Factors Associated With Persistence in Science and Engineering Majors: An Exploratory Study Using Classification Trees and Random Forests

GUILLERMO MENDEZ
Department of Mathematics and Statistics
Arizona State University

TRENT D. BUSKIRK
School of Public Health
Saint Louis University

SHARON LOHR
Department of Mathematics and Statistics
Arizona State University

SUSAN HAAG
Ira A. Fulton School of Engineering
Arizona State University

ABSTRACT

Many students who start college intending to major in science or engineering do not graduate, or decide to switch to a non-science major. We used the recently developed statistical method of random forests to obtain a new perspective of variables that are associated with persistence to a science or engineering degree. We describe classification trees and random forests and contrast the results from these methods with results from the more commonly used method of logistic regression. Among the variables available in Arizona State University data, high school and freshman year GPAs have highest importance for predicting persistence; other variables such as number of science and engineering courses taken freshman year are important for subgroups of the student population. The method used in this study could be employed in other settings to identify faculty practices, teaching methods, and other factors that are associated with high persistence to a degree.

Keywords: classification tree, logistic regression, random forest

I. INTRODUCTION

Many studies have shown a lack of persistence among U.S. students who complete a science and engineering degree (Besterfield-Sacre, Atman, and Shuman, 1997; Brainard and Carlin, 1997; Burtner, 2005; Grandy, 1998; May and Chubin, 2003; LeBold and Ward, 1998; Leslie, McClure, and Oaxaca, 1998; Levin and Wyckoff, 1991; Rayman and Brett, 1995; Seymour and Hewitt, 1997; White, 2005; Zhang, Anderson, Ohland, and Thorndyke, 2004). These studies have identified a number of variables such as high school GPA that are associated with persistence to a degree. Most previous work has identified factors related to persistence using standard statistical methods such as logistic regression. These methods work well for identifying simple relationships in the data. However, when predicting whether a student will graduate with an engineering degree, the relationships are often more complex. For example, female Hispanic students who participate in a mentorship program are more likely to persist to a degree, while some other groups of students in the program are less likely to persist. Such a relationship is easily missed when techniques such as logistic regression are used.

In this paper we use classification trees (Breiman, Friedman, Olshen, and Stone, 1984) to produce a new view of variables associated with persistence to earn a science, technology, engineering, or mathematics (STEM) degree. We also use the recently developed statistical method of random forest (Breiman, 2001), related to tree-based classification methods, to identify factors that may be related to persistence but that might not be identified by other statistical procedures such as logistic regression. The primary goal of this paper is to show how classification trees and random forests can be used to identify factors and interactions not found by other methods.

Zhang et al. (2004) suggested that high school GPA and SAT math scores predicted engineering student graduation. However, these two cognitive variables explained only a small fraction of the overall variability in student graduation persistence rates suggesting that more predictors are needed to fully understand the nature of persistence in science and engineering. A recent study by Burtner (2005) supports the use of non-cognitive variables, such as confidence in college-level math/science ability, in models to predict student persistence. Other studies (Besterfield-Sacre, Atman, and Shuman, 1997; Brainard and Carlin, 1997) have supported Burtner's assertions by demonstrating associations between graduation rates and attitudinal and belief factors such as self-confidence and perceived ability in engineering as well as other factors such as work status, high school ranking, and SAT scores. Levin and Wyckoff (1991) also reported that high school GPA, scores on college placement tests in Chemistry, along with grades in Calculus, Chemistry, and Physics courses were all strong predictors of persistence through the second year of engineering programs. LeBold and Ward (LeBold and Ward, 1988) found that first and second semester grades along with cumulative GPA were strong predictors of persistence for freshmen engineering majors.

The majority of studies investigating persistence in science and engineering have focused on engineering students. Enrollment and tracking of engineering majors may be two key factors related to the

restricted scope of these studies. Students enter university engineering programs early in their tenure (i.e., as freshmen) and their progress is directly tracked by the engineering school or college. Tracking students becomes more complicated for students majoring in STEM since students in these majors can change their area of study to another STEM field among and within colleges offering STEM degrees. While tracking students may differ between STEM and engineering, both groups of students have similar low retention rates. Cognitive and non-cognitive variables previously shown in models predicting graduation persistence in general STEM fields has been largely unexplored in the research literature up to this point.

In this article, we use logistic regression, classification trees, and random forests to study persistence among students who have already entered the "Freshman STEM Pipeline" (FSP) by either declaring a STEM major or by having an "undecided" major while enrolling in at least one STEM course as a freshman. Students who have entered the FSP from Arizona State University's (ASU) 1999–2000 freshman class served as the population of interest for this study. While interest is often given to factors related to enrollment in STEM programs, this paper (like Besterfield-Sacre, Atman, and Shuman, 1997; Levin and Wyckoff, 1991; Zhang, Anderson, Ohland, and Thorndyke, 2004), focuses on those factors that may be associated with the persistence of students who have already entered the pipeline at the university level. Variables in the data set were retrieved from ASU's institutional data and included demographic variables (race, sex, age), cognitive variables (such as High School GPA, SAT scores, etc.), and non-cognitive variables (including work-study status, number of courses, and financial aid support). No attitudinal or belief variables were available for study.

In the next section, we provide additional details of the variables used in the FSP data set along with the criteria used to identify freshmen in the STEM pipeline. In section III we describe logistic regression, classification trees, and the random forest method in more detail and highlight the advantages and disadvantages of each for studying persistence. In section IV we derive both a logistic regression model and complementary classification tree model for predicting persistence in engineering using a subset of the FSP sample comprised of only engineering students. These models were formed using only a subset of the variables to compare our results to those from Zhang et al. (2004). In the second part of section IV, we investigate the more general STEM persistence classification using all variables in our study. Here we construct both a logistic regression model, using stepwise variable selection, and a classification tree model, using the random forest method as the variable selection process. We compare the results of selecting the best subset of predictors using a stepwise logistic regression model to those obtained from random forest. We conclude the paper with a discussion of how researchers and educators may make use of the additional information available from both classification trees and random forests.

II. Data and Definition of Persistence

The data set used for studying STEM persistence consisted of students who enrolled at ASU as a freshman in the 1999–2000 academic year and either (1) declared a STEM major or (2) did not declare a major but enrolled in at least one of the introductory courses described below (n = 1,884). A student is considered to have persisted in STEM if he or she graduates with a degree in a STEM major in May 2005 or earlier. The student has not persisted if he or she has not graduated by that time or has graduated with a non-STEM major. The outcome is recorded after six years as is established practice (Zhang, Anderson, Ohland, and Thorndyke, 2004).

The data set used for studying engineering students consisted of students who enrolled at ASU as freshmen in the 1999–2000 academic year and declared an engineering major as a freshman (n = 348). A student is considered to have persisted in engineering if he or she graduates with an engineering degree in May 2005 or earlier. The student has not persisted if he or she has not graduated by that time or has graduated with a major in a field other than engineering.

A list of occupational categories from the National Science Foundation Scientists and Engineers Statistical Data System (SESTAT) (National Science Foundation, 2007) was used to determine whether a college major should be classified as STEM. All majors related to science and engineering occupations from SESTAT were included, as well as the technology/technical fields from non-science and engineering occupations. A full list of the specific STEM majors, over 150, is available upon request. College courses that freshmen are typically required to take in order to graduate with a STEM degree were labeled as STEM courses and are listed in Table 1.

We were somewhat limited in the data available for analysis by the variables found in the institutional data warehouse at ASU. Thus, for this study, we did not have information available on teaching methods employed by instructors, students' perception of the quality of instruction at ASU, interaction of students with other students or faculty, students' beliefs or attitudes about science, and other factors that may be associated with persistence (Astin, 1985; Haag, Garcia, and Hubele, 2007; Seymour and Hewitt, 1997). In total, 18 variables were collected from the institutional data warehouse at ASU for students who were freshmen in 1999–2000 and are described in Table 2. Only the categorical variables GENDER, ETHNIC and CITIZEN and the continuous variables HSGPA, SATQ and SATV were used in the analysis of engineering persistence. All of the variables in Table 2 were used in the analysis of STEM persistence (descriptive statistics for each variable is available upon request).

III. Statistical Methods for Studying Persistence

A. Multiple Logistic Regression

Multiple logistic regression is one technique commonly used to predict a dichotomous outcome with mutually exclusive categories such as in this study. For student i, let $Y_i = 1$ if the student persists to graduation and $Y_i = 0$ if the student does not persist. The response of interest is the probability that student i persists or $pi = P(Y_i = 1)$. The predicted values from a multiple logistic regression model are

$$\text{logit}(\hat{p}_i) = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_m x_{im}, \quad (1)$$

where $\text{logit}(\hat{p}_i)$, $\ln[\hat{p}_i/(1-\hat{p}_i)]$, $x_{i1},\ldots,x_{im}$ are explanatory variables for student i, $\hat{\beta}_0$ is the estimated intercept, and the estimates $\hat{\beta}_1,\ldots,\hat{\beta}_m$ (slopes) are maximum likelihood estimates (Neter, Kutner, Nachtscheim, and Wasserman, 1996). For variable j in the model,

Course ID	Course Description	Course ID	Course Description
BIO 187	General Biology I	MAT 271	Calculus w / Anal. Geometry II
BIO 188	General Biology II	MAT 272	Calculus w / Anal. Geometry III
CHM 113	General Chemistry	MAT 274	Elementary Differential Equations
CHM 114	General Chemistry for Engineers	MAT 290	Calculus I
CHM 115	General Chemistry w / Qual. Anal.	MAT 291	Calculus II
CHM 117	General Chemistry for Majors I	PHY 111	General Physics
CHM 118	General Chemistry for Majors II	PHY 112	General Physics
ECE 100	Intro to Engineering Design	PHY 113	General Physics Laboratory
ECE 201	Electrical Networks I	PHY 114	General Physics Laboratory
ECE 210	Engineering Mechanics I: Statics	PHY 121	University Physics I: Mechanics
GLG 101	Intro to Geology I (Physical)	PHY 131	University Physics II: Elec./Mag.
MAT 170	Precalculus	PHY 150	Physics I
MAT 270	Calculus w / Analytical Geometry I	PHY 151	Physics II

Table 1. A description of ASU classes that were considered STEM courses. These classes are introductory courses for the different STEM majors.

$\exp(\hat{\beta}_j)$ is the estimated odds ratio; if $x_{i1} = 1$ for males and 0 for females, then

$$\exp(\hat{\beta}_1) = \frac{\text{odds of persistence of males}}{\text{odds of persistance of females}}$$

for persons with the same values for all other variables. Estimated odds ratios along with 95 percent confidence intervals were calculated using SAS (version 9.1). Odds ratios can be better described as *adjusted* odds ratios since they control for all other variables in the model.

Most studies investigating factors related to persistence in engineering have used logistic regression models with stepwise variable selection (Besterfield-Sacre, Atman, and Shuman, 1997; Levin and Wyckoff, 1991; White, 2005; Zhang, Anderson, Ohland, and Thorndyke, 2004), or a classification model such as discriminant analysis (Burtner, 2005). While these standard statistical techniques are useful, they have restrictive assumptions about the data structure including linearity or additivity of factor effects. Furthermore, finding conditional relationships among the factors, for example, taking a higher number of credit hours may be associated with higher graduation persistence for some groups of students but not for others, may require numerous interaction terms in the logistic regression model. These terms usually require large amounts of data for proper estimation. In addition, the odds ratios that are obtained from logistic regression are useful for understanding the impact of "unit" changes in the factors on the likelihood of persistence. However, using the odds ratios alone will not identify important "ranges" of continuous factors or common clusters for categorical factors for persistence classification. Finally, important variables can be masked by other variables in a model, which is typical in all regression methods. For example, if HSGPA and CUMGPA were both included as predictors in the model and are highly correlated, each might mask the effect of the other and it is possible that neither variable would be statistically significant in the full model.

In our study, we collected values for 18 variables yet many students have missing values for some of the covariates. In a logistic regression model with HSGPA as a covariate, if a student's record does not contain the value of HSGPA, we cannot use the logistic regression equation to predict the probability that the student will persist. That student's record also cannot be used to build the logistic regression model unless a value is imputed for the missing covariate, but imputation requires additional and often unverifiable modeling assumptions. When many students are missing values for one or more of the covariates, serious depletion of the data set can occur especially as the number of variables increases.

B. Classification Trees

Classification trees take a different approach, compared to logistic regression, to predict categorical responses. A classification tree is constructed by partitioning the data into separate regions in which the predicted classification is constant within each region. An unlimited number of variables can be used to build the tree model since the tree building process inherently selects the best covariates by considering all possible binary splits on the variable's

Variable	Description
GENDER	Gender of student (F=Female, M=Male)
ETHNIC	Race / Ethnicity of Student A=Asian / Pacific Islander, B=Black, H=Hispanic, N=Native American / Alaskan Native, W=White / Caucasian
CITIZEN	Citizenship status of Student: Citizen=U.S. Citizen, ResAlien=U.S. resident but not a Citizen, NRAlien=Foreign student
HSGPA	High school cumulative grade point average ([0.0, 4.0] scale)
SATQ	SAT quantitative exam score (see note)
SATV	SAT verbal exam score (see note)
HALL	Did student live in residence hall freshman year? (N=No, Y=Yes)
WORKSTUDY	Was freshman in a work-study program? (N-No, Y=Yes)
SCHOLAR	Did freshman receive a scholarship? (N=No, Y=Yes)
LOAN	Did freshman receive a loan for education / living expenses? (N=No, Y=Yes)
ATHLETE	Was freshman an athlete? (N=No, Y=Yes)
AZRES	Was freshman a resident of Arizona? (N=No, Y=Yes)
FTF	Was student a first-time freshman? (N=No, Y=Yes)
FRESHCALC	Did student take calculus course in freshman year? (N=No, Y=Yes)
NUMSTEM	Number of STEM courses in freshman year (see Table I)
TOTHOURS	Total hours enrolled in freshman year
CUMGPA	Cumulative grade point average after freshman year ([0.0, 4.0] scale)
AGE	Age of student as freshman in years

Note: Some SATQ and SATV values are based on ACT to SAT concordance tables (Dorrans, 1999)

Table 2. Description of the 11 categorical and seven continuous variables that were used when modeling STEM persistence. Only the first six variables were used when modeling engineering persistence.

range. The modeling process then selects the best set of branches that minimizes the misclassification rate. Where a binary split occurs is called an internal node, whereas the end of a branch is a terminal node which contains a predicted category. The predicted response in that node is the category with the majority of cases. Classification trees can be constructed using commercial software such as SAS Enterprise Miner or CART by Salford Systems; in this paper, we construct trees using the rpart add-on package in the R system for statistical computing (version 2.4.0) (Maindonald and Braun, 2003; The R Project, 2007).

Before we discuss the details of how trees are formed, we first demonstrate how predictions are made. Given a tree model, a prediction results after answering a series of yes/no questions about an individual. If the answer is yes, the left path is taken, otherwise the right path is chosen. This continues down the tree model until a terminal node is reached and, hence, a prediction is made. For example, consider a data set recording CLASS status (1st, 2nd, 3rd, or Crew), SEX (M = Male or F = Female), and AGE (Child or Adult) for each person on board of the fateful voyage of the Titanic. The data are based on information originally collected by the British Board of Trade (Great Britain Parliament, 1990). Suppose we wanted to investigate which variables were associated with the survival status of passengers. In this example, the response of interest is whether a passenger on Titanic survived, "Y", or did not survive, "N". Figure 1 shows the classification tree that is formed from the data. Now, consider a male child passenger who happened to be traveling in 1st class. To predict whether he survived or not, we answer each question down the tree. Firstly, since the passenger is male, the left path is initially chosen; secondly, since he is not an adult, the right path is chosen; finally, his 1st class status lands him in terminal node 5 predicting that he survived. If he were traveling in 3rd class, then the prediction would have been that he did not survive (terminal node 4). The ease of predicting in this example demonstrates the advantage of interpretability that tree models possess.

To build a tree model from a given data set, two criteria are needed: one to select the best split at a node while building the tree model and another for "pruning" to find the right-sized tree. These criteria can be the same but in this study the Gini index was used to build the trees while the misclassification rate was used to prune the trees. For a binary response, such as predicting whether a student persists or does not persist, the Gini index at node m is

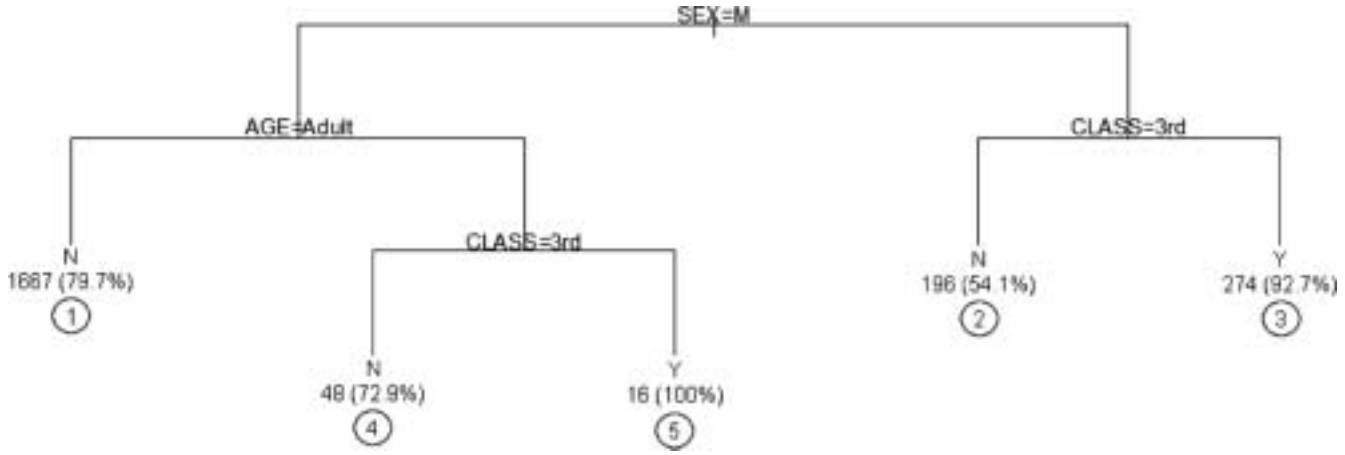

Figure 1. Example of a classification tree using the Titanic data set. Terminal nodes 1 through 5 are labeled by the appropriate encircled number and a "Y" and "N" in the node refers to whethere a passenger on Titanic survived or not, respectively. The larger number in the node description refers to the number of passengers who were classified at that point in the tree and the percent value refers to the percentage of those passengers who were correctly classified by the tree model.

defined by

$$\text{Gini index}(m) = 2\hat{p}_m(1 - \hat{p}_m) \quad (2)$$

where $\hat{p}_m$ is the proportion of students in node m who persist. The Gini index is 0 if all of the students in the node are in one category, either all persist or all do not persist. Such a node is called homogeneous. It attains its highest possible value of one-half if half of the students in the node are in each category. When building a tree, the process selects the variable and corresponding split point which gives the smallest value of the Gini index for each node. We selected the Gini index as the splitting criterion to build the tree because it is more sensitive to changes in the node probabilities than misclassification rate (Hastie, Tibshirani, and Friedman, 2001). Using this criterion, trees are built to maximum depth (e.g., very large trees where the terminal nodes are as homogeneous as possible). This process results in a tree with many nodes (split points) that is too data-specific. This is analogous to a linear model fit with all main effects and every possible interaction term. To fix this problem, the trees are "pruned" by using the misclassification rate and a tuning parameter $\alpha \geq 0$ that governs the tradeoff between tree size and its best fit to the data. The optimal tuning parameter is found by using a cross validation method (Breiman, Friedman, Olshen, and Stone, 1984; Ripley, 1996). If the data set is sufficiently large, part of the data may be used to build the tree with the remainder used for validating the predictions.

We noted in section III.A that logistic regression requires every student to have complete data. On the other hand, classification trees use the whole, possibly incomplete, data set and handle missing values by constructing surrogates (alternate variables) for each variable in the model. During the tree building process, the best (primary) variable and split point are selected at each node by using all available data. Then a list of surrogates and split points is found for each node. The first surrogate and corresponding split point is the one that best mimics the split of the data by the primary split. The second surrogate and corresponding split point is the second best at mimicking the action of the primary split on the data, and so on. In the end, each internal node has a set of t surrogates with sequentially declining mimicking ability with respect to the primary split. During prediction of a new observation, if a split depends on a variable whose value is missing, the best surrogate split is used to determine the path down the tree. If that value is missing, then the second best surrogate split is used, and so on. In essence, the surrogates exploit the correlation between the predictors to reduce the loss of information from missing values. The number of surrogates, t, can be set and should be at least the number of missing values that can be expected in future observations. Also, even though the surrogates are not the optimal variables, they should not be ignored. Sometimes they can detect an important variable which is being masked by the data structure or suggest an alternate branching of the tree model. Some evidence of this phenomenon is when a surrogate has a high concordance measure, or percent of agreement between the primary split and surrogate split. By analyzing the surrogates along with the overall tree model, a broader picture of the data structure can be achieved.

Along with handling missing values, classification trees have other advantages over traditional linear models. As shown in the Titanic example above, prediction using a tree model involves answering a series of yes/no questions in an intuitive and simple manner. Yet this efficient form makes powerful use of conditional information in handling non-homogeneous relationships. In other words, after the data have been split into two parts based on one predictor, the optimal split of another predictor, possibly the same, is searched individually within each part. Hence conditional relationships can be modeled using a tree. In the Titanic example, note that female passengers traveling in 3rd class are predicted to have not survived, whereas females in other classes are predicted to have survived. Modeling such relationships in logistic regression would require interaction terms for SEX and CLASS.

Another advantage of the tree model is the ability to estimate the probability of correct classification. This is done by dividing the number of correct classified observations at a terminal node m by the total number of cases at that node. This gives an idea of the accuracy of the prediction at node m. Note that the correct classification percent will be greater than or equal to 50 percent for each terminal node since it corresponds to the majority class for that node. In this paper, the tree diagrams show the number of observations and percent of correct

classifications in parentheses under each terminal node. Finally, tree models are also highly robust with respect to outliers or misclassified points since one essentially counts how many cases of each class go left or right. This is similar to the robustness property of median values (Breiman, Friedman, Olshen, and Stone, 1984).

C. Random Forest

As stated before, the tree building process inherently selects the best variables from the full set of covariates using the Gini index criterion. Still, ranking the variables with respect to predictive ability cannot easily be done. The importance of a variable at one branch is usually not the same at a different branch down the tree, since the data are recursively split into subsets at each node. In this paper, we used importance scores from random forest to help select the best subset of variables to build the STEM tree model.

A random forest (Breiman, 2001) is a model that consists of many classification trees. Typically, as many as 500 unpruned trees are built as described above. Each tree is constructed using a different with-replacement bootstrap sample from the data. To reduce the correlation between trees, only a few of the variables are randomly sampled as candidates at each split, usually the square root of the number of variables in the data set. From this random construction of multiple trees, the name is derived. The prediction of the random forest model is the category with a majority of votes across all trees in the forest. Because of the large number of trees, the predictions tend to be more accurate than those from a single classification tree, yet the random forest model can be more difficult to interpret than a single tree. Therefore, a random forest model is sometimes thought of as a "black box" without much to say about the relationship between the response and the explanatory variables. In this paper, the random forest method is not directly applied to model STEM persistence. Instead, a byproduct of the random forest model, variable importance scores, is used to help find the optimal subset of variables to build a single classification tree. In this sense, the random forest method provides an assessment of variable importance that does not depend on a specified model structure.

In logistic regression, when model selection methods are used to reduce the number of variables, interactions among the explanatory variables must be stated explicitly in the model. The random forest method, however, allows complex subsetting and interactions that are difficult to express in a traditional regression model. Also, this technique can list the variables in order of predictive ability or importance. By analyzing the importance scores, a large set of variables can be reduced to a working subset without making any model assumptions. We used this tree-based technique in a similar fashion as the stepwise selection procedure is used in logistic regression.

To calculate the importance score of variable *m*, first the sum of Gini indices over all nodes of a tree is calculated. Then the values of variable *m* are randomly scrambled among the observations and the sum of Gini indices is again calculated. The mean decrease in Gini index among all trees is the importance score of variable *m*. As stated before, the Gini index is apt to measure the impurity of nodes and therefore was chosen as the criterion to measure variable importance. Because multiple trees are used and each tree includes only some of the explanatory variables, the importance scores detect variables that are predictive of the response only for some of the subgroups of the data and variables that are highly correlated with other predictors. A researcher using logistic regression, by contrast, must include complex interactions in the model to detect such variables.

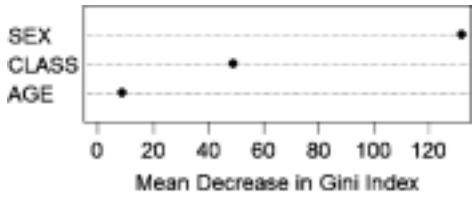

Figure 2. Importance scores for each variable with respect to survial status using the Titanic data set.

Figure 2 shows the importance scores for each variable in the Titanic data set. Note that the variable SEX is the most important variable with respect to predicting survival status, while CLASS and AGE are second and third, respectively. The importance ranking for this example roughly matches the order of splits (yes/no questions) of the Titanic classification tree in Figure 1.

IV. Analysis and Results

We study two responses using the methods outlined in section III: persistence of engineering students and persistence of students in STEM majors. We first look at engineering persistence to see how the tree models complement logistic regression on variables that were found important in (Zhang, Anderson, Ohland, and Thorndyke, 2004). We then compare the logistic regression analysis to tree-based analysis with respect to STEM persistence.

A. Engineering Persistence

In this section we illustrate how classification trees and the random forest method can be used as companion methods to the results obtained from the more traditional approach of stepwise logistic regression. In particular, we demonstrate how classification trees can yield information on important ranges of continuous variables or groups of categories for categorical variables. Moreover, we also show how classification trees can be useful for visually deciphering conditional relationships among the variables relating to persistence classification. This would otherwise be accomplished through a series of interaction terms in the logistic regression models.

While modeling engineering persistence, list-wise deletion was used. In other words, whenever a value was missing from any of the predictors, the student was excluded from the study. This type of deletion is used because the optimization technique in logistic regression cannot handle missing values. Although classification trees can handle missing values through the use of surrogates, as was explained in section III.B, excluding these students from the study allows us to accurately compare results using both methods.

To begin, we apply the approach of Zhang et al. (2004) to 1999 freshmen engineering students from Arizona State University using the first six variables described in Table 2 in a stepwise logistic regression (with a significance level for entry and retention both equal to 0.05) to identify variables that effectively predict graduation persistence in engineering. The exact levels of the categorical variables used in the models derived by Zhang et al. (2004) match the descriptions given in Table 2. Two of the categorical variables were modified for our model to avoid difficulties with estimation as well as to protect the confidentiality of the students in sparse categories. In our model ETHNIC has four levels: Asian (A), Black

(B), Hispanic (H) and White (W) which also includes Native American. We also reduced the CITIZEN variable from three to two categories: U.S. Citizen and Non-U.S. Citizen. All other variables and levels of variables concur with the definitions in Zhang et al. (2004). Altogether, complete information for the six variables of interest was available for 348 out of 684 freshmen engineering students from ASU in the 1999–2000 school year.

Using the stepwise selection process, only four of the six variables were chosen in the logistic regression model. The likelihood ratio test statistic for the test of the global null hypothesis of no dependence between graduation persistence in engineering and the variables HSGPA, ETHNIC, CITIZEN and SATQ was calculated to be 61.425 (p-value < 0.0001). The maximum rescaled R-squared for the model reported in Table 3 was 0.240. The statistics and odds ratio confidence intervals for the variables presented in our model along with the model fit indices are fairly consistent with those presented by Zhang et al. (2004) accounting for the total sample size. The final model from which the odds ratio estimates in Table 3 were derived is given by

$$\begin{aligned}\text{logit}(\hat{p}) = &-10.295 + 1.571(\text{HSGPA}) + 1.491(\text{ASIAN}) \\ &-0.056(\text{BLACK}) - 0.120(\text{HISPANIC}) \\ &+1.717(\text{CITIZEN}) + 0.005(\text{SATQ}). \qquad (3)\end{aligned}$$

The logistic regression model suggests that Asian engineering freshmen have about 4.4 times higher odds of persisting compared to White freshmen engineering students. Both Hispanic and Black engineering students have lower odds of persistence compared to White students, but these differences are not statistically significant. Thus, impact of ethnicity on the likelihood of persistence seems to be driven in large part by the differences in persistence odds between Asian and White students. Engineering students having a high-school GPA of 3.5 have almost five times higher odds of persisting compared to engineering students who have a high-school GPA of 2.5, for example. The odds of persistence in engineering increases by a factor of $1.051 = (1.005)^{10}$ for every ten point increase in the quantitative SAT score. Finally, students who are non-U.S. citizens have about 5.6 times higher odds of persistence when compared to engineering students who are U.S. citizens.

Additional information about the conditional relationships between these variables and engineering persistence classification may be explored by using a classification tree. We display one such classification tree in Figure 3 that was derived using the four variables that were identified by the stepwise logistic regression model described in Table 3 and in model (3).

Recall that the tree building process inherently selects the best predictors through the recursive splitting of the data. Figure 3 reveals that the data were initially split on HSGPA which, in part, agrees with the results of the logistic model in the sense that it is the strongest predictor of engineering persistence. From the tree model, we see that there are no additional predictors useful for classifying students as persistent in engineering if those students have high school GPA below 3.595. In fact, according to the tree model, the vast majority of engineering students with HSGPA < 3.595 (i.e., 158 out of 181 or 87.3 percent) do not persist in engineering (see terminal node 1 in Figure 3). This result suggests that even though students may do well in high school, the transition into and retention throughout an engineering program is difficult. The Gini index for terminal node 1 is 0.222 which suggests that the node is close to homogeneous.

If we only focus on students whose HSGPA ≥ 3.595, then we see that the next split is based on the student's ethnicity which was the second best predictor of engineering persistence in the logistic regression model. Specifically, given a HSGPA that is at least 3.595, Asian students persist with a relatively high estimated probability (80 percent, given in terminal node 2 in Figure 3) while the persistence of Blacks, Hispanics, and Whites is based on additional information about their HSGPA as well as their SATQ scores (see terminal nodes 3–7). Other than Asian students, the only other group of students with a large estimated probability of persistence (i.e., 71.4 percent) are non-Asian engineering students with a

Variable (Comparison)	Odds Ratio Estimate	95% CI for Odds Ratio	Wald χ^2	(p-value)
HSGPA	4.810	[2.182, 10.606]	15.160	(<0.0001)
ETHNIC (Overall)	–	–	12.666	(0.0054)
ETHNIC (Asian vs. White)	4.439	[1.919, 10.67]	–	–
ETHNIC (Black vs. White)	0.946	[0.187, 4.787]	–	–
ETHNIC (Hispanic vs. White)	0.887	[0.314, 2.507]	–	–
SATQ	1.005	[1.001, 1.009]	7.083	(0.0078)
CITIZEN (Non-US Citizen vs. US Citizen)	5.570	[1.229, 25.243]	4.962	(0.0259)

Table 3. Estimates of odds ratios with 95 percent confidence interval. Wald χ^2 and p-values calculated using SAS based on stepwise logistic regression modeling the probability of engineering persistence.

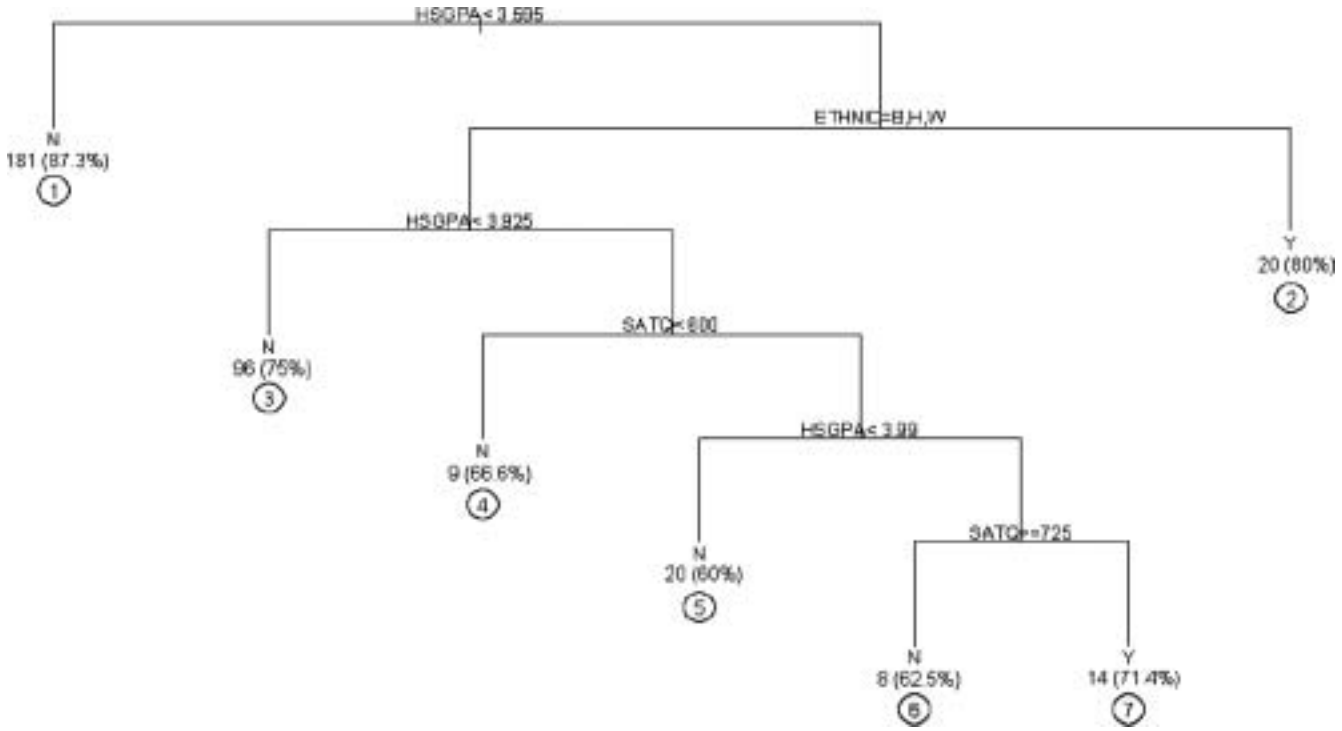

Figure 3. Classification tree formed using the variables identified as significant predictors of Engineering persistence based on the stepwise logistic regression model. Terminal nodes 1 through 7 are labeled by the appropriate encircled number and a "Y" and "N" in the node refers to persisting and non-persisting engeneering students, respectively. The larger number in the node description refers to the number of students who were classified at that point in the tree and the percent value refers to the percentage of those students who were correctly classified by the tree model.

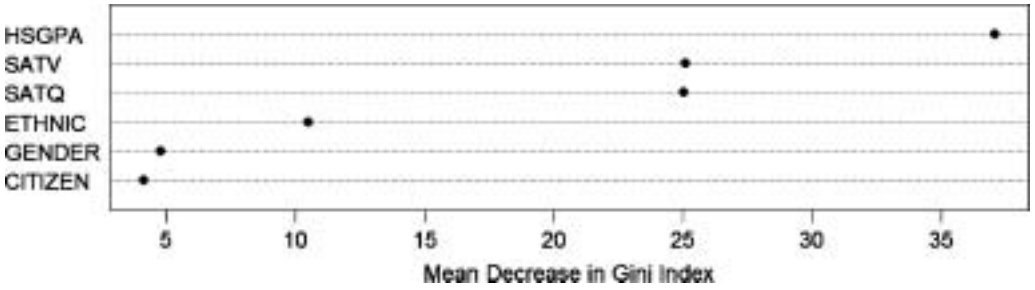

Figure 4. Importance scores per variable with respect to engineering persistence. An importance score is a measure of predictive ability and therefore variables with large scores should be kept in the classification tree model.

perfect HSGPA of 4.0 and an SATQ score between 600 and 725 as illustrated by terminal node 7.

An engineering student's quantitative SAT score was found to be a significant predictor of the probability of persistence in both the logistic regression model and the classification tree, specifically, terminal nodes 4, 6, and 7. However, the influence of SATQ is not limited to these nodes in the classification tree. In fact, SATQ was the best surrogate for HSGPA, with the highest concordance percent between any two splits based on HSGPA being 69.4 percent and occurring at HSGPA < 3.925. While both citizenship status and ethnicity were important predictors in the logistic regression model, only ethnicity appeared in the classification tree. However, the effect of citizenship was not absent from the classification tree as CITIZEN = U.S. Citizen was the best surrogate for the ETHNIC = B, H, W split having a concordance measure of 89.2 percent. In other words, if the ethnicity of a freshman is unknown, the citizenship status of the student can be used instead to decide whether to go right or left at this node with a high level of agreement.

The random forest analysis using all six variables gives additional information through the ranking of their importance scores displayed in Figure 4. The logistic regression model (3) did not include SATV because of its strong correlation with SATQ ($r = 0.5220$, p-value < 0.0001). The masking of the SATV by SATQ, ETHNIC, and CITIZEN might eliminate it from further investigation in a logistic regression analysis. However, the importance score for SATV is much higher than the variables ETHNIC and

CITIZEN that were selected by the stepwise logistic procedure. If the ranking of importance scores were used to select a subset of variables for the construction of a classification tree, only HSGPA, SATV, SATQ, and possibly ETHNIC would have been chosen.

In general, the results of the classification tree are consistent with those of logistic regression. However, the information obtained from the classification tree model and importance scores may further illuminate the results of the logistic regression. For example, while HSGPA is a strong predictor of persistence in the logistic regression model (as demonstrated by the large odds ratio estimate), little information is obtained about what portion of the range of HSGPA is most crucial for understanding persistence. It is not obvious from the logistic regression model that HSGPA's below 3.595, for example, should imply a risk factor for non-persistence. Moreover, without interaction terms in the logistic regression model, it is not clear that this range for non-persistence risk would increase to 3.595 through 3.925 for non-Asian students, for example. The classification trees provide a method for pinpointing important points in the ranges or groups of categories of predictors outright, conditional on levels of previous predictors.

B. STEM Persistence

We now turn to investigating general STEM persistence by first formulating a logistic regression model and then by constructing a classification tree. Whereas before we constructed a classification tree based on the variables of the logistic model, in this section we construct a tree model independent of the logistic regression analysis. We compare the results of (1) the subset of variables each method selects from those in Table 2, and (2) the information gained from each type of model.

Since ASU allows both SAT and ACT scores to be used for admissions, about half (47.1 percent) of students in the FSP data set did not take the SAT test, but many of these students did take the ACT test. Instead of deleting the cases with missing SAT scores, we used concordance tables to impute SAT verbal and SAT quantitative scores from ACT English and ACT math scores, respectively. These concordance tables are based on a study by Dorrans (1999), using 103,525 students from 14 universities and two states. After imputing the missing SAT scores, some students still had missing values for other variables. In the end, 1497 of the 1884 students had complete records and therefore only these were used to form the STEM logistic regression model. On the other hand, the STEM classification tree was constructed using all information from the FSP data set including the imputed SAT scores.

Using the same stepwise selection procedure as in the previous section, a logistic model was formulated with likelihood ratio test statistic of 326.180 (p-value < 0.0001) and maximum rescaled R-squared value of 0.303. The final model from which the odds ratio estimates in Table 4 were derived is given by

$$\begin{aligned}\text{Logit}(\hat{p}) = &-5.830 + 1.382(\text{CUMGPA}) + 0.421(\text{NUMSTEM}) \\ &+1.214(\text{ASIAN}) - 0.760(\text{BLACK}) \\ &-0.441(\text{HISPANIC}) \\ &-0.183(\text{NATIVE AMERICAN}) \\ &-0.038(\text{TOTHOURS}) + 0.004(\text{SATQ}) \\ &-0.003(\text{SATV}). \end{aligned} \quad (4)$$

According to this model, the variable most highly associated with STEM persistence was CUMGPA, the cumulative GPA after freshman year, with an odds ratio of approximately 4, indicating

Variable (Comparison)	Odds Ratio Estimate	95 percent CI for Odds Ratio	Wald χ^2	(p-value)
CUMGPA	3.983	[3.082, 5.148]	111.566	(<0.0001)
NUMSTEM	1.524	[1.346, 1.725]	44.301	(<0.0001)
ETHNIC (Overall)	–	–	31.589	(<0.0001)
ETHNIC (Asian vs. White)	3.368	[2.081, 5.449]	–	–
ETHNIC (Black vs. White)	0.468	[0.126, 1.731]	–	–
ETHNIC (Hispanic vs. White)	0.643	[0.377, 1.097]	–	–
ETHNIC (Native American vs. White)	0.833	[0.300, 2.316]	–	–
TOTHOURS	0.963	[0.943, 0.984]	12.220	(0.0005)
SATQ	1.004	[1.002, 1.006]	12.094	(0.0005)
SATV	0.997	[0.995, 0.999]	10.45.	(0.0012)

Table 4. Estimates of odds ratios with 95 percent confidence intervals, Wald χ^2 and p-values calculated using SAS based on stepwise logistic regression modeling the probability of STEM persistence.

that STEM students with high CUMGPA are more likely to persist. Among the other factors, SATQ and NUMSTEM are positively associated with persistence, while TOTHOURS and SATV have a negative association, after adjusting for other variables in the model. Note that if we fit a logistic regression model with only one variable, SATV, the odds ratio for SATV in that model is 1.003 (with 95 percent CI [1.002, 1.004]. The predicted model in (4) has SATV odds ratio less than one because of multicollinearity in the data, which can change the sign of coefficients (Neter, Kutner, Nachtscheim, and Wasserman, 1996, Chapter 7), and reflects more complex relationships in the data. A referee suggested that of two students with the same SATQ, the one with the higher SATV score may be more likely to drop out of a STEM field because that student may have more options in other non-STEM disciplines. Finally, the significance of ethnicity with respect to STEM persistence is driven primarily by the difference in persistence odds between Asian and White students, as in the analysis of engineering persistence. Specifically, Asian students are more likely to persist than White students while Blacks, Hispanics, and Native American students have no significant differences compared to White students.

We then constructed a classification tree modeling STEM persistence as was done in section IV.A except we reduced the number of variables by analyzing the importance scores from random forest and not stepwise selection. This step could have been skipped and a tree model could have been grown using all 18 variables but this process allowed us to estimate the ranking of the variables with respect to association with STEM persistence. Also, analyzing importance scores allowed us to detect variables which may be masked by other predictors in the final tree model. We note that we used all 1,884 observations in the FSP data set from now on since classification trees and random forests can handle missing covariates.

Figure 5 shows the importance scores for all 18 variables with respect to STEM persistence using the random forest method. The highest importance score belongs to CUMGPA, which is consistent with the results of stepwise selection for logistic regression. The other top performers are HSGPA, SATQ, SATV, TOTHOURS and NUMSTEM listed in decreasing order of importance. While there is no standard cutoff score or statistical test that we can use to gauge the importance of a variable compared to noise, we can see that these six variables should clearly be kept in the model, whereas the importance of ETHNIC appears to be ambiguous. Since the importance scores of AGE, FRESHCALC, LOAN, etc. decrease substantially, we discarded any variable that had an importance score of 20 or less and kept the top seven variables, which includes ETHNIC, to construct the STEM classification tree.

With the exception of HSGPA, the variables selected by comparing importance scores are the same as those selected using stepwise selection in Table 4. In the logistic regression analysis, HSGPA is being masked by other predictors in the model, such as CUMGPA which is highly correlated with HSGPA ($r = 0.4869$, p-value < 0.0001). Therefore HSGPA was not included in model (4) since it is not statistically significant. By contrast, the random forest method detects the importance of HSGPA and ranks it second with respect to predicting STEM persistence. Although one often would discard HSGPA in a regression model to avoid multicollinearity problems in the data (Neter, Kutner, Nachtscheim, and Wasserman, 1996), the construction of a tree model is not hindered by the correlation between these two predictors. Instead the correlation becomes a benefit for prediction with trees, since HSGPA is likely to become the best surrogate for any splits based on CUMGPA. Also, any conditional information from HSGPA, not explained by CUMGPA, may arise further down a tree after splitting on other variables.

Using the most important seven variables in Figure 5, a classification tree was constructed and is displayed in Figure 6. We see that the only variables used to split the data are CUMGPA, NUMSTEM, ETHNIC, and SATV. As explained in section III.B, other variables that were deemed important, HSGPA, SATQ, and TOTHOURS, are being masked by the primary variables and are used as surrogates. The fact that CUMGPA is the first variable at the top of the tree corresponds to it also having the highest importance score in the random forest analysis. As might be expected, the best surrogate for this split is HSGPA < 3.785

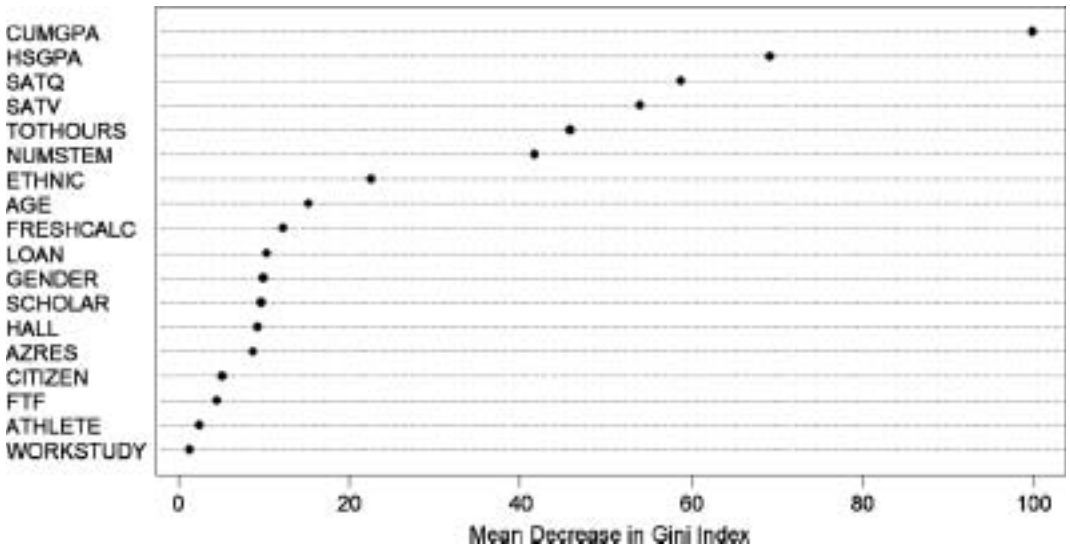

Figure 5. Importance scores per variable with respect to STEM persistence. An importance score is a measure of predictive ability and therefore variables with large scores should be kept in the classification tree model.

with a concordance percent of 71.3 percent. This very high HSGPA cut off point suggests that many students with cumulative GPAs less than 3.175 also had high school GPAs less than 3.785, and that the HSGPA variable is overshadowed as a predictor by CUMGPA. We can also see from terminal node 1 that students with CUMGPA < 3.175 do not persist in STEM with very high estimated probability (1102 out of 1245 or 88.5 percent). Terminal node 1 contains 1,245 of the 1,884 students in the data. Therefore the variables and their split points on the initial right branch of Figure 6 (terminal nodes 2–8) are based on just 34 percent of the data ($n = 639$). As we move down the tree, the subsets of data become even smaller yet more homogeneous. However, this recursive subsetting and homogenization of the data is what makes tree models so effective in detecting interactions.

We now discuss students with a CUMGPA of at least 3.175 who constitute terminal nodes 2–8. We can see that NUMSTEM < 1.5 is the predictor used to split this subset of students, although outcomes for Asian students are not dependent on this split. This is true because regardless of the number of STEM courses taken, the next split down the tree predicts Asian students will persist in STEM with estimated probability of 75 percent and 84.6 percent, in terminal nodes 3 and 4, respectively. Focusing on terminal node 3, among students who take at most one STEM course in freshman year (NUMSTEM < 1.5), only Asian students are likely to graduate with a STEM major.

Note from Figure 6 that Asian students with CUMGPA $\geq$ 3.175 are predicted to persist in a STEM major regardless of their value of NUMSTEM. For students in other ethnic groups, their classification depends on other variables including NUMSTEM, CUMGPA, and SATV. To detect this difference between Asian students and others using logistic regression, we would have needed to include four NUMSTEM_ETHNIC interaction terms in the model, one for each level of ETHNIC excluding White. When this was done, the interaction effect was not significant at the $\alpha = 0.05$ level suggesting the interaction may be too complex to be detected in logistic regression.

Focusing now on terminal nodes 5–8, we see that STEM persistence of non-Asians who enrolled in two or more STEM courses, is contingent on additional information of their CUMGPA, SATV scores, and finally the number of STEM courses. Of these, students with CUMGPA ≥ 3.705 persist in STEM with estimated probability 60.8 percent (terminal node 5). The last group of students that the tree model suggests will likely persist in a STEM major are non-Asian students whose CUMGPA is between 3.175 and 3.705, SATV < 595 and NUMSTEM ≥ 3 (70.3 percent, terminal node 8).

Conditional relationships among the variables are clear in the tree model in Figure 6. For example, only the number of STEM courses taken as a freshman, ethnicity and SAT verbal scores are relevant for classifying a student as persistent in STEM given that the student's cumulative GPA is at least 3.175; students with cumulative GPA's less than 3.175 are unlikely to persist to STEM graduation regardless of the values of other variables. Also, recall that the coefficient of SATV in the logistic regression model (4) was negative. The tree model gives a partial explanation for this anomaly; SATV is used as a predictor only for students who take at least 2 STEM courses, are non-Asian, and have CUMGPA between 3.175 and 3.705. For these students, high SATV leads to a prediction of non-persistence.

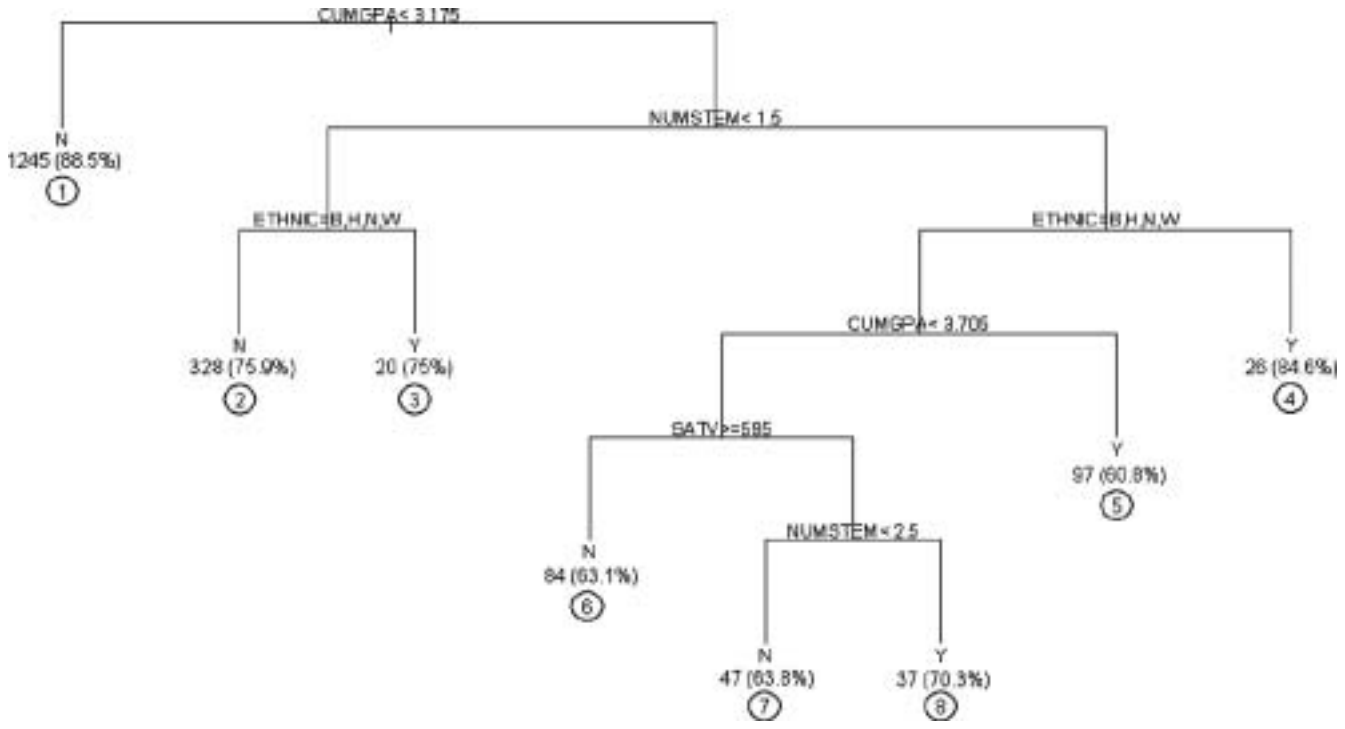

Figure 6. Classification tree formed using the variables identified as important with respect to STEM persistence based on the random forest importance scores. Terminal nodes 1 through 8 are labeled by the appropriate encircled number and a "Y" an "N" in the node refers to persisting and non-persisting STEM students, respectively. The larger number in the node description refers to the number of students who were classified at that point in the tree and the percent value refers to the percentage of those students who were correctly classified by the tree model.

Such students may decide they may have better opportunities or grades in non-STEM fields, and it would be interesting to interview such students to identify their reasons for non-persistence.

Although some variables with high importance scores are not represented in the tree model, they do play a surrogate role. For example, TOTHOURS is not used as a primary predictor yet it is the best surrogate for both NUMSTEM splits and also the CUMGPA < 3.705 split. Also, SATQ is not shown in the tree model but SATQ ≥ 645 was the best surrogate for the SATV ≥ 595 split although the surrogate's concordance percent was only 64.7 percent.

Note that the variable GENDER did not appear in the stepwise logistic regression model (4) nor was it selected as one of the variables to construct the STEM classification tree. The relatively low rank of GENDER in the random forest importance scores provides additional information, indicating that GENDER is not being masked by other variables. Xie and Shauman (Xie and Shauman, 2003) also found that rates of persistence for men and women were similar. Our analysis indicates that the persistence is similar within subgroups of the data defined by cumulative GPA and number of STEM courses taken. A recent National Research Council Report (2006) suggests that many students, particularly women, who could be successful in STEM fields have already decided not to major in science or engineering before they enter college. Our analyses using ASU data support that finding. In fact, the percentage of students entering the STEM pipeline at ASU has decreased since the early 1990's. In 1992, about 58 percent of male freshmen and 43 percent of female freshmen enrolled in at least one STEM course; in 1999, the year the freshmen in our data set started college, 54 percent of male freshmen and 45 percent of female freshmen enrolled in at least one STEM course. By 2005, those percentages had dropped to 39 percent for men and 34 percent for women. While the gender gap may be closing for students in the STEM pipeline, it is closing because the percentage of men in the pipeline has decreased more than the percentage of women. Clearly, students can not persist if they are not in the pipeline to begin with.

Adelman (2006), using national longitudinal survey data, reported that students who graduated from college (in any field, not necessarily in a STEM field) were much more likely to have completed college-level mathematics before the end of the second year. While we did not find that taking calculus as a freshman (FRESHCALC) had a large importance score with respect to STEM persistence classification, the number of STEM courses taken as a freshman was significantly associated with persistence when examined in conjunction with other factors. This finding, in conjunction with information in interviews with students reported in Haag et al. (2007), suggests that the freshman year experience is important for student persistence in STEM fields.

V. Conclusions

In this paper we presented a new method for studying persistence of students in engineering and STEM fields that can be used to complement standard methods such as logistic regression or used as a stand-alone analysis technique. Our analyses do not give prescriptives for increasing the number of students who persist in STEM majors. As stated in section II, we did not have information on student attitudes or perceptions on teaching methods; most of the variables that emerged as important in our models were demographic or related to high school or college GPA. If more detailed information were available, however, classification trees and importance scores using the random forest method could be used to identify factors that might be manipulated in an experimental setting to assess their effect on persistence.

We showed that by using importance scores from a random forest model, an analyst can distinguish important variables to reduce the number of covariates without specifying a model structure. Therefore, the random forest method allows us to identify important predictors of persistence that may be deemed not significant in logistic regression models because of their high correlation with other predictors. Afterwards, a model can be constructed, such as a classification tree, using the subset of variables selected as most important.

Classification trees are able to easily illustrate complex structures in the data that otherwise would take many interaction terms to find using traditional regression techniques. This allows us to look at conditional relationships among the factors much more simply. Also, meaningful ranges of continuous variables and common levels of categorical variables are highlighted in a tree model. Highlighting a particular cut-off level of a variable can help identify a subpopulation of students that might need attention. Also, classification trees can handle highly correlated predictors by taking advantage of the correlation through the creation of surrogates. Along with prediction using incomplete data, surrogate variables allow the analyst to get an idea of other possible branchings of the model. Perhaps more importantly, the full, perhaps incomplete, data set can be used with trees since they are able to handle missing values.

In our analyses, we found that classification trees and random forests identified factors and complex relationships not found by other statistical methods. We used a binary response, but all of these methods discussed in this paper can be used with more than two outcome categories. We believe classification trees and the random forest method shows great promise as an additional methodology for studying persistence to graduation for STEM fields.

Acknowledgments

This research was partially supported by grants EHR-0412537 and SES-0604373 from the National Science Foundation. The authors are grateful to the associate editor and reviewers for their helpful comments and suggestions.

References

Adelman, C. 2006. The toolbox revisited: Paths to degree completion from high school to college. US Department of Education Web site. www.ed.gov/rschstat/research/pubs/toolboxrevisit/toolbox.pdf.

Astin, A. 1985. Involvement: The cornerstone of excellence. *Change* (July/August).

Besterfield-Sacre, M., C. Atman, and L. Shuman. 1997. Characteristics of freshmen engineering students: Models for determining student attrition in engineering. *Journal of Engineering Education* 86 (2): 139–49.

Brainard, S., and L. Carlin. 1997. A longitudinal study of undergraduate women in engineering and science. Proceedings, Frontiers in Education Conference, Pittsburgh, PA.

Breiman, L., J. Friedman, R. Olshen, and C. Stone. 1984. *Classification and regression trees*. California: Wadsworth.

Breiman, L. 2001. Random forests. *Machine Learning* 45 (1): 5–32.

Burtner, J. 2005. The use of discriminant analysis to investigate the influence of non-cognitive factors on engineering school persistence. *Journal of Engineering Education* 94 (3): 335–38.

Dorrans, N. J. 1999. Correspondence between ACT and SAT I scores. *College board research report*, 99–1. New York: The College Board.

Grandy, J. 1998. Persistence in science of high-ability minority students: Results of a longitudinal study. *The Journal of Higher Education* 69 (6): 589–620.

Great Britain Parliament. 1990 (reprint). *Report on the loss of the S.S. Titanic: British Board of Trade inquiry report*. Gloucester, UK: Allan Sutton Publishing.

Haag, S., A. Garcia, and N. Hubele. 2007. Forthcoming. Engineering undergraduate attrition and contributing factors. *International Journal of Engineering Education*.

Hastie, T., R. Tibshirani, and J. Friedman. 2001. *The elements of statistical learning*. New York: Springer.

LeBold, W. K., and S. K. Ward. 1988. Engineering retention: National and institutional perspectives. Proceedings, 1988 American Society for Engineering Education Conference. 843–51.

Leslie, L., G. McClure, and R. Oaxaca. 1998. Women and minorities in science and engineering: A life sequence analysis. *The Journal of Higher Education* 69 (3): 239–76.

Levin, J., and J. Wyckoff. 1991. Predicting persistence and success in baccalaureate engineering. *Education* 111 (4): 461–68.

Maindonald, J., and J. Braun. 2003. *Data analysis and graphics using R*, England: Cambridge University Press.

May, G., and D. Chubin. 2003. A retrospective on undergraduate engineering success for underrepresented minority students. *Journal of Engineering Education* 92 (1): 1–13.

National Research Council. 2006. B*eyond bias and barriers: Fulfilling the potential of women in academic science and engineering.* Washington, D. C.: National Academies Press.

National Science Foundation (NSF). 2007. Scientists and engineers statistical data system (SESTAT). sestat.nsf.gov.

Neter, J., M. H. Kutner, C. J. Nachtscheim, and W. Wasserman. 1996. *Applied linear statistical models*, 4th Edition. Columbus, OH: McGraw-Hill Companies, Inc.

R Project for Statistical Computing, The. www.r-project.org. Accessed October 14, 2007.

Rayman, P., and B. Brett. 1995. Women science majors: What makes a difference in persistence after graduation? *The Journal of Higher Education* 66 (4): 388–414.

Ripley, B. D. 1996. *Pattern recognition and neural networks*, Cambridge, England: Cambridge University Press.

Seymour, E., and N. Hewitt. 1997. *Talking about leaving: Why undergraduates leave the sciences.* Boulder, CO: Westview Press.

White, J.L. 2005. Persistence of interest in science, technology, engineering and mathematics: An analysis of persisting and non-persisting students. Paper presented at the Annual Conference of the Mid-Western Educational Research Association, Columbus, OH.

Xie, Y., and K.A. Shauman. 2003. *Women in science: Career processes and outcomes*. Cambridge, MA: Harvard University Press.

Zhang, G., T. Anderson, M. Ohland, and B. Thorndyke. 2004. Identifying factors influencing engineering student graduation: A longitudinal and cross-institutional study. *Journal of Engineering Education* 93 (4): 313–20.

AUTHORS' BIOGRAPHIES

Guillermo Mendez is a Ph.D. candidate in Statistics in the Department of Mathematics and Statistics at Arizona State University. He received an M.S. in Statistics at Arizona State University and a B.S. in Mechanical Engineering at Purdue University. His research interests include modeling dependent data, data mining and tree-based ensemble methods and models.

Address: Department of Mathematics and Statistics, Arizona State University, Tempe, AZ 85287; telephone: (+1) 480.727.8751; e-mail: guillermo.mendez@asu.edu.

Trent Buskirk recently joined the Division of Biostatistics with the School of Public Health at St. Louis University. His main research foci include survey sampling statistics, survey research methods, multivariate statistical techniques, and health intervention design and analysis. Prior to joining St. Louis University, Dr. Buskirk served on the research staff of the NSF-funded "Project Pathways," a teacher-based intervention for improving student achievement and participation in mathematics and science. Dr. Buskirk has also served as a principal investigator, collaborator, and statistical consultant for various health related studies including, most notably, the American Cancer Society's national quality of life of cancer survivors studies.

Address: School of Public Health, St. Louis University, St. Louis, MO 63104; telephone: (+1) 314.977.8127; e-mail: tbuskirk@slu.edu.

Sharon Lohr is the Thompson Industries Dean's Distinguished Professor of Statistics, Department of Mathematics and Statistics, Arizona State University. She obtained her Ph.D. in statistics from the University of Wisconsin-Madison in 1987 and has been at ASU since 1990. Professor Lohr's research focuses on survey sampling, design of experiments, and applications of statistics in the social sciences and education. She is the author of the widely used text *Sampling: Design and Analysis*, and has published numerous articles in a wide variety of journals including *The Annals of Statistics, Journal of the American Statistical Association, Journal of the Royal Statistical Society, Biometrika, Journal of Quantitative Criminology, Wisconsin Law Review, and The American Statistician*. She has served as chair of the Survey Research Methods Section of the American Statistical Association, member of the Census Advisory Committee of Professional Associations, member of the Statistics Canada Advisory Board on Statistical Methodology, and president of the Arizona Chapter of the American Statistical Association. She is a Fellow of the American Statistical Association, an Elected Member of the International Statistical Institute, and was selected to receive the inaugural Gertrude M. Cox Award from the Washington Statistical Society in 2003 for contributions to statistical practice.

Address: Department of Mathematics and Statistics, Arizona State University, Tempe, AZ 85287; telephone: (+1) 480.965.4440; e-mail: sharon.lohr@asu.edu.

Susan Haag is the director of Evaluation and Assessment for the Ira A. Fulton School of Engineering at ASU. Over the past 11 years, she has conducted research on institutional reform and learner persistence in STEM fields; implemented technological advancements and cognitive applications to improve performance outcomes for learners in STEM education; developed and evaluated e-learning contexts;

examined industry-university collaboration; and studied the recruitment and persistence of underrepresented populations longitudinally. Dr. Haag is currently working on a second Ph.D. in Cognitive Psychology investigating recognition memory employing visuo-haptic crossmodal conditions and transfer. Her current study aims to introduce visual-haptic events as a paradigm, with ecological validity, which may yield new insights into the nature of perceptual interactions and the generalizability of existing theories, both for perception and recognition. In addition, a current study includes examination of the occipitotemporal region verifying that this region is a multimodal (visuo-haptic) information processing region devoted to object recognition, which participates in haptic, auditory, and possibly linguistic tasks. Dr. Haag functioned as a national evaluator for the NSF Foundation Coalition grant from 1998 to 2004; she implemented and evaluated multiple curricular reform strategies; examined student performance nationally (7 university sites); developed manuscript documentation; and disseminated results locally and nationally.

Address: Ira A. Fulton School of Engineering, Arizona State University, Tempe, AZ 85287; telephone: (+1) 480.965.7219; e-mail: susan.haag@asu.edu.