Analysis of Enrolment Data to Predict Undergraduate Engineering Recruitment

[1]E.E. Anderson, [2]J.G.Surles, [3]H. Wu
Department of Mechanical Engineering, Texas Tech University, Lubbock, Texas, ed.anderson@ttu.edu[1];
Department of Mathematics, Texas Tech University, Lubbock, Texas, james.surles@ttu.edu [2];
Department of Mathematics, Texas Tech University, Lubbock, Texas, Haifeng.wu@ttu.edu [3]

Abstract

Engineering student recruiters often need to target student groups to optimize their time and resources. This study uses the logistic regression, decision tree, and random forest statistical classification methods to examine typical college application variables and to identify those variables that have the most impact upon students electing to study engineering. The entire Texas Tech University Fall 2008 entrance class of 4436 students (693 of whom were engineers) was included in this study. The independent classification variables selected for this study included major, gender/ethnicity, family income, entrance examination scores, parent education level, high school class rank, high school size, and high school senior college bound size. The most significant discriminator was gender followed by entrance examination mathematics scores by all three classification methods. Family income, parent education level, high school class size, and high school college bound size were the last four significant classification variables. The order of the last four variables varied depending upon the classification method. Logistic regression indicated parent education level was more significant among the last four variables and the remaining three were approximately equivalent. Similar discrimination patterns were also found by decision tree and random forest analysis. Female students only were isolated from the general population and analyzed using the random forest method. Entrance examination mathematics scores were found to be the second most significant predictor for this subset of students.

1.Introduction

The need to increase the number of students enrolled in and graduating from engineering programs by increasing either retention or recruitment or both has been clearly pointed out in a number of national studies and reports. In view of the downturn in engineering freshman enrollments starting in 2001, it is clear that this shortage will become even more severe in the future. Improvements in retention and graduation rates will certainly reduce this problem somewhat, but the problem cannot be solved without increases in engineering admissions. We definitely need to do a better job of getting students to enroll in engineering while we simultaneously improve retention and persistence to graduation to properly address this problem.

College entrance data are used for many different purposes such as: demographic studies, university resource planning, marketing, predicting retention and graduation rates, success of programs like “No Child Left Behind”, correlations between variables, and many other purposes.

Studies of enrolment trends, impact of special programs to encourage STEM studies, demographic characteristics, and similar objectives are plentiful. The investigators were unable to find any studies that analyze this data from a classification perspective with a view of identifying those factors that lead a student to select one course of study over another at the time of admission. This type of study needs to be performed using statistical classification techniques such as discriminate analysis, logistic regression, decision tree statistics, and random forests specifically designed to address classification problems. These techniques are designed to determine the impact of the various independent variables on the dependent variable (student selection of engineering at the time of admission in this project). They are capable of simultaneously considering many independent variables, flawed data, continuous and binary data with reasonable computational effort unlike correlation and other techniques. This research was conducted to apply statistical classification techniques to the questions "What elements enter into the decision of a student to study engineering and what is the significance of these elements?"

2. Background

In this project, the statistical classification methods, logistic regression, decision trees, and random forests, were applied to Texas Tech University (TTU) admissions data to study the relationship between students' engineering major preference and several cognition and demographic variables. We were unable to find any studies that employed statistical classification methods as applied to college admission data for the purpose of predicting college major preference. We did, however, find several papers that applied statistical classification to predict engineering student retention and persistence.

Schiller and Muller (2003) used a hierarchical linear model to explore the association between the taking of college level mathematics courses, high school graduation requirements and assessment policies. They found that the higher the high school graduation requirements were, the more likely the student will take and persist in college mathematics courses.

Ali, et al. (1992) applied discriminate function modeling to students admitted to Beirut University College to build a predictive model of academic success as a function of admission data. The most important factor for natural science students was found to be a variable composed of high school GPA and the type of high school attended. They also found gender and entrance examination scores to be significant in predicting student success.

Biel, et al. (1999) employed logistic regression analysis to study the impact of commitment, academic integration, and social integration on retention. Their results show that commitment is the strongest predictor of retention and that integration is driven in part by commitment.

Wohlgemuth, et al. (2007) applied logistic regression to retention and graduation rates of students in a Midwestern, research extensive university and developed models for students in seven different colleges/departments, including engineering. They found that admission examination scores are a strong predictor of retention and graduation rates and that financial assistance is significant for predicting the retention of first year students. An interesting finding of theirs was that female and first generation students were most likely to be retained the first year, but not during latter years.

Mendez, et al. (2008) applied three statistical classification techniques to investigate engineering student persistence in a six year longitudinal study. They found that the random forest method of Breiman (2001) and the classification tree method (see, for example, Breiman, et al., 1984) can identify important predictors without identifying a model structure as required by more traditional techniques, such as logistic regression.

In each of these studies, the dependent variable is a binary variable while the independent variables were either categorical (such as classification as a freshman, sophomore, etc.) or numerical (such as GPA or the score on a standardized exam). In this study, we used statistical classification methodology as applied to college admission data for the purpose of predicting engineering major preference. The admissions dataset for the entire freshman class at TTU in the fall of 2008 was obtained. Data for a total of 4436 students were available.

In this study, the dependent variable is the selection of engineering as a major (with two possible outcomes: they either did or did not select engineering. The independent variables come from the admission data of the students, which includes the following:

- Major
- Gender and Ethnicity
- Family Income
- Entrance examination scores (ACT or SAT)
- Mother/Father education level
- High school class rank
- High school size
- High school senior class size

Three methods were used to analyze the data to study which of the factors listed above are significant predictors for choice of engineering as a major: Logistic regression, decision trees, and random forests.

2.1 Logistic Regression

Suppose that Y is a dependent variable whose only possible values are 0 and 1, and that the probabilities that Y takes on the value of 0 or 1 depend on p independent variables $x_1, x_2, \ldots, x_p$. In logistic regression, the following equation is used to model $E[Y|x_1, x_2, \ldots, x_p]$ (where $E[\cdot]$ represents expectation):

$$\pi(x_1, x_2, \ldots, x_p) = E[Y|x_1, x_2, \ldots, x_p] = P(Y = 1|x_1, x_2, \ldots, x_p) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2 \ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2 \ldots+\beta_p x_p}} \quad (1)$$

Applying the logit function, $\text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$, to $\pi(x_1, x_2, \ldots, x_p)$ results in:

$$g(x_1, x_2, \ldots, x_p) = ln\left(\frac{\pi(x_1, x_2, \ldots, x_p)}{1-\pi(x_1, x_2, \ldots, x_p)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_p x_p \quad . \quad (2)$$

That is, we model the logit of the probability that Y takes on the value of 1 as a linear model of the independent variables. Hosmer and Lemeshow (1989) is a standard reference for logistic regression, and we refer the reader there for further details on logistic regression. It is worth mentioning, however, that the method of maximum likelihood was used in calculating the parameter estimates, the stepwise variable selection method was used to build the model, and lastly the goodness-of-fit test proposed by Hosmer and Lemeshow (1980, 1982) was used to assess model adequacy.

2.2 Decision trees

What follows is a brief description of decision trees. Generally speaking, there are two basic types of decision trees: classification trees and regression trees. Classification trees are used when the response variable is categorical (as in our case), and regression trees are used when the variable is continuous. Collectively, they are known as CART or C&RT, which stands for Classification and Regression Trees. The reader is referred to Brieman, et al. (1984) for further details regarding decision trees.

A decision tree is a set of binary questions (those with a "yes" or "no" answer) that divides the whole dataset into smaller groups in which the predicted probability of classification is a constant within each group. This is most often represented as a tree diagram where each node represents a question. The observations that answer "yes" to a question travel to the right descendant node and those answering "no" travel to the left descendant. The basic structure of decision trees can be described by the following:

- Each split depends on the value of only a single variable.
- If x_m is a numeric variable, the questions are of the form {is $x_m \leq c$?} for all c over the range of x_m.
- x_m is a categorical variable taking values in $\{v_1, v_2, \ldots, v_L\}$, then the questions are of the form {is $x_m \in S$?} as S ranges over all possible subsets of $\{v_1, v_2, \ldots, v_L\}$.

For each node, the tree algorithm searches through the p variables one-by-one, finding the best split for each. It then compares the p best splits and selects the best of the best.

A criterion function is used to define what is meant by "best" which has the basic property that each split of a node results in two descendant nodes that are more pure than the parent node. Such a criterion function should have the following properties:

- The function reaches its maximum when all classes are equally mixed together in the node, and
- The function reaches its minimum when the node contains only one class.

There are many different criteria that might be used, but the most commonly used, and the one employed here, is the Gini index. For a given node m, the Gini index is given by $\text{Gini}(m) = 2p_m(1-p_m)$, where p_m is the proportion of the observations in the node of class 1. The Gini index has a maximum of 1/2 when $p_m = 1/2$ and a minimum of 0 when $p_m = 0$ or 1, meaning that there is no impurity in the node. The split that we select at a node is the one which reduces the Gini index by the largest amount in the immediate descendant nodes.

When a tree is found, we do not generally want it to be such that every terminal node is completely pure. Significant over fitting occurs, and the tree has poor predictive performance (Breiman, et al., 1984). The way that over fitting is avoided is to grow the tree until each terminal node is completely pure, and then prune it upwards according to a certain rule. The rule employed here is known as *minimal cost-complexity pruning*. For a tree with T terminal nodes, the cost-complexity measure is given by $R_\alpha(T) = R(T) + \alpha T$, where $R(T)$ is the misclassification rate, and α is a positive real number called the complexity parameter. If α is small, then the cost for having a large number of nodes will be relatively small and the trees produced will tend to be larger. For a specific α, T can be found such that $R_\alpha(T)$ is a minimum. Thus α is a tuning parameter, and its value is selected by the researcher, generally in an arbitrary fashion.

2.3 Random Forests

Appropriately named, a random forest has the basic idea that it is composed of many individual classification trees. Each classification tree in the forest is constructed from a data set composed of observations that have been sampled, with replacement, from the original data set. Each tree is grown completely until each terminal node is pure. To classify an observation, each tree in the forest is used to give a classification, known as a vote. Generally speaking, the classification with the largest number of votes is the predicted classification.

In practice, building a random forest can be computationally prohibitive if the data set is large and/or there are a large number of variables. To make the random forest feasible in those

cases, measures such as building the tree using a re-sampled data set that is smaller than the original (e.g. two-thirds the size) and randomly selecting a subset of k of the p variables for each tree ($k < p$) can be employed. The reader is referred to Breiman (2001) for further details on Random Forests.

3. Data Analysis

Over all observations, 693 students self-declared as engineering majors. There were 2113 females in the data, of which 80 self-declared as engineering majors. Thus, 2324 males were represented, of which 613 were declared as engineering majors. The data set was randomly split in such a way that two-thirds of the data were taken to be the training data set, and the remaining one-third was the validation data set. The proportion of engineering students in each data set was also maintained to be the same as in the original data set (about 15.6%). The variables studied for each of the classification procedures are detailed in Table 1. The training data set was used to build classification models using each of the techniques described above and then each model was applied to the validation data set to diagnose effectiveness. The results of these are given next.

Table 1: Detailed Variable Descriptions

Variable	**Description/Possible Values**
MAJOR	Engineer, Non-Engineer
GENDER	Male, Female
ETHNICITY	White, Asian or Pacific Islander, Black or African-American, Hispanic or Latino, Indian American or Alaska Native, Non-Resident Alien, Other
FAMILY INCOME	< 20,000, 20,000-39,999, 40,000-59,999, 60,000-79,999, > 79,999
MOTHER EDUCATION LEVEL	Bachelor's/4-Year Degree, Graduate/Professional Degree, High School Diploma/GED/Home School, No High School, Some College-No Degree/Certificate, Some High School-No Diploma, Unknown/Unavailable
FATHER EDUCATION LEVEL	Same as MOTHER EDUCATION LEVEL
HIGH SCORE	Continuous variable. Highest score on ACT or SAT standardized exams. Converted to SAT scale if ACT by matching percentiles.
HIGH SCHOOL CLASS RANK	Continuous variable, percentage.
HIGH SCHOOL CLASS SIZE	Continuous variable.
HIGH SCHOOL PERCENTILE	Continuous variable.
HIGH SCHOOL PERCENT COLLEGE BOUND	25% or fewer, 26%-50%, 51%-75%, 75%-90%, over 90%, Unknown.
HIGH SCHOOL SENIOR CLASS SIZE	100 or fewer, 101-500, 501-1000, more than 1000
MAXIMUM MATH PERCENTILE	Continuous variable. Highest score (as a percentile) on the math portion of the SAT or ACT.

3.1 Data Analysis: Logistic Regression

SAS's PROC LOGISTIC was used to carry out this analysis. MAJOR was used as the dependent variable, and all other variables outlined in Table 1 were used as the independent variables to fit the model. Using stepwise selection with $p_E = 0.3$ (the p-value required for a variable to enter the model), and $p_R = 0.35$ (the p-value required for removal from the model), six of the variables were chosen in the logistic regression model. They are given in SAS output form in Table 2. It is clear that GENDER is the "most important" variable in predicting the choice of engineering, followed by MAXIMUM MATH PERCENTILE. FATHER EDUCATION LEVEL and HIGH SCHOOL CLASS SIZE are both significant, but neither to the degree that GENDER and MAXIMUM MATH PERCENTILE are. Note that care must be taken when interpreting these variables due to co-linearity (correlation) that exists between the independent variables. For example, MAXIMUM MATH PERCENTILE and HIGH SCORE are highly correlated, so only one appears in the model. The Hosmer and Lemeshow (1980) goodness-of-fit test reveals no lack-of-fit for logistic regression applied to this data (p-value = 0.9484, see Table 3).

Table 2: Logistic Regression Variable Summary

Step	Entered	DF	Number In	Score Chi-Square	Pr > ChiSq
1	Gender	1	1	217.2051	<.0001
2	MaxMathP	1	2	85.8391	<.0001
3	FatherEduc	5	3	21.2476	0.0007
4	HSClassS	1	4	7.5025	0.0062
5	HSPercentCollegeB	4	5	7.8374	0.0977
6	FamilyIncome	4	6	5.3917	0.2494

Table 3: Logistic Regression Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
2.7621	8	0.9484

When this model is applied to the validation data set, the result is the predicted probability of having engineering for each of the observations in the validation data set. The value 0.1562 (693 engineering majors / 4437 total students) is used as the cut-point to "predict" a student as an engineering major or not. A predicted probability above this, and they are predicted to be more likely to be an engineering major than the general population. Using this criterion, the confusion matrix for logistic regression is given in Table 4. This is compared to a random classifier which simply classifies a student as an engineering major with probability 0.1562 for all students. The expected confusion matrix for the random classifier is given in Table 5.

Table 4: Confusion Matrix for Logistic Regression

		Actual	
		Engineering	Non-Engineering
Predicted	Engineering	138	305
	Non-Engineering	36	521

Table 5: Expected Confusion Matrix for Random Classifier

		Actual	
		Engineering	Non-Engineering
Predicted	Engineering	27.2	129.0
	Non-Engineering	146.8	697.0

The false positive rate (classifying a student as being more likely to be an engineering major when they are, in fact, not an engineering major) for the logistic regression is thus 0.369 (305/826), while the false negative rate (classifying a student as being less likely to be an engineering major when they are, in fact, an engineering major) is 0.207 (36/174). The false positive rate for the random classifier is 0.156, while the false negative rate is 0.844. At the cost of doubling the false positive rate compared to the random classifier, the logistic regression lowers the false negative rate more than four times.

3.2 Data Analysis: Decision Tree

The same training data set used to build the logistic regression was also used to build the decision tree. The resulting tree is given in Figure 1. Non-terminal nodes list the variable that the data is split on, while terminal nodes contain information regarding the students represented by that node. The first is the proportion of students in that node that are engineering majors (which would thus be the estimated probability of a student in that node having engineering as their major), followed by the total number of students in that node, then the decision as to whether they are predicted to be more likely engineering majors or not.

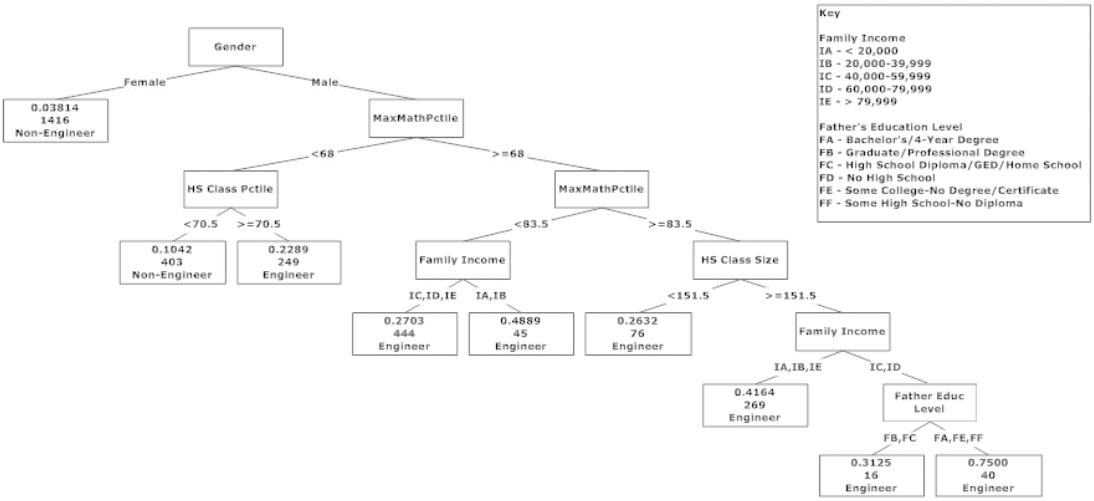

Figure 1: Decision Tree

As can be seen, GENDER and measures of aptitude (MAXIMUM MATH PERCENTILE and HIGH SCHOOL PERCENTILE) are the most important. Every student in the top one-third percentile would be more likely to select engineering than the 0.1562 for the entire population. This decision tree was applied to the validation data set, which results in the confusion matrix given in Table 6. (Note that the reason that more students are represented in Table 6 than in Tables 4 and 5 is because the decision tree is able to use more of the data than logistic regression, as any variable containing a missing value will cause that entire data point to be excluded.) The false positive rate for the decision tree is 0.339 and the false negative rate is 0.242, both comparable to the logistic regression.

Table 6: Confusion Matrix for the Decision Tree

		Actual	
		Engineering	Non-Engineering
Predicted	Engineering	175	423
	Non-Engineering	56	824

3.3. Data Analysis: Random Forests

Again using the same training data set that was used in the logistic regression and decision tree analysis, the R package "randomForest" was used to generate 500 trees. As described previously, predictions are made using the random forest by running a data point through each tree in the forest and determining the number of "votes" for each of engineering and non-engineering. The percentage of those that are engineering is taken to be the predicted probability of selecting engineering. Using the same cut-point of 0.1562 as before, Table 7 gives the confusion matrix for the random forest. The false positive rate is 0.400 and the false negative rate is 0.213.

Table 7: Confusion Matrix for the Random Forest

		Actual	
		Engineering	Non-Engineering
Predicted	Engineering	139	330
	Non-Engineering	35	496

Determining variable importance is a non-trivial task when considering random forests. Each tree can split on different variables at different levels. In order to investigate variable importance, a random forest was manually built by constructing 500 decision trees, with pruning. The pruning was done so as to make the task of compiling the information feasible. In every single one of the 500 trees, GENDER was the first split. Of those that were female, there were only 122 of the 500 that had an additional split, and 84 of those were split on the MAXIMUM MATH PERCENTILE. Of those that were male, 490 of the 500 were next split on MAXIMUM MATH PERCENTILE. The remaining 10 were split on HIGH SCORE. On the third level, 414 of the 500 trees *again* split on MAXIMUM MATH PERCENTILE, while 200 split on HIGH SCHOOL CLASS SIZE. Bear in mind that the total on the third level within the males can be greater than 500, as you can have two nodes that are split at this point.

4. Discussion

GENDER is clearly the single most significant factor in determining if a potential student is a strong candidate for studying engineering as demonstrated by all three classification analysis. In fact there were so few females declaring engineering as their major that tree analysis remove them from consideration immediately and did not include them in determining the role of the other variables. Of the 2113 females in the sample, only 3.8% self-declared as engineers. Hence it is self-evident that recruiting which targets males has a higher probability of success. That is not to say that females should not be targeted, in fact, they should. But, we should expect a lower rate of success recruiting females as males. In order to increase the female populations studying engineering the recruiting effort directed to females will need to be several factors greater than that directed towards males. However, each of the models can be used to provide estimates of the probability of that student selecting engineering (if left to their own

devices), which provides a recruiter a parameter to compare/rank students. This applies to sub-setting the data by groups (e.g. sex or race). For example, if one female student has a predicted probability of selecting engineering equal to 12% while another has a predicted probability of 6%, then the recruiter is reasonable in expecting to be more likely to be successful in recruiting the first student.

Beyond GENDER, the next most significant variable, according to all three analyses, was MAXIMUM MATH PERCENTILE from ACT and SAT tests. This is also consistent with researchers who have studied retention and persistence. This is certainly logical in that one would expect those factors that lead a student to electing engineering would be the same as those factors that carry them through to graduation.

HIGH SCHOOL CLASS SIZE, FATHER EDUCATION LEVEL, and FAMILY INCOME are third-order variables with significantly less impact on a potential student's decision to study engineering.

Thus it is clear that GENDER is the most important, followed by MAXIMUM MATH PERCENTILE. HIGH SCHOOL CLASS SIZE comes in at a distant third. Note that all three of these were highly significant in the logistic regression. FATHER EDUCATION LEVEL was highly significant in the logistic regression, but appears to be of less importance in both the decision tree and the random forest.

When making a decision to recruit a given student or class of students, one can either make a random choice or adopt and use one or all of the three models reported here. When making a dichotomous classification, the probability of a "false positive" decision for a random decision is 15.6% according to Table 5 and the probability for a "false negative" decision is 84.4%. Hence, 84.4% of the candidates who actually are engineers will be misclassified. Logistic regression lowers the "false negative" rate to 20.7. This is a significant reduction and considerable savings in non-productive effort. Similarly, the "false negative" rates for decision tree and random forest analysis were 24.2 and 21.3, respectively. All three classification analyses then demonstrate a gain in recruiting efficiency. The price of this gain is missing engineering candidates among those believed to be non-engineering candidates as shown in the "false positive" probabilities. But more than this, each model provides a recruiter with a method to rank students according to their predicted probabilities of selecting engineering as a major. It is, of course, up to the recruiter to decide how this information is to be best used in their department or school, but it is certainly available and is potentially a very valuable resource. Note, however, that logistic regression and random forests offer predicted probabilities that are fine, in the sense that they are, for all practical intents and purposes, continuous, while the decision tree's predicted probabilities would be very coarse (indeed, in the tree found in this research, there are only 9 possible predicted probabilities).

5. Conclusions

This study has applied three classification analysis, logisitic regression, decision trees, and random forests, to the student admission data of a research-intensive, Midwestern, public university to explore the question "what factors identify a newly admitted student as an engineering major?" Of the twelve independent variables considered, two were found to be very significant, gender and academic success (entrance examination mathematics scores, high school percentile, and high school class rank). Gender is clearly the first discriminator which dominates a student's decision to study engineering. Academic success is the second discriminator. Interestingly enough, academic success appears to be less of a factor for females who have chosen engineering than males. It is suspected that this is due to some other consideration that is not reflected in the dataset selected for this investigation. Further study is recommended. Four other variables; father's education level, high school size,

percentage of graduation class that were college bound, and family income; were found to play a role in this decision to study engineering. These variables were much less significant as compared to gender and academic performance. But father's education level was more significant than the other three secondary variables.

It was found that the probability of misclassifying an engineering candidate as a non-engineer was lowered by a factor of four when these classification techniques are compared to simple random guessing. The probability of misclassifying a non-engineer as an engineer was found to double by using these techniques.

6. Acknowledgments

This research was sponsored by the United States National Science Foundation under grant No. EEC-0836028

References

1. H.F. Ali, A. Charbaji,, and N.K. Haji, (1992). "A Discriminant Function Model for Admission at Undergraduate University Level". *International Review of International Education*, **38**, 505-518.
2. C. Beil, C.A. Reisen, M.C. Zea, and R.C. Caplan, (1999). "A Longitudinal Study of the Effects of Academic and Social Integration and Commitment on Retention". *NASPA Journal*, **37**, 376-385.
3. L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, (1984). Classification and Regression Trees. Wadsworth International Group, Belmont, CA.
4. L. Breiman, (2001). "Random Forests". *Machine Learning*, **45**, 5-32.
5. D.W. Hosmer, and S. Lemeshow, (1980). "A Goodness-of-Fit Test for the Multiple Logistic Regression Model". *Communications in Statistics*, **A10**, 1043-1069.
6. D.W. Hosmer, and S. Lemeshow, (1982). "The Use of Goodness-of-Fit Statistics in the Development of Logistic Regression Models". *American Journal of Epidemiologyi,* **115**, 92-106.
7. D.W. Hosmer, and S. Lemeshow, (1989). Applied Logistic Regression. John Wiley & Sons, New York, NY.
8. G. Mendez, T.D. Buskirk, S. Lohr, and S. Haag, (2008). "Factors Associated with Persistence in Science and Engineering Majors: An Exploratory Study Using Classification Trees and Random Forests". *Journal of Engineering Education*, **97**, 57-70.
9. D. Wohlgemuth, D. Whalen, J. Sullivan, C. Nading, M. Shelley, and Y. Wang, (2007). "Financial, Academic, and Environmental Influences on the Retention and Graduation of Students". *Journal of College Student Retention*, **8**, 457-475.