, 2007). The data set was randomly split into a 5,000-record set to train the models, and a 5,488-record set to test the models. In each case probabilities greater than 0.5 were considered positive predictions (Y_i=1), and those less than or equal to 0.5 were considered negative predictions (Y_i=0). Accuracy was measured as the number of correct predictions divided by the total number of predictions.

Using the student data, the procedure above produced the fitted logistic regression model described in table 3. This model predicted student graduation within six years with 54.5% accuracy. The two strongest positive predictors of graduation in the model are GPA and not receiving Pell grant support, respectively. Receiving UA grant support is not significantly associated with graduation. Given that every student falls into one of the preparedness

Coefficient	Estimate	P-value
Prepared	**-5.233**	**3.12*10^-12**
Unprepared_English	**-5.632**	**8.71*10^-14**
Unprepared_math	**-5.506**	**1.03*10^-13**
Unprepared	**-5.706**	**9.14*10^-15**
Ave. Scholarship	0.00000462	0.8687
Pell, no w/application	**0.2491**	**0.0436**
Pell, no data	0.2999	0.206
UA grant, no w/application	0.7628	0.2666
UA grant, no data	0.7704	0.2661
UA grant, unavailable	0.5325	0.4352
GPA	**1.244**	**< 2*10^-16**
Unprepared*Ave. Scholarship	**-0.00009451**	**0.0279**
Unprepared_English*Ave. Scholarship	0.0001037	0.1077
Unprepared_Math*Ave. Scholarship	**0.0001031**	**0.0211**

Table 3. Parameter estimates for the logistic regression model fitted on University of Alaska student data. Significant parameters in bold.

categories, their strongly negative coefficient estimates can be seen to represent the low baseline probability of graduation among all University of Alaska first-time freshman in the data set, which is less than 0.30. Note that the preparedness coefficients follow the same relative pattern seen in the model described in Table 1. Despite the fact that average scholarship support was not significant, its interaction with levels of preparedness was.

The model fitted by the gradient boosting machine predicted student graduation within six years with 74.6% accuracy. The description in Section 2.3 should make clear that there is no easy way to summarize a gradient boosting model in the way a logistic regression model can be summarized, but examination of the relative influence and partial dependence plots of the variables in the model can provide some insight.

Figure 5 is a plot of the relative influence of the variables in the model. Relative influence is a function of the number of times a variable is selected for splitting, weighted by the improvement to the model as a result of each split, and averaged across all trees. The influences are scaled to sum to 100 (Elith et al 2008). Figure 5 clearly shows the importance of grade point average in predicting graduation. Approximately tied for second are preparedness, already identified as a strong predictor of success, and scholarship support, which is of interest to this study.

Note that the hierarchy of importance of predictors in the gradient boosting model is different from that of the logistic model. While both identify GPA as an important predictor of graduation, gradient boosting model then lists preparedness, scholarships and Pell support, while the logistic regression has lack of Pell support followed by interactions between preparedness and scholarship support. This can be explained by the fact that the relative influence plot in Figure 5 is a measure of the number of times a predictor is involved in decisions, which takes into account interaction 'terms,' as well as simple 'terms.'

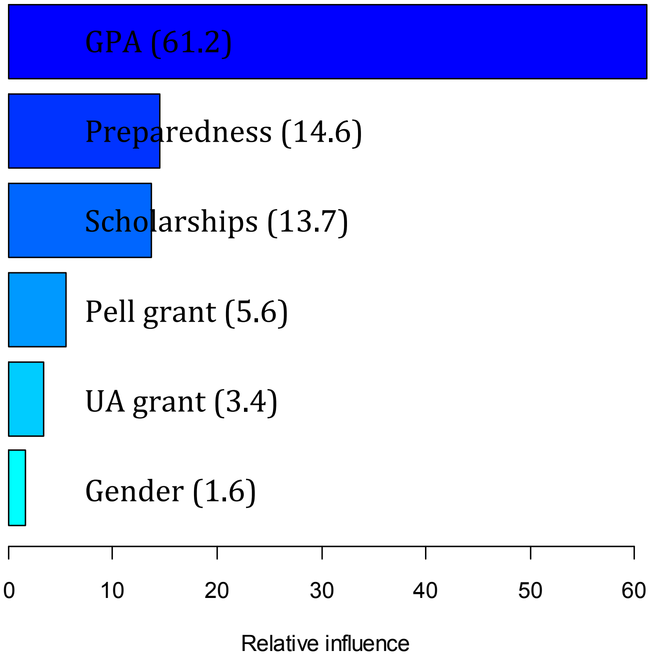

Figure 5. Relative influence of the predictor variables in the GBM model.

Partial dependence plots show a variable's effect on the gradient boosting model after the mean effects of all other variables in the model have been accounted for. Figures 6a and 6b show the partial dependence plots for the variables in the fitted gbm model. The vertical axes show the logit of the probability of graduation.

Figure 6a contains partial dependency plots of the three most influential variables in the gradient boosting model. We see a strong positive relationship between grade point average and probability of graduation, which is intuitively obvious and supports other findings. We also see that being prepared in both math and English is positive predictor, relative to the other categories of preparedness. Being prepared in English but unprepared in math also is associated with an increased probability of graduation, but to a lesser degree, while being unprepared in English or in both math and English is not. First-year scholarship support also appears to be a positive predictor of graduation, at least for the first $5,000. Beyond that, the predicted probability falls.

While Pell grant support was relatively uninfluential in the model, the partial dependence plot does show a relationship that has been observed in the logistic regression model and elsewhere (Mortenson, Brunt 2009), which is that Pell grant support is negatively correlated with probability of graduation. The final two variables, UA Grant status and gender, have so little influence that commenting on their partial dependence plots is not meaningful.

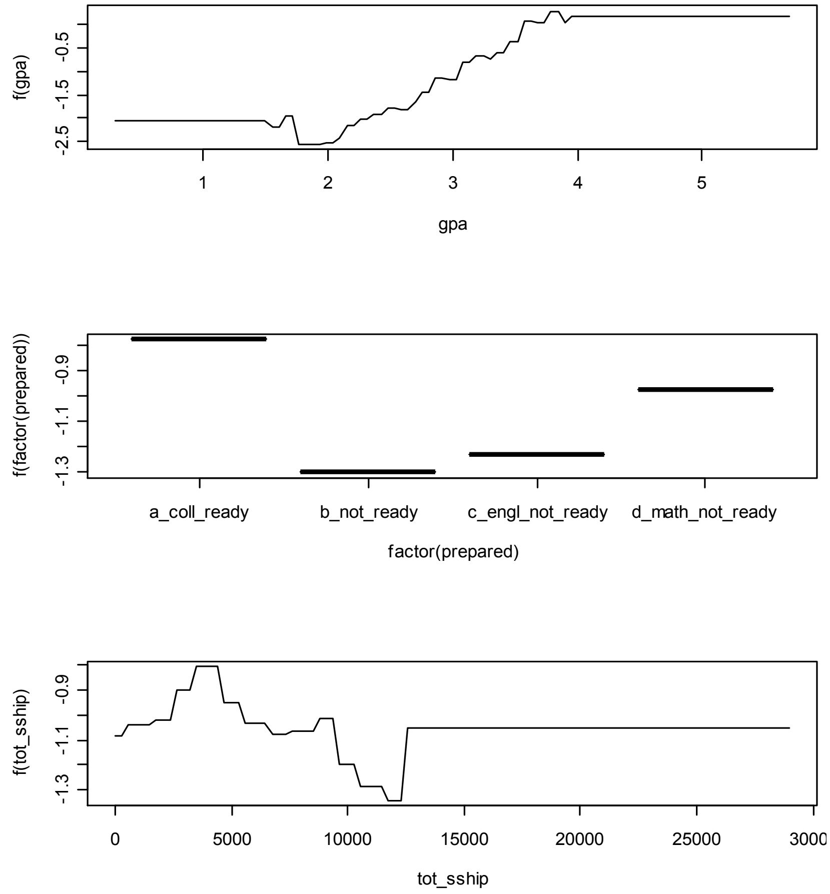

Figure 6a. Partial dependence plots of variables in the fitted gbm model.

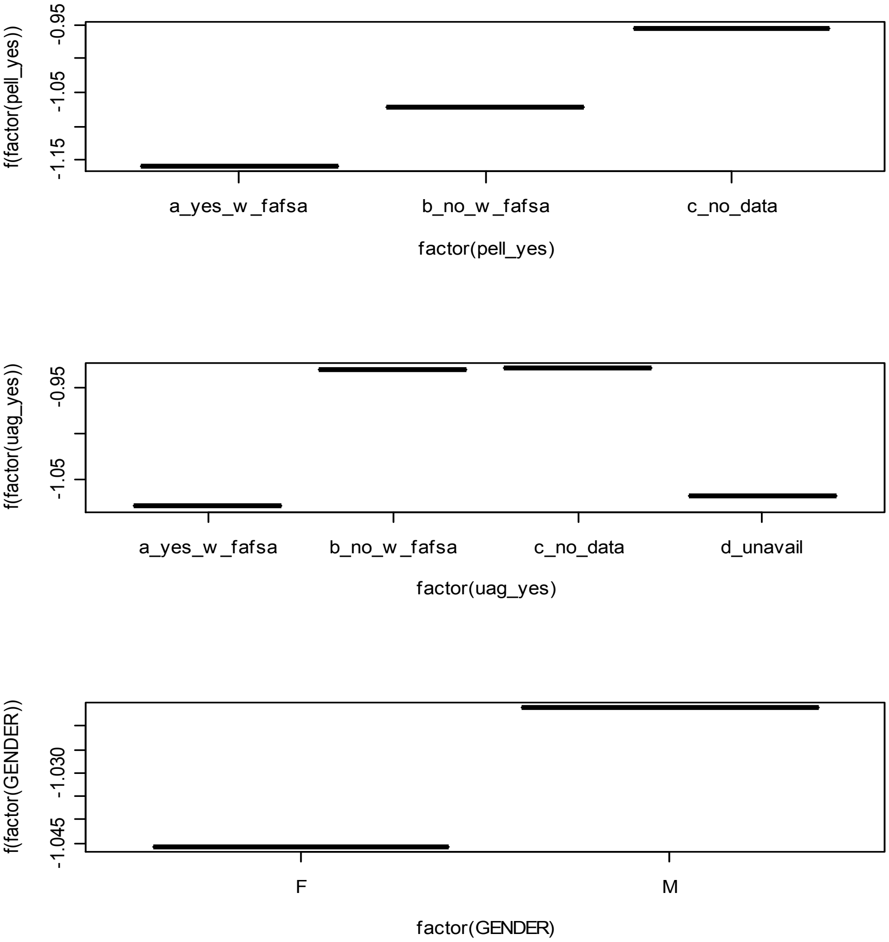

Figure 6b. Partial dependence plots of variables in the fitted gbm model.

4. CONCLUSIONS

This study outlined the theory behind logistic regression and gradient boosting. It reported on preliminary logistic regression model fitting intended to explore the relationship between financial aid, measures of academic preparedness and student success, measured primarily by graduation within six years, and secondarily by retention. These preliminary models suggested that financial aid support, and specifically scholarship support, is associated with an increase in University of Alaska students' probability of staying in school and graduating.

A comparison between the predictive abilities of logistic regression and gradient boosting revealed that when using simulated data with controlled levels of problematic characteristics, logistic regression performed slightly better than did gradient boosting, except when data were missing, when gradient boosting markedly outperformed logistic regression.

It is of interest to examine the historical relationship between predictors, especially those involving money, and measures of success in University of Alaska student data. This allows us to identify successful and unsuccessful, and potentially wasteful, efforts to improve student success. It would also be beneficial to be able to use current data to predict the probability that UA students will be successful, and so identify those most (and least) in need of assistance. UA student data can be problematic. Particularly troublesome is the fact that high school grade point average, a powerful predictor of student success, is missing from large numbers of student records, especially those of the many older, non-traditional students. This study demonstrated that gradient boosting, which markedly outperformed logistic regression in predicting student success when data were missing, has the potential to be a useful tool in attempting to deal with these challenges. This suggests the benefits of further investigation of gradient boosting techniques, especially in the area of model interpretation.

5. APPENDIX

5.1. R code to simulate data.

```
N<-2000
alpha<-0.34; B1 <- 1.45; B2 <- -2.44;
prob<- function(alpha, B1, B2, X1, X2) {
          (exp(alpha+B1*X1 +B2*X2))/(1+exp(alpha+B1*X1 +B2*X2))
                                }
X1<- runif(N, min=-1, max=1)*5
X2<- runif(N, min=-1, max=1)*10
p<- prob(alpha, B1, B2, X1, X2)
YY<- rbinom(N,1,p)
data<-data.frame(y=YY, x1=X1, x2= X2, p=p)
```

The function prob() returns a probability (p) from the logistic response function (7).

YY<- rbinom(N,1,p) assigns a binary response based on the probability generated by the logistic function above.

5.2. R code to create a vector of matching length and defined correlation.

```
corr.vect<- function(var1, rho) {
          N<- length(var1)
                    newvar<-runif(N, min=round(min(var1),0), max=round(max(var1),0) )
          X<-cbind(var1, newvar)
          dim(X) <- c(N, 2)
          M<-array(rho, dim=c(2,2))
          diag(M)<-1
          cF<- chol(M)
          Y<-X%*%cF
          Y[,2]
          }

correlation<- c(0.0,0.1,0.2,0.4,0.75,0.9, 0.95)

X1<- runif(N, min=-1, max=1)*5
X2<- corr.vect(X1, correlation[ii])
p<- prob(alpha, B1, B2, X1, X2)
YY<- rbinom(N,1,p)

data<-data.frame(y=YY, x1=X1, x2= X2, p=p)
```

Note that in the corr.vect() function, M is the desired variance-covariance matrix, with variances 1 on the diagonal, and the off-diagonal covariances equal to rho. A square-root of M is generated by chol(M), the Cholesky decomposition of M. Multiplying X, the n by 2 matrix consisting of [,1], the original vector, and [,2], the randomly generated vector of the same length and amplitude, by the Cholesky decomposition of M returns an n-by-2 matrix of [,1], the original vector, and [,2], a new vector with correlation rho with [,1]. The correlation between variables generated by this function is fairly accurate with large N, say N> 500. With smaller N, and especially with small absolute values of rho, the response becomes more stochastic.

REFERENCES

Elith, J., Leathwick, J.R., Hastie, T. (2008) *A Working Guide to Boosted Trees.* Journal of Animal Ecology, 77, 802-813

Hastie, T., Tibshirani, R., Friedman, J., (2008) *The Elements of Statistical Learning: Data Mining, Inference and Prediction, 2nd Edition.* Springer-Verlag, New York.

Kutner, M.H., Nachtsheim, C.J., Neter, J., Li, W. 2005. *Applied Linear Statistical Models.* McGraw-Hill/ Irwin, Boston, MA.

Mortenson T.G., Brunt, N., *Access to What? Restricted Educational Opportunity for Students from Low and Lower-Middle Income Families – FY1974 to FY2009*, Postsecondary Education Opportunity 207, September 2009.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Ridgeway, G. (2007). *Generalized Boosted Models: A Guide to the GBM Package.* URL http://cran.r-project.org

Ridgeway, G. (2010). *Package 'GBM': Generalized Boosted Regression Models.* URL http://cran.r-project.org

University of Alaska Statewide Budget and Institutional Research. 2011. *University of Alaska in Review.* UA SWBIR, Fairbanks, AK.